16S rRNA Accreditation Exercise Report - *icipe*

Participants: Caleb Kibet, Festus Nyasimi, Ruth Nanjala, Winfred Gatua, Rose Wambui, Kauthar Omar, Asatsa Nabwire, Hebrew Simon, Eric Kariuki, Wilson Mudaki

Submitted: 9th April 2021

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 the most common method targeting housekeeping genes to study bacterial phylogeny and genus/species classification. 16S rRNA gene sequencing is used 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.

The 16S rRNA gene amplicon sequencing has gained popularity for microbial surveys within the environmental and biomedical sciences. At the International Center of Insect Physiology and Ecology, 16S is widely used in insect gut microbiome and microbial symbionts. Notably, research within the center uses microbiome research to develop strategies to reduce disease transmission and control crop pests. This includes endosymbionts' role in making insects more resistant to pathogens and preventing disease transmission, such as in the African honey bees.

We received a total of 124 paired-end reads samples with an average read length of 238bp. We also received the metadata file for each sample with the following fields: Sample, Inflammation, BV, Age, BMI. Further research and analysis revealed that the data was derived from a study on bacterial vaginosis. Our analysis and results interpretation is informed by that information, as we did not receive any description of the data.

Description of approach

Two teams conducted the accreditation exercise, each testing a different pipeline: Qiime2 and Dada2. We also developed a workflow for the Qiime2 pipeline using Nextflow. The needs and experience of the team informed the choice of pipelines. We chose Nextflow due to the detailed reports it provides on memory and space usage and the tools' runtime. The team has experience in using Snakemake as well.

The exercise was undertaken using the HPC available at the center. It is composed of a master node and two worker nodes: hpc01 and hpc02. Hpc01 has 256GB RAM, 22TB storage space, and 64 computer cores, while hpc02 contains 64GB RAM, 10TB storage, and 64 compute cores. For the exercise, the analysis was undertaken using hpc01. The data was stored in a shared accreditation folder, accessible to the members of the accreditation exercise.

Further, we used Git and GitHub for collaboratively creating pipelines, sharing results, and discussing the output. Finally, for quick communication, we created a slack channel within our slack workspace. These tools allowed us to collaborate effectively.

The two pipelines are described in two separate reports. We highlight our rationale for the parameter and options chosen and our interpretation of the results obtained.

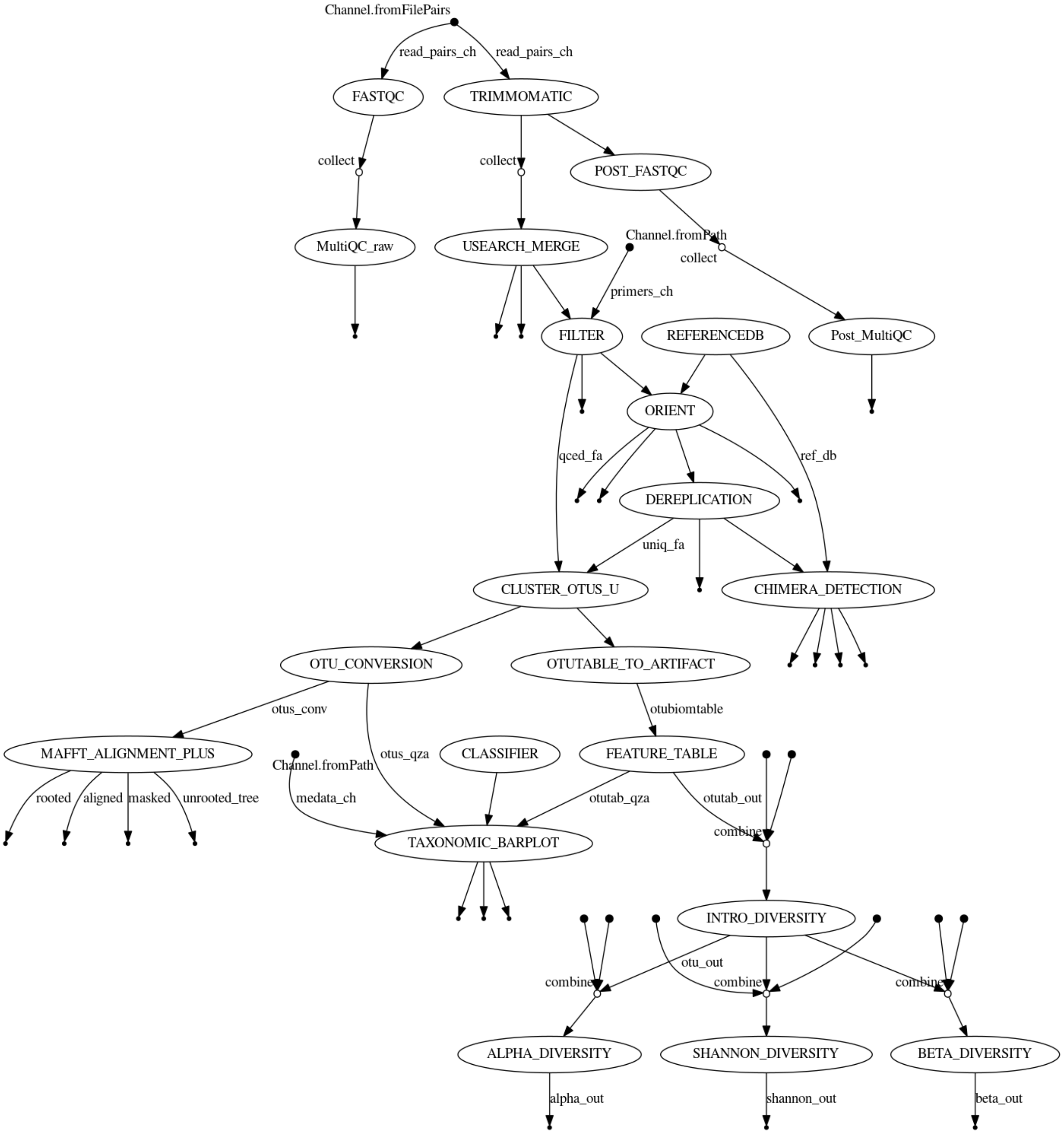
Dada2 Pipeline

The report for Dada2 Pipeline describes all the steps undertaken in the DADA2 pipeline.

Dada2 pipeline written in R is available here.

Qiime2 pipeline

The report for Qiime Nextflow Pipeline describes the steps and results obtained when using Qiime2 pipeline. A reproducible Qiime2 pipeline written in Nextflow is available here



Summary Answers to Key Questions

Although these questions have already been captured within the report, we summarize the responses below.

Table 2: Summary of responses to accreditation questions.

Accreditation Question	Response
Were the number, length, and quality of the reads obtained in line with what would be expected for the sequencing platform used?	We received 124 paired-end sequences of an average length of 238, and Phred scores quality of over 25 for most reads, except at the beginning. These are expected from reads generated using Illumina.
Was the input dataset of sufficiently good quality to perform the analysis?	No. The data was characterized by low per base sequence content, high sequence duplication, and overrepresentation.
How did the reads' quality and GC content affect the way analysis was run?	The quality of the reads informed the trimming parameters
What percentage of the reads were removed during the quality trimming step? Did all samples have similar number of reads after the preprocessing of reads steps? What was the median, maximum and minimum read count per sample? How many reads were discarded due to ambiguous bases?	Approximately 11.6% of the reads were lost after trimming. After trimming the sequences did not have the same length. Median=235 bp; Mean=234bp; min=187 bp. No ambiguous bases were found.
What percentage of reads could not be stitched? Were unstitched reads retained or discarded?	Using usearch merge, we were able to stitch 91.76%. Unstitched reads were discarded. Using Dada2 mergePairs, 1.83% of the reads were not merged
How many chimeras were detected?	- **Qiime:** We found 114201 (15.1%) chimeras, 636866 (84.0%) non-chimeras, and 7098 (0.9%) borderline sequences in 758165 unique sequences. Taking abundance information into account, this corresponds to 378564 (2.1%) chimeras, 17362178 (97.5%) non-chimeras, and 64439 (0.4%) borderline sequences in 17805181 total sequences. - **Dada2**: 95.8% of the reads were retained. Chimera detection led to the identification of 7928 bimeras out of 11807 input sequences, therefore retaining 3879 ASVs
How does the trimming or filtering strategy affect the number of OTUs picked and the classification and phylogenetic analysis of the OTUs?	Trimming removes the primer set from our sequences while filtering removes the low-quality reads. A total of 755318 sequences were discarded.
How many OTUs were picked? What percentage of the OTUs could be classified to the genus and species level? What percentage of OTUs could only be assigned to taxonomic ranks higher than genus? What is the confidence threshold for the classifications?	**Dada2:** A total of 3879 ASVs were picked. Of these, 40.71% and 8.25 were assigned to genus and species level, respectively. 51% could only be assigned to rank higher than genus. Qiime: Found 1493, which is lower than Dada2 pipeline due to the stringency of usearch unoise3 algorithm, and possibly the database used (GreenGenes is smaller).
Does the use of a different 16S rRNA database for classification affect the results (e.g. were a lesser or greater number of OTUs classified to lower taxonomic ranks (genus, species))? Were any OTUs classified differently?	Yes. Each of the databases has a different number of nodes, with SILVA having the largest and GreenGenes the lowest, and these affect the classification, especially to lower ranks.
Did the samples have enough sequence depth to capture the diversity? Did the rarefaction curve flatten? Should any samples be excluded because of low read count?	Yes, it did. However, a few samples did not and were excluded.

Accreditation Question	Response
Were there any differences in the alpha diversity between the samples in the different metadata categories (e.g. higher phylogenetic diversity in treatment 1 vs. treatment 2)?	Yes, in particular, we observed a significant reduction in the genus lactobacillus for samples that were positive for BV.
When groups of samples were compared (e.g. treatment 1 vs. treatment 2) based on distance metrics, such as unifrac, was there any particular clustering pattern observed?	Yes, using Bray Curtis Principal Coordinate Analysis (PcoA), we observed a clustering based on BV, inflammation and BMI
Were any of the OTUs significantly correlated to any of the treatments or other metadata?	Yes, there were OTUs significantly correlated to the BV and Inflammation status.

Challenges

Commercial tools

Usearch We had an issue using usearch the 32-bit version simply because it is capped at 4GB maximum. Our data was huge when running the test analysis; thus, we couldn't run the pipeline. We had to find an alternative, vsearch, and find the equivalent commands on the vsearch tool: orientation, chimera detection, and dereplication.

Vsearch The vsearch version that ships with the Qiime 2020.8 environment is v2.7.0. This version does not have all the features we need. Hence we set up vsearch v2.16.0, which has orient and discarded the vsearch in the Qiime2 environment. We used vsearch for most of the steps.

Internet Outage

During the exercise, the center experienced an internet outage and cut off access to the HPC we used for our analysis. This derailed the progress and time it took for the analysis. The current setup of the HPC is restricted for access outside the center.

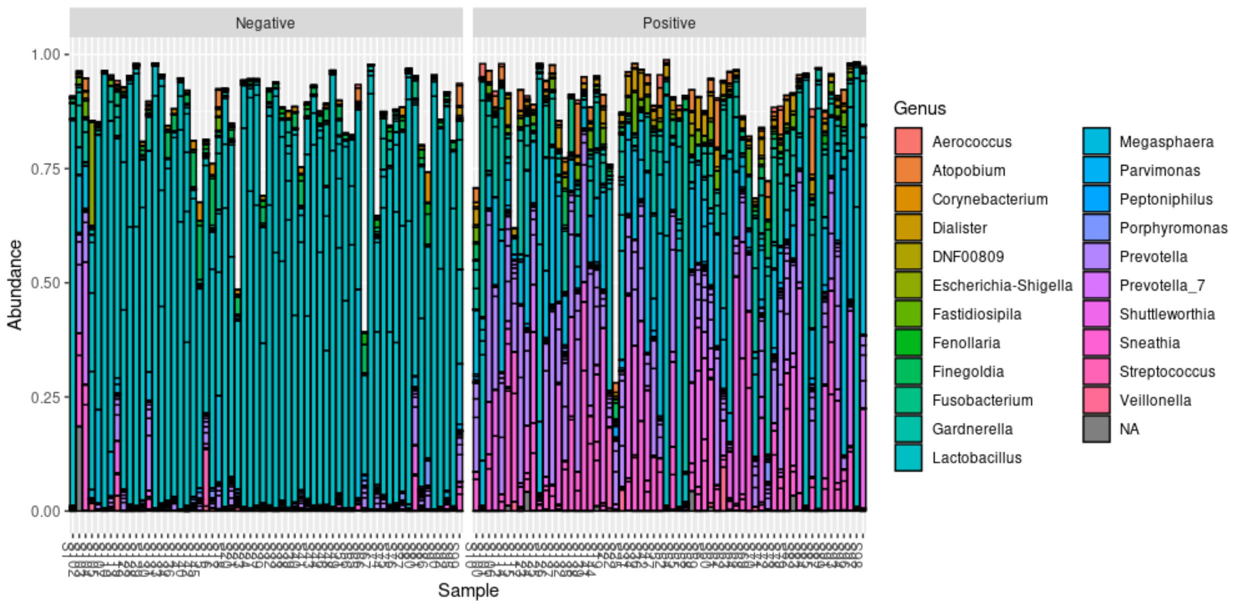
Due to the third COVID wave and the requirement to work from home, the curfews imposed by the government slowed down some of the runs. We worked around this by having some members access the center within the provided protocols.

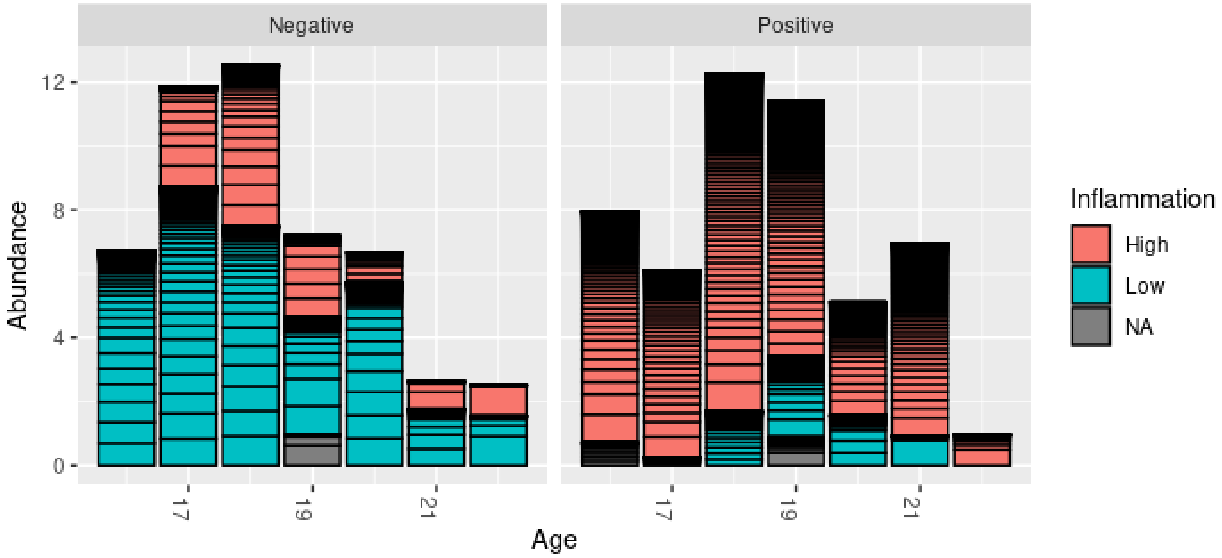
Qiime2 GUI and restricted to online visualization

The Qiime2 pipeline has advanced GUI visualization, which is excellent for users with limited compute ability. However, the use of proprietary file format, the need to visualize online, and lack of text files for easy parsing and export of figures. Therefore, although we developed a reproducible workflow using Nextflow, the look, feel and setings of the figures have to be done manually, hence reducing reproducibility at that stage. On this challenge, we found Dada2 to provide more flexibility and control by the user.

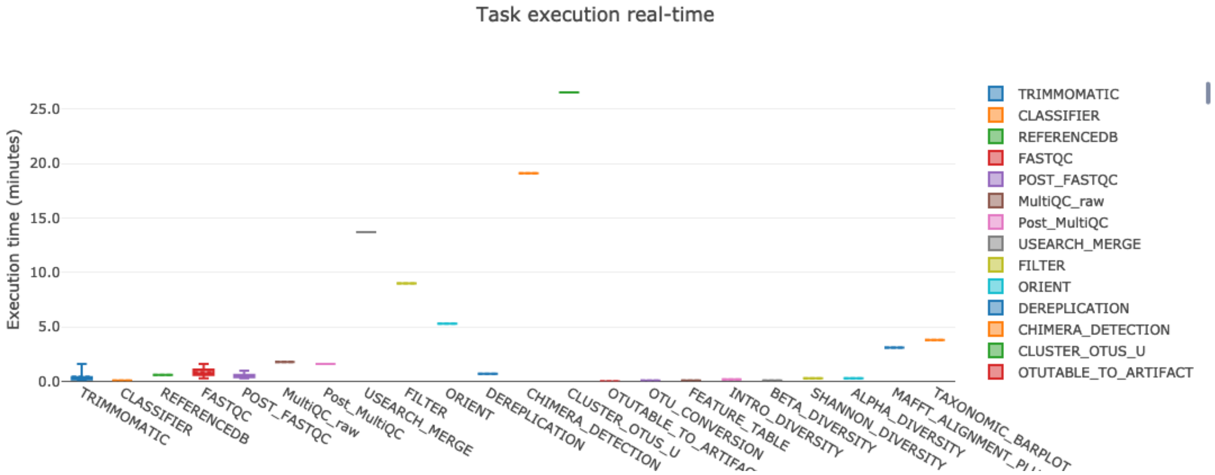
Conclusion

We have successfully analyzed microbial 16S rRNA amplicon sequencing using Qiime2 and Dada2 pipelines, some of the most commonly used. From the ASVs picked, we observed that the diversity of the microbiota is altered by BV, which leads to increased inflammation in the positive samples, consistent with literature.





We note that the quality of the ASVs identified is determined by the pipeline, the database, and the algorithm used in OTU/ASV picking. It is essential to compare the output for concordance.



The computational resources available at *icipe* are sufficient to perform the analysis of such scale. During the analysis, the HPC was still capable of undertaking other work, including whole-genome alignment and ChIP-seq peak calling.