QIIME2 End-to-End 16S rRNA Analysis Workflow

Introduction

The 16S ribosomal RNA gene codes for the RNA component of the 30S subunit of the bacterial ribosome. It is widely present in all bacterial species (E. Stackebrandt and B. M. Goebel, 1994). Different bacterial species have one to multiple copies of the 16S rRNA gene. 16S rRNA gene sequencing is by far one of the most common methods targeting housekeeping genes to study bacterial phylogeny and genus/species classification. DNA–DNA hybridization is the gold standard for identifying bacterial species (E. Stackebrandt and B. M. Goebel, 1994). Due to the complexity of DNA–DNA hybridization, 16S rRNA gene sequencing is used as a tool to identify bacteria at the species level and assist with differentiating between closely related bacterial species .

Many clinical laboratories rely on this method to identify unknown pathogenic strains. During 2001–2007, 16S rRNA gene sequencing identified 215 novel bacterial species, 29 of which were classified to novel genera (E. Stackebrandt and B. M. Goebel, 1994). Identification of these novel bacterial species is accomplished through 16S rRNA analysis after sequencing is done. There are several bioinfrmatcs tools used in 16S rRNA analysis which subsequently depend on various programming languages such as Python, R and Bash. Based on the sequences generated, a bioiformatician can use this data to determine diffeerent aspects of the study in queston such as phylogeny and efficacy of treatment.

Objectives

1. To generate a QIIME2 End-to-End 16S rRNA analysis pipeline.

2. To determine the efficiency of different treatments on the sample data provided using the generated pipeline.

QIIME2 End-to-End 16S rRNA Analysis is straightforward and has a simple syntax. For it to work effectively, the working dataset needs to be converted into a QIIME2 Artefact (Bolyen E, Rideout JR, Dillon MR, et al, 2019). Each step allows you to visualize what is happening and assess if the step has completed successfully. At the end of each step, there is a visualization feature which generates a detailed report of the completed step. The report or visualization feature is stored and can be viewed through dragging and dropping the results of that step into a QIIME2 View website (Bolyen E, Rideout JR, Dillon MR, et al, 2019). The wesbite also allows you to save your visualization in pictoral format for better presenattion.

Data samples used in this project were obtained from the H3eaBionet website. This pipeline was developed through following the QIIME2 moving pictures tutorial and 16S rRNA workflow example. The sample dataset was used to test the pipeline whence the results were generated. To follow the script used in the entire pipeline, click here. The results from the pipeline are found here.

Method

QIIME2 End-to-End 16S rRNA Analysis outlines the following steps in analysing any 16S rRNA data from sequencing platforms:

Step	Process
1. Data preprocessing	Reformat the barcodes of the paired-end reads, renaming the original.

	Format the metadata file into a Qiime-compatible object. Perform a quality check to assess the quality of your original reads. Import your data into a QIIME2 artefact. This requires a manifest file which can also be generated automatically based on your data set.
2. Demultiplex	Demultiplex your dataset to remove barcodes from the sequence reads.
3. Quality control and Q stats	QIIME2 quality control to check for and trim bad quality reads from the sequnces.
4. OTU Picking	(Using either QIIME2 Deblur or DADA2 feature tools) Trim the length of your reads to capture quality sequences. Remove chimeras Perform denoising Feature table creation to end up with OTU feature table with representative sequences.
5. Classification and Phylogeny	Alignment of the OTUs Construction of phylogeny tree based on the similarite and the dissimilarities of the OTUs.
6. Core Metrics analyses	Alpha diversity analysis. Beta diversity analysis. Rarefaction.
7. Taxonomic assignment	(Through training your own classifier from one of the three databases available; SILVA, GreenGenes, Ribosomal Database Project (RDP). Trainclassifier Assign taxonomy.
8. Other analyses.	Differential abundance analysis through; (a) ANCOM analysis. (b) Gneiss analysis.

QIIME2 End-to-End 16S rRNA Analysis workflow assessment

Input data assessment

The practice dataset comes from Illumina sequencing platform. However, a MultiQc report of the initial FastQc summary of all indicates that a large section of sequences are duplicates. This can be due to similarities of the different regions found on the 16S rRNA gene (Bokulich, Nicholas A., et al, 2013). The GC% content of the samples is above 50% for most samples indicating the stability of the reads. DNA with low GC-content is less stable than DNA with high GC-content; however, the hydrogen bonds themselves do not have a particularly significant impact on molecular stability, which is instead caused mainly by molecular interactions of base stacking.

In spite of the higher thermostability conferred to a nucleic acid with high GC-content, it has been observed that at least some species of bacteria with DNA of high GC-content undergo autolysis more readily, thereby reducing the longevity of the cell per se (Bokulich, Nicholas A., et al, 2013). Because of the thermostability of GC pairs, it was once presumed that high GC-content was a necessary adaptation to high temperatures, but this hypothesis was refuted in 2001. Even so, it has been shown that there is a strong correlation between the optimal growth of prokaryotes at higher temperatures and the GC-content of structural RNAs such as ribosomal RNA, transfer RNA, and many other non-coding RNAs (Bokulich, Nicholas A., et al, 2013). The AU base pairs are less stable than the GC base pairs, making high-GC-content RNA structures more resistant to the effects of high temperatures.

These statistics thus informs that the sequences are of good quality and can be used in dowstream analysis to identify the diversity of samples in different treatments.

Sample Name	% Dups	% GC	Length	% Failed	M Seqs
Dog1_R1	97.4%	52%	300 bp	36%	0.1
Dog1_R2	95.6%	52%	300 bp	45%	0.1
Dog2_R1	97.2%	50%	300 bp	36%	0.1
Dog2_R2	95.3%	51%	300 bp	45%	0.1
Dog3_R1	97.6%	51%	300 bp	36%	0.1
Dog3_R2	96.0%	52%	300 bp	45%	0.1
Dog8_R1	97.3%	52%	300 bp	36%	0.1
Dog8_R2	95.3%	51%	300 bp	45%	0.1
Dog9_R1	97.4%	51%	300 bp	36%	0.1
Dog9_R2	95.6%	51%	300 bp	45%	0.1
Dog10_R1	97.0%	51%	300 bp	36%	0.1
Dog10_R2	95.6%	51%	300 bp	45%	0.1
Dog15_R1	97.9%	51%	300 bp	36%	0.1
Dog15_R2	95.5%	51%	300 bp	45%	0.1
Dog16_R1	97.5%	51%	300 bp	27%	0.1
Dog16_R2	95.5%	51%	300 bp	45%	0.1
Dog17_R1	97.7%	51%	300 bp	36%	0.1
Dog17_R2	95.7%	51%	300 bp	45%	0.1
Dog22_R1	98.1%	51%	300 bp	36%	0.1
Dog22_R2	96.7%	51%	300 bp	45%	0.1
Dog23_R1	98.0%	51%	300 bp	36%	0.2
Dog23_R2	96.6%	51%	300 bp	45%	0.2
Dog24_R1	98.1%	51%	300 bp	36%	0.2
Dog24_R2	96.4%	51%	300 bp	45%	0.2

Dog29_R1	98.0%	51%	300 bp	36%	0.1
Dog29_R2	96.8%	52%	300 bp	45%	0.1
Dog30_R1	97.8%	51%	300 bp	36%	0.1
Dog30_R2	96.6%	51%	300 bp	45%	0.1
Dog31_R1	98.2%	51%	300 bp	36%	0.2
Dog31_R2	96.6%	52%	300 bp	45%	0.2

%GC – Percentage GC
%Dups – Percentage Duplicates
% Failed – Percentage of modules failed in FastQc report
Length – Average sequence length (base paires – bp)
M Seqs – Total Sequences (millions)

Operational assessment

Demultiplexed filtered statistics

QIIME2 offers an interactive visualization of demultiplexed filtered statistics of the quality check done in QIIME2 so as to work with good quality reads (Bolyen E, Rideout JR, Dillon MR, et al, 2019). This is done using QIIME2 Deblur or DADA2 feature tools. QIIME2 End-to-End 16S rRNA Analysis despite running unsupervised cannot be paused and then contnued, unless each step is run independently.

QIIME2 feature tools have various options for each of the steps and these usually depend on personal choice of the user and available options for data (Bolyen E, Rideout JR, Dillon MR, et al, 2019). The tool has a readily available help option to help navigate different options for each of the processes depending on hypotheses and what would work best for the available data. Most of the options for this setup were left on default. Once installed, QIIME2 does not require heavy computational resources. It should however be noted that the better the computational resources, the more efficient it works. QIIME2 End-to-End 16S rRNA Analysis runtime depends on the computational resources availablea and the size of the data.

Larger sizes of sequenced data can be run on High Performance Computers (HPC) with just one single script. Caution must however be exercised when scripting to ensure that the right steps are called in the right order. The steps depend on one another in succession, which means that the output of one step is the input for the next step. Should one step fail, the rest of the steps will fail to execute. Production of reports for each step of the process ensures that one can assess and modify each individual steps depending on the options or parameters given.

Demux-filter-stats.qzv

sample-id	total-input-reads	total-retained-reads	reads-truncated	reads-too-short-after-truncation	reads-exceeding-maximum-ambiguous-bases
#q2:types	numeric	numeric	numeric	numeric	numeric
Dog1	118343	118214	24	24	105
Dog10	79342	79244	24	24	74
Dog15	131483	131302	40	40	141
Dog16	114424	114327	32	32	65
Dog17	99610	99514	20	20	76
Dog2	108679	108562	44	44	73

Dog22	145029	144879	46	46	104
Dog23	193158	192960	47	47	151
Dog24	162487	162320	44	44	123
Dog29	122776	122634	32	32	110
Dog3	101482	101392	20	20	70
Dog30	137315	137165	34	34	116
Dog31	150613	150446	45	45	122
Dog8	108731	108614	36	36	81
Dog9	109500	109382	29	29	89

Runtime analysis assessment

The dataset used in this case was 2.6 Gigabytes of data ran from start to end in a little under two hours.

Before the diversity of the sample data in different treatments was determined, it was important that Operational Taxonomic Units were picked using the *De novo* strategy. This ultimately created representative sequences that would be useful in classification, phylogeny and assigning taxonomy to the different samples (Callahan, B. J., et al, 2016). This therefore necessitated that chimeras and ambiguous sequences are removed and unique sequence variants with quality picked for use. The following represents a summary table of the steps to ascertain the quality of the sequence variants used in the diversity analyses.

Deblur statistics of quality of variants picked

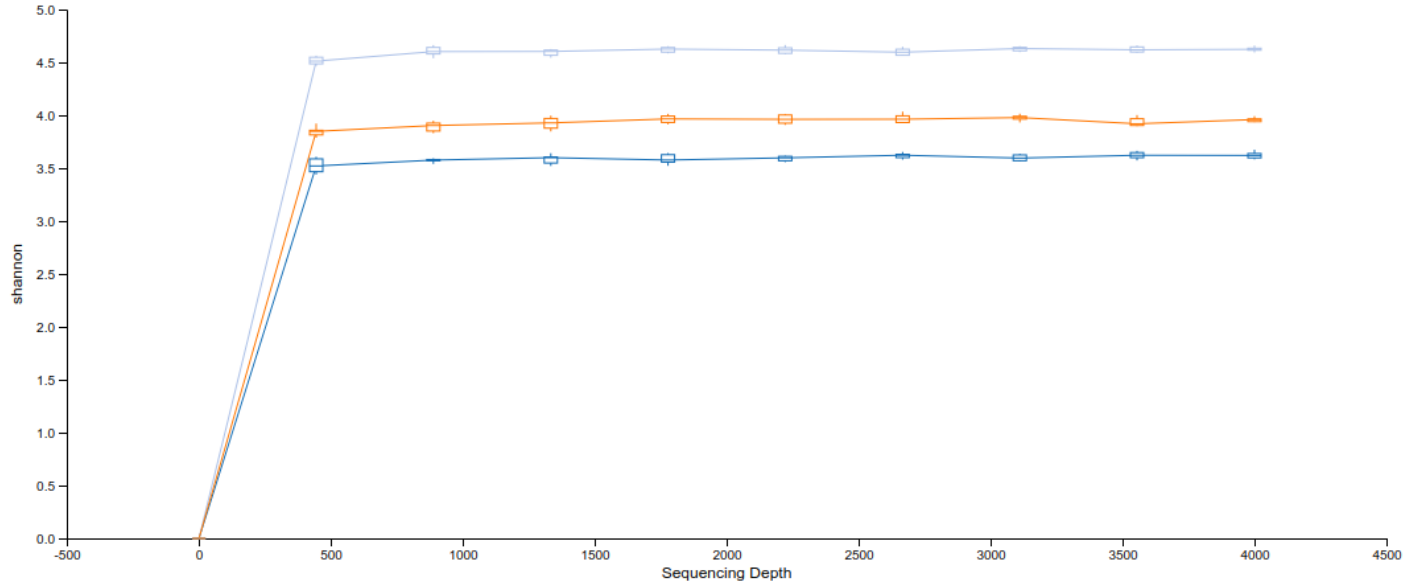
	sample-id	reads-raw	fraction-artifact-with-minsize	fraction-artifact	fraction-missed-reference	unique-reads-derep	reads-derep	unique-reads-deblur	reads-deblur	unique-reads-hit-artifact	reads-hit-artifact	unique-reads-chimeric	reads-chimeric	unique-reads-hit-reference	reads-hit-reference	unique-reads-missed-reference	reads-missed-reference
0	Dog2	108562	0.072613	0.0	0.000000	3041	100679	270	77202	0	0	103	721	149	76453	0	0
1	Dog9	109382	0.071547	0.0	0.000000	3034	101556	297	77163	0	0	113	1070	146	76021	0	0
2	Dog8	108614	0.071105	0.0	0.000000	2821	100891	244	79373	0	0	102	384	130	78969	0	0
3	Dog10	79244	0.070024	0.0	0.000000	2341	73695	199	58400	0	0	69	403	117	57976	0	0
4	Dog3	101392	0.068556	0.0	0.000026	2513	94441	214	76945	0	0	92	420	104	76487	1	2
5	Dog15	131302	0.067493	0.0	0.000081	2691	122440	204	99441	0	0	67	335	110	99051	1	8
6	Dog16	114327	0.066590	0.0	0.000000	3055	106714	238	80807	0	0	67	2345	148	78417	0	0
7	Dog1	118214	0.066227	0.0	0.000000	3228	110385	279	88052	0	0	118	716	128	87278	0	0
8	Dog17	99514	0.058404	0.0	0.000000	2251	93702	201	78049	0	0	53	237	122	77761	0	0
9	Dog24	162320	0.057861	0.0	0.000000	3663	152928	289	122155	0	0	124	890	138	121197	0	0
10	Dog23	192960	0.057675	0.0	0.000000	4748	181831	362	133743	0	0	152	2891	154	130747	0	0
11	Dog29	122634	0.057260	0.0	0.000000	2552	115612	233	93458	0	0	85	671	113	92717	0	0
12	Dog30	137165	0.056713	0.0	0.000000	3625	129386	353	95927	0	0	158	2918	159	92939	0	0
13	Dog22	144879	0.056157	0.0	0.000055	3029	136743	310	109324	0	0	136	1128	123	108078	1	6
14	Dog31	150446	0.049652	0.0	0.000000	2892	142976	295	117654	0	0	141	994	123	116594	0	0

Given that the OTU picking step assigns similar sequences to Operational Taxonomic Units (OTUs) by clustering sequences based on a 97% similarity threshold (default). Sequences which are similar at or above the threshold level are taken to represent the presence of a taxonomic unit in the sequence collection. At this threshold, 209 OTUs are picked from these samples which can then be classified up to the sub-species level. The quality OTUs generated are enough to infer abundaces within the three different treatments used as well as classification.

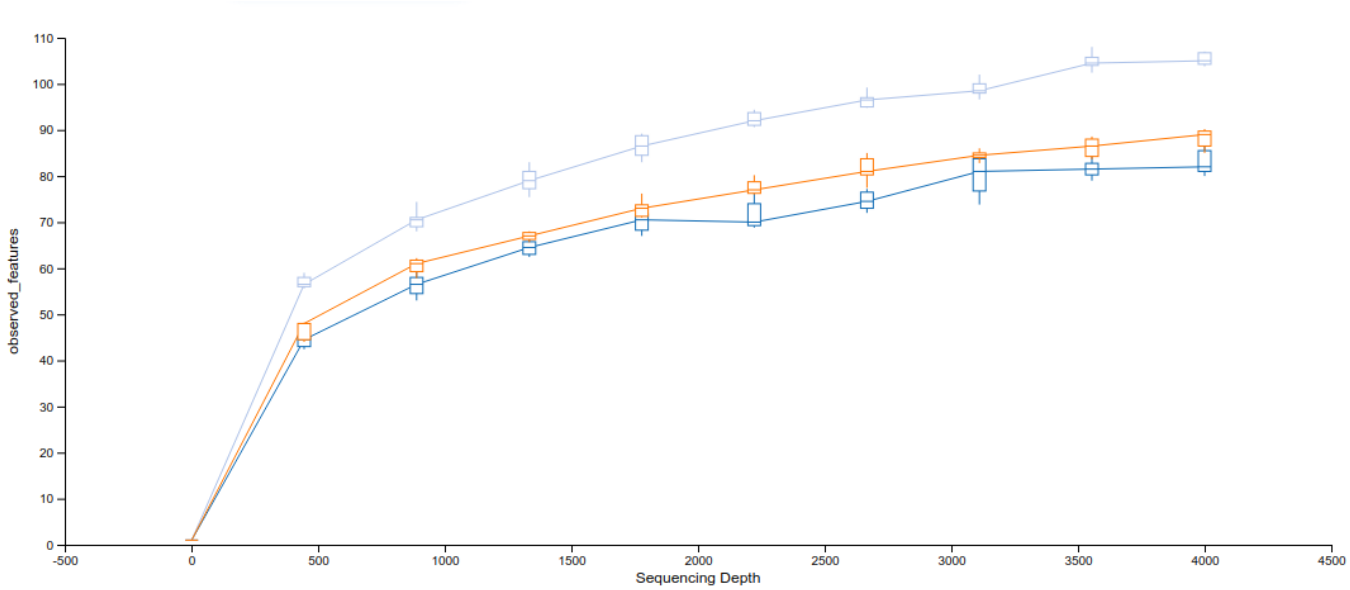
Abundance/Richness Estimation using Rarefaction

Based on the number of OTUs picked from this method, the rarefaction curve flattens giving a comparison of abundances of samples found in each of the three treatments (B, K and G from the metadata).

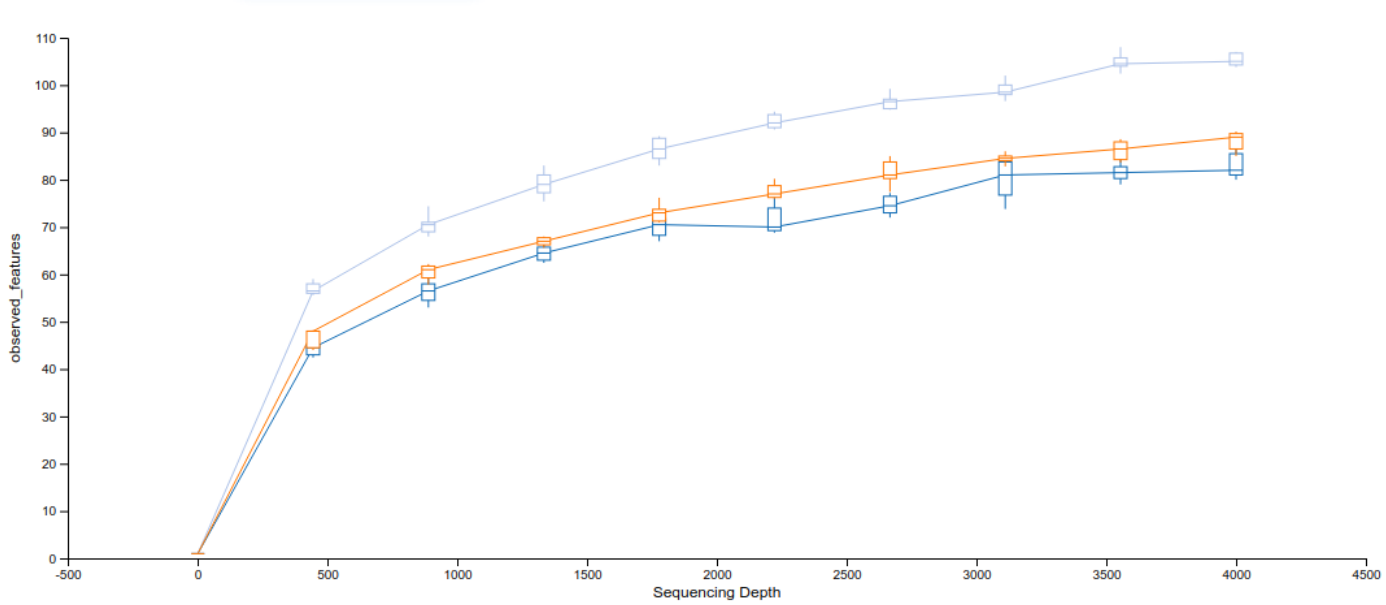
Rarefaction_shannon_diversity



Rarefaction_faithpd_diversity



Rarefaction_observed_features



In all the sub-sampling, the sampling depth used was 4000

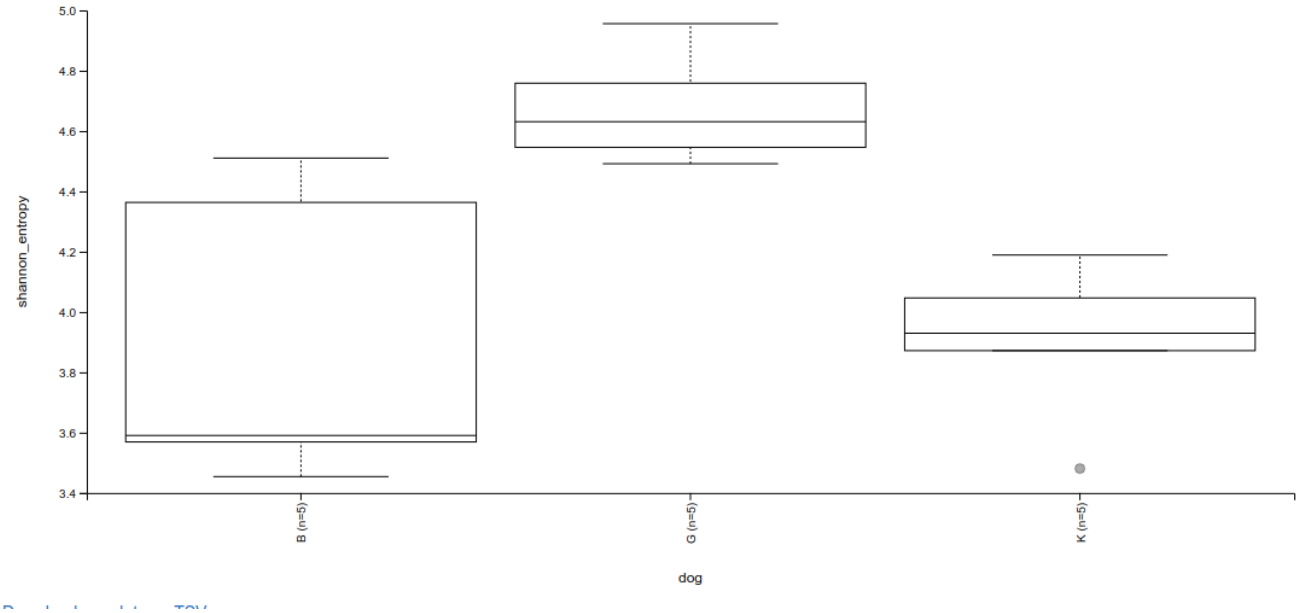
Legend

Core Metrics Analyses

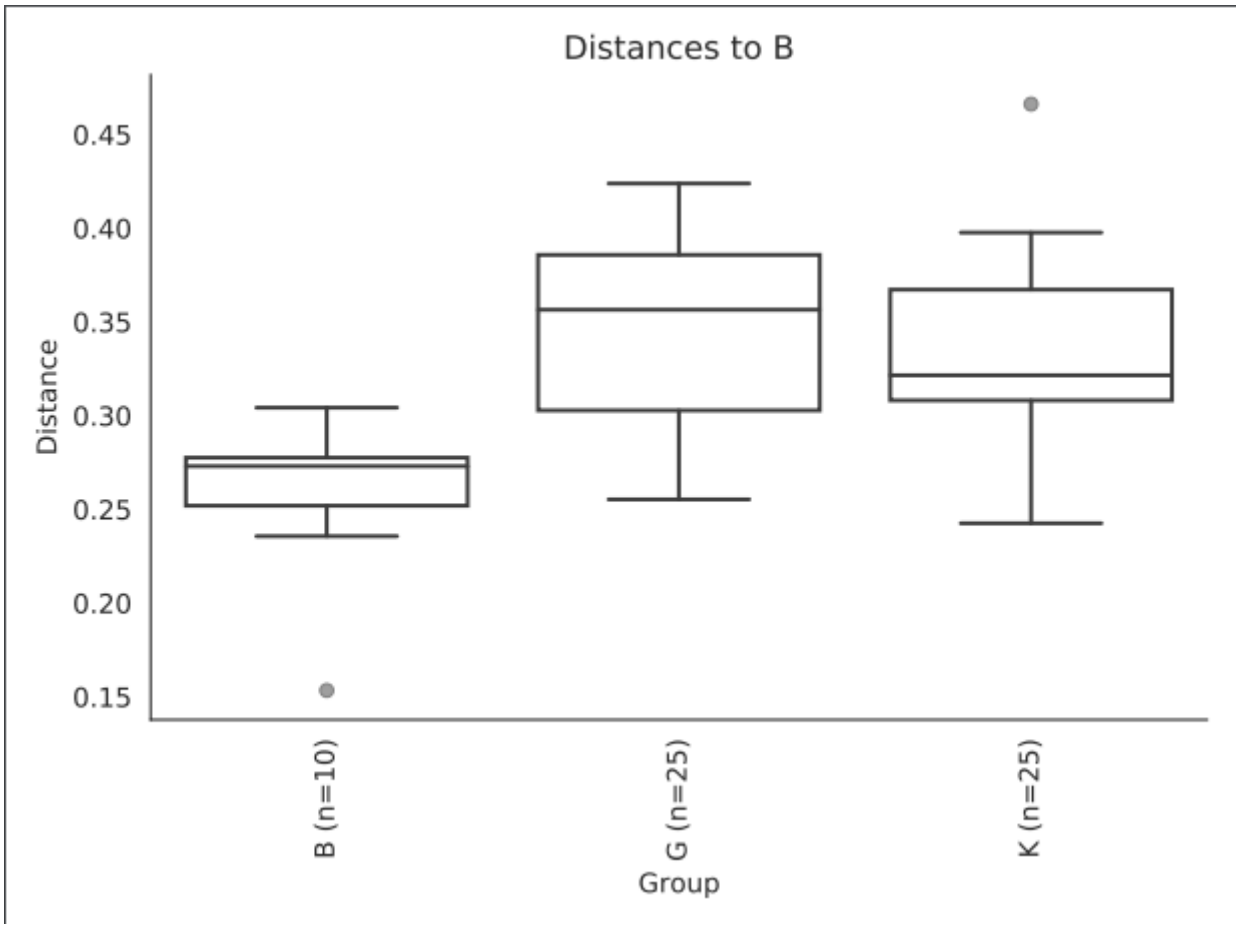
Alpha Diversity and Beta Diversity

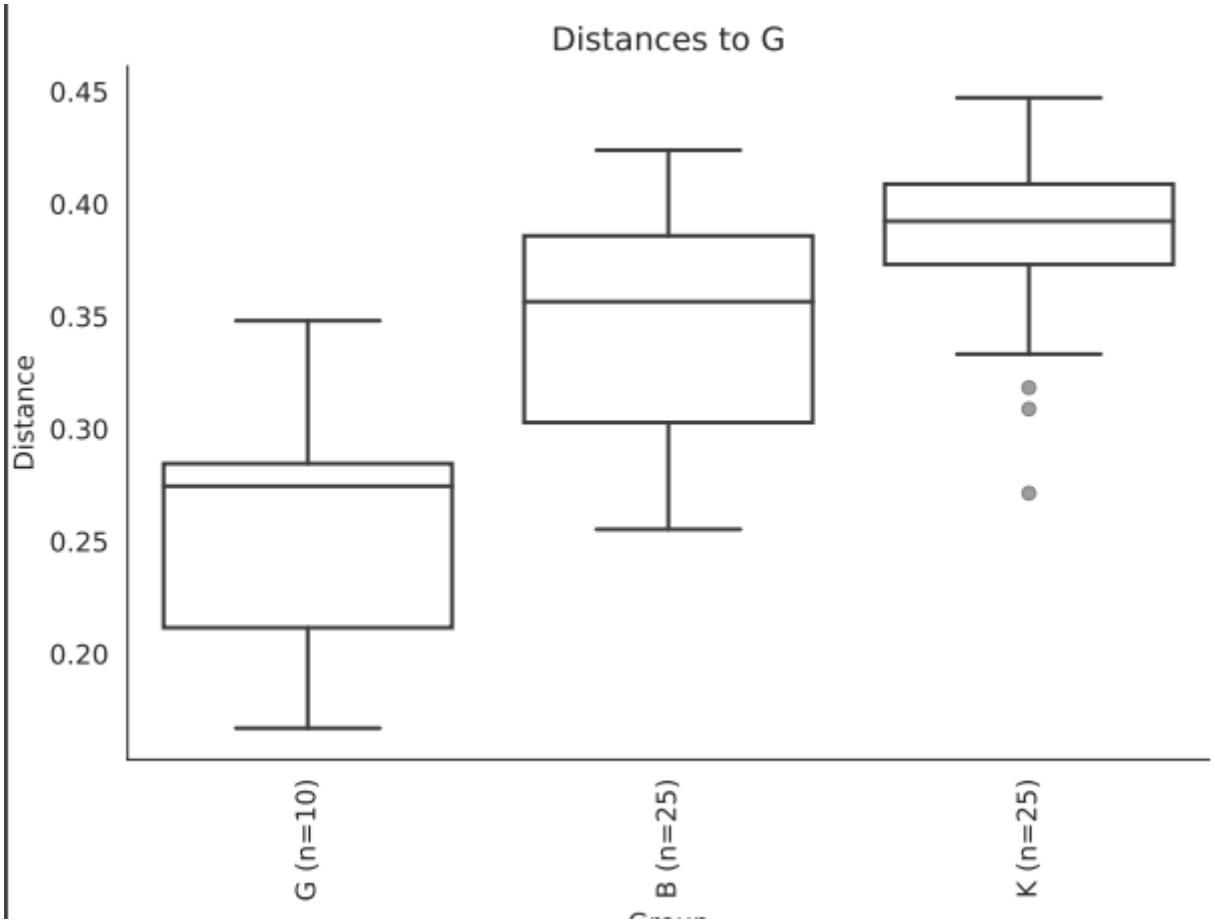
There were more diverse populations found within samples treated with treatment B compared to K and G. If the three treatments were antibiotics used on different dog patients displaying a common symptom, then treatment G offers a more viable solution. Inferred from the results, samples exposed to treatment G had the least diverse populations implying that most of the bacterial populations were killed when exposed to it. In this case, G would be more specific while B would be generic with K in between.

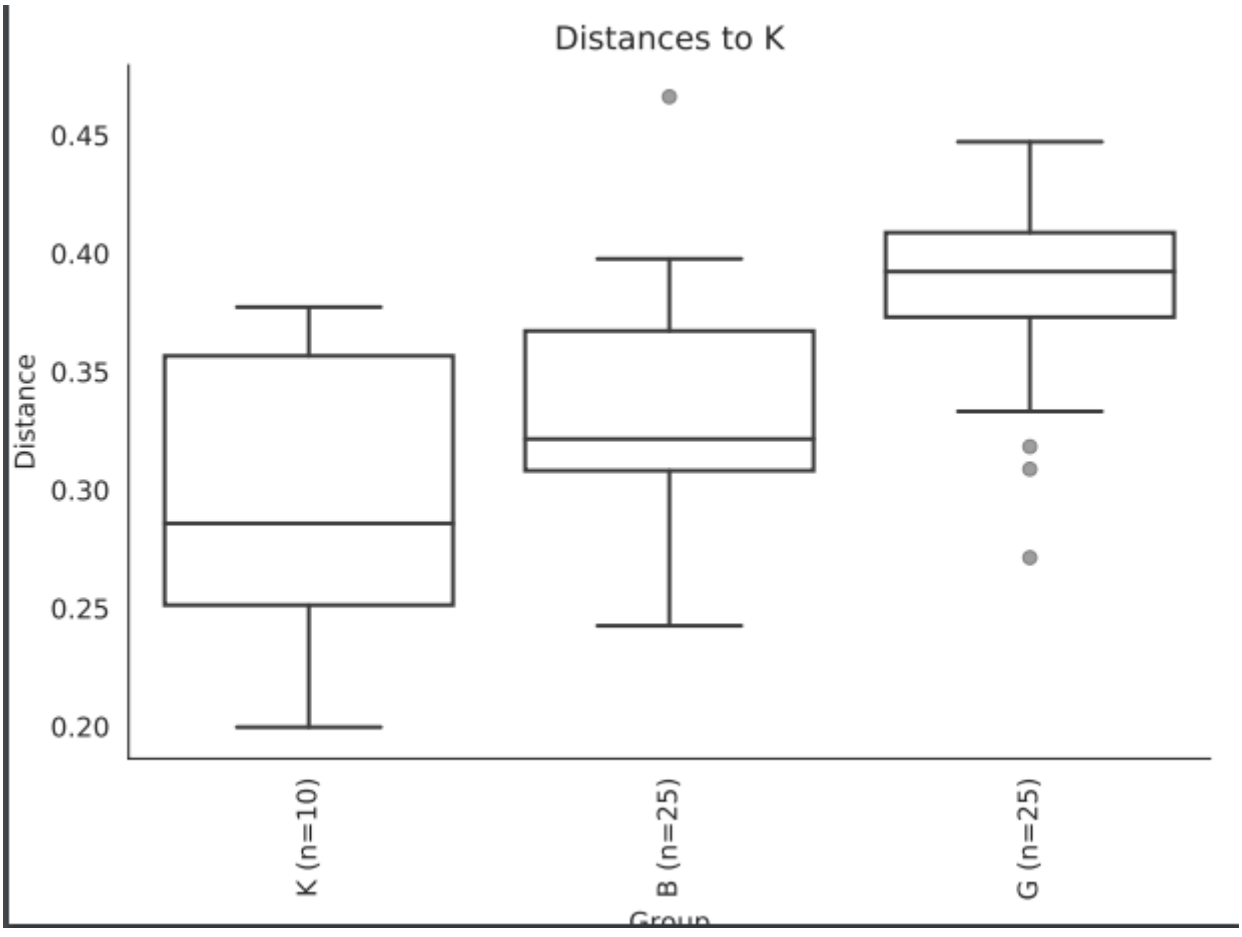
Alpha Diversity Shannon Box Plots



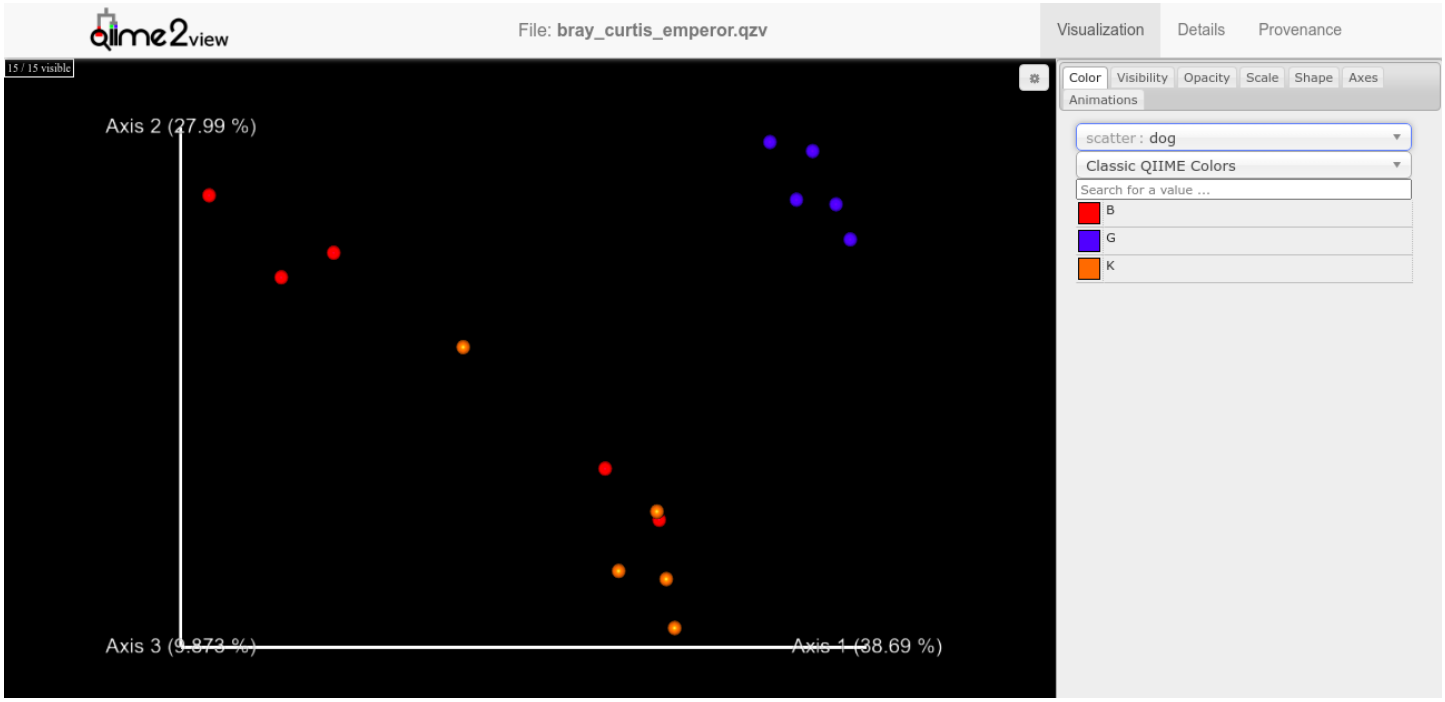
Unweighted unifrac dog significance




Distances to G
Distance
0.45
0.40
0.35
0.30
0.25
0.20
G (n=10)
B (n=25)
K (n=25)


Distances to K
Distance
0.45
0.40
0.35
0.30
0.25
0.20
K (n=10)
B (n=25)
G (n=25)
Group

Bray Curtis Emperor Plot

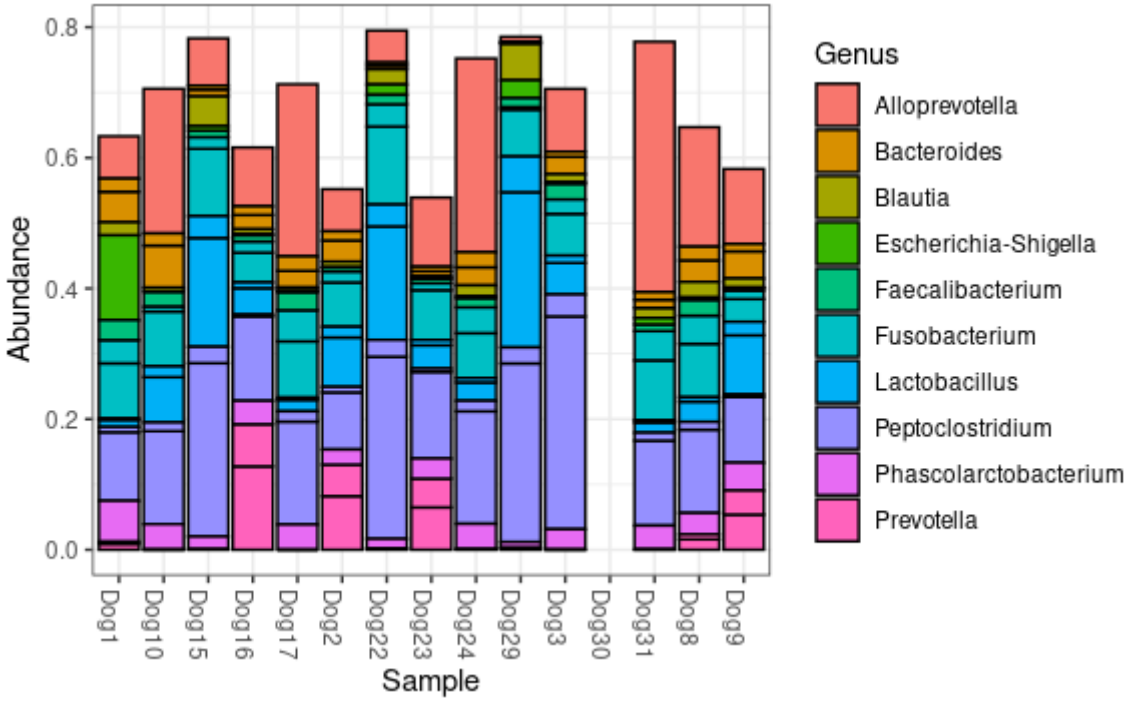


Samples treated with treatment G clustered away from those samples treated with B and K. This infers that populations found within G were distantly related to those found within samples treated with B and K.

Taxonomic Assignment

QIIME2 feature tools were also able to assign taxonomies up to the sub-species level (Bolyen E, Rideout JR, Dillon MR, et al, 2019). From the samples collected and treated with various treatments, B, G and K, their taxonomies were assigned.

Taxonomy Bar-Plot Genus Level



Details of more specific sub-speccies taxonomic assignment can be viewed from here on the QIIME2 View Website.

QIIME2 Feature Tools Stats

(based on the H3eaBionet dataset provided)

Process/Strategy	Purpose	Time Taken	Pros	Cons
Demultiplexing	Removes barcodes	NA	NA	NA
Quality control & Q stats	Checks quality of reads and removes bad quality reads.	4 minutes	1. Automatic 2. Fast	NA
OTU Picking (using either Deblur or Dada2)	Select features/sequence variants: 1. Chimera detection. 2. Dereplication. 3. Denoising. 4. Feature table creation.	35 minutes (using Deblur)	1. Many processes in one command. 2. Generates statistics tables to visualize what is happening. 3. Stats for each step.	Dada2 needs more computational resources eg HPCs.
Classification/Phylogeny (using Mafft and Fasttree	Align observed features and establishes their similarities and their dissimilarities.	15 minutes	1. Fast	Limited to Mafft and Fasttree.
Core metrics (for both Alpha and Beta Diversity tests)	Assess efficacy of the three treatments on the bacterial populations found within the 15 dog samples.	25 minutes	1. All diversity tests in one command. 2. Automatic	NA
Taxonomic Assignment	Assign taxonomies to reads of the selected variants	50 minutes	Highly specific	Involves training a classifier which can be resource intensive.
Abundance analyses (ANCOM/Gneiss)	NA	NA	NA	NA

Conclusion

QIIME2 is able to do an end to end 16S rRNA analysis both on small and large datasets. The pipeline uses QIIME2 artefacts whereby the results from each step in the process can be tabulated and visualized. The QIIME2 View is an online interactive website feature integrated to visualize

QIIME2 artefacts for the different features created from the pipeline. QIIME2 apart from a 16S rRNA analysis workflow, has different protocols and tutorials which can easily be followed and adapted different sequence reads from different sequencing platforms.

In this project; there were more diverse populations found within samples treated with treatment B compared to K and G. The abundance of populations found in samples treated with B was more than that observed in K and G. Samples treated with treatment G also clustered away from those samples treated with B and K. Hypothetically, treatment G is observed to be more effective compared to K and B if the three treatments represented forms of antibiotics.

Bibliography

1. Bokulich, Nicholas A., et al. “Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing.” Nature methods 10.1 (2013): 57.

2. Bolyen E, Rideout JR, Dillon MR, et al. “Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2.” Nature Biotechnology 37: 852–857. https://doi.org/10.1038/s41587-019-0209-9

3. Callahan, B. J., et al. “DADA2: High-resolution sample inference from Illumina amplicon data.” Nature methods, 13.7 (2016), 581-3.

4. Callahan, B. J., et al. “Exact sequence variants should replace operational taxonomic units in marker-gene data analysis.” The ISME journal, 11(12), 2639-2643.

5. E. Stackebrandt and B. M. Goebel. “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 Volume 44, Issue 4. First Published: 01 October 1994 https://doi.org/10.1099/00207713-44-4-846.