Omics based approaches to study *Campylobacter jejuni* pathogenesis

Ozan Gundogdu

2020

This research dissertation is submitted for the MSc in Bioinformatics at Queen Mary, University of London

Table of Contents

Abstract 3

Acknowledgements 5

1. Introduction 6

1.1 Campylobacter 6
1.1 Next-generation sequencing 6
1.3 16S metagenomics 7
1.4 Operational Taxonomic Units (OTUs) 8
1.5 Amplicon Sequence Variants (ASVs) 9
1.6 Data analysis 10
1.6 Bacterial transcriptomics 11
1.7 Bacterial transcriptomics data analysis 12
1.8 Aims and objectives 13
2. Methods 15

2.1 File organisation 15
2.2. Qiime2 bioinformatics pipeline with DADA2 15
2.3. Qiime2 bioinformatics pipeline with VSEARCH 16
2.4. Statistical analysis of 16S microbiome data 17
2.5. RNA-Seq bioinformatics pipeline 20
2.6. Experimental details and statistical analysis of RNA-Seq transcriptomics data 21
3. Results 23

3.1 Microbia1 community survey using DADA2 pipeline 23
3.2 Microbia1 community survey using VSEARCH pipeline 34
3.3 VSEARCH vs. DADA2 34
3. RNA-Seq analysis using StringTie pipeline 34
3.5 RNA-Seq analysis using bedtools pipeline 37
3.6 StringTie vs bedtools 37
4. Discussion 38

5. Conclusions 41

6. References 42

Appendix I – Bioinformatics Steps for Qiime2 analysis with DADA2 49

Appendix II – Bioinformatics Steps for Qiime2 analysis with VSEARCH 57

Appendix III –16S microbiome results using VSEARCH pipeline 67

Appendix IV – Significant gene list obtained from RNA-Seq StringTie pipeline 76

Appendix V –RNA-Seq results using bedtools pipeline 77

Abstract

Background: *Campylobacter jejuni* is present within the chicken gut and is the leading cause of bacterial foodborne gastroenteritis within humans worldwide. Infection can lead to secondary sequelae such as Guillain-Barré syndrome and stunted growth in children from low-resource areas. Despite the microaerophilic nature of the bacterium, we do not understand how *C. jejuni* can survive the atmospheric oxygen conditions in the environment. We also do not understand how chickens can tolerate $>10^6$ *C. jejuni* and not display any overt disease, yet only 500 *C. jejuni* cause severe disease in human hosts.

Methods: 16S metagenomics bioinformatics and statistical analyses were performed on genomic DNA isolated from chicken ceca under normal and test conditions where chicken were exposed to different concentrations of the natural antimicrobial carvacrol within feed. A comparison between bioinformatics methods whether driven purely by assigning threshold at 97% similarity for species discretisation (VSEARCH OTUs) or based on single nucleotide variants (DADA2 ASVs) was assessed to view the best realisation of underlying community structure and its biological relevance. In addition, RNA-Seq bioinformatics and statistical methods were performed to investigate the expression levels of *C. jejuni* genes when comparing stress conditions against normal growth conditions from RNA isolated at late log phase (16 hours). For comparison purposes, we have explored alternative choices for obtaining transcripts abundances based on different software choices i.e. StringTie and bedtools.

Results: For 16S metagenomics, alpha diversity metrics displayed an increasing microbial diversity as the concentration of carvacrol was increased over time. Subset regression (identifying a subset of confounders to play a role in community statistics) identified day 21 displaying a decreasing effect on microbial diversity, whereas day 35 displays an increasing effect. Also, day 10 seems to shift the microbial community structure away from the average beta diversity status. In addition, increasing the concentration of carvacrol seems to shift the microbial population structure away from the average beta diversity status. Functional analysis using PICRUSt2 identified a range of genes involved in biofilm and sporulation. For comparison, Procrustes analysis (consolidation of patterns between multivariate datasets) was performed on the VSEARCH and DADA2 abundance tables. A Procrustes correlation of 0.787 with a p-value of 0.0001 was obtained, indicating significant similarity in obtaining underlying microbial community structures using either of the methods.

For RNA-Seq, significant up and down regulated genes were obtained using StringTie and bedtools (here used as different methods to ascertain what could be logically defined as a transcript) and then correlating the results specifically with genes involved in the oxidative

stress response. For comparative purposes, Procrustes analysis was performed on the multivariate tables generated from StringTie and bedtools pipelines to ascertain how much they correlate. A Procrustes correlation of 0.8148 with a p-value of 0.0001 was obtained. This again validates conformation between the two methods.

Conclusions: This study investigated different transcriptomics and 16S metagenomics pipelines analysing the pathogenesis of *C. jejuni*. For 16S metagenomics, certain differences were observed between the carvacrol treated versus normal samples. Irrespective of the bioinformatics tools used, the underlying patterns were more or less conserved. This was also observed for RNA-Seq pipelines.

Acknowledgements

I would like to thank Dr Umer Zeeshan Ijaz and Professor Conrad Bessant for providing me with this opportunity to work on the project. Many thanks for the guidance from Professor Bessant. I would like to thank Professor Nicolae Corcionivoschi for providing carvacrol data and always on hand to discuss research topics. A special thanks to Dr Umer Zeeshan Ijaz who has provided incredible guidance during the project, dedicating time for teaching (http://www.tinyurl.com/JCBioinformatics3) and going through all of the data with practical demonstrations. Umer, I am truly grateful for all of your time and generosity. As you have said many times, “a cat never teaches a lion to climb a tree”.

1. Introduction

1.1 Campylobacter

Campylobacter jejuni is the most common bacterial cause of human gastroenteritis worldwide with an estimated 400 million human infections per year (Ruiz-Palacios, 2007). The symptoms of campylobacteriosis are malaise, fever, severe abdominal pain, and diarrhea (Brondsted et al., 2005). *C. jejuni* infection has also been associated with postinfectious sequelae, including septicemia and neuropathies such as Guillain-Barré syndrome (GBS) (Nachamkin et al., 1998). Despite specific microaerobic growth requirements, *C. jejuni* is ubiquitous in the aerobic environment and appears capable of withstanding different stresses, including suboptimal carbon source growth, temperature changes, and exposure to atmospheric oxygen (Fields and Thompson, 2008). A more complete understanding of the regulation of *C. jejuni* response mechanisms to the diverse stresses encountered both during the infection cycle and within the natural environment is required to facilitate appropriate intervention strategies to reduce the burden of *C. jejuni*-associated disease (Pittman et al., 2007).

C. jejuni are also widely found in avians and so the main route of transmission is via the consumption and handling of poultry products. The predominance of *C. jejuni* can be attributed to the ability to survive in the environment as well as within avian and mammalian hosts despite the microaerophilic nature of this bacterium (Byrne et al., 2007). How avians tolerate trillions of *C. jejuni* cells without having overt disease, yet only 500 cells cause severe disease in human hosts remains unknown. Intervention and control strategies against *C. jejuni* have been limited to biosecurity at poultry farms and consumer education in terms of cooking practices. The ban on antibiotics as growth promoters and consumer desire for antibiotic-free chickens has led to exploring natural alternatives to reduce pathogens and simultaneously achieve performance enhancement of chickens. As an example, carvacrol is a natural plant-derived antimicrobial (from oregano) that targets the outer membrane integrity and biofilm formation of certain bacterial species, and so use of carvacrol may be an option to reducing *C. jejuni* levels (Ultee et al., 1999). Thus, all these reasons serve as a basis to explore *C. jejuni* pathogenesis in the context of this study as well as the role of microbiota if implicated towards this end.

1.1 Next-generation sequencing

The arrival of next-generation sequencing (NGS) during the first decade of the 21st century revolutionised science, and this has substantially advanced microbial ecology. Traditional Sanger sequencing technology typically returns 500-600 nucleotides per run. NGS transformed the field by producing large volumes of read data simultaneously at significantly reduced prices (Bonetta, 2010). NGS moved away from the electrophoretic sequencing (utilised by the Sanger methodology), however the key change was multiplexing, where instead of a single tube per reaction, a complex library of DNA templates could be immobilized onto a 2D surface, with all templates accessible to a single reagent volume (Shendure et al., 2017). Bacterial cloning was replaced with *in vitro* amplification which generates copies of each template to be sequenced. Finally, instead of measuring fragment lengths, sequencing includes cycles of biochemistry (e.g. polymerase-mediated incorporation of fluorescently labelled nucleotides) and imaging, also known as 'sequencing-by-synthesis' (SBS) (Shendure et al., 2017, Goodwin et al., 2016). The Illumina platform is the market leader with unprecedented accuracy (Shendure et al., 2017, Goodwin et al., 2016). In addition to the low error rate and relative cost, the Illumina set up has great flexibility allowing applicability to different omics methodologies.

More recently, a third generation of sequencing technologies has been developed, largely referred to as real-time single molecule sequencing where the market leaders are Oxford Nanopore Technologies (ONT) and PacBio. These methodologies, albeit fast and/or cheaper, and/or comprehensive, may provide sequencing in real-time in some cases. However they are far from perfect and issues including copying errors, sequence-dependent biases and information loss (e.g. methylation), along with time and complexity are prevalent (Goodwin et al., 2016). However, the major drawback is that there is a trade-off between read length and accuracy where relative error rates can be as high as 10% for longer reads.

The revolution in genome sequencing has led to a deluge of genomes being sequenced. In addition, the development of NGS also had an important impact on alternative omics-based methodologies, particularly to metagenomics and transcriptomics.

1.3 16S metagenomics

Microbial community profiling using 16S ribosomal RNA (rRNA) is a high-throughput methodology utilised as a *de facto* approach for microbial community surveys, and allows us to gain insights into their spatial and temporal dynamics (Hamady et al., 2008). Microbial community surveys can be broken down to two methodologies: i) targeted amplicon sequencing, typically the 16S rRNA region; or ii) whole-genome shotgun metagenomics

sequencing where the complete DNA is sequenced using a PCR-free methodology. In prokaryotes, the 16S rRNA genes are essential and occur in at least one copy in a genome (Acinas et al., 2004) and the variant regions, V1–V9, are used for species identification (Lane et al., 1985). These characteristics allow the 16S rRNA gene to provide accurate taxonomic classification. Limitations of the 16S methodology are: i) amplification biases exist for various hypervariable regions (Hong et al., 2009, Sharpton et al., 2011); ii) genomic loci other than the 16S region vary in differential strength at distinguishing taxa (Schloss, 2010, Jumpstart Consortium Human Microbiome Project Data Generation Working, 2012); iii) amplicon sequencing elucidates what is there, but not necessarily the biological functions associated with the organisms identified (Quince et al., 2017); iv) comparative studies have highlighted the limitations of delineation to strain level (Johnson et al., 2019); v) given the pervasiveness of horizontal gene transfer, microbial profiling with amplicon sequencing can result in inflated estimates of the community diversity (Acinas et al., 2004); and vi) with the sequencing data and its similarity able to define species boundaries, what is the appropriate threshold to discretise species.

1.4 Operational Taxonomic Units (OTUs)

The use of the 16S rRNA gene sequence to profile microbial communities has inherent problems as the taxonomic resolution of sequence variation across a marker region is limited both biologically and technically, as sequence divergence may not represent wider biological divergence between taxa (Stackebrandt and Goebel, 1994). In addition, sequencing errors introduce artificial divergence (Huse et al., 2010). Thus, enumeration of all sequences is impractical, especially given that many unique reads are often present. To disentangle this issue, reads within a 16S rRNA dataset are typically collapsed into what is referred to as operational taxonomic units (OTUs). OTUs were initially used in the context of numerical taxonomy, where an "Operational Taxonomic Unit" defines the group of organisms being studied (Sneath and Sokal, 1973). In essence, OTUs are a practical representative for species at different taxonomic levels. OTUs are common units of diversity, especially for 16S marker gene datasets. During the 16S bioinformatics pipeline, sequences are clustered according to their similarity to one another, and OTUs are typically clustered on a similarity threshold of 97% (based on majority consensus that exists in the literature). Considerable debate exists whether OTUs can summarise and encapsulate true microbial species phylogeny or ecology, and what exactly constitutes a bacterial species (Jackson et al., 2016).

A range of methods enable collapsing of 16S data to OTUs (Schloss and Handelsman, 2005, Edgar, 2013, Jackson et al., 2016, Rognes et al., 2016) which are often implemented within software wrappers such as QIIME and Mothur (Schloss et al., 2009, Caporaso et al., 2010, Bolyen et al., 2019).A discussion point in choice of method is whether experimental sequences should be clustered against a reference database of sequences (closed reference clustering) (Liu et al., 2008, Navas-Molina et al., 2013), or simply clustered based on the experimental data producing what are termed *de novo* OTUs based on the choice of a threshold in terms of sequence similarity (Schloss and Handelsman, 2005, Navas-Molina et al., 2013). Closed reference clustering has issues in that users are limited to the sequences within the database. *De novo* clustering does not have this limitation and includes all experimental reads in resultant OTUs, which may better represent rare and novel taxa (Navas-Molina et al., 2013). A third method is termed open-reference clustering which aims to utilise the best of both worlds, by first clustering experimental sequences against a reference, followed by *de novo* clustering of discarded sequences (Navas-Molina et al., 2013).

Based on the reference or *de novo* approach selected, different algorithms have their own analytical procedures (Schloss and Handelsman, 2005, Caporaso et al., 2010, Edgar, 2013, Rognes et al., 2016). Linkage based methods can be used to calculate pairwise distances between all sequences allowing hierarchical clustering to OTUs (Schloss and Handelsman, 2005). Greedy algorithms are also usable which aim to reduce computation time via heuristic approaches to finding optimal groups without calculating all possible distances (Edgar, 2013, Rognes et al., 2016). Furthermore, there have been a number of methods proposed to summarise 16S data without using a predetermined global similarity threshold. These include simply using de-replicated sequences (reads collapsed by 100% similarity), defining OTUs by inherent separation within the dataset using local rather than global cut-offs (Mahé et al., 2014), and splitting reads into groups based on sequence entropy at each position in aligned reads (Eren et al., 2015).

1.5 Amplicon Sequence Variants (ASVs)

Whilst thresholding-based methods (OTUs) is widely established, there is still a lot of debate as to what constitutes as a perfect threshold to define species boundaries. As an alternative, amplicon sequence variant (ASVs) have been popularised recently, which relaxes the assumption and gives us the distribution of species by fitting a probability distribution function and recover variants based on underlying noise models (Callahan et al., 2017,

Needham et al., 2018). Modern methods control errors sufficiently such that ASVs can be resolved exactly down to the level of single-nucleotide differences over the sequenced gene region (Eren et al., 2015, Tikhonov et al., 2015, Callahan et al., 2016, Edgar, 2016, Amir et al., 2017). This circumvents the use of arbitrary dissimilarity thresholds typically used for the aforementioned OTUs (Callahan et al., 2017). Of note, each of the methods have denoted their own naming convention (ASV (Callahan et al., 2017, Needham et al., 2018); ESV (Callahan et al., 2017); sub-OTUs (sOTUs) (Amir et al., 2017); zero Radius OTUs (zOTUs) (Edgar, 2016)), however they are all essentially referring to the same measurement: amplicons from NGS following a discretisation based on an underlying analytical approach. ASV methods infer the biological sequences in the sample prior to the introduction of amplification and sequencing errors, and distinguish sequence variants differing by as little as one nucleotide (Callahan et al., 2017). This is achieved by generating an error model tailored to an individual sequencing run and employing algorithms that use the model to distinguish between true biological sequences and those generated by error (Callahan et al., 2017). Because ASVs represent actual biological material, in theory they can be directly compared between different studies, and this is one of the proposed major strengths of the methodology. ASVs have also demonstrated sensitivity and specificity as good or better than OTU methods and better discriminate ecological patterns (Eren et al., 2015, Callahan et al., 2016, Needham et al., 2018). Popular methods for resolving ASVs include DADA2 (Callahan et al., 2016), Deblur (Amir et al., 2017) and MED (Eren et al., 2015).

There exists an ongoing discussion as to which of ASVs and OTUs is more accurate and should be used for microbiological community surveys (Callahan et al., 2017). Arguments in favour of ASVs focus on the utility of finer sequence resolution and the advantage of being able to easily compare sequences between different studies. In contrast, others have argued that ASVs are not ideal as references for linking microbial identity with functions (Dueholm et al., 2019) as: i) ASVs do not contain sufficient evolutionary information to confidently resolve their phylogeny (Yarza et al., 2014), which makes it impossible to report and infer how microbial traits are conserved at different phylogenetic scales (Dueholm et al., 2019); ii) comparison of ASVs between different projects is only possible if the ASVs are produced and processed the same way (Dueholm et al., 2019). Nonetheless, they are gaining increasing importance as the *de facto* standard for mitigating any biases associated with having similar thresholds for all species.

1.6 Data analysis

Microbial communities can be understood at both fine scales (species that are differentially abundant) and at coarse scales (1D-realisation of the community structure). For the latter analyses, diversity measures are typically applied and they fall within three categories (Ley et al., 2008b, Magurran, 2013): i) "alpha diversity" (the number of taxa or lineages within a specific sample) and "beta diversity" (how taxa or lineages are shared among samples); ii) analysis can be either "qualitative" (presence/absence) or "quantitative" (taking into account relative abundance); iii) analysis can be either "phylogenetic" (making use of a phylogenetic tree to relate the sequences) or "taxon-based" (treating all taxa at a given rank as phylogenetically equivalent) (Ley et al., 2008b, Magurran, 2013, Kuczynski et al., 2010). With phylogenetic methods, differences in abundances that involve closely related species are given lower weights, on the assumption that closely related species have similar genetic capabilities. One example is UniFrac (unique fraction), which has been reported to correlate well with the biological properties of samples (Navas-Molina et al., 2013) and measures the amount of "unique evolution" of a community in comparison to others (Lozupone and Knight, 2005, Lozupone et al., 2006). In general, taxon-based and phylogenetic methods provide alternative views, but equally insightful and are used in tandem. Taxon-based analyses are useful for analysing: i) how many different "species" (or taxonomic units) are most likely found within a sample (Schloss and Handelsman, 2005); ii) for comparing which OTUs are shared among subsets of samples (Schloss and Handelsman, 2006); iii) for building networks that relate species and samples to one another (Ley et al., 2008a). Phylogenetic methods are predominantly used for associating drivers of community assembly, mainly due to taxon-based methods not being free of assumptions about phylogeny (that all taxa are equally related to one another) (Hamady and Knight, 2009). This assumption is problematic because it ignores the correlation between evolutionary relatedness and ecological similarity (Stackebrandt and Goebel, 1994). Although errors in phylogenetic reconstruction can affect the clustering results, regardless of reconstruction method, a tree will provide a more accurate model of evolution than the taxon-based method (Lozupone et al., 2007).

1.6 Bacterial transcriptomics

As with 16S rRNA microbial community surveys, NGS has also left its mark on bacterial transcriptomics, specifically with the method RNA-Seq (Sorek and Cossart, 2010). Before the use of omics-based approaches, in the 20th century, study of transcription was largely based on northern blotting, RT-PCR and qPCR where these methods investigated

transcription at the individual gene level. At the turn of the century microarrays provided a means to ascertain transcription profiles at a genome wide level. RNA-Seq methodologies have provided further advances allowing researchers to capture data that was not previously possible. Alternative transcripts within operons have challenged the classical operon definition, and many small RNAs involved in regulation of transcription, translation and pathogenesis have been discovered (Güell et al., 2011, Creecy and Conway, 2015). In 2008, RNA-Seq was developed which performed deep sequencing of cDNA generated from RNA preparations (Wilhelm et al., 2008). This technology has overcome some of the drawbacks of tiling arrays: providing single-base resolution, a better signal-to-noise ratio owing to a reduced background and a higher dynamic range (Vivancos et al., 2010).

1.7 Bacterial transcriptomics data analysis

The methodology of RNA-Seq has many applications and every experimental scenario could potentially have different optimal methods for gene or transcript quantification, normalization, and ultimately differential expression analysis (Conesa et al., 2016). The general steps for typical RNA-Seq analysis involve quality control, read alignment with or without a reference genome, obtaining metrics for gene or transcript expression, and approaches for detecting differential gene expression (Conesa et al., 2016, Sorek and Cossart, 2010). For transcript alignment, if a genome sequence is available for the studied organism, it should be possible to identify transcripts by mapping RNA-Seq reads onto the genome. By contrast, for organisms without sequenced genomes, quantification would be achieved by first assembling reads *de novo* into contigs and then mapping these contigs onto the transcriptome (Conesa et al., 2016). The read coverage can then be used to quantify transcript expression levels.

Estimating gene and transcript expression is principally based on the number of reads that map to each transcript sequence. The simplest method to quantification is to combine raw counts of mapped reads. A typical program that does this is HTSeq-count (Anders et al., 2015). This methodology is typically referred to as a gene-level (as opposed to transcript-level) quantification methodology which utilises a gene transfer format (GTF) file that contains all of the coordinates of genes and other features of interest (Conesa et al., 2016). Other methods include StringTie, an assembler of RNA-Seq alignments into potential transcripts, which uses a novel network flow algorithm as well as an optional *de novo* assembly step to assemble and quantitate full-length transcripts representing multiple splice variants for each gene locus (Pertea et al., 2015). As a sanity check, bedtools, a powerful

toolset for genome arithmetic can be used with the multicov command which reports the count of alignments from multiple position-sorted and indexed BAM files that overlap intervals in a BED file (Quinlan, 2014).

It is important to note that raw counts alone are not adequate to compare expression levels between samples as these values are affected by factors such as transcript length, total number of reads, sequencing biases, and sequencing depth (Conesa et al., 2016, Soneson and Delorenzi, 2013). Thus, normalisation methods such as RPKM (reads per kilobase of exon model per million reads) and FPKM (fragments per kilobase of exon model per million mapped reads) which are within-sample normalisation methods (the latter for transcripts) are utilised (Mortazavi et al., 2008). Different methodologies exist to estimate transcript-level expression in order to circumvent the problem of related transcripts sharing similar amplicons. A popular method is Cufflinks which estimates transcript expression from a mapping to the genome obtained from mappers such as TopHat using an expectation-maximization approach that estimates transcript abundances (Trapnell et al., 2010). Differential gene expression analysis allows the comparison of gene expression values among samples. Normalisation methods described above such as RPKM, FPKM and TPM remove the biases associated with different sampling depths. These methods rely on total or effective counts and usually perform inefficiently when samples have heterogeneous transcript distributions (Bullard et al., 2010). If only normalisation methods were used, the sum of the normalized counts across all genes would not necessarily be equal between samples, thus the aim is instead to make the normalised counts for non-differentially expressed genes similar between the samples (Soneson and Delorenzi, 2013). Thus, methods have been developed to consider these factors and the most well-known is DESeq2 (Anders and Huber, 2010).

1.8 Aims and objectives

The aim of this study is to explore, utilise, and compare different 16S rRNA pipelines as well as transcriptomics pipelines by focusing on understanding the pathogenesis of the foodborne pathogen *C. jejuni*. These are delineated in terms of objectives as follows: -

Objectives:

- Investigate the impact of different concentrations of carvacrol on the chicken cecal microbiome and compare different metagenomics bioinformatics pipelines. The comparison will be between DADA2 (ASV) and VSEARCH (OTU) methodologies.

- Investigate the impact of oxidative stress on the *C. jejuni* transcriptome and compare different RNA-Seq bioinformatics pipelines. The comparison will be between StringTie and bedtools, both used at the read count stage in this study.

2. Methods

All relevant pipelines were established by Dr Umer Zeeshan Ijaz (http://userweb.eng.gla.ac.uk/umer.ijaz/) at the University of Glasgow and training was provided accordingly (http://www.tinyurl.com/JCBioinformatics3). FASTQ files used within this project were obtained by Ozan Gundogdu previously at the LSHTM using Illumina guidelines from existing projects.

2.1 File organisation

FASTQ files were organised to allow functioning and flexibility in downstream pipelines. To achieve this, the following set up was created within a Linux environment: -

```
mkdir Carvacrol2020 #make the project folder
cd Carvacrol2020 #move into the project folder
cp /home/ozan/Documents*.fastq . #copy all FASTQ files to the Carvacrol folder:-
for i in $(awk -F"_" '(Lindqvist et al.)' <(ls *.fastq) | sort | uniq); do mkdir $i; mkdir $i/Raw; mv $i*.fastq $i/Raw/.; done #Place all FASTQ files within a folder (filename designates folder name), and within it, a 'Raw' folder which contains the FASTQ files are placed.
```

2.2. Qiime2 bioinformatics pipeline with DADA2

The DADA2 v1.14 (Callahan et al., 2016) software was used to produce the abundance table by constructing Amplicon Sequence Variants (ASVs), a higher-resolution analogue of the traditional OTU table, which records the number of times each exact amplicon sequence variant was observed in each sample. The DADA2 bioinformatics section of the pipeline is based on the version developed by Dr Ijaz (https://github.com/umerijaz/tutorials/blob/master/qiime2_tutorial.md). A complete methodology is provided in Appendix I. Briefly, fictitious barcodes were generated and saved for each sample. Forward and reverse reads were assembled together and loaded into Qiime2 (Bolyen et al., 2019). Samples were demultiplexed and exported for viewing in Qiime2 viewer (https://view.qiime2.org/). DADA2 was performed with truncating forward reads above 280 and reverse reads above 220 after visually inspecting where the qualities

dropped significantly on the length of the reads, and for sufficient nucleotide overlap between the truncated amplicons (https://benjjneb.github.io/dada2/tutorial.html). A phylogenetic tree within Qiime2 was created with align-to-tree-mafft-fasttree using MAFFT (v7.310) (Katoh and Standley, 2013) and FastTree (v2.1.10) (Price et al., 2010). The phylogenetic tree was exported to NEWICK format. Taxonomy was created to represent the ASVs against the SILVA SSU Reference NR database release v132 database (Quast et al., 2013). ASV table (table.qza) was exported to a BIOM format. ASV sequences (rep-seqs.qza) were exported to FASTA format. The taxonomy was exported to a TSV file. The ASV abundance table (feature_table.txt) was merged with the taxonomy (taxonomy.tsv) to create a final BIOM file for compatibility to R and phyloseq (McMurdie and Holmes, 2013). Finally, the ASV table was generated by matching the original barcoded reads against clean ASVs.

2.3. Qiime2 bioinformatics pipeline with VSEARCH

The VSEARCH v2.3.4 pipeline was used to produce the abundance table by constructing OTUs, a representation of species, as described in http://github.com/torognes/vsearch/wiki/VSEARCH-pipeline). The VSEARCH section of the pipeline is based on the publication by Ijaz et al. with modifications (Ijaz et al., 2018). A complete methodology is provided in Appendix II. Briefly, paired-end reads were pre-processed based on guidelines from recent publications (Schirmer et al., 2015, D'Amore et al., 2016). Reads were trimmed (average Phred quality score of 20 using a sliding window approach) and reads were filtered using Sickle v1.33 (Joshi and Fass, 2011). BayesHammer (Nikolenko et al., 2013) was used from Spades v3.1.1 assembler to perform error-correction on paired-end reads. Then, PANDAseq (v2.11) (Masella et al., 2012) was used to assemble the forward and reverse reads into a single sequence spanning the entire V4 region with a minimum overlap of 10 bp. Reads were then pooled, dereplicated, sorted in order of decreasing abundance and singletons were discarded. Following this, the reads were clustered based on 97% similarity and removal of clusters was performed with chimeric models built from more abundant reads (--uchime_denovo option in vsearch). In addition, a reference-based chimera filtering step (--uchime_ref option in vsearch) was performed using a gold database (https://www.mothur.org/w/images/f/f1/Silva.gold.bacteria.zip). Finally, the OTU table was generated by matching the original barcoded reads against clean OTUs. After creating a tab-delimited version of the OTU table (otus.fa; representing all of the OTU sequences) an otu_table.txt (representing the OTU abundance table) was created. The sequence file was then loaded into Qiime2 (Bolyen et al., 2019), followed by assigning a

taxonomy to represent the OTUs against the SILVA SSU Ref NR database release v132 database (Quast et al., 2013). The taxonomy was exported to a TSV format. A phylogenetic tree within Qiime2 was created with align-to-tree-mafft-fasttree (Katoh and Standley, 2013) using MAFFT (v7.310) (Katoh and Standley, 2013) and FastTree (v2.1.10) (Price et al., 2010). The phylogeny was exported to NEWICK format. A BIOM file was created within Qiime2 merging the otu_table.txt (abundance table) and the taxonomy.tsv (taxonomy) for compatibility to R and phyloseq (McMurdie and Holmes, 2013).

2.4. Statistical analysis of 16S microbiome data

Statistical analyses was performed as described by Ijaz et al., 2018 with R using output files generated from the DADA2 (section 2.2) or VSEARCH (section 2.3) bioinformatic pipelines and the associated metadata (Table 1) predominantly employing the vegan package for diversity measures (Oksanen et al., 2015). We have employed both diversity within samples (alpha diversity) and between samples (beta diversity) as is the norm. For alpha diversity measures, *Richness* is an estimated measure of species/features per rarefied sample (rarefied to minimal library size i.e. read numbers); *Shannon* entropy is a commonly used index to measure the balance of a community within a sample (higher the index, the more balanced the community is); *Pielou's* index represents the evenness of a community; *Simpson* measures evenness of the community from 0 to 1; and *Fisher alpha* is an alternative diversity index. All of these diversity indices are based on different analytical procedures and highlight different aspects of diversities, whether focussing on rare species, or on predominant species.

Table 1. Metadata table for 16S carvacrol study. Table includes dummified (original categorical labelling converted to presence and absence) data used for subset regression analysis.

	Day	Sample_Type	CarvacrolCon	Status_C	Status_T1	Status_T2	Status_T3	Day_10	Day_21	Day_35
Day21C	21	Control	0	1	0	0	0	0	1	0
Day21T1	21	T1	120	0	1	0	0	0	1	0
Day21T2	21	T2	200	0	0	1	0	0	1	0
Day21T3	21	T3	300	0	0	0	1	0	1	0
Day35C	35	Control	0	1	0	0	0	0	0	1
Day35T1	35	T1	120	0	1	0	0	0	0	1
Day35T2	35	T2	200	0	0	1	0	0	0	1
Day35T3	35	T3	300	0	0	0	1	0	0	1
Day10T1	10	T1	120	0	1	0	0	1	0	0
Day10T2	10	T2	200	0	0	1	0	1	0	0
Day10T3	10	T3	300	0	0	0	1	1	0	0
Day10C	10	Control	0	1	0	0	0	1	0	0

For beta diversity, three alternative distance matrices were utilised: i) *Bray-Curtis* which only focusses on OTUs/ASVs abundances as a dissimilarity measure; ii) unweighted *UniFrac* which is a phylogenetic distance metric calculating the distances between samples by taking the proportion of the sum of unshared branch lengths in the sum of all of the branch lengths of the phylogenetic tree for the OTUs/ASVs observed between two samples (without considering their abundances); and iii) weighted *UniFrac* which is a phylogenetic distance metric combining phylogenetic distance and relative abundances, thus emphasising dominant OTUs/ASVs or taxa. Unifrac distances were calculated using the phyloseq package.

Next, Local Contribution to Beta Diversity (LCBD) analysis (Legendre and De Caceres, 2013) was performed using LCBD.comp() from adespatial package (Dray et al., 2018). This was performed to identify outliers in beta diversity space. The *Hellinger distance* (abundances), unweighted *Unifrac* (phylogenetic distance) and weighted *Unifrac* (phylogenetic distance weighted by abundance) dissimilarities were used. LCBD provides the sample-wise local contributions to beta diversity that could be derived as a proportion of the total beta diversity. In the context of this study, it provides a mean to show how markedly different the microbial community structure of a single sample is from the average (with extreme LCBD values from the average representing outliers), and also provides a method to demonstrate when the community structure has stabilized in a temporal setting.

To investigate environment filtering (phylogenetic overdispersion versus clustering), phylogenetic distances within each sample were further characterised by calculating the nearest taxa index (NTI) and net relatedness index (NRI). Here, the aim was to determine whether the community structure was stochastic (overdispersion in phylogenetic tree and driven by competition), or deterministic (phylogenetic clustering and driven by strong environmental pressure). The NTI was calculated using mntd() and ses.mntd(), and the mean phylogenetic diversity (MPD) and NRI were calculated using mpd() and ses.mpd() functions from the picante package (Kembel et al., 2010). NTI and NRI represent the negatives of the output from ses.mntd() and ses.mpd(), respectively. Additionally, they quantify the number of standard deviations that separate the observed values from the mean of the null distribution (999 randomisation using null.model-'richness' in the ses.mntd() and ses.mpd() functions and only considering taxa as either present or absent regardless of their relative abundance). As opposed to authors' recent work (Ijaz et al., 2018), OTUs collated at genera were used for the calculations.

Next we wanted to see what is the minimal subset of species that can explain roughly the same beta diversity as compare to utilising all of the OTUs/ASVs in the sample space. For

this purpose, the “BVSTEP” routine (Clarke and Ainsworth, 1993) was used to search for the highest correlation, in a Mantel test, by imploding the abundance table at genera level to absolute minimal set of genera that preserve the beta diversity between samples. To run this algorithm, bvStep() (from the sinkr package) (Taylor, 2014) was used as considered in author’s recent publication (Ijaz et al., 2018).

To allow identification of genera that are significantly different between categories, DESeqDataSetFromMatrix() function from DESeq2 (Love et al., 2014) package was used with the adjusted p-value significance cut-off of 0.05 and log2 fold change cut-off of 2. This function uses negative binomial GLM to obtain maximum likelihood estimates for OTUs/ASVs log fold change between two conditions. Then, Bayesian shrinkage was applied to obtain shrunken log fold changes subsequently employing the Wald test for obtaining adjusted p-values for multiple comparisons. DESeq2 identified changes on a local scale (in conjunction with beta diversity analysis) to identify genera that are causing the shift in microbial communities. Also, to see what the dominant species distributions are, we have given a visual representation of the proportions of the top-25 abundant species at a particular level along with binning everything else as “others” category in taxa bars.

Next, we wanted to get the directionality out by focussing on microbiome characteristics and the sources of variations that were captured for each sample. For this purpose, subset regression of different microbiome metrics were performed against a comprehensive set of explanatory variables ("CarvacrolConc", "Status_C", "Status_T1", "Status_T2", "Status_T3", "Day_10", "Day_21", "Day_35"), by selecting the best model (a subset of these variables) according to some statistical criteria (fit of regression etc), with recommendations given in (Kassambara, 2018) and code available at http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/155-best-subsets-regression-essentials-in-r/. The R function regubsets() from leaps (Lumley and Miller, 2009) package was used to identify different best models of different sizes, by specifying the option nvmax, set to the maximum number of predictors to incorporate the model. Having obtained the best possible subsets, the k-fold cross-validation consisting of first dividing the data into k subsets. Each subset (10%) served successively as test data set and the remaining subset (90%) as training data. The average cross-validation error is then computed as the model prediction error. This was all done using a custom function utilising R’s train() function from the caret package (Kuhn, 2008). Finally R’s tab_model() function from sjPlot package (Lüdecke, 2018) was used to obtain the statistics for each model.

To find the core microbiome, we have used R’s microbiome package (Lahti et al., 2017) and the recommendations given in (Shetty et al., 2017) to find OTUs/ASVs that are consistently

prevalent and have a reasonable detection limit in terms of abundances to ascertain that they are core. To investigate association with environmental ontology terms, seqenv (Sinclair et al., 2016) was utilised. Seqenv performs similarity searches of short sequences against the "nt" nucleotide database provided by NCBI and from every match extract (if available) the textual metadata field. This was performed to incorporate historical data about where these sequences were previously observed in (habitat, environmental microbiomes etc). Alpha and beta diversity metrics were also applied to the results. After collecting all of the isolation sources (habitat) from all the search results, we ran a text mining algorithm to identify and parse words that are associated with the Environmental Ontology (EnvO) controlled vocabulary. This, in turn, enabled us to determine both in which environments individual sequences or taxa have previously been observed and, by weighted summation of those results, to summarize complete samples.

PICRUSt2 (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States) is a software for predicting functional abundances based only on marker gene sequences (Douglas et al., 2020). Although predicting functional and metabolic potential of microbial communities based on reference based genomic potential, is still a prediction, ~20K genome representation in PICRUSt2 databases makes it a very strong contender to profile the functional dynamics. Alpha and beta diversity metrics were also applied to the results. Here, "Function" usually refers to gene families such as KEGG orthologs and Enzyme Classification numbers, but predictions can be made for any arbitrary trait.

For comparisons between two 16S microbiome analysis methodologies, the abundance tables from each pipeline were compared using Procrustes, which was used to demonstrate similarity between different configurations (Peres-Neto and Jackson, 2001).

In the majority of the figures displaying boxplots, pair-wise ANOVA was performed taking two categories at a time, and where significant ($p \leq 0.05$), the categories were joined together by a line and the significance was plotted on top (*: $0.01 \leq p < 0.05$; **: $0.05 \leq p < 0.001$; ***: $p \leq 0.001$).

2.5. RNA-Seq bioinformatics pipeline

For exploring *Campylobacter* pathogenesis, we performed laboratory scale experiments on *C. jejuni* isolates with bioinformatic steps given as follows and details of these experiments following this section. Once the sequences were obtained, paired-end reads were trimmed and filtered using Sickle v1.200 (Joshi and Fass, 2011) by using a sliding window approach and trimming the reads where the average base quality drops below 20. The NCTC11168

reference genome sequence (FASTA format) and annotation (GFF format) were downloaded (https://www.ebi.ac.uk/ena/data/view/Taxon:192222). Bowtie2 (Langmead and Salzberg, 2012) was used to map the reads against the reference sequence. This generated the mapping in SAM format which were later converted to BAM format using Samtools (Li et al., 2009) and were index sorted. gffread from cufflinks suite (Trapnell et al., 2010) was used to convert annotations from GFF to GTF format. These were then used in bedtools (with multicov-bams option) (Quinlan and Hall, 2010) or StringTie (Pertea et al., 2015) (with –e –B –o option) with the mapped reads to generate transcript counts per samples. A shell utility created by Dr Ijaz (http://userweb.eng.gla.ac.uk/umer.ijaz/bioinformatics/GENERATEtable.sh) was then used to collate all these transcripts into a transcripts abundance table for the 11168H strain. Statistical analysis on these abundance tables were performed in R.

2.6. Experimental details and statistical analysis of RNA-Seq transcriptomics data

Hydrogen peroxide (H_2O_2) are a type of Reactive Oxygen Species (ROS) where accumulation can lead to damage of nucleic acids, proteins and membrane structures (D'Autreaux and Toledano, 2007, Atack and Kelly, 2009). Statistical analysis of RNA-Seq data was performed for two separate comparisons where physiological concentrations of H_2O_2 were utilised: -

- 5 mins 5 mM H_2O_2 stress vs. Control (no stress)
- 15 mins 5 mM H_2O_2 stress vs. Control (no stress)

where RNA was initially isolated from *C. jejuni* grown to late-log phase (16 hr).

Ordination of abundance tables in reduced space (beta diversity) was performed using Principal Coordinate Analysis (PCoA) plots of transcripts *Bray-Curtis* distance in Vegan's cmdscale() function (Oksanen et al., 2015). This was performed to find if there were any clustering on categorical basis (clustering based on samples originating from the same categories).

The DESeqDataSetFromMatrix() function from DESeq2 (Love et al., 2014) package was used with the adjusted p-value significance cut-off of 0.05 and log fold change cut-off of 2. The abundances for significant transcriptomes were then visualised using RPKM representations. For comparisons between two transcriptomics analysis methodologies, the

abundance tables from each pipeline were compared using Procrustes, which was used to demonstrate similarity between different configurations (Peres-Neto and Jackson, 2001).

3. Results

3.1 Microbia1 community survey using DADA2 pipeline

Whilst we obtained the tables for both OTUs and ASVs separately, we followed the ASVs route (with details on OTUs given in the appendices, and not followed further based on Procrustes and other analyses, and ASVs based methodologies explained henceforth). The ASV abundance table used for statistical pipelines within R contained a total of 3485 ASVs for n=12 samples, with summary read statistics for samples as follows: [Min: 12125 1st Quantile: 20138 Median: 24980, Mean: 27692, 3rd Quantile: 36980, Max: 47429]. The breakdown of taxa (top 25 most abundant taxa) between the different grouping categories are represent in Figure 1. This was followed up with taxa differential analysis whereby ASVs that were significantly modified between different groups were identified. As an example, Table 2 displayed the comparison between the groups Control vs. T1.

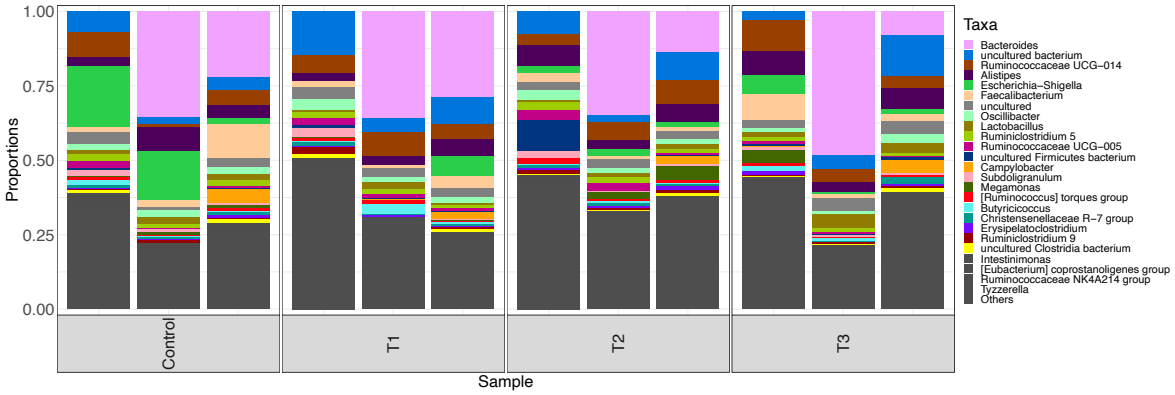


Figure 1. The top 25 most abundant Genera representative of different sample categories. Carvacrol concentration of different groups (Control = 0 mg/ml, T1 = 120 mg/ml, T2 = 200 mg/ml and T3 = 300 mg/ml).

Table 2. Taxa differential of ASVs statistically modified when comparing groups Control vs. T1. These are log 2-fold different and statistically significant.

	baseMean	log2FoldChange	pvalue	padj	Upregulated
Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Ruminococcaceae UCG-005	77.67058989	6.812718644	3.80E-06	0.006497947	T1
Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Rikenellaceae;Alistipes	107.3880923	-7.320853924	2.28E-05	0.012990479	Control
Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Rikenellaceae;Alistipes	110.3142592	-7.359950905	2.01E-05	0.012990479	Control
Bacteria;Bacteroidetes;Bacteroidia;Bacteroidales;Rikenellaceae;Alistipes	95.39154466	-7.148913045	3.19E-05	0.013670854	Control
Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group;uncultured bacterium;uncultured bacterium	99.7618011	7.175734532	7.59E-05	0.023113469	T1
Bacteria;Firmicutes;Erysipelotrichia;Erysipelotrichales;Erysipelotrichaceae;uncultured;Clostridiales bacterium 60-7e	76.16165355	-6.821556394	8.10E-05	0.023113469	Control
Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group;uncultured bacterium;uncultured bacterium	63.98696157	6.529311162	0.000199769	0.048857883	T1

To connect microbiome and extrinsic parameters together, we implode microbiome multivariate datasets to a single dimensional realisation, which is mainly diversity measures

either diversity within a sample (alpha), or between samples (beta). Initially we investigated at the alpha diversity level how diversity was influenced between the different sample categories (Figure 2). Though not significant, it was observed that increasing the concentration of carvacrol (Control = 0 mg/ml, T1 = 120 mg/ml, T2 = 200 mg/ml and T3 = 300 mg/ml) resulted in an increased microbial diversity. Using certain diversity metrics, T2 demonstrated the greatest microbial diversity. These measures are based on the number of samples that we acquired; the p-values may become statistically significant if we increased the sample size.

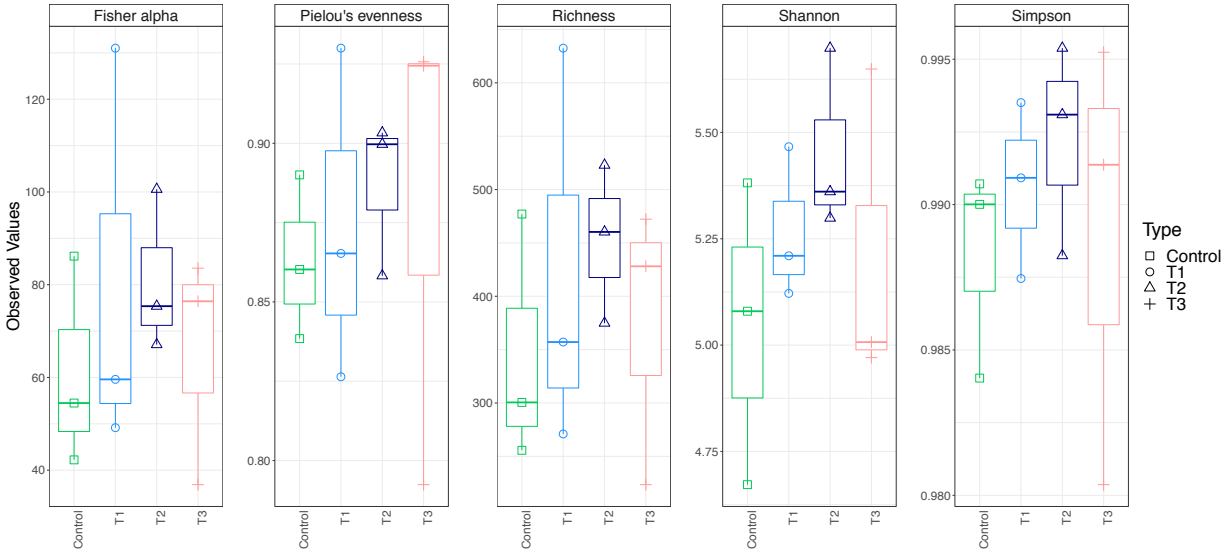


Figure 2. Alpha diversity metrics for Control, T1, T2 and T3 samples using DADA2 pipeline. *Richness* is an estimated number of species/features per sample; *Shannon* entropy measured the balance of a community within a sample; *Pielou's* index represents the evenness of a community; *Simpson* measures evenness of the community from 0 to 1, and *Fisher alpha* is an alternative diversity index. Carvacrol concentration of different groups (Control = 0 mg/ml, T1 = 120 mg/ml, T2 = 200 mg/ml and T3 = 300 mg/ml).

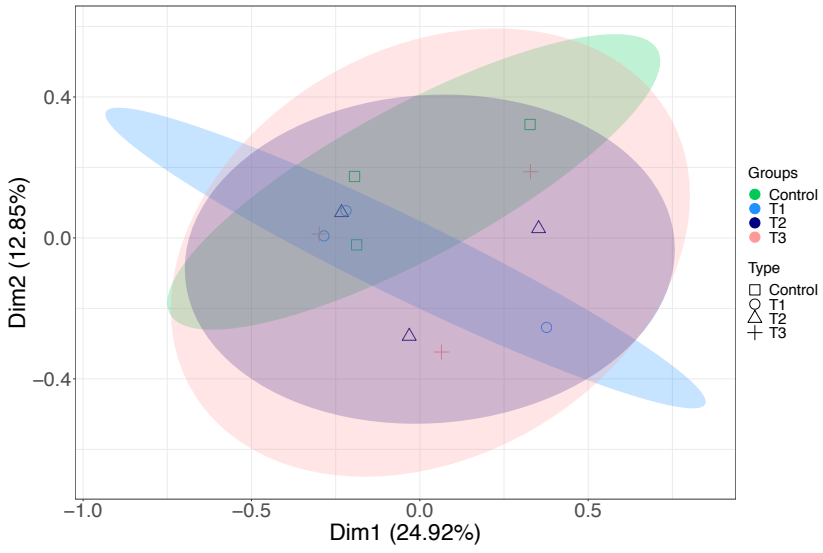
Beta diversity was performed with all the different measures as highlighted in the methods. Beta diversity with *Bray-Curtis* displayed considerable overlap between samples in terms of bacterial numbers. Sample T1 displayed a more stringent dispersion between biological replicates (Figure 3).

In parallel, to see if there is any difference in microbiome community structure between these groups, we performed PERMANOVA (full list is provided in an accompanied file) (Table 3). Here, for *Bray-Curtis* distance, 16% of the variability in community structure is explained by the day groupings.

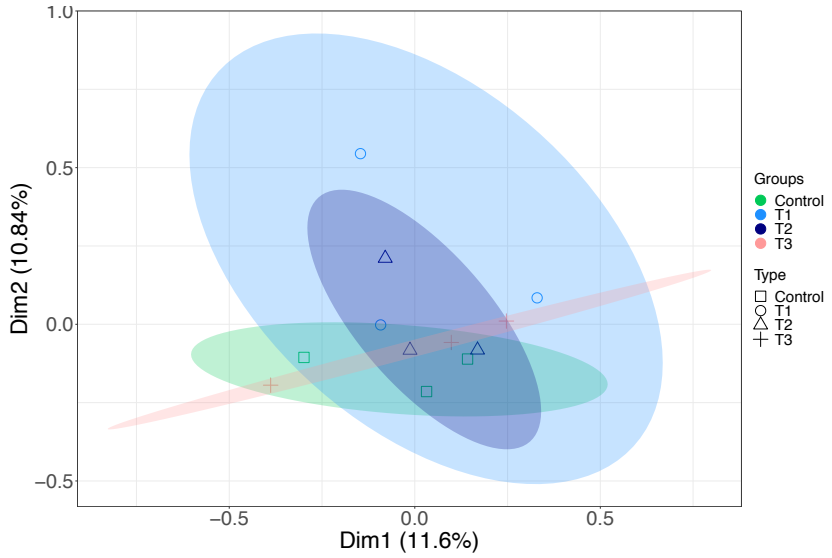
Table 3. To understand the impact of different parameters on microbial community structure, we performed PERMANOVA. Here, using beta diversity (*Bray-Curtis*) distance metrics, 16% (R2 in the table given below) of the microbiome structure is explained.

```
Call:
adonis(formula = as.formula(paste("dist ~", paste(PERMANOVA_variables,
collapse = "+"))), data = meta_table[rownames(otu_table(physeq)),      ])

Permutation: free
Number of permutations: 999

Terms added sequentially (first to last)

            Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
Sample_Type  3    0.7651 0.25503 0.88194 0.22925  0.743
Day          1    0.5481 0.54815 1.89559 0.16424  0.020 *
Residuals    7    2.0242 0.28917         0.60651
Total       11    3.3374                 1.00000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(A)



(B)



(C)

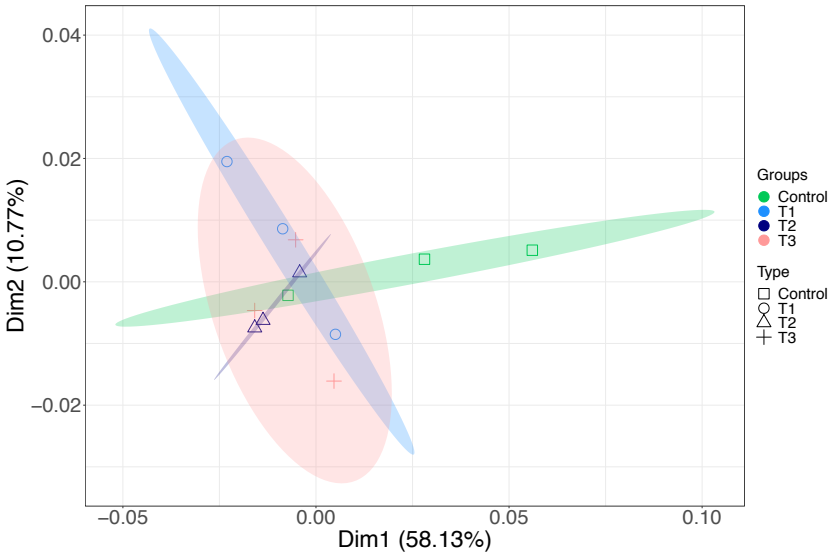


Figure 3. PCoA of beta diversity was measured for all groups using: (A) *Bray-Curtis*; (B) unweighted *UniFrac*; (C) weighted *UniFrac*. Two samples if similar lie very close to each other. The ellipses represent the standard error in terms of grouping variations.

The core microbiome where genera persist in 85% of the samples for different sample groups (Control, T1, T2 and T3) was assessed (Figure 4). In Figure 4, the ASVs are sorted by their abundances with those on the left being low abundant prevalent ASVs, whereas those at the right are highly abundant prevalent ASVs. Prominent genera identified include *Ruminococcaceae UCG-05*, *Ruminiclostridium 5*, *Lactobacillus*, *Escherichia-Shigella*, *Oscilibacter*, *Faecalibacterium*, *Alistipes* and *Ruminococcaceae UCG-014* at varying levels of abundance.

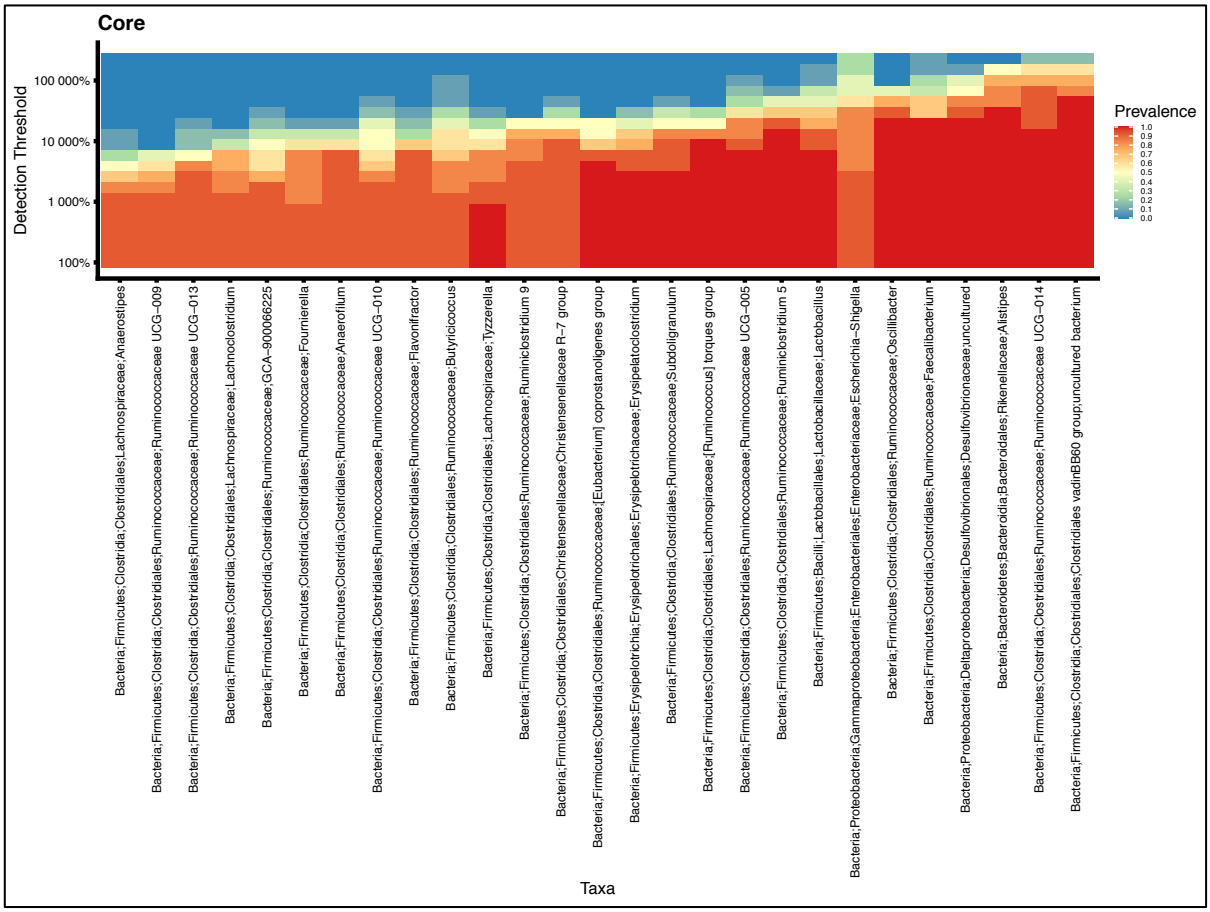


Figure 4. Core microbiome analysis of all the samples. The figure shows at least 85% prevalent ASVs ordered by lower abundance (left) to higher abundance (right). The y-axis (right) represents limit of detection in terms of relative abundance.

Next, we explored ecological drivers of microbial community to determine whether there was any environmental pressure (deterministic influence contingent upon external environment/any other extrinsic parameter). A value above +2 signifies a strong phylogenetic clustering at both global (NRI) and local (NTI) levels, whilst a value below -2 signifies a strong overdispersion in phylogenetic tree. The more clustered the phylogenetic tree is, the more evidence there is of environment/any other parameter playing a role in its assembly. NTI and NRI displayed relatively low values indicating a reduced impact of environmental influence on microbial community structure (Figure 5). Since chicken ceca are already a constrained environment to begin with (as opposed to real environmental datasets), the values <0 (traditionally this implies stochasticity) may not be feasible to ascertain randomness/stochasticity/competitive exclusion principle, and hence the values should be taken relatively with an increasing value implying increasing host environmental pressure. For NTI, though no statistical significance was observed, it is interesting to note the observation that addition of carvacrol did lead to a reduced environmental influence.

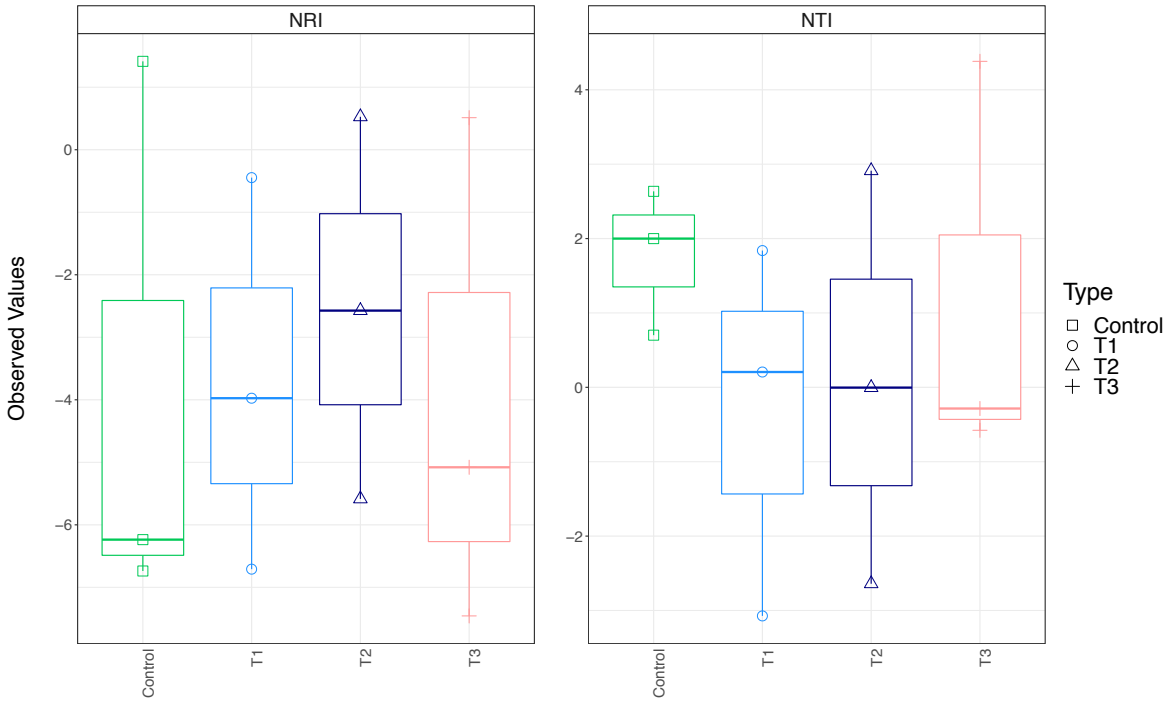
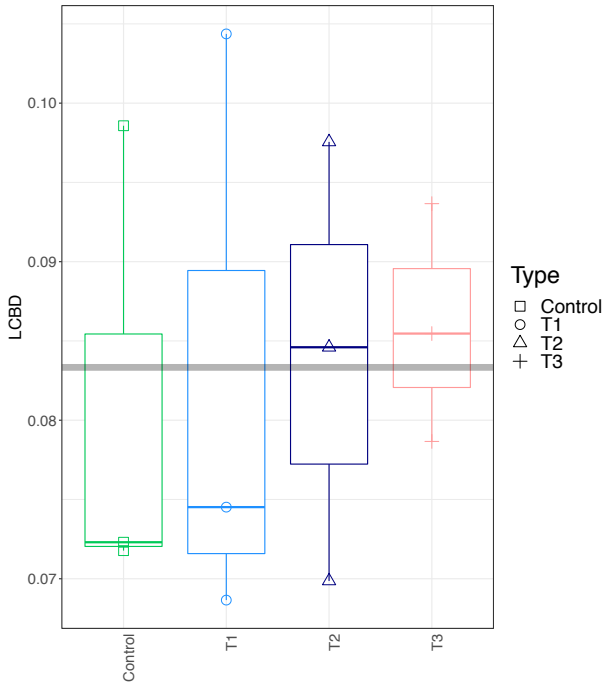


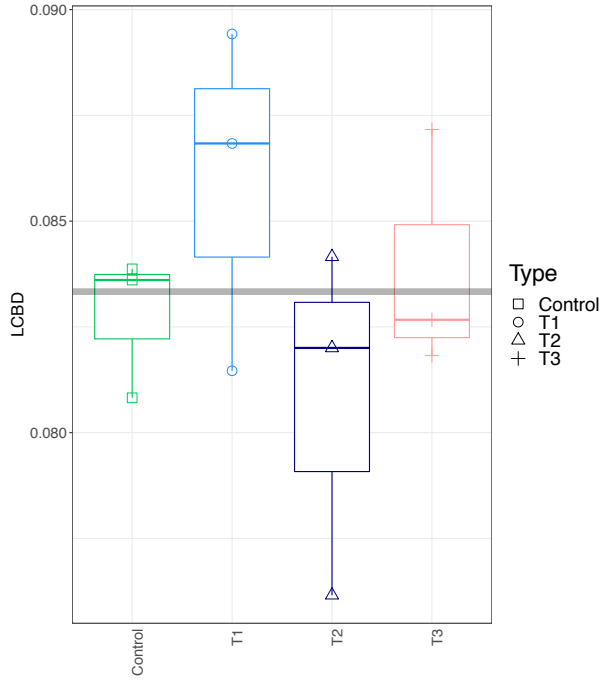
Figure 5. Investigating the environmental pressure on microbial community structure using NRI/NTI.

To further analyse the data, we performed Local Contributions to Beta Diversity (LCBD) which shows how markedly different the microbial community structure of a single sample is from the average (with LCBD values differing from the mean LCBD values representing outliers). LCBD analysis was performed by using: (A) The *Hellinger* distance (abundances); (B) unweighted *UniFrac* (phylogenetic distance); and (C) weighted *UniFrac* (phylogenetic distance weighted by abundance) dissimilarities (Figure 6). Interestingly, when considering abundances alone (*Hellinger* distance), though not significant, we observe a gradual trend away from the average when increasing the level of carvacrol. For unweighted *UniFrac* distance (phylogeny without abundance), results are inconclusive, however when measuring weighted *UniFrac* (phylogeny with abundance), results though not statistically significant, display the Control as being furthest away from the average.

(A)


0.10
0.09
0.08
0.07
LCBD
Control
T1
T2
T3
Type
Control
T1
T2
T3

(B)


0.090
0.085
0.080
LCBD
Control
T1
T2
T3
Type
Control
T1
T2
T3

(C)

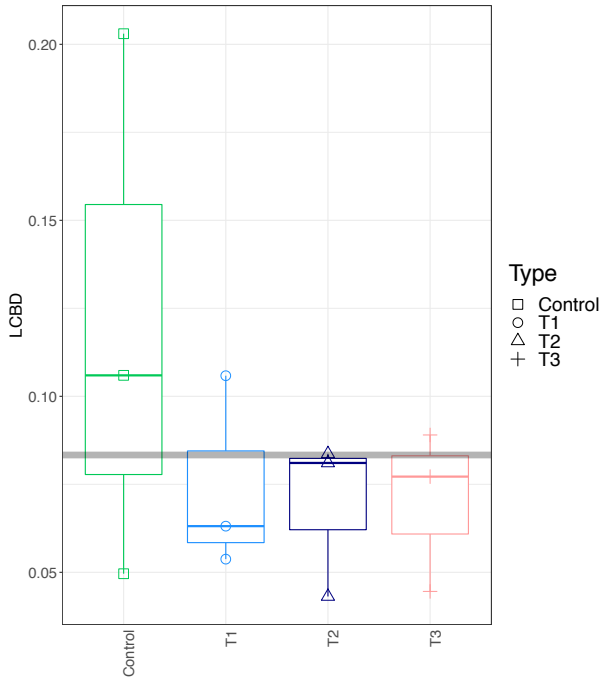


Figure 6. Local contribution to beta diversity (LCBD) calculated by using sample wise proportional diversities: (A) *Hellinger* transform on the microbial counts; (B) unweighted *UniFrac* dissimilarity (phylogenetic distances only); and (C) weighted *UniFrac* dissimilarity (phylogenetic distances weighted with abundance counts), with all values summing up to 1.

Subset analysis was performed identifying a subset of ASVs that explain roughly the same beta diversity between samples as all of the ASVs (Table 4). Essentially, we have obtained a reduced feature set (ASVs) in the sample space that is deriving the change. In the interest of space, only one group comparison is displayed, with remaining results present within the accompanied files. At the same time, after imploding to the subset of genera, PERMANOVA analysis was performed to see if the resulting subset still has the discriminatory power (in terms of grouping). Analysis of all possible comparisons though identified possible ASVs that may differentiate between the group comparisons, were not statistically significant.

Table 4. Subset analysis showing top 3 subsets of ASVs along with the correlation of the beta diversity distances between these subsets and full ASV table. The last column shows PERMANOVA statistics for these subsets highlighting their discriminatory power. R^2 is the percentage variability of these subsets in terms of groups. In the interest of space, only one group comparison is displayed, with remaining results within the accompanied files.

Group Comparison	Subset No	Subset	Correlation of Subset with Full Table (R)	PERMANOVA Subsets (Groups)
Control, T1	S1	Ruminococcaceae UCG-005 + Ihubacter massiliensis + Faecalibacterium + Clostridiales vadinBB60 group	0.04888	R^2 = 0.7139 (p > 0.05)
	S2	Ruminococcaceae UCG-005 + Ihubacter massiliensis + Faecalibacterium	0.06937	R^2 = 0.6639 (p > 0.05)
	S3	Ruminococcaceae UCG-005 + Ihubacter massiliensis	0.05506	R^2 = 0.7000 (p > 0.05)

To assess the impact of extrinsic parameters on microbial community structure, whilst PERMANOVA analysis show the extent of influence on the microbiome structure in terms of variability, to obtain the directions to whether an increase or decrease in these parameters cause an increase or decrease in the properties of microbiome metrics, subset regressions on one dimensionality realisation of microbiome (alpha diversity – *Richness*, *Shannon, Pielou's* index*, Simpson, Fisher alpha*; LCBD beta diversity – *Bray-Curtis*, unweighted *UniFrac*, weighted *UniFrac*) were performed. Subset regression against different sources of variation ("CarvacrolConc", "Status_C", "Status_T1", "Status_T2", "Status_T3", "Day_10", "Day_21", "Day_35") were performed by testing all the combination of all these variations and then selecting the best model according to some statistical criteria (adjusted R^2 etc) (Figure 7). These subset regressions permuted through all possible subsets of explanatory variables (extrinsic parameters considered in this study) by ranking them in terms of quantitative fit after performing cross-validation. Note that red and blue backgrounds represent whether the predictors have a positive or a negative influence, respectively in the regression model. In addition, all categorical variables highlighted yellow were "dummified" (a standard procedure) to represent as present/absent as a tag and were used in the regression model to see whether their inclusion/exclusion has an effect on the final model. Day 21 displays a clear negative effect on microbial diversity within the chicken cecum, whilst day 35 displays a positive impact on microbial diversity. Interestingly, increasing the concentration of carvacrol (ControlConc) led to a shift of microbial diversity, away from the

norm. Day 10 also has a positive impact on influencing the microbial community structure away from the average.

	Richness	FisherAlpha	Simpson	Pielou	Shannon Entropy		LCBD (Bray-Curtis Distance)	LCBD (Unweighted UniFrac)	LCBD (Weighted UniFrac)
ControlConc									+*
Status_C									
Status_T1									
Status_T2					+*				
Status_T3									
Day_10							+***	+*	+*
Day_21			-***	-*					
Day_35	+*	+**			+***				

Figure 7. Subset regression where red and blue represent the significant positive and negative beta coefficients that were consistently selected in different regression models. The categorical variables are represented with a yellow highlight (coded as 1 (present) or 0 (absent)) and if selected is interpreted as the samples belonging to those categories having positive/negative influence on the respective microbiome metrics.

Seqenv was performed to investigate association with environmental ontology terms (Figure 8). A single significant environmental ontology was identified; travertine, a calcareous rock deposited from mineral springs.

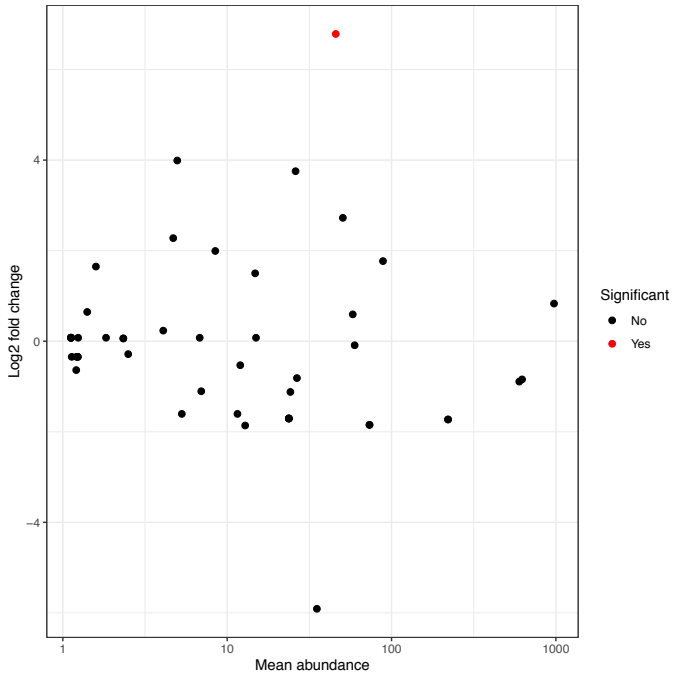


Figure 8. Seqenv identified certain EnvO ontological terms associated with the sequences and in general with the sampling space. On performing differential analysis, weighted with abundances, the above figure is obtained with red highlighting any significant changes.

PICRUSt2 (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States) was performed to predict functional abundances. Using the search results from Table 5, genes within specific KEGG pathways were identified. These were K16044 (iolW; scyllo-inositol 2-dehydrogenase (NADP+) [EC:1.1.1.371]), K06348 (KapD; sporulation inhibitor), K14259 (KdxD, 2-dehydro-3-deoxy-D-arabinonate dehydratase [EC:4.2.1.141]), K19449 (sinR; XRE family transcriptional regulator, master regulator for biofilm formation) and K13533 (two-component system, sporulation sensor kinase E [EC:2.7.13.3]). As an example, K16044 is part of the inositol phosphate metabolism and the specific enzyme can be observed within the pathway (Figure 9).

Table 5. PICRUSt2 search results in relation to KEGG pathways.

	baseMean	log2FoldChar	pvalue	padj	Upregulated
K16044	26.8690609	2.52928993	3.36E-09	2.58E-05	T3
K06348	223.7907	2.07878817	5.06E-07	0.00194132	T3
K14259	40.9530141	3.29154038	8.72E-06	0.02229411	T2
K19449	128.668425	2.93952365	1.99E-05	0.03814717	T3
K13533	105.951965	3.5023712	3.07E-05	0.0471199	T2

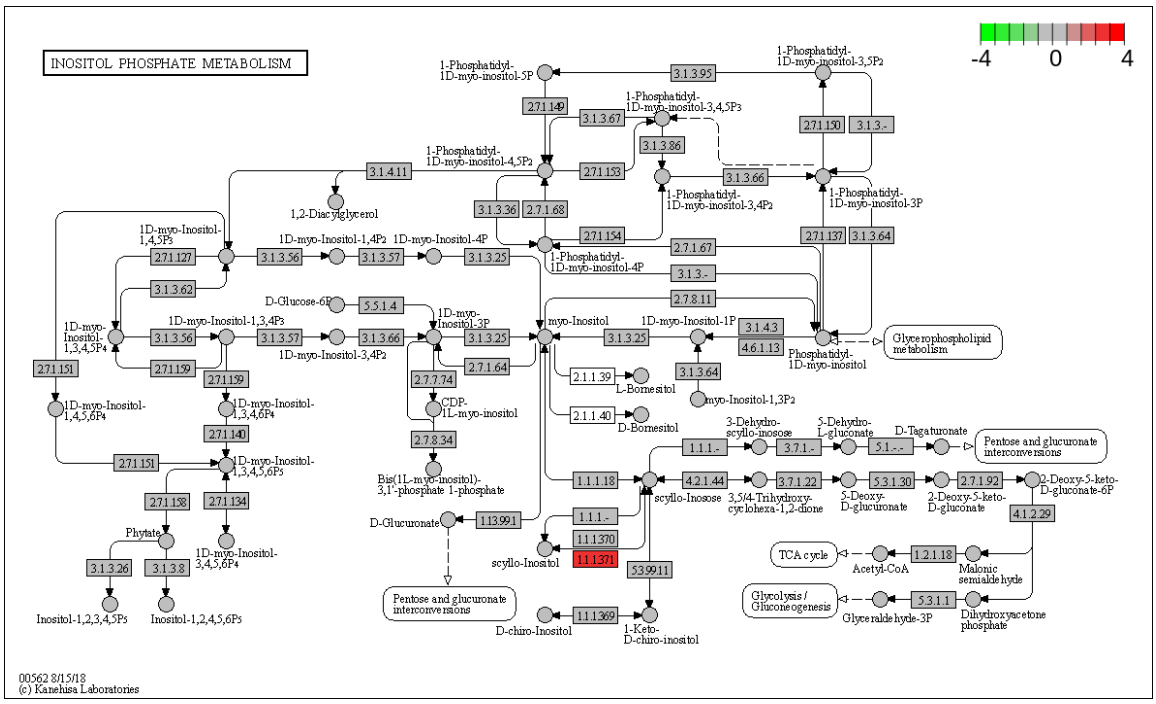


Figure 9. Example of KEGG pathway K16044 (inositol phosphate metabolism) pathway and specific pathway identified as upregulated for T3.

3.2 Microbia1 community survey using VSEARCH pipeline

The OTU abundance table used for statistical pipelines within R contained a total of 2993 OTUs for n=12 samples, at 97% similarity (a proxy for species-level separation) with summary read statistics for samples as follows: [Min: 77263 1st Quantile: 152109, Median: 180338, Mean: 208278, 3rd Quantile: 298742, Max: 339291]. In the interest of space, all VSEARCH pipeline data is available within Appendix III.

3.3 VSEARCH vs. DADA2

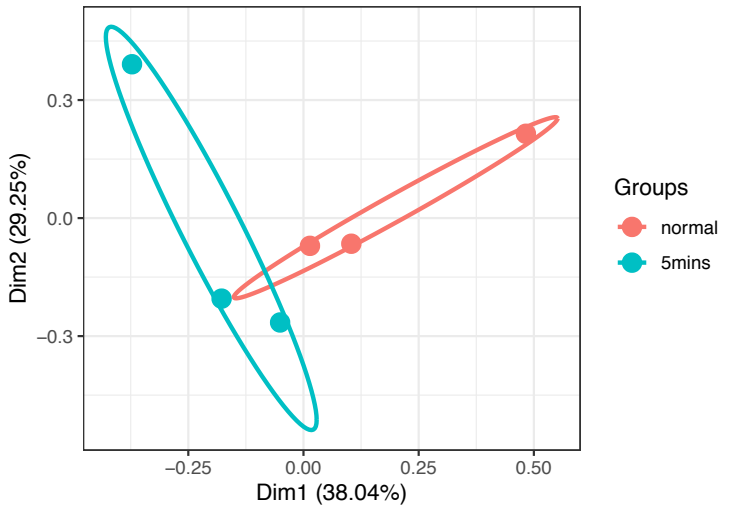
A comparison of the similarity between VSEARCH and DADA2 was performed using Procrustes analysis. The results identified a correlation of 0.787 with a p-value of 0.0001 between the two data sets. This indicates a significant similarity between the methods.

3. RNA-Seq analysis using StringTie pipeline

The StringTie transcript abundance table used for statistical pipelines within R contained a total of 1628 transcripts for n=9 samples, with summary read statistics for samples as follows: [Min: 58780 1st Quantile: 241717 Median: 279600, Mean: 318816, 3rd Quantile:

347288, Max:773386]. Ordination of abundance tables in reduced space (beta diversity) was performed using Principal Coordinate Analysis (PCoA) plots of transcripts *Bray-Curtis* distance in Vegan's cmdscale() function (Figure 10). Here, though groups were visually distinct, statistical significance was not observed (data not shown), most likely due to a lack of replicates.

(A)



(B)

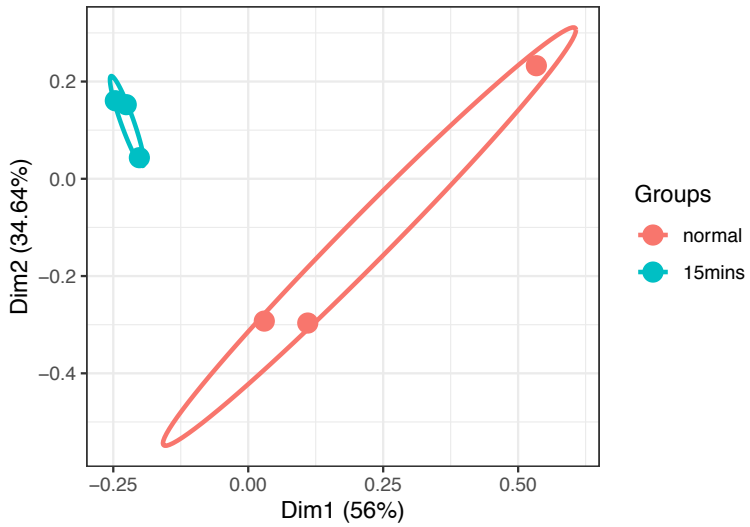
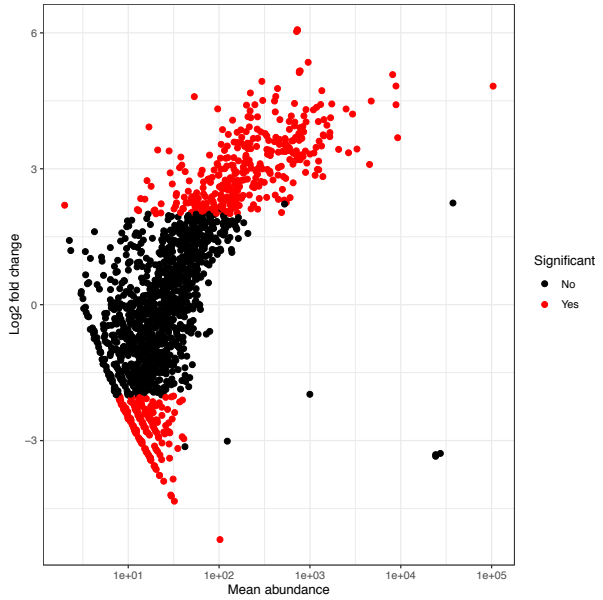


Figure 10. PCoA plots of transcripts *Bray-Curtis* distance comparing the different categorical variables, 5 mins 5 mM H_2O_2 stress (A) or 15 mins 5 mM H_2O_2 stress (B). RNA-Seq analysis was performed using StringTie pipeline.

DESeq2 was then used to identify transcripts that were differentially abundant (Figure 11). A subset of up and down gene lists is displayed in Appendix IV for normal vs 5 mins 5 mM

H_2O_2 stress (full list is provided in an accompanied file). The up and down gene lists for normal vs. 15 mins 5 mM H_2O_2 stress is also provided as an accompanied file due to space restrictions.

(A)



(B)

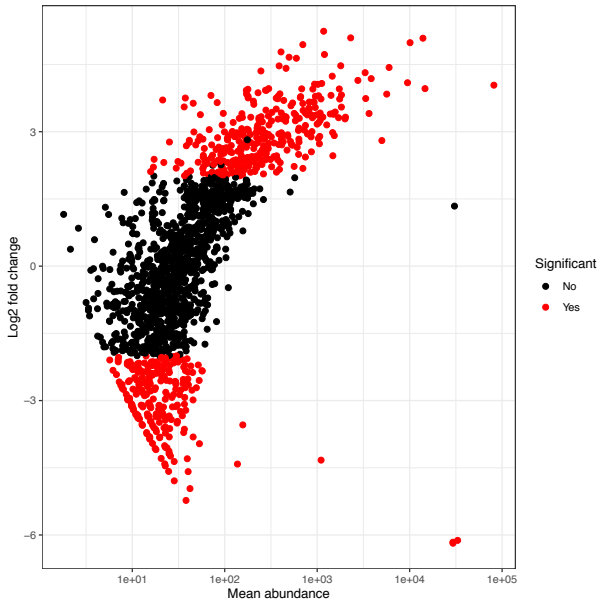


Figure 11. The transcript found to be differentially expressed in terms of log 2 fold changes from mean are shown in red. These correspond to two conditions: (A) 5 mins 5 mM H_2O_2 stress; and (B) 15 mins 5 mM H_2O_2 stress. RNA-Seq analysis performed using StringTie pipeline.

For both 5 mins 5 mM H_2O_2 stress and 15 mins 5 mM H_2O_2 stress against the control, at a preliminary level, key genes that were involved in oxidative stress were identified as

significantly upregulated within the test conditions. For example, for 5 mins 5 mM H_2O_2 stress, *sodB* was 3.8-fold upregulated and *ahpC* was 3.50-fold upregulated. Interestingly, the main gene involved in a response to peroxide stress, the gene *katA* (catalase) was not statistically upregulated. The results were similar for 15 mins 5 mM H_2O_2 stress; *sodB* was 4.14-fold upregulated and *ahpC* was 3.55-fold upregulated, with the increase was slightly higher for the increased peroxide stress. (Appendix IV and accompanied files).

3.5 RNA-Seq analysis using bedtools pipeline

The bedtools transcript abundance table used for statistical pipelines within R contained a total of 1628 transcripts for n=9 samples, with summary read statistics for samples as follows: [Min: 731478 1st Quantile: 2714702 Median: 3445795, Mean: 3093608, 3rd Quantile: 3783061, Max: 4330536]. In the interest of space, all RNA-Seq bedtools pipeline data is available within Appendix V. For significant up and down data using bedtools, due to space limitations were not displayed, but available as accompanied files.

3.6 StringTie vs bedtools

A comparison of the similarity between StringTie and bedtools was performed using Procrustes analysis. The results identified a correlation of 0.8148 with a p-value of 0.0001 between the two methods. This indicates a significant similarity between the methods.

4. Discussion

Through incursions into multiple pipelines, for both 16S rRNA amplicon analysis and RNA-Seq, we have been comprehensive in choosing the best strategy to delineate the underlying biological relevance. *C. jejuni* is the most common bacterial cause of human gastroenteritis worldwide. *C. jejuni* are widely found in avians and so the main route of transmission is via the consumption and handling of poultry products. How avians tolerate trillions of *C. jejuni* cells without having overt disease, yet only 500 cells cause severe disease in human hosts remains unknown. The European Union (EU) ban on antimicrobial growth promoters in 2006 has created an increased need to devise alternative methods to improve performance and potentially reduce numbers of pathogenic bacteria. Examples include use of natural plant derived products such as carvacrol (Kelly et al., 2017), addition of dietary prebiotics (Sethiya, 2016) and administration of live probiotic bacteria (Gadde et al., 2017). The impact of carvacrol on the chicken microbiome and the potential reduction of pathogen such as *Campylobacter* is unknown.

To investigate the impact of carvacrol on the chicken gut microbiome, data from a previously performed experiment was analysed using two different bioinformatics pipelines (DADA2 ASVs vs. VSEARCH OTUs, and a range of statistical analyses). Here, a comparison of samples from day 10, 21 and 35 were performed. Carvacrol concentrations within the different day groups were as follows, Control = 0 mg/ml; T1 = 120 mg/ml; T2 = 200 mg/ml and T3 = 300 mg/ml. For statistical analysis, sample groupings were based on sample type (Control, T1, T2 and T3). Further delineation of samples was ideally preferred, specifically, days and specific concentrations, however arguably the greatest criticism of this study was the low sample number, essentially only a single sample was available per day and per concentration. Thus, we have grouped the samples where we believe there is little or no change in community structure, and in future studies it would be desirable to have multiple data sets for multiple categories. Dominant bacteria (from top-25 most abundant taxa; Figure 1) include *Bacteroides*, *Ruminococcaceae UCG-014* and *Alistipes*. Statistically significant bacteria were identified (taxa differential) between the different groups (Table 2). As an example, *Alistipes* was significant for Control vs T1, although again, this analysis compared all days simultaneously.

Alpha diversity metrics (Figure. 2) using a range of diversity measures identified an increasing microbial diversity over time, though no statistical significance was observed between groups. Of note, T3 did show a trend of decreased microbial diversity when compared to T2. Beta diversity (Figure. 3) was performed and using distance matrix *Bray-*

Curtis (sample numbers), a large overlap was observed between all samples, with Control and T1 displaying greater similarity. For unweighted *UniFrac* (phylogeny), an increasing similarity of data points was observed from T1 to T3 with all data points overlapping. For weighted *UniFrac* (phylogeny with abundance) T3 displayed the greatest dispersion, with all samples again overlapping. Core microbiome (Figure. 4) identified key genera that were present within all samples. Interestingly, bacteria that were selected via taxa differential as varying between samples may still be observed within core microbiome e.g. *Alistipes*. Thus, there is not necessarily a case of genera being present or absent, just varying in prevalence. No statistical significance was observed for environmental filtering (Figure. 5) and LCBD (Figure. 6), though LCBD did observe an increasing trend of environmental influence as carvacrol concentration was increased. Subset analysis (Table 4) identified key genera influencing differences between Control and T1. Subset regression (Figure. 7) identified day 21 displaying a decreasing effect on microbial diversity, whereas day 35 displays an increasing effect. Also, day 10 seems to shift the microbial community structure away from the average diversity status. In addition, increasing the concentration of carvacrol seems to shift the microbial population structure away from the norm. Seqenv (Figure. 8) was performed although with limited success as most likely the environmental ontologies were less associated with a chicken microbiome, although it did show some relevant patterns. PICRUSt2 (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States) was performed (Table 5) to predict functional abundances. Interestingly a number of genes involved in sporulation and biofilm were identified. *C. jejuni* can also persist in the environment through survival in biofilms (Gundogdu et al., 2016). Clearly the varying concentration of carvacrol has an association with microbiome, most likely based on the antimicrobial effect of carvacrol. Overall, carvacrol is having an impact on the microbial community structure within the chicken microbiome, however, due to a lack of sample number within the day category, it was difficult to tease the differences out. In addition, the range of statistical analysis has produced a plethora of data that if more time was permitting, could have been performed on an extensive dataset, if available.

In terms of bioinformatics, a comparison between VSEARCH and DADA2 was performed using Procrustes analysis. The results identified a correlation of 0.787 with a p-value of 0.0001 indicating that OTU and ASV methods produced statistically significant results in terms of the abundance table. Interestingly, the DADA2 pipeline produced a total of 3485 ASVs whereas the VSEARCH pipeline produced a total of 2993 OTUs. Additionally, even though Procrustes did identify a similarity between the methodologies, downstream applications using for example subset regression did display some differences as to what was

identified as extrinsic parameters that were influencing the microbial community structure (Figure 7, DADA2 ASVs vs. Appendix III Figure 7, VSEARCH OTUs). As has been described elsewhere, a more detailed comparison of correlations between alpha and beta diversity could also have been a viable option (Glassman and Martiny, 2018). In terms of the biological output, here I only discuss DADA2 pipeline. Alternative steps could have been investigated such as Deblur (Amir et al., 2017) and MED (Eren et al., 2015).

C. jejuni is classified as microaerophilic, and this should imply a restricted niche for the growth of the bacterium, yet the opposite is true (Ugarte-Ruiz et al., 2018, Gundogdu and Wren, 2020). *C. jejuni* is omnipresent within the environment and demonstrates a genetic architecture that clearly allows it to survive within the ambient environment (Liaw et al., 2019). To investigate this further, RNA-Seq experimental data were obtained from previous studies which compared 5 mins 5 mM H_2O_2 stress against normal grown *C. jejuni*. This was also repeated for 15 mins.

Beta diversity of the different samples (Figure. 10) clearly identified differences between the different conditions. This was demonstrated also for differential expression analysis with DESeq2 where clear changes in gene expression were produced. One of the criticisms of the experimental design is that there were only two biological replicates for the respective datasets. The implication of this is that during statistical analyses the up and down-regulated gene lists are large and difficult to investigate. This point has been highlighted in a number of studies (Butcher et al., 2015, Conesa et al., 2016) where greater biological replicates significantly reduces the gene list numbers. Time permitting, a full analysis of the gene lists would have been performed and investigated in terms of biological importance. At a preliminary level, key genes that were involved in oxidative stress were identified as significantly modified for the test conditions e.g. *sodB*, *ahpC*. Interestingly, the main gene involved in a response to peroxide stress, the gene *katA* (catalase) was not identified in any of the test conditions. Further replicates would be required to fully ascertain the reasons for this.

In terms of bioinformatics, a comparison of the similarity between StringTie and bedtools was performed using Procrustes analysis. The results identified a correlation of 0.8148 with a p-value of 0.0001. This indicates a significant similarity between the methods. In this study, the focus was on selecting a range of tools for counting transcripts. In addition, variations in normalisation methods (e.g. TPM (transcripts per million) or FPKM (fragments per kilobase of transcript per million mapped reads)) could be investigated, and also differential expression methods; here only DESeq2 was performed, and so other methodologies could

have been attempted such as Cuffdiff (Trapnell et al., 2012), edgeR (Robinson et al., 2010), PoissonSeq (Li et al., 2012), and baySeq (Hardcastle and Kelly, 2010).

5. Conclusions

The aim of this study was to apply, utilise and compare different strategies to understand *C. jejuni* pathogenesis as well as the microbiome structure that is implicated in the underlying categorical representation of these samples, including both the active (RNA-Seq) and the passive (16S rRNA) components. Although we have utilised several strategies, our analytical understanding is limited by sample size and therefore the conclusions drawn should be treated with caution. Future studies that are comprehensively directed by a large sample size, and extensive methodology is recommended.

6. References

ACINAS, S. G., MARCELINO, L. A., KLEPAC-CERAJ, V. & POLZ, M. F. 2004. Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *Journal of bacteriology,* 186**,** 2629-2635.

AMIR, A., MCDONALD, D., NAVAS-MOLINA, J. A., KOPYLOVA, E., MORTON, J. T., XU, Z. Z., KIGHTLEY, E. P., THOMPSON, L. R., HYDE, E. R. & GONZALEZ, A. 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. *MSystems,* 2.

ANDERS, S. & HUBER, W. 2010. Differential expression analysis for sequence count data. *Nature Precedings***,** 1-1.

ANDERS, S., PYL, P. T. & HUBER, W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics,* 31**,** 166-169.

ATACK, J. M. & KELLY, D. J. 2009. Oxidative stress in Campylobacter jejuni: responses, resistance and regulation. *Future Microbiol,* 4**,** 677-90.

BOLYEN, E., RIDEOUT, J. R., DILLON, M. R., BOKULICH, N. A., ABNET, C. C., AL-GHALITH, G. A., ALEXANDER, H., ALM, E. J., ARUMUGAM, M., ASNICAR, F., BAI, Y., BISANZ, J. E., BITTINGER, K., BREJNROD, A., BRISLAWN, C. J., BROWN, C. T., CALLAHAN, B. J., CARABALLO-RODRIGUEZ, A. M., CHASE, J., COPE, E. K., DA SILVA, R., DIENER, C., DORRESTEIN, P. C., DOUGLAS, G. M., DURALL, D. M., DUVALLET, C., EDWARDSON, C. F., ERNST, M., ESTAKI, M., FOUQUIER, J., GAUGLITZ, J. M., GIBBONS, S. M., GIBSON, D. L., GONZALEZ, A., GORLICK, K., GUO, J., HILLMANN, B., HOLMES, S., HOLSTE, H., HUTTENHOWER, C., HUTTLEY, G. A., JANSSEN, S., JARMUSCH, A. K., JIANG, L., KAEHLER, B. D., KANG, K. B., KEEFE, C. R., KEIM, P., KELLEY, S. T., KNIGHTS, D., KOESTER, I., KOSCIOLEK, T., KREPS, J., LANGILLE, M. G. I., LEE, J., LEY, R., LIU, Y. X., LOFTFIELD, E., LOZUPONE, C., MAHER, M., MAROTZ, C., MARTIN, B. D., MCDONALD, D., MCIVER, L. J., MELNIK, A. V., METCALF, J. L., MORGAN, S. C., MORTON, J. T., NAIMEY, A. T., NAVAS-MOLINA, J. A., NOTHIAS, L. F., ORCHANIAN, S. B., PEARSON, T., PEOPLES, S. L., PETRAS, D., PREUSS, M. L., PRUESSE, E., RASMUSSEN, L. B., RIVERS, A., ROBESON, M. S., 2ND, ROSENTHAL, P., SEGATA, N., SHAFFER, M., SHIFFER, A., SINHA, R., SONG, S. J., SPEAR, J. R., SWAFFORD, A. D., THOMPSON, L. R., TORRES, P. J., TRINH, P., TRIPATHI, A., TURNBAUGH, P. J., UL-HASAN, S., VAN DER HOOFT, J. J. J., VARGAS, F., VAZQUEZ-BAEZA, Y., VOGTMANN, E., VON HIPPEL, M., WALTERS, W., et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol,* 37**,** 852-857.

BONETTA, L. 2010. Whole-genome sequencing breaks the cost barrier. *Cell,* 141**,** 917-9.

BRONDSTED, L., ANDERSEN, M. T., PARKER, M., JORGENSEN, K. & INGMER, H. 2005. The HtrA protease of *Campylobacter jejuni* is required for heat and oxygen tolerance and for optimal interaction with human epithelial cells. *Appl Environ Microbiol,* 71**,** 3205-12.

BULLARD, J. H., PURDOM, E., HANSEN, K. D. & DUDOIT, S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics,* 11**,** 94.

BUTCHER, J., HANDLEY, R. A., VAN VLIET, A. H. & STINTZI, A. 2015. Refined analysis of the Campylobacter jejuni iron-dependent/independent Fur- and PerR-transcriptomes. *BMC Genomics,* 16**,** 498.

BYRNE, C. M., CLYNE, M. & BOURKE, B. 2007. Campylobacter jejuni adhere to and invade chicken intestinal epithelial cells in vitro. *Microbiology,* 153**,** 561-569.

CALLAHAN, B. J., MCMURDIE, P. J. & HOLMES, S. P. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME journal,* 11**,** 2639.

CALLAHAN, B. J., MCMURDIE, P. J., ROSEN, M. J., HAN, A. W., JOHNSON, A. J. A. & HOLMES, S. P. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. *Nature methods,* 13**,** 581.

CAPORASO, J. G., KUCZYNSKI, J., STOMBAUGH, J., BITTINGER, K., BUSHMAN, F. D., COSTELLO, E. K., FIERER, N., PENA, A. G., GOODRICH, J. K. & GORDON, J. I. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature methods,* 7**,** 335.

CLARKE, K. R. & AINSWORTH, M. 1993. A Method of Linking Multivariate Community Structure to Environmental Variables. *Marine Ecology Progress Series,* 92**,** 205-219.

CONESA, A., MADRIGAL, P., TARAZONA, S., GOMEZ-CABRERO, D., CERVERA, A., MCPHERSON, A., SZCZEŚNIAK, M. W., GAFFNEY, D. J., ELO, L. L. & ZHANG, X. 2016. A survey of best practices for RNA-seq data analysis. *Genome biology,* 17**,** 13.

CREECY, J. P. & CONWAY, T. 2015. Quantitative bacterial transcriptomics with RNA-seq. *Current opinion in microbiology,* 23**,** 133-140.

D'AMORE, R., IJAZ, U. Z., SCHIRMER, M., KENNY, J. G., GREGORY, R., DARBY, A. C., SHAKYA, M., PODAR, M., QUINCE, C. & HALL, N. 2016. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. *BMC Genomics,* 17**,** 55.

D'AUTREAUX, B. & TOLEDANO, M. B. 2007. ROS as signalling molecules: mechanisms that generate specificity in ROS homeostasis. *Nat Rev Mol Cell Biol,* 8**,** 813-24.

DOUGLAS, G. M., MAFFEI, V. J., ZANEVELD, J., YURGEL, S. N., BROWN, J. R., TAYLOR, C. M., HUTTENHOWER, C. & LANGILLE, M. G. I. 2020. PICRUSt2: An improved and customizable approach for metagenome inference. *bioRxiv***,** 672295.

DRAY, S., BLANCHET, G., BORCARD, D., CLAPPE, S., GUENARD, G., JOMBART, T., LAROCQUE, G., LEGENDRE, P., MADI, N. & WAGNER, H. H. 2018. adespatial: Multivariate Multiscale Spatial Analysis. 0.0–9 ed. https://CRAN.R-project.org/package=adespatial.

DUEHOLM, M. S., ANDERSEN, K. S., PETRIGLIERI, F., MCILROY, S. J., NIERYCHLO, M., PETERSEN, J. F., KRISTENSEN, J. M., YASHIRO, E., KARST, S. M. & ALBERTSEN, M. 2019. Comprehensive ecosystem-specific 16S rRNA gene databases with automated taxonomy assignment (AutoTax) provide species-level resolution in microbial ecology. *Biorxiv***,** 672873.

EDGAR, R. C. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature methods,* 10**,** 996.

EDGAR, R. C. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv***,** 081257.

EREN, A. M., MORRISON, H. G., LESCAULT, P. J., REVEILLAUD, J., VINEIS, J. H. & SOGIN, M. L. 2015. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *The ISME journal,* 9**,** 968-979.

FIELDS, J. A. & THOMPSON, S. A. 2008. *Campylobacter jejuni* CsrA mediates oxidative stress responses, biofilm formation, and host cell invasion. *J Bacteriol,* 190**,** 3411-6.

FOSTER, Z. S., SHARPTON, T. J. & GRÜNWALD, N. J. 2017. Metacoder: an R package for visualization and manipulation of community taxonomic diversity data. *PLoS computational biology,* 13**,** e1005404.

GADDE, U., KIM, W., OH, S. & LILLEHOJ, H. S. 2017. Alternatives to antibiotics for maximizing growth performance and feed efficiency in poultry: a review. *Animal health research reviews,* 18**,** 26-45.

GLASSMAN, S. I. & MARTINY, J. B. H. 2018. Broadscale ecological patterns are robust to use of exact sequence variants versus operational taxonomic units. *MSphere,* 3.

GOODWIN, S., MCPHERSON, J. D. & MCCOMBIE, W. R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet,* 17**,** 333-51.

GÜELL, M., YUS, E., LLUCH-SENAR, M. & SERRANO, L. 2011. Bacterial transcriptomics: what is beyond the RNA horiz-ome? *Nature Reviews Microbiology,* 9**,** 658-669.

GUNDOGDU, O., DA SILVA, D. T., MOHAMMAD, B., ELMI, A., WREN, B. W., VAN VLIET, A. H. & DORRELL, N. 2016. The Campylobacter jejuni Oxidative Stress Regulator RrpB Is Associated with a Genomic Hypervariable Region and Altered Oxidative Stress Resistance. *Front Microbiol,* 7**,** 2117.

GUNDOGDU, O. & WREN, B. W. 2020. Microbe Profile: Campylobacter jejuni–survival instincts. *Microbiology,* 166**,** 230-232.

HAMADY, M. & KNIGHT, R. 2009. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome research,* 19**,** 1141-1152.

HAMADY, M., WALKER, J. J., HARRIS, J. K., GOLD, N. J. & KNIGHT, R. 2008. Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods,* 5**,** 235-7.

HARDCASTLE, T. J. & KELLY, K. A. 2010. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics,* 11**,** 422.

HONG, S., BUNGE, J., LESLIN, C., JEON, S. & EPSTEIN, S. S. 2009. Polymerase chain reaction primers miss half of rRNA microbial diversity. *The ISME journal,* 3**,** 1365-1373.

HUSE, S. M., WELCH, D. M., MORRISON, H. G. & SOGIN, M. L. 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental microbiology,* 12**,** 1889-1898.

IJAZ, U. Z., SIVALOGANATHAN, L., MCKENNA, A., RICHMOND, A., KELLY, C., LINTON, M., STRATAKOS, A. C., LAVERY, U., ELMI, A., WREN, B. W., DORRELL, N., CORCIONIVOSCHI, N. & GUNDOGDU, O. 2018. Comprehensive Longitudinal Microbiome Analysis of the Chicken Cecum Reveals a Shift From Competitive to Environmental Drivers and a Window of Opportunity for Campylobacter. *Front Microbiol,* 9**,** 2452.

JACKSON, M. A., BELL, J. T., SPECTOR, T. D. & STEVES, C. J. 2016. A heritability-based comparison of methods used to cluster 16S rRNA gene sequences into operational taxonomic units. *PeerJ,* 4**,** e2341.

JOHNSON, J. S., SPAKOWICZ, D. J., HONG, B.-Y., PETERSEN, L. M., DEMKOWICZ, P., CHEN, L., LEOPOLD, S. R., HANSON, B. M., AGRESTA, H. O. & GERSTEIN, M. 2019. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature communications,* 10**,** 1-11.

JOSHI, N. & FASS, J. 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33). 1.33 ed.

JUMPSTART CONSORTIUM HUMAN MICROBIOME PROJECT DATA GENERATION WORKING, G. 2012. Evaluation of 16S rDNA-based community profiling for human microbiome research. *PloS one,* 7.

KASSAMBARA, A. 2018. *Machine Learning Essentials: Practical Guide in R*, sthda.

KATOH, K. & STANDLEY, D. M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol,* 30**,** 772-80.

KELLY, C., GUNDOGDU, O., PIRCALABIORU, G., CEAN, A., SCATES, P., LINTON, M., PINKERTON, L., MAGOWAN, E., STEF, L., SIMIZ, E., PET, I., STEWART, S., STABLER, R., WREN, B., DORRELL, N. & CORCIONIVOSCHI, N. 2017. The In Vitro and In Vivo Effect of Carvacrol in Preventing Campylobacter Infection,

Colonization and in Improving Productivity of Chicken Broilers. *Foodborne Pathog Dis,* 14**,** 341-349.
KUCZYNSKI, J., LIU, Z., LOZUPONE, C., MCDONALD, D., FIERER, N. & KNIGHT, R. 2010. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nature methods,* 7**,** 813.
KUHN, M. 2008. Building predictive models in R using the caret package. *Journal of statistical software,* 28**,** 1-26.
LAHTI, L., SHETTY, S., BLAKE, T. & SALOJARVI, J. 2017. microbiome R package. *Tools Microbiome Anal R*.
LANE, D. J., PACE, B., OLSEN, G. J., STAHL, D. A., SOGIN, M. L. & PACE, N. R. 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci U S A,* 82**,** 6955-9.
LANGMEAD, B. & SALZBERG, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods,* 9**,** 357-9.
LEGENDRE, P. & DE CACERES, M. 2013. Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. *Ecol Lett,* 16**,** 951-63.
LEY, R. E., HAMADY, M., LOZUPONE, C., TURNBAUGH, P. J., RAMEY, R. R., BIRCHER, J. S., SCHLEGEL, M. L., TUCKER, T. A., SCHRENZEL, M. D. & KNIGHT, R. 2008a. Evolution of mammals and their gut microbes. *Science,* 320**,** 1647-1651.
LEY, R. E., LOZUPONE, C. A., HAMADY, M., KNIGHT, R. & GORDON, J. I. 2008b. Worlds within worlds: evolution of the vertebrate gut microbiota. *Nature Reviews Microbiology,* 6**,** 776-788.
LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & GENOME PROJECT DATA PROCESSING, S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.
LI, J., WITTEN, D. M., JOHNSTONE, I. M. & TIBSHIRANI, R. 2012. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics,* 13**,** 523-538.
LIAW, J., HONG, G., DAVIES, C., ELMI, A., SIMA, F., STRATAKOS, A., STEF, L., PET, I., HACHANI, A. & CORCIONIVOSCHI, N. 2019. The Campylobacter jejuni Type VI Secretion System Enhances the Oxidative Stress Response and Host Colonization. *Frontiers in Microbiology,* 10.
LIU, Z., DESANTIS, T. Z., ANDERSEN, G. L. & KNIGHT, R. 2008. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res,* 36**,** e120.
LOVE, M. I., HUBER, W. & ANDERS, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol,* 15**,** 550.
LOZUPONE, C., HAMADY, M. & KNIGHT, R. 2006. UniFrac–an online tool for comparing microbial community diversity in a phylogenetic context. *BMC bioinformatics,* 7**,** 371.
LOZUPONE, C. & KNIGHT, R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol,* 71**,** 8228-35.
LOZUPONE, C. A., HAMADY, M., KELLEY, S. T. & KNIGHT, R. 2007. Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.,* 73**,** 1576-1585.
LÜDECKE, D. 2018. sjPlot: Data visualization for statistics in social science. *R package version,* 2.
LUMLEY, T. & MILLER, A. 2009. Leaps: regression subset selection. R package version 2.9. *See http://CRAN. R-project. org/package= leaps*.
MAGURRAN, A. E. 2013. *Measuring biological diversity*, John Wiley & Sons.

MAHÉ, F., ROGNES, T., QUINCE, C., DE VARGAS, C. & DUNTHORN, M. 2014. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ,* 2**,** e593.

MASELLA, A. P., BARTRAM, A. K., TRUSZKOWSKI, J. M., BROWN, D. G. & NEUFELD, J. D. 2012. PANDAseq: paired-end assembler for illumina sequences. *BMC bioinformatics,* 13**,** 31.

MCMURDIE, P. J. & HOLMES, S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One,* 8**,** e61217.

MORTAZAVI, A., WILLIAMS, B. A., MCCUE, K., SCHAEFFER, L. & WOLD, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods,* 5**,** 621.

NACHAMKIN, I., ALLOS, B. M. & HO, T. 1998. *Campylobacter* species and Guillain-Barre syndrome. *Clin Microbiol Rev,* 11**,** 555-67.

NAVAS-MOLINA, J. A., PERALTA-SÁNCHEZ, J. M., GONZÁLEZ, A., MCMURDIE, P. J., VÁZQUEZ-BAEZA, Y., XU, Z., URSELL, L. K., LAUBER, C., ZHOU, H. & SONG, S. J. 2013. Advancing our understanding of the human microbiome using QIIME. *Methods in enzymology.* Elsevier.

NEEDHAM, D. M., FICHOT, E. B., WANG, E., BERDJEB, L., CRAM, J. A., FICHOT, C. G. & FUHRMAN, J. A. 2018. Dynamics and interactions of highly resolved marine plankton via automated high-frequency sampling. *ISME J,* 12**,** 2417-2432.

NIKOLENKO, S. I., KOROBEYNIKOV, A. I. & ALEKSEYEV, M. A. 2013. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics,* 14 Suppl 1**,** S7.

OKSANEN, J., BLANCHET, F., KINDT, R., LEGENDRE, P., MINCHIN, P., O'HARA, R., SIMPSON, G., SOLYMOS, P., STEVENS, M. & WAGNER, H. 2015. Vegan: Community Ecology Package, R Package version 2.2-1. version 2.2-1 ed.

PERES-NETO, P. R. & JACKSON, D. A. 2001. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia,* 129**,** 169-178.

PERTEA, M., PERTEA, G. M., ANTONESCU, C. M., CHANG, T.-C., MENDELL, J. T. & SALZBERG, S. L. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology,* 33**,** 290.

PITTMAN, M. S., ELVERS, K. T., LEE, L., JONES, M. A., POOLE, R. K., PARK, S. F. & KELLY, D. J. 2007. Growth of *Campylobacter jejuni* on nitrate and nitrite: electron transport to NapA and NrfA via NrfH and distinct roles for NrfA and the globin Cgb in protection against nitrosative stress. *Mol Microbiol,* 63**,** 575-90.

PRICE, M. N., DEHAL, P. S. & ARKIN, A. P. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One,* 5**,** e9490.

QUAST, C., PRUESSE, E., YILMAZ, P., GERKEN, J., SCHWEER, T., YARZA, P., PEPLIES, J. & GLOCKNER, F. O. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res,* 41**,** D590-6.

QUINCE, C., WALKER, A. W., SIMPSON, J. T., LOMAN, N. J. & SEGATA, N. 2017. Shotgun metagenomics, from sampling to analysis. *Nature biotechnology,* 35**,** 833.

QUINLAN, A. R. 2014. BEDTools: the Swiss-army tool for genome feature analysis. *Current protocols in bioinformatics,* 47**,** 11-12.

QUINLAN, A. R. & HALL, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* 26**,** 841-2.

ROBINSON, M. D., MCCARTHY, D. J. & SMYTH, G. K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* 26**,** 139-140.

ROGNES, T., FLOURI, T., NICHOLS, B., QUINCE, C. & MAHÉ, F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ,* 4**,** e2584.

RUIZ-PALACIOS, G. M. 2007. The health burden of Campylobacter infection and the impact of antimicrobial resistance: playing chicken. The University of Chicago Press.
SCHIRMER, M., IJAZ, U. Z., D'AMORE, R., HALL, N., SLOAN, W. T. & QUINCE, C. 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res,* 43**,** e37.
SCHLOSS, P. D. 2010. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS computational biology,* 6.
SCHLOSS, P. D. & HANDELSMAN, J. 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl. Environ. Microbiol.,* 71**,** 1501-1506.
SCHLOSS, P. D. & HANDELSMAN, J. 2006. Introducing SONS, a tool for operational taxonomic unit-based comparisons of microbial community memberships and structures. *Appl. Environ. Microbiol.,* 72**,** 6773-6779.
SCHLOSS, P. D., WESTCOTT, S. L., RYABIN, T., HALL, J. R., HARTMANN, M., HOLLISTER, E. B., LESNIEWSKI, R. A., OAKLEY, B. B., PARKS, D. H. & ROBINSON, C. J. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.,* 75**,** 7537-7541.
SETHIYA, N. K. 2016. Review on natural growth promoters available for improving gut health of poultry: an alternative to antibiotic growth promoters. *Asian Journal of Poultry Science,* 10**,** 1-29.
SHARPTON, T. J., RIESENFELD, S. J., KEMBEL, S. W., LADAU, J., O'DWYER, J. P., GREEN, J. L., EISEN, J. A. & POLLARD, K. S. 2011. PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. *PLoS computational biology,* 7.
SHENDURE, J., BALASUBRAMANIAN, S., CHURCH, G. M., GILBERT, W., ROGERS, J., SCHLOSS, J. A. & WATERSTON, R. H. 2017. DNA sequencing at 40: past, present and future. *Nature,* 550**,** 345-353.
SHETTY, S. A., HUGENHOLTZ, F., LAHTI, L., SMIDT, H. & DE VOS, W. M. 2017. Intestinal microbiome landscaping: insight in community assemblage and implications for microbial modulation strategies. *FEMS microbiology reviews,* 41**,** 182-199.
SINCLAIR, L., IJAZ, U. Z., JENSEN, L. J., COOLEN, M. J. L., GUBRY-RANGIN, C., CHRONAKOVA, A., OULAS, A., PAVLOUDI, C., SCHNETZER, J., WEIMANN, A., IJAZ, A., EILER, A., QUINCE, C. & PAFILIS, E. 2016. Seqenv: linking sequences to environments through text mining. *PeerJ,* 4**,** e2690.
SNEATH, P. H. A. & SOKAL, R. R. 1973. *Numerical taxonomy. The principles and practice of numerical classification*.
SONESON, C. & DELORENZI, M. 2013. A comparison of methods for differential expression analysis of RNA-seq data. *BMC bioinformatics,* 14**,** 91.
SOREK, R. & COSSART, P. 2010. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nature Reviews Genetics,* 11**,** 9-16.
STACKEBRANDT, E. & GOEBEL, B. M. 1994. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *International journal of systematic and evolutionary microbiology,* 44**,** 846-849.
TAYLOR, M. 2014. *sinkr: A collection of functions featured on the blog 'me nugget'* [Online]. Available: https://github.com/menugget/sinkr.
TIKHONOV, M., LEACH, R. W. & WINGREEN, N. S. 2015. Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. *The ISME journal,* 9**,** 68-80.

TRAPNELL, C., ROBERTS, A., GOFF, L., PERTEA, G., KIM, D., KELLEY, D. R., PIMENTEL, H., SALZBERG, S. L., RINN, J. L. & PACHTER, L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols,* 7**,** 562-578.

TRAPNELL, C., WILLIAMS, B. A., PERTEA, G., MORTAZAVI, A., KWAN, G., VAN BAREN, M. J., SALZBERG, S. L., WOLD, B. J. & PACHTER, L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology,* 28**,** 511.

UGARTE-RUIZ, M., DOMÍNGUEZ, L., CORCIONIVOSCHI, N., WREN, B. W., DORRELL, N. & GUNDOGDU, O. 2018. Exploring the oxidative, antimicrobial and genomic properties of Campylobacter jejuni strains isolated from poultry. *Research in Veterinary Science,* 119**,** 170-175.

ULTEE, A., KETS, E. P. W. & SMID, E. J. 1999. Mechanisms of action of carvacrol on the food-borne pathogen Bacillus cereus. *Appl. Environ. Microbiol.,* 65**,** 4606-4610.

VIVANCOS, A. P., GÜELL, M., DOHM, J. C., SERRANO, L. & HIMMELBAUER, H. 2010. Strand-specific deep sequencing of the transcriptome. *Genome research,* 20**,** 989-999.

WILHELM, B. T., MARGUERAT, S., WATT, S., SCHUBERT, F., WOOD, V., GOODHEAD, I., PENKETT, C. J., ROGERS, J. & BÄHLER, J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature,* 453**,** 1239-1243.

YARZA, P., YILMAZ, P., PRUESSE, E., GLÖCKNER, F. O., LUDWIG, W., SCHLEIFER, K.-H., WHITMAN, W. B., EUZÉBY, J., AMANN, R. & ROSSELLÓ-MÓRA, R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology,* 12**,** 635-645.

Appendix I – Bioinformatics Steps for Qiime2 analysis with DADA2

All relevant pipelines were established by Dr Umer Zeeshan Ijaz (http://userweb.eng.gla.ac.uk/umer.ijaz/) at the University of Glasgow and training was provided accordingly (http://www.tinyurl.com/JCBioinformatics3).

Step - 1 Create fictitious barcodes

The file structure was organised as described in 2.1. Initially, fictitious barcodes were created and saved as a sample_metadata.tsv file: -

```
d="/home/ozan/Documents/Nic_meta/Carvacrol/sequences"; #Sets path to folder variable
t=$(ls $d | wc -l); #Sets a one-line code to view contents of folder
paste <(ls $d) <(perl -le 'sub p{my $l=pop @_;unless(@_){return map [$_],@$l;}return map { my $ll=$_; map [@$ll,$_],@$l} p(@_);} @a=[A,C,G,T]; print join("", @$_) for p(@a,@a,@a,@a,@a,@a,@a,@a);' | awk -v k=$t 'NR<=k{print}') | awk 'BEGIN{print "sample-id\tbarcode-sequence\n#q2:types\tcategorical"}1' > sample_metadata.tsv
#Creates fictitious 8 bp barcodes and saves them in a .tsv file.
```

The output produces an sample_metadata.tsv file: -

```
sample-id       barcode-sequence
#q2:types       categorical
Day10C  AAAAAAAA
Day10T1 AAAAAAAC
Day10T2 AAAAAAAG
Day10T3 AAAAAAAT
```

Step - 2 Generate barcodes for each read

The next step was to generate the respective barcodes for each read within our respective folder structure: -

```
(for i in $(ls $d); do bc=$(awk -v k=$i '$1==k{print $2}' sample_metadata.tsv); bioawk -cfastx -v k=$bc '{print "@"$1" "$4"\n"k"\n+";for(i=0;i< length(k);i++){printf "#"};printf "\n"}' $d/$i/Raw/*_R1.fastq ; done) > barcodes.fastq
```

The output produces an barcodes.fastq.gz file: -

```
barcodes.fastq.gz | head -n 10
@M01637:109:000000000-AM9RR:1:1101:14441:1282 1:N:0:21
AAAAAAAA
+
########
@M01637:109:000000000-AM9RR:1:1101:9590:1353 1:N:0:21
AAAAAAAA
+
########
```

Step 3 - Assemble the forward reads

This was then followed by assembling the forward reads: -

```
(for i in $(ls $d); do bioawk -cfastx '{print "@"$1"
"$4"\n"$seq"\n+\n"$qual}' $d/$i/Raw/*_R1_001.fastq ; done) >
forward.fastq
```

The output produces an forward.fasta.gz file: -

```
forward.fastq.gz | head -n 10
@M01637:109:000000000-AM9RR:1:1101:14441:1282 1:N:0:21
CCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGGGAAACCCTGATGCAGCGACGCC
GCGTGAGTGATGAAGTACTTCGGTATGTAAAGCTCTATCAGCAGGGAAGTAAGTGACGGTA
CCTGAGTAAGAAGCCCCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGGGCAAG
CGTTATCCGGCTTTACTGGGTGTAATGGGAGTTTAGTCGGCGATGCAAGTCTGGCGTGAAA
TCCCCGTGCTCAACACCCGGTCTTCTTTGGACCCTTTTATGCTGGCTTTTCGGGTGG
+
8BCCCECAAB5:0==;;=6;;8,8C6EF<@,C,6CEF@@E@,6;;FGF<C@D,FEC+8C@F
GGGGG,=BF<,C,,CF,5CFGCF?4BEF9,59,AAB<B<,<<,BDC+,@,,,5A,;>+@@+
?AF:,,73,,,,,==D=7+CB8,86,@6D6C8EED6ACDD>9>C**=,599?F*==55458
:3CB3;:;=7(:8A+;@:)409D)*).1*.53(9)*).18)*),)-)*)-)-
40),4(())).(0(((((--)),)((((4(((()-6:6)(,)((((-4*)--
*)(.(()),((.((
```

Step 4 - Assemble the reverse reads

This was then followed by assembling the reverse reads respectively: -

```
(for i in $(ls $d); do bioawk -cfastx '{print "@"$1"
"$4"\n"$seq"\n+\n"$qual}' $d/$i/Raw/*_R2_001.fastq ; done) >
reverse.fastq
```

The output produces an reverse.fastq.gz file: -

```
reverse.fastq.gz | head -n 10
@M01637:109:000000000-AM9RR:1:1101:14441:1282 2:N:0:21
GACTACTCGGGTATCTAATCCTGTTTGCTCCCCACGCTTTCGAGCCTCCACGTCAGTTACC
GTCCAGTAAGCCGCCTTCGCCACTGGTGTTCTTCCTAATATCTACGCATTTCCCCGCTACC
CTAGGAATTCCGCTTACCTCTCCGGCACTCCAGCTCTACAGTTTCCAACGCAGTCCCGGTG
TTGAGCCCCGGGCTTTCCCTCCAGACTTGCCTCGCCGTCTACACTCCCTTTCCCCCCCGTA
CATCCGGATAACGCTTGCCCCCTACCTCTTACCCCGGCTGCCTGCCCCTACTTCAC
+
8-
ABBFFFFB@E7EFF;,EFGF@EFECDFGGGG?C:CCCFGF,=BFGF,4;BFG,@EF9@@=A
?=,@C,,=49@CECEEDEE6>@AAF>EGFDFFGG9,==DDE7E@C9DCDG7FE?8588)@3
)96,+=?DD=?8A5DFFFEFF5)85=DFD)DF5)0*7:D=D>DD***:=))1)385*0*<@
5*=1,:5)))-05;BEB55@5(.658).()/(26((.(((/)5)/27026)7),(1(-
29<6)43-((,().(4(-(-5266((4(.)4))-69>)(2((((()-183((,()))).
```

At this stage it was possible to check if as expected, the forward, reverse and barcodes FASTQ files were the same size: -

```
bioawk -cfastx 'END{print NR}' forward.fastq.gz
bioawk -cfastx 'END{print NR}' reverse.fastq.gz
bioawk -cfastx 'END{print NR}' barcodes.fastq.gz
```

Step 5 - Further file organisation

All files were then zipped and moved to a new folder (emp-paired-end-sequences), distinct from sequencing folder (../sequences; which holds the FASTQ files respectively).

```
gzip *.fastq
mkdir emp-paired-end-sequences; mv *.gz emp-paired-end-
sequences/.
```

Step 6 - Load into Qiime2 format

The created files were then ready to be loaded into the Qiime2 platform as follows: -

```
qiime tools import \
```

```
--type EMPPairedEndSequences \
--input-path emp-paired-end-sequences \
--output-path emp-paired-end-sequences.qza
```

Step 7- Demultiplexed

The samples were then demultiplexed as follows: -

```
qiime demux emp-paired --p-no-golay-error-correction --i-seqs emp-paired-end-sequences.qza --m-barcodes-file sample_metadata.tsv --m-barcodes-column barcode-sequence --o-per-sample-sequences demux.qza --o-error-correction-details demux-details.qza
```

Step 8 - Create visualisable demultiplexed files

The sequences were then converted from .qza files into visualisable .qzv format. To view, .qzv files were directly loaded into the Qiime2 viewer website (https://view.qiime2.org/) (Figure 1A,B and C). To convert to .qzv: -

```
qiime demux summarize --i-data ./demux.qza --o-visualization ./demux.qzv
qiime tools export --input-path demux.qzv --output-path output
```

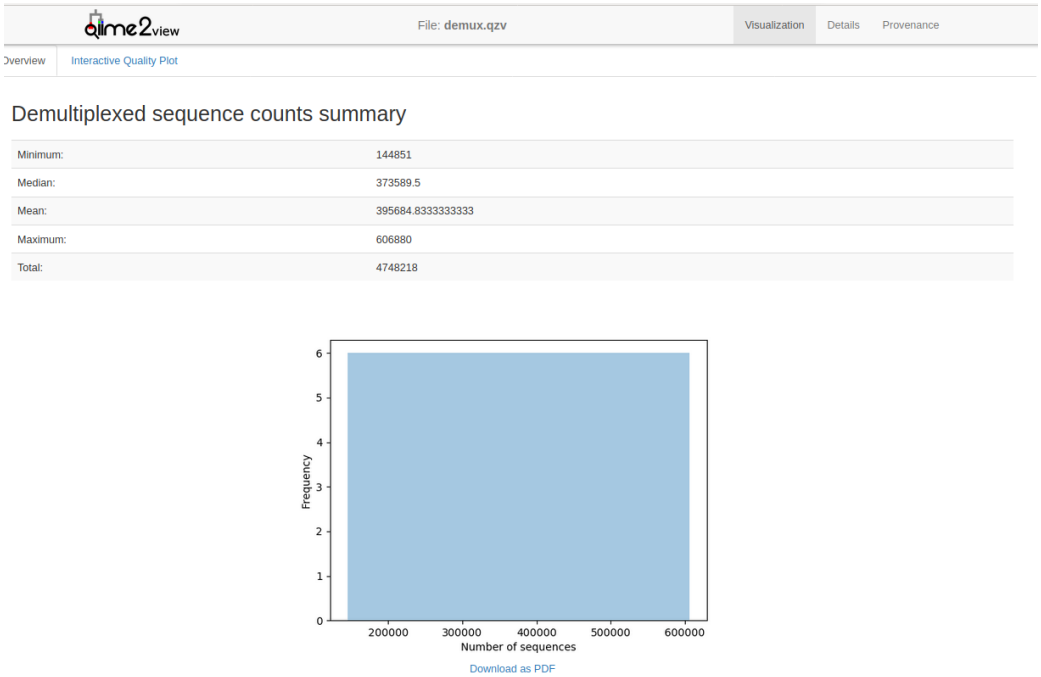
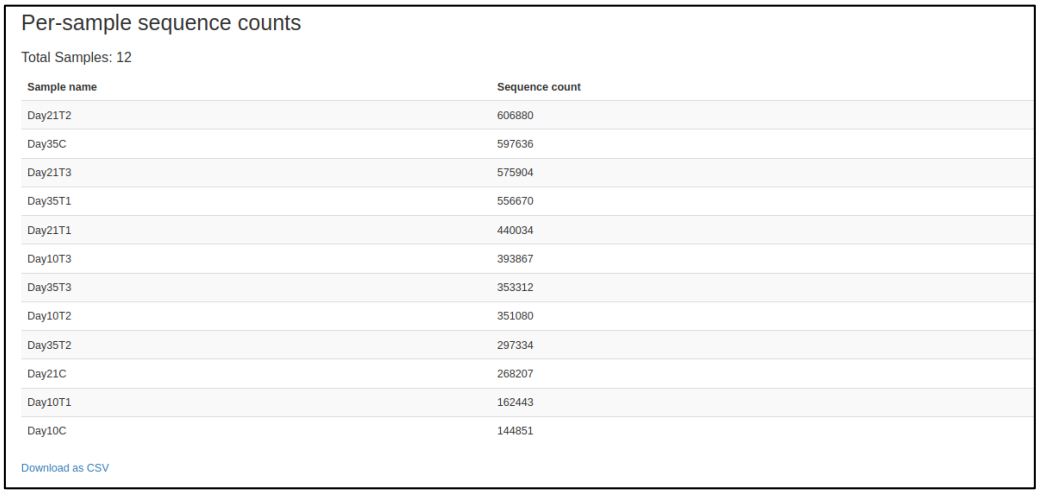


Figure 1A. Demultiplexed sequence counts summary of demux.qzv loaded into Qiime2 viewer.



Per-sample sequence counts

Total Samples: 12

Sample name	Sequence count
Day21T2	606880
Day35C	597636
Day21T3	575904
Day35T1	556670
Day21T1	440034
Day10T3	393867
Day35T3	353312
Day10T2	351080
Day35T2	297334
Day21C	268207
Day10T1	162443
Day10C	144851

Download as CSV

Figure 1B. Per-sample sequencing counts of each sample from demux.qzv loaded into Qiime2 viewer.

Figure 1C. Quality plot of forward and reverse paired-end sequencing reads from demux.qzv loaded into Qiime2 viewer.

Step 9 - Perform DADA2 search

The next step was to perform DADA2 search as follows: -

```
qiime dada2 denoise-paired --i-demultiplexed-seqs demux.qza --p-trim-left-f 0 --p-trim-left-r 0 --p-trunc-len-f 240 --p-trunc-len-r 200 --p-n-threads 0 --o-table table.qza --o-
```

```
representative-sequences rep-seqs.qza --o-denoising-stats denoising-stats.qza --verbose
```

Step 10 - Create a phylogeny

A phylogenetic tree was created as follows: -

```
qiime phylogeny align-to-tree-mafft-fasttree --i-sequences rep-seqs.qza --o-alignment aligned-rep-seqs.qza --o-masked-alignment masked-aligned-rep-seqs.qza --p-n-threads 0 --o-tree unrooted-tree.qza --o-rooted-tree rooted-tree.qza
```

Step 11 - Create a taxonomy

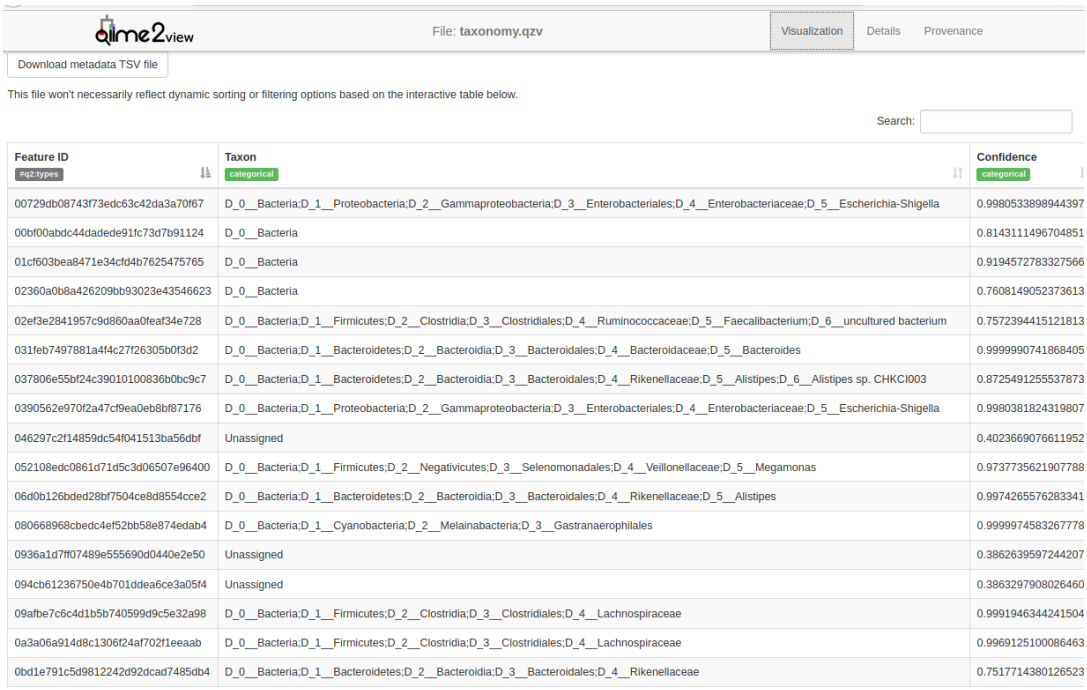
A taxonomy was created as follows: -

```
qiime feature-classifier classify-sklearn --i-classifier ~/Downloads/silva-132-99-nb-classifier.qza --i-reads rep-seqs.qza --o-classification taxonomy.qza
```

To visualise the output, the .qza file was converted into a .qzv as follows: -

```
qiime metadata tabulate --m-input-file taxonomy.qza --o-visualization taxonomy.qzv
```

taxonomy.qzv was placed in Qiime viewer (https://view.qiime2.org/) (Figure 2).



qiime2view File: taxonomy.qzv Visualization Details Provenance

Download metadata TSV file

This file won't necessarily reflect dynamic sorting or filtering options based on the interactive table below.

Search:

Feature ID (#q2:types)	Taxon (categorical)	Confidence (categorical)
00729db08743f73edc63c42da3a70f67	D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Enterobacteriales;D_4__Enterobacteriaceae;D_5__Escherichia-Shigella	0.9980533898944397
00bf00abdc44dadede91fc73d7b91124	D_0__Bacteria	0.8143111496704851
01cf603bea8471e34cfd4b7625475765	D_0__Bacteria	0.9194572783327566
02360a0b8a426209bb93023e43546623	D_0__Bacteria	0.7608149052373613
02ef3e2841957c9d860aa0feaf34e728	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Ruminococcaceae;D_5__Faecalibacterium;D_6__uncultured bacterium	0.7572394415121813
031feb7497881a4f4c27f26305b0f3d2	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Bacteroidaceae;D_5__Bacteroides	0.9999990741868405
037806e55bf24c39010100836b0bc9c7	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Rikenellaceae;D_5__Alistipes;D_6__Alistipes sp. CHKCI003	0.8725491255537873
0390562e970f2a47cf9ea0eb8bf87176	D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Enterobacteriales;D_4__Enterobacteriaceae;D_5__Escherichia-Shigella	0.9980381824319807
046297c2f14859dc54f041513ba56dbf	Unassigned	0.4023669076611952
052108edc0861d71d5c3d06507e96400	D_0__Bacteria;D_1__Firmicutes;D_2__Negativicutes;D_3__Selenomonadales;D_4__Veillonellaceae;D_5__Megamonas	0.9737735621907788
06d0b126bded28bf7504ce8d8554cce2	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Rikenellaceae;D_5__Alistipes	0.9974265576283341
080668968cbedc4ef52bb58e874edab4	D_0__Bacteria;D_1__Cyanobacteria;D_2__Melainabacteria;D_3__Gastranaerophilales	0.9999974583267778
0936a1d7ff07489e555690d0440e2e50	Unassigned	0.3862639597244207
094cb61236750e4b701ddea6ce3a05f4	Unassigned	0.3863297908026460
09afbe7c6c4d1b5b740599d9c5e32a98	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae	0.9991946344241504
0a3a06a914d8c1306f24af702f1eeaab	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae	0.9969125100086463
0bd1e791c5d9812242d92dcad7485db4	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Rikenellaceae	0.7517714380126523

Figure 2. Taxonomy representation of each samples.

Step 12 - Export files

All relevant files were then exported so to be useable in R with the package Phyloseq.

```
qiime tools export --input-path table.qza --output-path output
```

table.qza represents the ASV table. This command creates the feature-table.biom file (converted later to a .tsv file).

```
qiime tools export --input-path rep-seqs.qza --output-path output
```

rep-seqs.qza represents the ASV sequences. This command creates the dna-sequences.fasta file.

```
head dna-sequences.fasta
>78269ac3aa118a0e28b713d456b4bb64
CCTACGGGGGGCTGCGGGTGCCTCTGCAGTGCAGACTGGAGGCCAGCTGGCAGCTGCTGGC
TGTGATGTGTCCCTCTGCATCCTGCACAAAAGCTGGTCAGCCAGAGTGTCTGGCGTGAAGT
CAGGCCTTAAGGGAATCCTTCACTTATCCATGCGGCTGGAAAACCCTCCACTAAGGGACAG
CAACTGAAAAGTCAAAGACTAGAGGAAGTAAGGATGCCCTGTGTACAAGTTTAATAAGTAA
TCCCTTGTTATCGCATGCCTTGTTGAGTGATTGCAAGATACCCTAGTAGTC
>138ee4a5a0d2a54d491b1988b217f5cf
CCTACGGGTGGCTGGATTTCACAGAATCATAGAATGTTTGAGACTGGAAGGTAGCTCTGGA
GTCATCCAGTGCAGCTCCACTGCTCATACAGGGCATACGATTTATAGGATATCTGCAAAGT
CTCTGGGCAATCAAGCACTGTTCTAGCACTTGGTCGCCCACATAGTAAAGAAGTGTTTCCT
GATAATCAAATGGGAAGGGAATTTAGAGACCGCTTCTCACCTGAAGAGATTCTCCCATCAA
ATGGTGGAAGCAGCCTGTTCTGATACCCCAGTAGTC
```

```
qiime tools export --input-path rooted-tree.qza --output-path output
```

rooted-tree.qza represents the tree file. This command creates the tree.nwk file.

```
qiime tools export --input-path taxonomy.qza --output-path output
```

taxonomy.qza represents the taxonomy. This command creates the taxonomy.tsv file.

```
head taxonomy.tsv
FeatureID       Taxon   Confidence
```

```
78269ac3aa118a0e28b713d456b4bb64        Unassigned      0.64449069240967
138ee4a5a0d2a54d491b1988b217f5cf        Unassigned      0.4249239542234641
200a42907ed7172aff752af65fb96389        Unassigned      0.46796058992048906
5b361e44aa27e0f844c43b97aeaeb42f        D_0__Eukaryota  0.8664500589483518
e735a996bd77843a5f1105642b0585d0        D_0__Bacteria   0.8951819508985015
```

Step 13 - Create compatible biom file

To create the biom file compatible with R and Phyloseq, the biom file is initially converted to a .tsv file: -

```
biom convert -i feature-table.biom -o feature-table.tsv –to-tsv
```

```
head feature-table.tsv
```

```
# Constructed from biom file
#OTU ID Day10C  Day10T1 Day10T2 Day10T3 Day21C  Day21T1 Day21T2 Day21T3
78269ac3aa118a0e28b713d456b4bb64        0.0     0.0     0.0     0.0     0.0     0.0     86.0    0.0
138ee4a5a0d2a54d491b1988b217f5cf        0.0     0.0     0.0     0.0     0.0     0.0     70.0    0.0
200a42907ed7172aff752af65fb96389        0.0     0.0     0.0     0.0     0.0     0.0     69.0    0.0
```

A modification of the taxonomy.tsv column names was performed: -

```
sed -i s/Taxon/taxonomy/ taxonomy.tsv | sed -i s/Feature\ ID/FeatureID/ taxonomy.tsv
```

```
head taxonomy.tsv
```

```
FeatureID       Taxon   Confidence
78269ac3aa118a0e28b713d456b4bb64      Unassigned      0.64449069240967
138ee4a5a0d2a54d491b1988b217f5cf      Unassigned      0.4249239542234641
200a42907ed7172aff752af65fb96389      Unassigned      0.46796058992048906
5b361e44aa27e0f844c43b97aeaeb42f      D_0__Eukaryota 0.8664500589483518
e735a996bd77843a5f1105642b0585d0      D_0__Bacteria  0.8951819508985015
edfc9bdabae68a14e4c34703a735d831      Unassigned      0.3139517223515642
```

Finally, the ASV feature table was merged with the taxonomy: -

```
biom add-metadata \
  -i feature-table.tsv \
  -o feature_w_tax.biom \
  --observation-metadata-fp taxonomy.tsv \
  --observation-header FeatureID,taxonomy,Confidence \
  --sc-separated taxonomy --float-fields Confidence
```

Appendix II – Bioinformatics Steps for Qiime2 analysis with VSEARCH

All relevant pipelines were established by Dr Umer Zeeshan Ijaz (http://userweb.eng.gla.ac.uk/umer.ijaz/) at the University of Glasgow and training was provided accordingly (http://www.tinyurl.com/JCBioinformatics3).

Step 1 - Size checking

The file structure was organised as described in 2.1. Initially, before performing the VSEARCH pipeline, quality control steps were performed to significantly reduced error rates. FASTQ files were checked to view if quality trimming was required: -

```
for i in $(ls -d *); do echo $i":";bioawk -cfastx
'{i[length($seq)]++}END{for(j in i)print j","i[j]}'
${i}/Raw/*_R1_001.fastq | sort -nrk2 -t",";done
```

For each sample, the number of sequences of a given length was obtained. The output is given as [LENGTH],[FREQUENCY].

```
Day35T3:
301,326963
297,22489
300,1500
Day35T2:
301,278601
297,14540
300,1922
```

Step 2 - Trimming of poor quality reads

Sickle was used to perform quality trimming where a 20 bp long window was considered to trim the reads, when average quality score drops below 20 as well as read length below 10 bp.

```
for i in $(ls -d *); do cd $i;cd Raw; R1=$(ls *_R1_*.fastq);
R2=$(ls *_R2_*.fastq); cd .. ; sickle pe -f Raw/$R1 -r Raw/$R2
```

```
-o ${R1%.*}_trim.fastq -p ${R2%.*}_trim.fastq -s
${R1%.*}_singlet.fastq -q 20 -l 10 -t "sanger";cd ..; done
```

The distribution was checked: -

```
for i in $(ls -d *); do echo $i":";bioawk -cfastx
'{i[length($seq)]++}END{for(j in i)print j","i[j]}'
$i/*_R1_*trim.fastq | sort -nrk2 -t",";done | head -30
```

```
Day35T3:
301,112713
297,11050
Day35T2:
301,104473
297,7648
228,4521
```

Step 3 - Error correction

SPAdes assembler was used to perform error-correction for the paired-end reads: -

```
for i in $(ls */ -d); do cd $i; spades.py -1 *_R1_*trim.fastq
-2 *_R2_*trim.fastq -o . --only-error-correction --careful --
disable-gzip-output ; cd ..; done
```

To check if all relevant files were created, the following code was used: -

```
for i in $(ls -d *); do echo $i:; ls -1 $i; done
```

The expected folder structure was as follows: -

```
Day35T3:
Day35T3_S18_L001_R1_001_singlet.fastq
Day35T3_S18_L001_R1_001_trim.fastq
Day35T3_S18_L001_R2_001_trim.fastq
corrected
input_dataset.yaml
params.txt
Raw
spades.log
```

Step 4 - Overlap sequences

To complete the error correction, PANDAseq was used to overlap the paired-end reads using a minimum overlap of 10. SPAdes (used within PANDAseq) has a technical issue when using for error correction, due to the reads not having the identifier to distinguish the forward and reverse reads. Thus, a fictitious identifier using the command below: -

```
for i in $(ls */ -d); do cd $i; awk 'NR % 4==1{$0=$0"
1:N:0:GGACTCCTGTAAGGAG"}1'     corrected/*R1*.cor.fastq     >
corrected/forward_corrected.fastq;  awk  'NR  %  4==1{$0=$0"
2:N:0:GGACTCCTGTAAGGAG"}1'     corrected/*R2*.cor.fastq     >
corrected/reverse_corrected.fastq;pandaseq                 -f
corrected/forward_corrected.fastq                          -r
corrected/reverse_corrected.fastq    -B    -d    bfsrk    -A
simple_bayesian  -o  10  >  $(basename  ${i})".overlap.fasta";
cd ..; done
```

The output produces an overlap file: -

```
>M01637:109:000000000-AM9RR:1:1101:15566:1526:18
CCTACGGGGGGCAGCAGTGAGGAATATTGGTCAATGGACGCAAGTCTGAACCAGCCATGCC
GCGTGCAGGATGACGGCTCTATGAGTTGTAAACTGCTTTTGTACGAGGGTAAACGCAGATA
CGAGTATCTGTCTGAAAGTATCGTACGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCCT
GGTAATGCCTCCAAAACTGTTTGGCTAGAGAGTAGTTGCGGTAGGCGGAATGTATGGTGTA
GCGGTGAAATGCTTAGAGATCATGCAGAACACCGATTGCGAAGGCAGCTTACCAAACTATA
TCTGACGTTGAGGCACGAAAGCGTGGGGAGCAAACAGGATTAGATACCCCAGTAGTC
>M01637:109:000000000-AM9RR:1:1101:15961:1558:18
CCTACGGGAGGCAGCAGTCGGGAATATTGCGCAATGGAGGAAACTCTGACGCAGTGACGCC
GCGTGCAGGAAGAAGGTTTTCGGATTGTAAACTGCTTTAGACAGGGAAGAAAAAAGACAGT
ACCTGTAGAATAAGCTCCGGCTAACTACGTGCCAGCAGCCGCGGTGCTGGAGAGGGAGGTG
GAATTCCTAGTGTAGCGGTGAAATGCGTAGATATTAGGAGGAACACCAGTGGCGAAGGCGA
CTTTCTGGACAGTAACTGACGTTGAGGCACGAAAGTGTGGGGAGCAAACAGGATTAGATAC
CCCTGTAGTC
```

Of note, SPAdes can be run with simple bayesian, pear, rdp_mle or stitch options. For the purpose of the MSc project, I have used simple bayesian.

Step 5 - VSEARCH

The next step is to create a vsearch_tutorial folder at the same level as the sequences folder. Within the vsearch_tutorial folder, initially the .overlap.fasta files were combined in VSEARCH/USEARCH format: -

```
for i in $(ls -d ../sequences/*/); do awk -v k=$(basename ${i})
'/^>/{$0=">barcodelabel="k";S"(++i)}1' < $i/*.overlap.fasta;
done > multiplexed.fasta
```

The final labels were in USEARCH format: >barcodelabel=FolderName;SID. The sequences in each sample were given internal identifiers starting with S1, S2, and so on: -

The output produces an multiplexed.fasta file: -

```
>barcodelabel=Day10C;S1
CTACGGGGGGCAGCAGTGGGGAATATTGGGCAATGGGGGAAACCCTGACCCAGCAACGCCG
CGTGAAGGAAGAAGGCCTTCGGGTTGTAAACTTCTTTTACCAGGGACGAAGGACGTGACGG
TACCTGGAGAAAAAGCAACGGCTAACTATGTGCCAGCAGCCGCGGTAATACGTAGGTGGCA
AGCGTTGTCCGGATTTACTGGGTGTAAAGGGCGTGTAGGCGGAGCTGCAAGGTAGCGGTGA
AATCCGTAGGTATTAGGAGGAACACCAGTGGCGAAGGCGGCTTGCTGGACGACAACTGACG
CTGAGGCGCGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTTGTAGTC
>barcodelabel=Day10C;S2
ACTACGGGCGGCTGCAGTGGGGAATATTGCACAATGGAGGAAACTCTGATGCAGCGATGCC
GCGTGAGGGAAGAAGGTTTTCGGATTGTAAACCTCTGTCTTTGGGGACGAGAATGACGGTA
CCCAAGGAGGAAGCTCCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGAGCAGT
GGCGATGCGGACTTACTGGGCTTTAACTGACGCTGAGGCTCGAACGCGTGGGGAGCAAACA
GGATTAGATACCCCGGTAGTC
```

At this point, all unique reads from all of the multiplexed samples were within a single file. This file will be used at a later point to blast against the created list of OTUs.

Step 6 - Linearising the file

The next step was to linearise the FASTA file: -

```
awk 'NR==1 {print ; next} {printf /^>/ ? "\n"$0"\n" : $1} END
{print}' multiplexed.fasta > multiplexed_linearized.fasta
```

Step 7 - Dereplicate, sort and remove singletons

This was then followed by dereplication, sorting, and removing singletons: -

```
vsearch --threads 20 --derep_fulllength
multiplexed_linearized.fasta --minuniquesize 2 --sizein --
sizeout --fasta_width 0 --uc
multiplexed_linearized_dereplicated_vsearch_min2.uc --output
multiplexed_linearized_dereplicated_vsearch_min2.fasta
```

The output produces a multiplexed_linearized_dereplicated_vsearch_min2.fasta file: -

```
>barcodelabel=Day21C;S1157;size=5702
CCTACGGGAGGCAGCAGTGAGGAATATTGGTCAATGGGCGAGAGCCTGAACCAGCCAAGTA
GCGTGAAGGATGACTGCCCTATGGGTTGTAAACTTCTTTTATAAAGGAATAAAGTCGGGTA
TGGATACCCGTTTGCATGTACTTTATGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGC
GGTAATACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGATGGA
TGTTTAAGTCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGGATAT
CTTGAGTGCAGTTGAGGCAGGCGGAATTCGTGGTGTAGCGGTGAAATGCTTAGATATCACG
AAGAACTCCGATTGCGAAGGCAGCCTGCTAAGCTGCAACTGACATTGAGGCTCGAAAGTGT
GGGTATCAAACAGGATTAGATACCCCAGTAGTC
>barcodelabel=Day21C;S1249;size=5292
CCTACGGGAGGCAGCAGTGAGGAATATTGGTCAATGGGCGAGAGCCTGAACCAGCCAAGTA
GCGTGAAGGATGACTGCCCTATGGGTTGTAAACTTCTTTTATAAAGGAATAAAGTCGGGTA
TGGATACCCGTTTGCATGTACTTTATGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGC
GGTAATACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGATGGA
TGTTTAAGTCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGGATAT
CTTGAGTGCAGTTGAGGCAGGCGGAATTCGTGGTGTAGCGGTGAAATGCTTAGATATCACG
AAGAACTCCGATTGCGAAGGCAGCCTGCTAAGCTGCAACTGACATTGAGGCTCGAAAGTGT
GGGTATCAAACAGGATTAGATACCCTGGTAGTC
```

Step 8 - Clustering

The next step was to perform clustering of the sequences at 97% similarity.

```
vsearch --threads 20 --cluster_size
multiplexed_linearized_dereplicated_vsearch_min2.fasta --id
0.97 --strand both --sizein --sizeout --fasta_width 0 --uc
multiplexed_linearized_dereplicated_vsearch_min2_preclustered
.uc --centroids
multiplexed_linearized_dereplicated_vsearch_min2_preclustered
.fasta
```

The output produces an mmultiplexed_linearized_dereplicated_vsearch_min2_preclustered.fasta file: -

```
>barcodelabel=Day21C;S1157;size=296516
CCTACGGGAGGCAGCAGTGAGGAATATTGGTCAATGGGCGAGAGCCTGAACCAGCCAAGTA
GCGTGAAGGATGACTGCCCTATGGGTTGTAAACTTCTTTTATAAAGGAATAAAGTCGGGTA
TGGATACCCGTTTGCATGTACTTTATGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGC
GGTAATACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGATGGA
TGTTTAAGTCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGGATAT
CTTGAGTGCAGTTGAGGCAGGCGGAATTCGTGGTGTAGCGGTGAAATGCTTAGATATCACG
AAGAACTCCGATTGCGAAGGCAGCCTGCTAAGCTGCAACTGACATTGAGGCTCGAAAGTGT
GGGTATCAAACAGGATTAGATACCCCAGTAGTC
>barcodelabel=Day10C;S8597;size=111114
CCTACGGGAGGCAGCAGTGGGGAATATTGCACAATGGGGGAAACCCTGATGCAGCGACGCC
GCGTGGAGGAAGAAGGTCTTCGGATTGTAAACTCCTGTTGTTGGGGAAAAGAAGGATGGTA
CCCAACAAGGAAGTGACGGCTAACTACGTGCCAGCAGCCGCGGTAAAACGTAGGTCACGAG
CGTTGTCCGGAATTACTGGGTGTAAAGGGAGCGCAGGCGGGTATGCAAGTTGGGAGTGAAA
TACATGGGCTCAACCCATGAACTGCTCTCAAAACTGTGTATCTTGAGTAGTGCAGAGGTAG
GCGGAATTCCCGGTGTAGCGGTGGAATGCGTAGATATCGGGAGGAACACCAGTGGCGAAGG
CGGCCTACTGGGCACCAACTGACGCTGAGGCTCGAAAGTGTGGGTAGCAAACAGGATTAGA
TACCCCAGTAGTC
```

Step 9 - *De novo* chimera removal

Once clustering was completed, *de novo* chimera removal was performed: -

```
vsearch --threads 20 --uchime_denovo
multiplexed_linearized_dereplicated_vsearch_min2_preclustered
.fasta --sizein --sizeout --fasta_width 0 --nonchimeras
multiplexed_linearized_dereplicated_vsearch_min2_preclustered
_nonchimeras.fasta
```

Step 10 - Reference based chimera removal

Reference based chimera removal was also performed as there are typically chimeras missed from the previous step (if they have parents that are absent from the reads or are present with very low abundance). Thus, a reference database is used called the gold database: -

```
vsearch --threads 20 --uchime_ref
multiplexed_linearized_dereplicated_vsearch_min2_preclustered
_nonchimeras.fasta --db ~/Downloads/gold.fasta --sizein --
```

```
sizeout --fasta_width 0 –nonchimeras
multiplexed_linearized_dereplicated_vsearch_min2_preclustered
_nonchimeras_ref.fast
```

Step 11 - Renaming of sequences

Sequences were renamed to begin with “OTU_”

```
python ~/bin/fasta_number.py
multiplexed_linearized_dereplicated_vsearch_min2_preclustered
_nonchimeras_ref.fasta OTU_ > otus.fa
```

The output produces an otus.fa file: -

```
>OTU_1
CCTACGGGAGGCAGCAGTGAGGGATATTGGTCAATGGGGGAAACCCTGAACCAGCAACGCC
GCGTGAGGGATGACGGCCTTCGGGTTGTAAACCTCTGTCCTCTGTGAAGATAATGACGGTA
GCAGAGGAGGAAGCTCCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGAGCAAG
CGTTGTCCGGATTTACTGGGTGTAAAGGGTGCGTAGGCGGTTTGGTAAGTCAGAAGTGAAA
TCCATGGGCTTAACCCATGAACTGCTTTTGAAACTATCGAACTTGAGTGAAGTAGAGGTAG
GCGGAATTCCCGGTGTAGCGGTGAAATGCGTAGAGATCGGGAGGAACACCAGTGGCGAAGG
CGGCCTACTGGGCTTTAACTGACGCTGAGGCACGAAAGCATGGGTAGCAAACAGGATTAGA
TACCCCAGTAGTC
>OTU_2
CCTACGGGAGGCAGCAGTGAGGAATATTGGTCAATGGACGCAAGTCTGAACCAGCCATGCC
GCGTGCAGGATGACGGCTCTATGAGTTGTAAACTGCTTTTGTACGAGGGTAAACGCAGATA
CGAGTATCTGTCTGAAAGTATCGTACGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGC
GGTAATACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGTGCGTAGGCGGT
TCGATAAGTTGGAGGTGAAATGTTAGGGCTTAACCCTGAAACTGCCTCCAATACTGTTGGG
CTAGAGAGTAGTTGCGGTAGGCGGAATGTATGGTGTAGCGGTGAAATGCTTAGAGATCATA
CAGAACACCGATTGCGAAGGCAGCTTACCAAACTATATCTGACGTTGAGGCACGAAAGCGT
GGGGAGCAAACAGGATTAGATACCCCAGTAGTC
```

Step 12 - Search against newly created OTUs

The original multiplexed sequences were searched against the OTUs just created: -

```
vsearch --threads 20 --usearch_global
multiplexed_linearized.fasta --db otus.fa --strand both --id
0.97 --uc map.uc
```

Step 13 - Create OTU tab-delimited format

An OTU table in a tab-delimited format is created: -

```
python ~/bin/uc2otutab.py map.uc > otu_table.txt
```

The output produces an otus_table.txt file: -

OTUId	Day10C	Day10T1	Day10T2	Day10T3	Day21C	Day21T1	Day21T2	Day21T3
OTU_127	575	229	6	335	2	99	159	131
OTU_278	346	7	0	0	0	0	0	16
OTU_23	834	552	0	9276	0	604	0	769
OTU_364	267	341	1394	1950	1568	570	397	822

Step 14 - Minor formatting of OTUs

At this point, we have obtained the otus.fa (sequences) and otu_table.txt (abundance table). The following steps can be performed within Qiime2. For this purpose, we convert all the sequences to their uppercase representation to function in Qiime2: -

```
bioawk -cfastx '{print ">"$name"\n"toupper($seq)}' otus.fa > otus_upper.fa
```

Step 15 - Import into Qiime2

The sequence file was imported Qiime2's qza format: -

```
qiime tools import --type 'FeatureData[Sequence]' --input-path otus_upper.fa --output-path otus.qza
```

Step 16 - Generate taxonomy

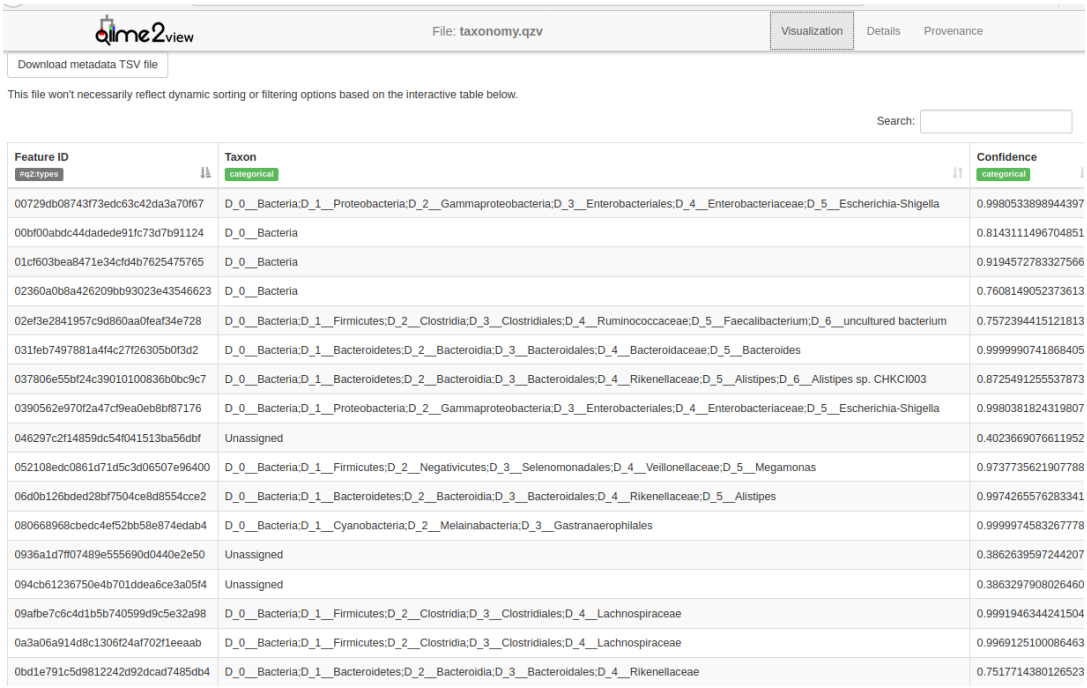
The taxonomy was generated using Qiime2 and Silva132 database: -

```
qiime feature-classifier classify-sklearn --i-classifier /software/qiime2_databases/silva-132-99-nb-classifier.qza --i-reads otus.qza --o-classification taxonomy.qza
```

To visualise the output, the .qza file was converted into a .qzv as follows: -

```
qiime metadata tabulate --m-input-file taxonomy.qza --o-visualization taxonomy.qzv
```

taxonomy.qzv was placed in Qiime viewer (https://view.qiime2.org/) (Figure 3).



qiime2view File: taxonomy.qzv Visualization Details Provenance

Download metadata TSV file

This file won't necessarily reflect dynamic sorting or filtering options based on the interactive table below.

Search:

Feature ID #q2:types	Taxon categorical	Confidence categorical
00729db08743f73edc63c42da3a70f67	D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Enterobacteriales;D_4__Enterobacteriaceae;D_5__Escherichia-Shigella	0.9980533898944397
00bf00abdc44dadede91fc73d7b91124	D_0__Bacteria	0.8143111496704851
01cf603bea8471e34cfd4b7625475765	D_0__Bacteria	0.9194572783327566
02360a0b8a426209bb93023e43546623	D_0__Bacteria	0.7608149052373613
02ef3e2841957c9d860aa0feaf34e728	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Ruminococcaceae;D_5__Faecalibacterium;D_6__uncultured bacterium	0.7572394415121813
031feb7497881a4f4c27f26305b0f3d2	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Bacteroidaceae;D_5__Bacteroides	0.9999990741868405
037806e55bf24c39010100836b0bc9c7	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Rikenellaceae;D_5__Alistipes;D_6__Alistipes sp. CHKCI003	0.8725491255537873
0390562e970f2a47cf9ea0eb8bf87176	D_0__Bacteria;D_1__Proteobacteria;D_2__Gammaproteobacteria;D_3__Enterobacteriales;D_4__Enterobacteriaceae;D_5__Escherichia-Shigella	0.9980381824319807
046297c2f14859dc54f041513ba56dbf	Unassigned	0.4023669076611952
052108edc0861d71d5c3d06507e96400	D_0__Bacteria;D_1__Firmicutes;D_2__Negativicutes;D_3__Selenomonadales;D_4__Veillonellaceae;D_5__Megamonas	0.9737735621907788
06d0b126bded28bf7504ce8d8554cce2	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Rikenellaceae;D_5__Alistipes	0.9974265576283341
080668968cbedc4ef52bb58e874edab4	D_0__Bacteria;D_1__Cyanobacteria;D_2__Melainabacteria;D_3__Gastranaerophilales	0.9999974583267778
0936a1d7ff07489e555690d0440e2e50	Unassigned	0.3862639597244207
094cb61236750e4b701ddea6ce3a05f4	Unassigned	0.3863297908026460
09afbe7c6c4d1b5b740599d9c5e32a98	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae	0.9991946344241504
0a3a06a914d8c1306f24af702f1eeaab	D_0__Bacteria;D_1__Firmicutes;D_2__Clostridia;D_3__Clostridiales;D_4__Lachnospiraceae	0.9969125100086463
0bd1e791c5d9812242d92dcad7485db4	D_0__Bacteria;D_1__Bacteroidetes;D_2__Bacteroidia;D_3__Bacteroidales;D_4__Rikenellaceae	0.7517714380126523

Figure 3. Taxonomy representation of each samples.

Step 17 - Generate phylogeny

The phylogenetic tree was generated using Qiime2

```
qiime phylogeny align-to-tree-mafft-fasttree --i-sequences otus.qza --o-alignment aligned-otus.qza --o-masked-alignment masked-aligned-otus.qza --p-n-threads 0 --o-tree unrooted-tree.qza --o-rooted-tree rooted-tree.qza
```

Step 18 - Qiime export

The data files were converted from Qiime2 format to native format in an "output" folder: -

```
qiime tools export --input-path rooted-tree.qza --output-path output
```

rooted-tree.qza represents the tree file. This command creates the tree.nwk file.

```
qiime tools export --input-path taxonomy.qza --output-path output
```

taxonomy.qza represents the taxonomy. This command creates the taxonomy.tsv file.

```
head taxonomy.tsv
```

```
FeatureID       Taxon   Confidence
78269ac3aa118a0e28b713d456b4bb64        Unassigned      0.64449069240967
138ee4a5a0d2a54d491b1988b217f5cf        Unassigned      0.4249239542234641
200a42907ed7172aff752af65fb96389        Unassigned      0.46796058992048906
```

```
5b361e44aa27e0f844c43b97aeaeb42f        D_0__Eukaryota  0.8664500589483518
e735a996bd77843a5f1105642b0585d0        D_0__Bacteria   0.8951819508985015
```

Step 19 - Create compatible biom file

Within the output folder, a biom file was created using the otu_table created from VSEARCH.

A modification of the taxonomy.tsv column names was performed: -

```
sed -i s/Taxon/taxonomy/ taxonomy.tsv | sed -i s/Feature\ ID/FeatureID/ taxonomy.tsv
```

```
head taxonomy.tsv
```

```
FeatureID       Taxon   Confidence
78269ac3aa118a0e28b713d456b4bb64      Unassigned      0.64449069240967
138ee4a5a0d2a54d491b1988b217f5cf      Unassigned      0.4249239542234641
200a42907ed7172aff752af65fb96389      Unassigned      0.46796058992048906
5b361e44aa27e0f844c43b97aeaeb42f      D_0__Eukaryota 0.8664500589483518
e735a996bd77843a5f1105642b0585d0      D_0__Bacteria  0.8951819508985015
edfc9bdabae68a14e4c34703a735d831      Unassigned      0.3139517223515642
```

```
biom add-metadata -i otu_table.txt -o feature_w_tax.biom --observation-metadata-fp taxonomy.tsv --observation-header FeatureID,taxonomy,Confidence --sc-separated taxonomy --float-fields Confidence
```

Appendix III –16S microbiome results using VSEARCH pipeline

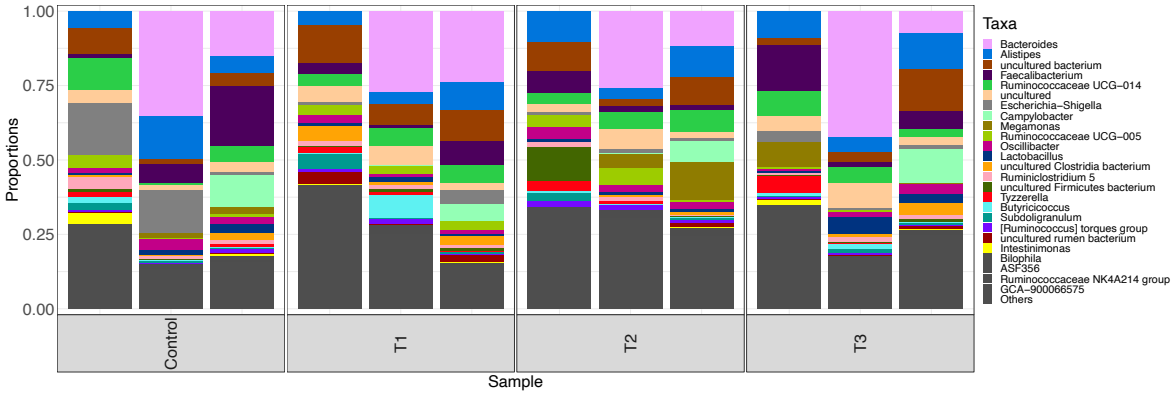


Figure 1. The top 25 most abundant genera representative of different sample categories. Carvacrol concentration of different groups (Control = 0 mg/ml, T1 = 120 mg/ml, T2 = 200 mg/ml and T3 = 300 mg/ml).

Table 1. Taxa differential of OTUs statistically modified when comparing groups Control vs. T1. These are log 2-fold different and statistically significant.

	baseMean	log2FoldChange	pvalue	padj	Upregulated
OTU_46 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group	359.4870054	8.836228407	4.75E-12	2.96E-09	T1
OTU_181 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae	757.5490905	10.38596537	1.08E-10	3.37E-08	T1
OTU_128 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group;uncultured bacterium;uncultured bacterium	365.2228583	-9.631611317	3.91E-10	8.11E-08	Control
OTU_59 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group;uncultured bacterium;uncultured bacterium	700.2597346	10.25753932	5.54E-09	8.63E-07	T1
OTU_124 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group;uncultured bacterium;uncultured bacterium	254.7923111	8.817858331	1.36E-08	1.69E-06	T1
OTU_26 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group;uncultured bacterium;uncultured bacterium	703.2553546	8.706264144	2.32E-07	2.41E-05	T1
OTU_162 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Ruminococcaceae UCG-014	78.86194554	6.654208237	6.76E-07	6.02E-05	T1
OTU_1879 Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae	176.7917132	8.271974134	1.32E-06	9.14E-05	T1
OTU_146 Bacteria;Cyanobacteria;Melainabacteria;Gastranaerophilales;uncultured bacterium;uncultured bacterium;uncultured bacterium	168.1420102	7.494701387	1.31E-06	9.14E-05	T1
OTU_78 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group	179.4709694	6.907459294	1.53E-06	9.53E-05	T1
OTU_537 Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae	555.55533	6.880969495	3.76E-06	0.000193517	T1
OTU_261 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Candidatus Soleaferrea;uncultured bacterium	74.89620811	7.044708783	4.04E-06	0.000193517	T1
OTU_69 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group;uncultured bacterium;uncultured bacterium	113.9978212	5.50028705	3.96E-06	0.000193517	T1
OTU_197 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group;uncultured bacterium;uncultured bacterium	55.96346664	5.54695372	7.20E-06	0.000320605	T1
OTU_208 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Ruminococcaceae UCG-014	98.82994652	7.429867818	1.23E-05	0.000478754	T1
OTU_246 Bacteria;Tenericutes;Mollicutes;Mollicutes RF39	48.06300818	-6.695434152	1.19E-05	0.000478754	Control
OTU_3 Bacteria;Firmicutes;Negativicutes;Selenomonadales;Veillonellaceae;Megamonas	1034.015558	-7.120764239	1.59E-05	0.000582223	Control
OTU_524 Bacteria;Firmicutes;Negativicutes;Selenomonadales;Veillonellaceae;Megamonas	63.63193725	-6.221310504	2.06E-05	0.000712105	Control
OTU_256 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae	112.4286997	6.124492464	3.36E-05	0.001101554	T1
OTU_18 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group;uncultured bacterium;uncultured bacterium	2757.065407	5.408838121	4.86E-05	0.001352398	T1
OTU_1140 Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae	39.76178545	6.122199986	4.57E-05	0.001352398	T1
OTU_635 Bacteria;Firmicutes;Negativicutes;Selenomonadales;Veillonellaceae;Megamonas;unidentified	61.61797005	-5.867269926	4.99E-05	0.001352398	Control
OTU_273 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group;uncultured bacterium;uncultured bacterium	97.2037865	7.398533461	4.63E-05	0.001352398	T1
OTU_2074 Bacteria;Firmicutes;Negativicutes;Selenomonadales;Veillonellaceae;Megamonas;unidentified	27.95811034	-5.915859256	5.73E-05	0.001487408	Control
OTU_2832 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group;uncultured bacterium;uncultured bacterium	43.86077561	6.265155508	7.96E-05	0.001984404	T1
OTU_203 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group;uncultured bacterium;uncultured bacterium	29.72108105	5.712583249	9.17E-05	0.002198399	T1
OTU_86 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Ruminococcaceae UCG-014;uncultured bacterium	40.5765647	4.070354164	0.000115025	0.002654097	T1
OTU_623 Bacteria;Firmicutes;Negativicutes;Selenomonadales;Veillonellaceae;Megamonas	55.44706986	-5.456210877	0.00014128	0.003143471	Control
OTU_179 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group;uncultured bacterium;uncultured bacterium	30.22928853	5.728821134	0.000160418	0.003446216	T1
OTU_95 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Hydrogenoanaerobacterium;uncultured bacterium	161.3732772	4.859019595	0.000180759	0.003699548	T1
OTU_164 Bacteria;Cyanobacteria;Melainabacteria;Gastranaerophilales;uncultured bacterium;uncultured bacterium;uncultured bacterium	197.8493072	6.234527413	0.000184087	0.003699548	T1
OTU_64 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Ruminococcaceae UCG-014;uncultured bacterium	133.9475113	5.247937683	0.000191039	0.003719283	T1
OTU_180 Bacteria;Firmicutes;Clostridia;Clostridiales;Defluviitaleaceae;Defluviitaleaceae UCG-011;uncultured bacterium	103.0240596	-4.344260525	0.000207126	0.003910297	Control
OTU_188 Bacteria;Firmicutes;Erysipelotrichia;Erysipelotrichales;Erysipelotrichaceae;uncultured;Clostridiales bacterium 60-7e	105.1290648	-4.689277038	0.000247864	0.004063673	Control
OTU_161 Bacteria;Cyanobacteria;Melainabacteria;Gastranaerophilales;uncultured rumen bacterium;uncultured rumen bacterium;uncultured rumen bacterium	113.089815	4.653689764	0.000231031	0.004063673	T1
OTU_379 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group;uncultured bacterium;uncultured bacterium	56.80060266	5.919754045	0.000236291	0.004063673	T1
OTU_763 Bacteria;Firmicutes;Negativicutes;Selenomonadales;Veillonellaceae;Megamonas;unidentified	29.77207145	-5.113598094	0.000233762	0.004063673	Control
OTU_368 Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae;[Eubacterium] hallii group;uncultured bacterium	149.4122344	-6.314204911	0.000246073	0.004063673	Control
OTU_371 Bacteria;Firmicutes;Erysipelotrichia;Erysipelotrichales;Erysipelotrichaceae	26.27167321	4.476113404	0.000293391	0.004686733	T1
OTU_42 Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae;ASF356;uncultured bacterium	1500.636474	6.551234131	0.000311675	0.004854345	T1
OTU_493 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;[Eubacterium] coprostanoligenes group;uncultured bacterium	31.6275143	5.77451258	0.00033498	0.005090067	T1
OTU_1167 Bacteria;Firmicutes;Negativicutes;Selenomonadales;Veillonellaceae;Megamonas;unidentified	25.04446518	-4.860312092	0.000397153	0.005891098	Control
OTU_448 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;uncultured	25.0876437	5.436251753	0.000449218	0.006508443	T1
OTU_286 Bacteria;Tenericutes;Mollicutes;Mollicutes RF39	9.790690318	-4.369935252	0.000832705	0.011790339	Control
OTU_227 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group;uncultured bacterium;uncultured bacterium	27.25410647	5.560618936	0.000892141	0.012351197	T1
OTU_289 Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae	17.06889556	-3.484828989	0.00091875	0.012443067	Control
OTU_856 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Ruminococcaceae UCG-014	10.92796346	4.226154915	0.00095237	0.012623963	T1
OTU_329 Bacteria;Tenericutes;Mollicutes;Mollicutes RF39;uncultured bacterium;uncultured bacterium;uncultured bacterium	32.77844256	3.891800272	0.001025877	0.01278243	T1
OTU_2054 Bacteria;Firmicutes;Negativicutes;Selenomonadales;Veillonellaceae;Megamonas	18.54788296	-4.414073785	0.001000253	0.01278243	Control
OTU_2066 Bacteria;Firmicutes;Negativicutes;Selenomonadales;Veillonellaceae;Megamonas;unidentified	14.0913368	-4.405198346	0.001010796	0.01278243	Control
OTU_172 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;uncultured;Gorbachella massiliensis	23.45421835	-4.928384882	0.001082415	0.013222448	Control
OTU_62 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Ruminococcaceae UCG-014	311.3910454	-4.00484041	0.001134124	0.013587682	Control
OTU_98 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae	131.4835288	4.842290223	0.001355939	0.015938683	T1
OTU_218 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group;uncultured bacterium;uncultured bacterium	13.98796626	4.583466906	0.001509821	0.017418865	T1
OTU_654 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;uncultured	29.70024246	-3.36531838	0.001595393	0.01774875	Control
OTU_353 Bacteria;Firmicutes;Clostridia;Clostridiales;Lachnospiraceae	9.541140466	-4.33175531	0.001579869	0.01774875	Control
OTU_136 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Ruminococcus 2;bacterium ic1277	135.7489279	3.842573392	0.001859487	0.02032387	T1
OTU_232 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group;gut metagenome;gut metagenome	13.98430468	-4.855638863	0.001959101	0.021043446	Control
OTU_567 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group;gut metagenome;gut metagenome	10.25566404	-4.09589687	0.002485428	0.026244436	Control
OTU_347 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Ruminococcaceae UCG-014	9.802070793	4.088152394	0.002652304	0.02753976	T1
OTU_29 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Butyricicoccus	2942.52745	4.663560073	0.002775397	0.028345449	T1
OTU_137 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Ruminococcaceae UCG-014	107.8584013	3.70333989	0.003363502	0.033797766	T1
OTU_698 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Ruminiclostridium 9	7.890297161	-3.504753349	0.003518945	0.034798454	Control
OTU_311 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae	8.362676932	-3.841546722	0.003614542	0.035185304	Control
OTU_82 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Ruminococcaceae UCG-014	322.100365	3.987748106	0.004703048	0.045076908	T1
OTU_2463 Bacteria;Cyanobacteria;Melainabacteria;Gastranaerophilales;uncultured rumen bacterium;uncultured rumen bacterium;uncultured rumen bacterium	97.78205813	3.450998477	0.004884835	0.046109877	T1
OTU_135 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group;uncultured bacterium;uncultured bacterium	130.556335	3.525710661	0.005064061	0.046659033	T1
OTU_187 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Ruminococcus 1	56.24839127	3.157938672	0.0050928	0.046659033	T1
OTU_456 Bacteria;Firmicutes;Clostridia;Clostridiales;Clostridiales vadinBB60 group;uncultured bacterium;uncultured bacterium	11.89353485	4.349956832	0.00529393	0.047798816	T1
OTU_376 Bacteria;Firmicutes;Clostridia;Clostridiales;Ruminococcaceae;Ruminococcaceae UCG-014	24.64506117	3.887079898	0.005512374	0.049060132	T1
OTU_2800 Bacteria;Cyanobacteria;Melainabacteria;Gastranaerophilales;uncultured rumen bacterium;uncultured rumen bacterium;uncultured rumen bacterium	19.7603462	3.18033276	0.005628307	0.049386411	T1

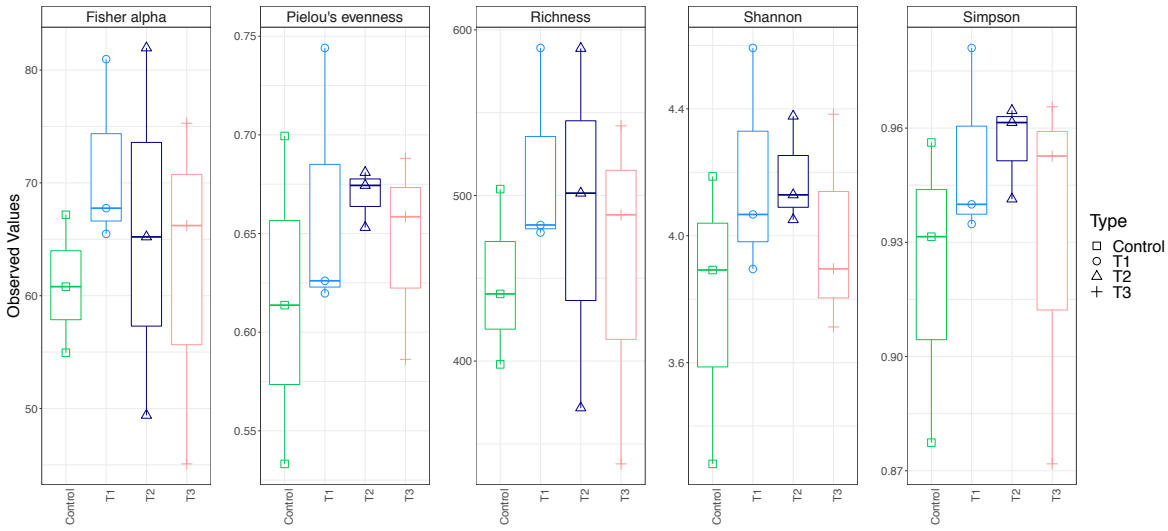
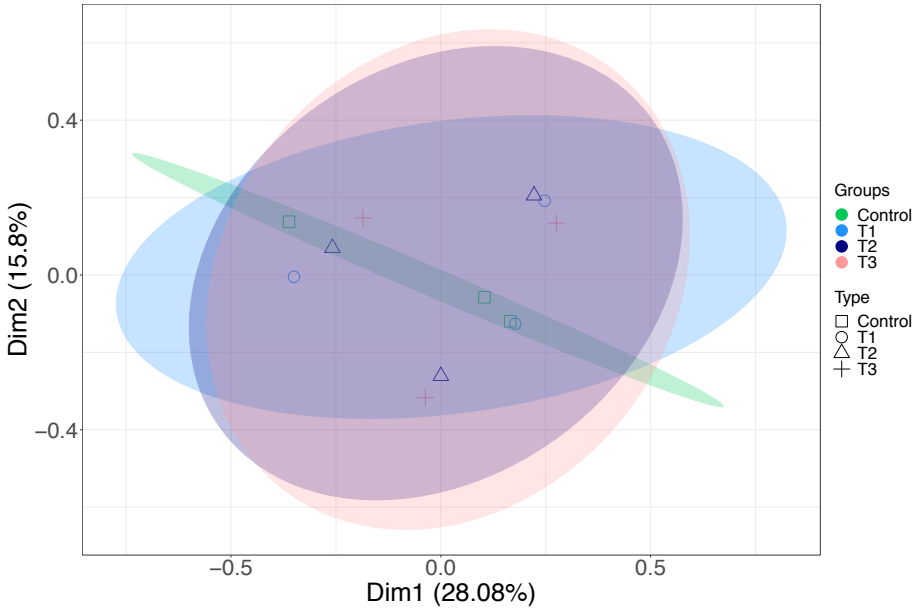
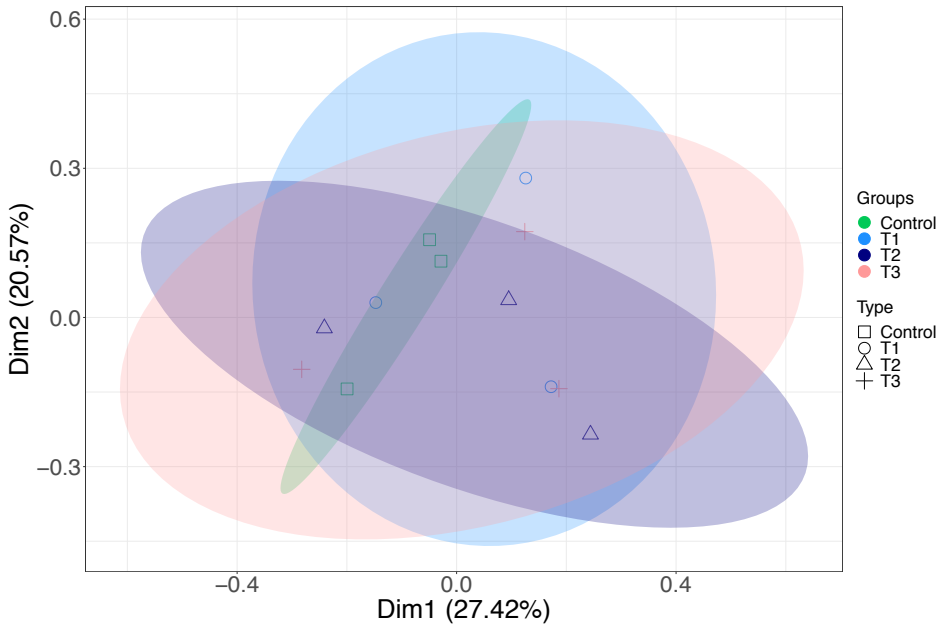


Figure 2. Alpha diversity metrics for Control, T1, T2 and T3 samples using VSEARCH pipeline. *Richness*, estimated number of species/features per sample; *Shannon* entropy measured the balance of a community within a sample; *Pielou's* index represents the evenness of a community; *Simpson* measures evenness of the community from 0 to 1, and *Fisher alpha* an alternative diversity index. Carvacrol concentration of different groups (Control = 0 mg/ml, T1 = 120 mg/ml, T2 = 200 mg/ml and T3 = 300 mg/ml).

(A)



(B)



(C)

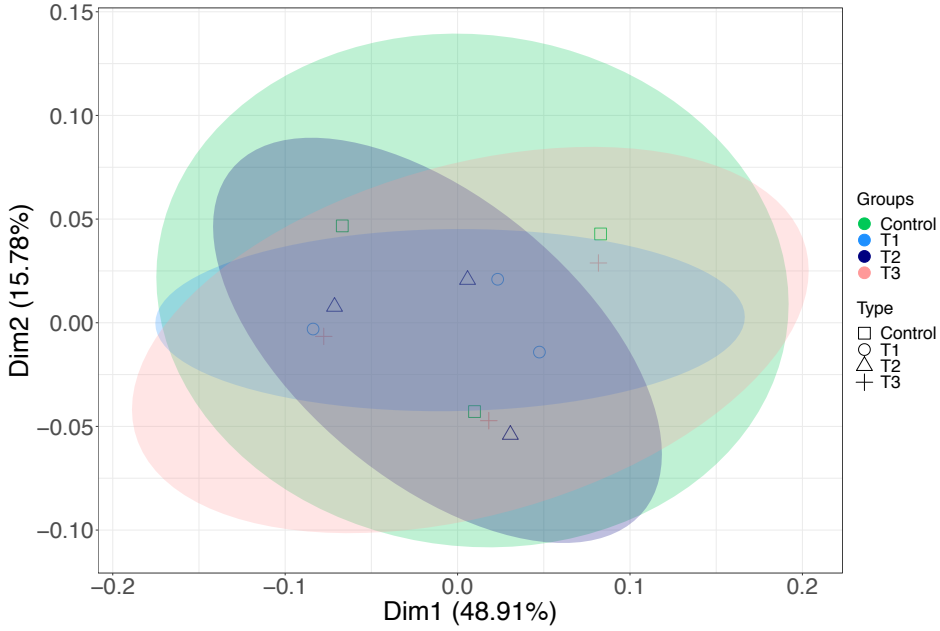


Figure 3. PCoA of beta diversity was measured for all groups using: (A) *Bray-Curtis*; (B) unweighted *UniFrac*; (C) weighted *UniFrac*. Two samples if similar lie very close to each other. The ellipses represent the standard error in terms of grouping variations.

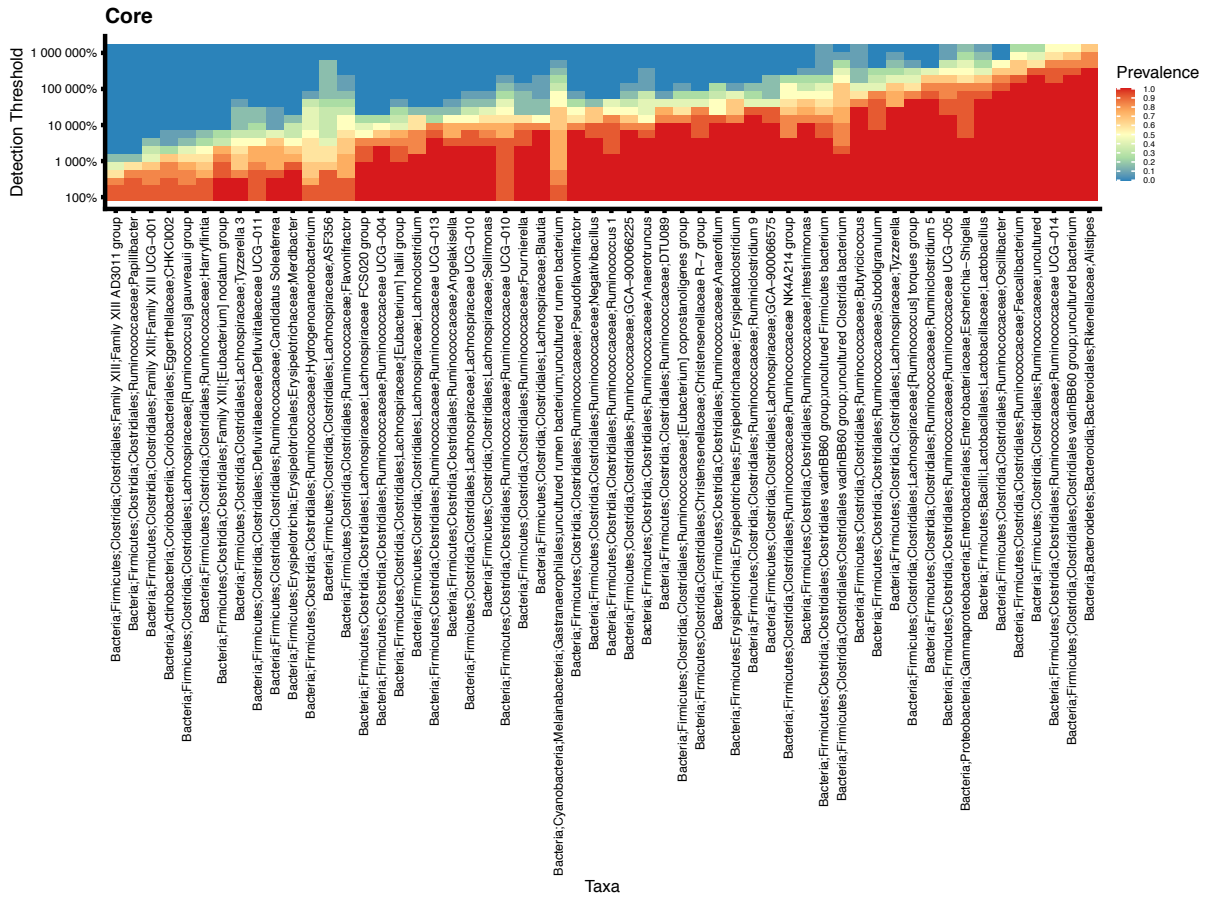


Figure 4. Core microbiome analysis of all the samples. The figure shows at least 85% prevalent OTUs ordered by lower abundance (left) to higher abundance (right). The y-axis (right) represents limit of detection in terms of relative abundance.

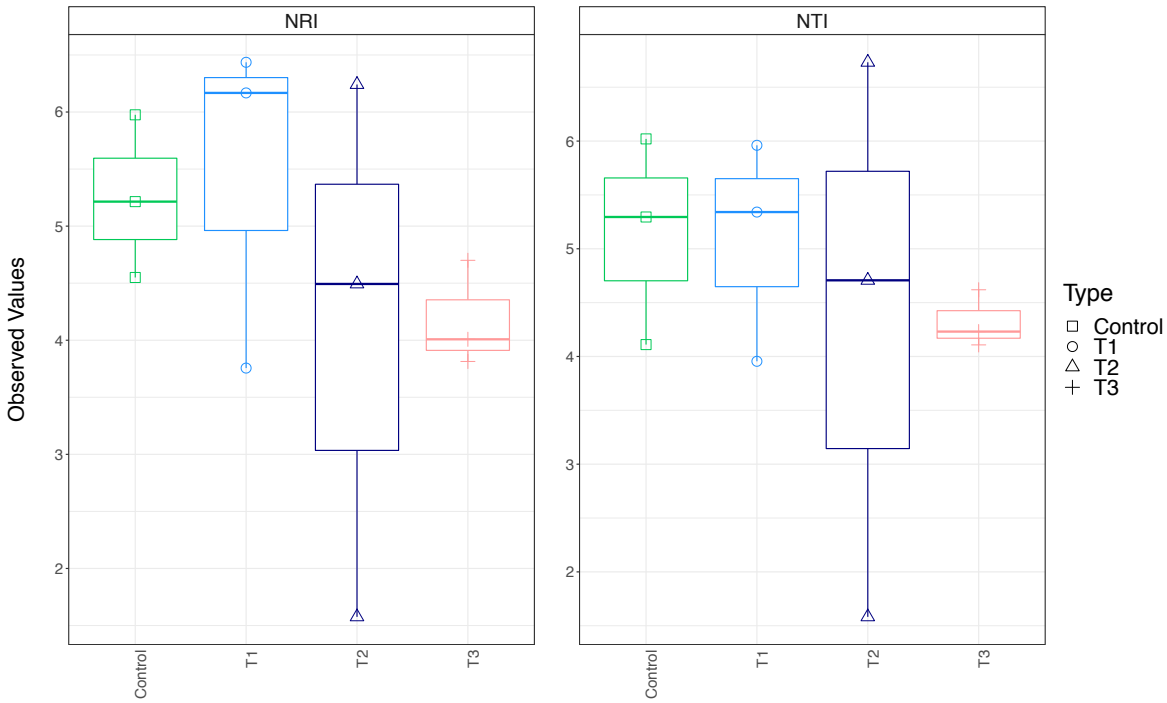
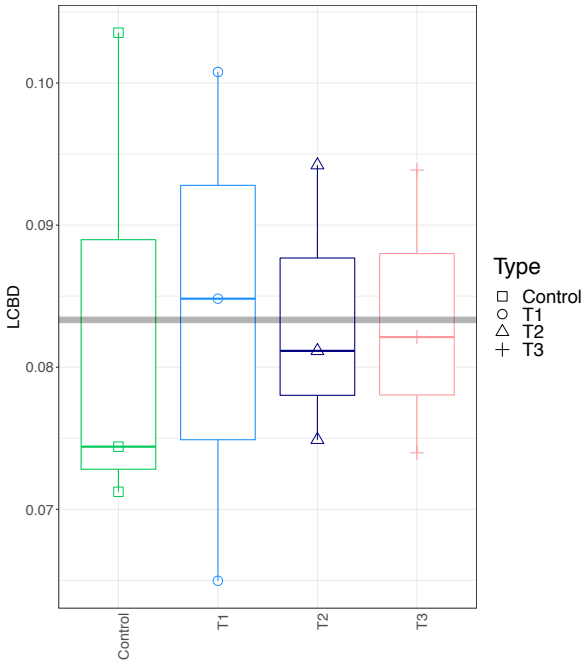
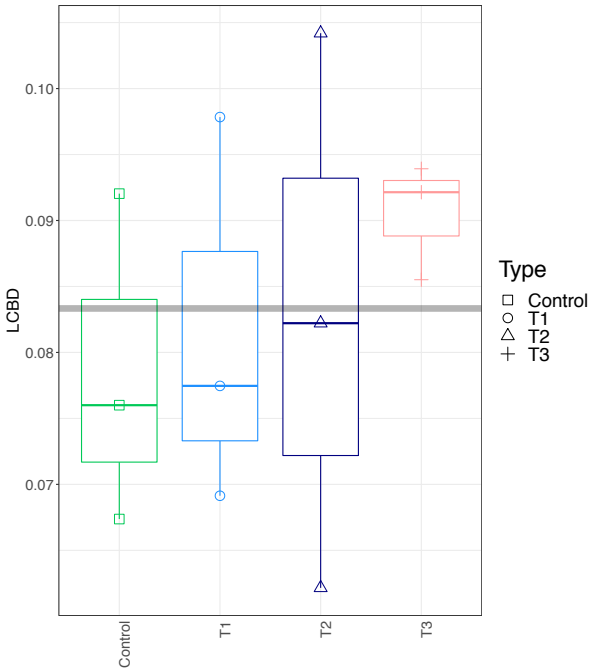


Figure 5. Investigating the environmental pressure on microbial community structure using NRI/NTI. NRI reflects the phylogenetic clustering in a broad sense (whole phylogenetic tree) with the lower values representing evenly spread community. NTI focuses more on the tips of the tree with positive values of NTI indicating that species co-occur with more closely related species than expected, and lower values indicating that closely related species do not co-occur by chance.

(A)



(B)



(C)

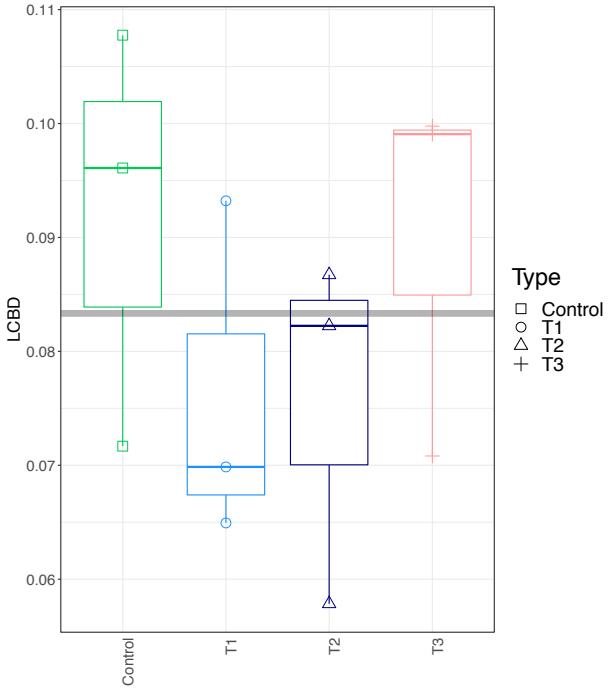


Figure 6. Local contribution to beta diversity (LCBD) calculated by using sample wise proportional diversities: (A) *Hellinger* transform on the microbial counts; (B) unweighted *UniFrac* dissimilarity (phylogenetic distances only); and (C) weighted *UniFrac* dissimilarity (phylogenetic distances weighted with abundance counts), with all values summing up to 1.

Table 2. Subset analysis showing top two subsets of OTUs along with the correlation of the beta diversity distances between these subsets and full OTU table. The last column shows PERMANOVA statistics for these subsets highlighting their discriminatory power. R^2 is the

percentage variability of these subsets in terms of groups. In the interest of space, only one group comparison is displayed, with remaining results within the accompanied files.

Group Comparison	Subset No	Subset	Correlation of Subset with Full Table (R)	PERMANOVA Subsets (Groups)
Control, T1	S1	Escherichia-Shigella + Ruminiclostridium 5	0.18041	R^2 = 0.4528 ($p > 0.05$)
	S2	Escherichia-Shigella + Ruminiclostridium 5 + Ruminococcaceae UCG-014	0.17201	R^2 = 0.4694 ($p > 0.05$)

	Richness	FisherAlpha	Simpson	Pielou	Shannon Entropy		LCBD (Bray-Curtis Distance)	LCBD (Unweighted UniFrac)	LCBD (Weighted UniFrac)
ControlConc									
Status_C									
Status_T1	+*								
Status_T2									
Status_T3									
Day_10	-***	-*					+***		
Day_21	-*		-**	-*	-*				
Day_35		+**							

Figure 7. Subset regression where red and blue represent the significant positive and negative beta coefficients that were consistently selected in different regression models. The categorical variables are represented with a yellow highlight (coded as 1 (present) or 0 (absent)) and if selected is interpreted as the samples belonging to those categories having positive/negative influence on the respective microbiome metrics.

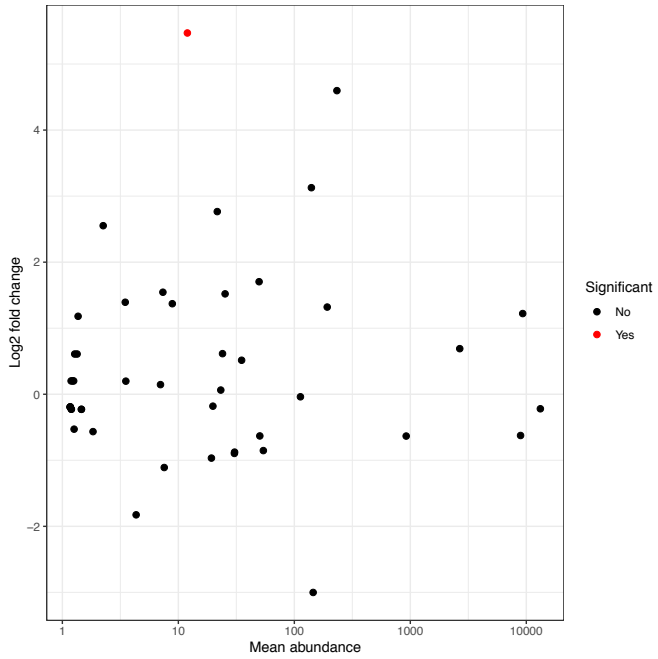


Figure 8. Seqenv identified certain EnvO ontological terms associated with the sequences and in general with the sampling space. On performing differential analysis, weighted with abundances, the above figure is obtained with red highlighting any significant changes.

Table 3. PICRUSt2 search results in relation to KEGG pathways.

	baseMean	log2FoldChai	pvalue	padj	Upregulated
K17335	35.5709471	3.30775151	1.45E-05	0.04170186	T3

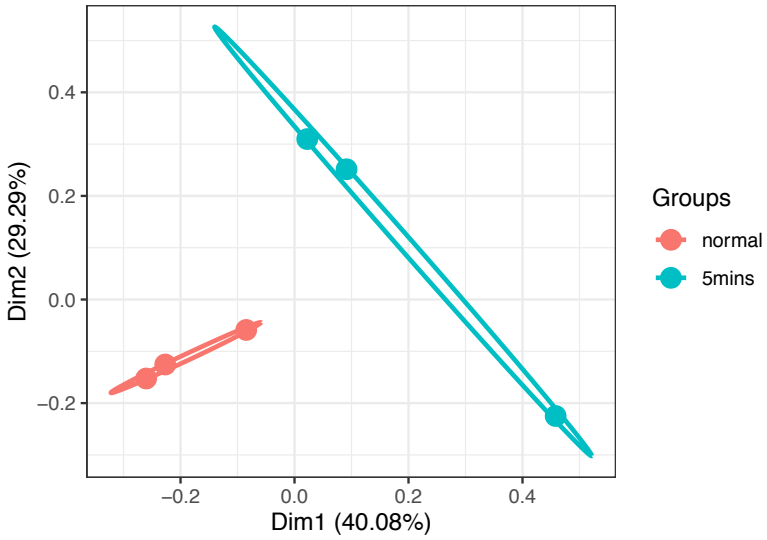
Appendix IV – Significant gene list obtained from RNA-Seq StringTie pipeline

Table 1. Up and down regulated genes when comparing normal vs 5 mins 5 mM H_2O_2 stress conditions using DESeq2. In the interest of space, only a subset is displayed. In the interest of space, normal vs 15 mins 5 mM H_2O_2 stress is provided as accompanied file.

	baseMean	log2 Fold Change	pvalue	padj	Upregulated	GeneName
gene-Cj1177c	173.66	3.86	2.14E-39	3.49E-36	5mins	gmk
gene-Cj1419c	591.15	4.04	6.43E-38	5.23E-35	5mins	Cj1419c
gene-Cj0062c	220.93	4.67	2.54E-36	1.38E-33	5mins	Cj0062c
gene-Cj0426	835.45	3.93	5.27E-35	2.15E-32	5mins	Cj0426
gene-Cj0639c	335.68	3.43	5.41E-34	1.76E-31	5mins	adk
gene-Cj0427	1185.44	4.36	4.99E-30	1.35E-27	5mins	Cj0427
gene-Cj1228c	276.02	3.03	3.32E-28	7.71E-26	5mins	htrA
gene-Cj0331c	207.84	4.15	4.54E-28	9.23E-26	5mins	Cj0331c
gene-Cj0664c	415.18	4.26	1.56E-27	2.82E-25	5mins	rplI
gene-Cj0102	247.82	4.41	1.07E-26	1.74E-24	5mins	atpF'
gene-Cj1487c	449.60	3.37	2.20E-26	3.26E-24	5mins	ccoP
gene-Cj0193c	288.03	3.81	6.13E-26	8.32E-24	5mins	tig
gene-Cj0912c	154.85	2.59	1.47E-25	1.84E-23	5mins	cysM
gene-Cj1110c	858.67	4.17	2.06E-24	2.39E-22	5mins	Cj1110c

Appendix V –RNA-Seq results using bedtools pipeline

(A)



(B)

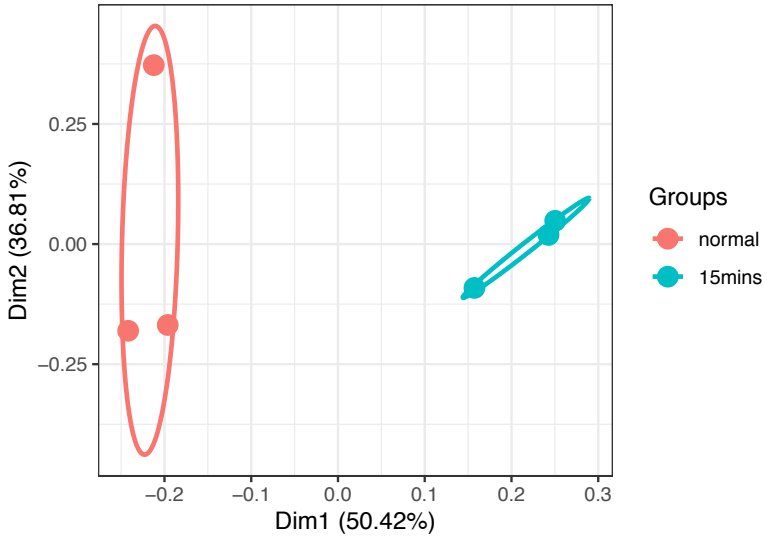
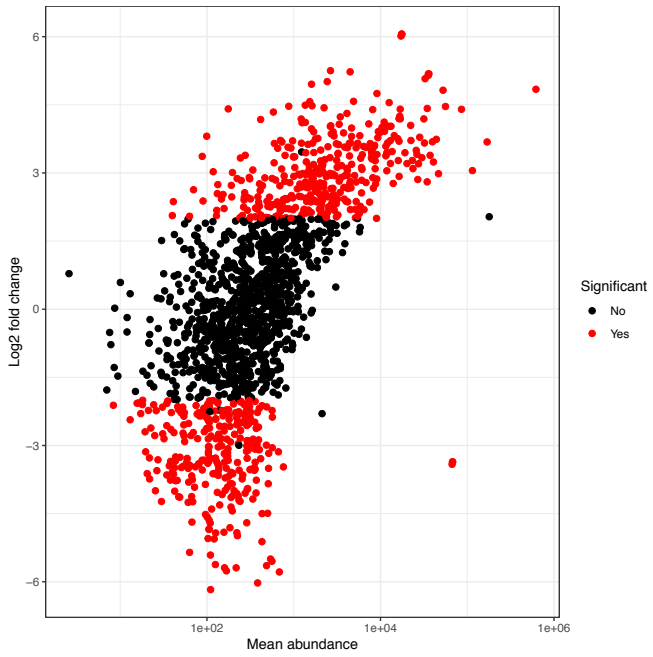


Figure 1. PCoA plots of transcripts *Bray-Curtis* distance comparing the different categorical variables, 5 mins 5 mM H_2O_2 stress (A) or 15 mins 5 mM H_2O_2 stress (B). RNA-Seq analysis performed using bedtools pipeline.

(A)



(B)

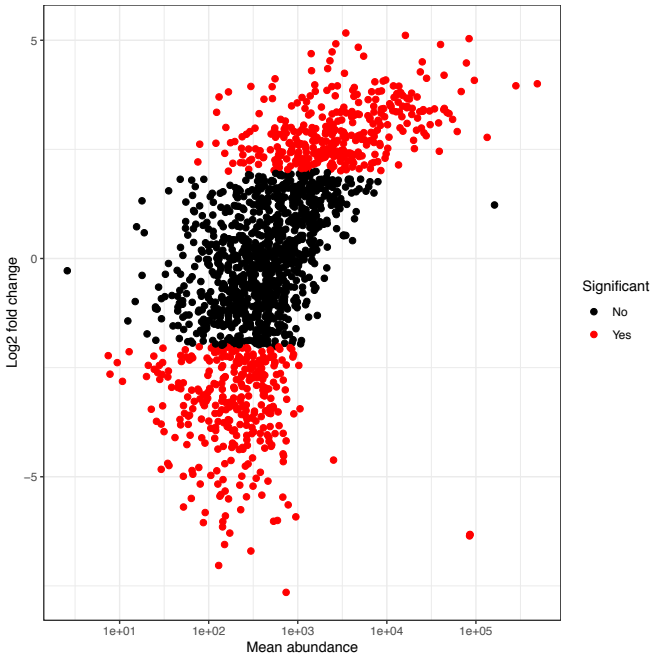


Figure 2. The DESeqDataSetFromMatrix() function from DESeq2 package was used with the adjusted p-value significance cut-off of 0.05 and log fold change cut-off of 2. This function uses negative binomial GLM (generalised linear model) to obtain maximum likelihood estimates for the transcripts log fold change between the two conditions. Then Bayesian shrinkage was applied to obtain shrunken log fold changes subsequently employing the Wald test for obtaining significances for 5 mins 5 mM H_2O_2 stress (A) and 15 mins 5 mM H_2O_2 stress (B). RNA-Seq analysis performed using bedtools pipeline.