WC_RUN2_QIIME2_processing

Jess Diaz

2023-09-07

Step 1: Import sequencing files

In order to properly use the sequencing files in QIIME2, I first need to download all the sequencing files from the Kohl Lab BaseSpace account and upload them into the cluster into one folder. The sequences from Run1 were already in the cluster in the main/seqs folder, so I copied this into main2/seqs. (NOTE: I did remove the Blank 5 files as those were from a different project).

Downloading the run2 files was done by Jose and the files were deposited into the shared OneDrive. I then uploaded these files into the cluster using the ondemand user interface into main2/seqs along with the run1 files. Jose had already defoldered these.

In the end, there were 306 files from 153 samples (151 samples including redos plus 2 blanks) for Run 1 and 120 files from 60 samples (57 samples plus 3 blanks) from Run 2.

Step 2: Generate manifest file

To import the files I needed to have a manifest file which associates each fastq.gz file with a sample name and whether it is a forward or reverse file. Instructions for this were in Elizabeth's *BaseSpace Protocol* Word Doc.
The final file name is called *WCRUN2-pe-33-manifest* and saved as a .csv. I then uploaded it to the cluster directory where the sequence files are.

Using QIIME2 in the cluster

The following code is reference for loading qiime2 in the cluster and running an interactive job. I used Qiime2 version 2023.5.

```
# load qiime2
module load qiime2/2023.5

# check it's working
qiime --help

# run interactive job
srun --pty bash

# submit normal job
sbatch [job script]
```

Step 3: Import samples to QIIME2

Instructions for this were in Elizabeth's *BaseSpace Protocol* Word Doc.

```
# first, cd to file directory
# import samples using manifest file
qiime tools import \
  --type 'SampleData[PairedEndSequencesWithQuality]' \
  --input-path WCRUN2-pe-33-manifest.csv \
  --output-path paired-end-demux.qza \
  --input-format PairedEndFastqManifestPhred33

# make sure it ran correctly
qiime tools peek paired-end-demux.qza
```

These files have been imported into QIIME2 as Sample Data (paired end with quality info), of the data format SingleLanePerSamplePairedEndFastqDirFmt.

Step 4: Sequence trimming

Instructions for this were in Elizabeth's *BaseSpace Protocol* Word Doc.

Now that the sequences have been uploaded into QIIME2 in a useable format, we can look at the sequence length and quality in order to know how to trim the data.

```
# visualize reads
qiime demux summarize \
  --i-data paired-end-demux.qza \
  --o-visualization paired-end-demux.qzv

# download .qzv file and view through view.qiime2.org
```

Based on the sequence quality and the tendencies of this region of the 16S rRNA gene, I trimmed using the following parameters. I did this through a bash script instead of an interactive job. Also note it should be run outside the working file directory.

```
#!/bin/bash
#SBATCH --partition=htc
#SBATCH --job-name=dada2.X
#SBATCH --output=outs/dada2.X.out
#SBATCH --error=errs/dada2.X.err
#SBATCH --time=0-10:00:00
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=4
#SBATCH --mail-type=END,FAIL
#SBATCH --mail-user=jed213@pitt.edu

module load qiime2/2023.5

cd WhoopingCranes/seqs

# trim reads based on sequence length and quality
```

```
qiime dada2 denoise-paired \
  --i-demultiplexed-seqs paired-end-demux.qza \
  --o-table table.qza \
  --o-representative-sequences rep-seqs.qza \
  --p-n-threads 0 \
  --verbose \
  --p-trunc-len-f 220 \
  --p-trunc-len-r 220 \
  --p-trim-left-f 19 \
  --p-trim-left-r 20 \
  --o-denoising-stats denoising-stats.qza

# visualize number of reads that passed each filter at each step
qiime metadata tabulate \
  --m-input-file denoising-stats.qza \
  --o-visualization denoising-stats.qzv

# download .qzv file and view through view.qiime2.org
```

Running this gives two important file outputs: table.qza and rep-seqs.qza. It also gives a stats output. At this point we can look at stats from this project, and visualize the number of reads that passed each filter at each step.

Note: There were two samples that had no reads pass the filters: Blank-6-1 and F383-50-1. The blank is fine and the other sample is one they redid anyway.

Overall the tracheal samples had high read counts but much lower percentages of reads passing through the filters than the fecal samples. I am still pretty happy with the number of reads left. Also worth noting the blanks in run2 had more reads than in run1 - not a ton but decontam will be interesting.

Step 5: Set up metadata file

The next steps will require a QIIME2-compatible metadata file. Instructions on proper format are here https://gregcaporaso.github.io/q2book/using/metadata.html. Note that sample names should be in the exact format as they were in the manifest file. Also, any samples not included here will be discarded from the rest of the analysis (so include only the desired copy of any samples that were sequenced twice, and make sure to include blanks). My file is *WC_qiime_metadata.txt*. There are 184 samples total since duplicates are not represented.

Step 6: Remove samples from sequencing redos

This step was required because many of my samples were sequenced twice due to quality control at the sequencing facility. I used the metadata file to exclude any repeated samples.

```
# remove sample duplicates based on what is in the metadata file
qiime feature-table filter-samples \
  --i-table table.qza \
  --m-metadata-file WC_qiime_metadata.txt \
  --o-filtered-table table-filtered.qza
```

More visualization

We can now visualize number of features (reads) per sample. This should match what was in denoising-stats.qzv.

```
# visualize number of features per sample
qiime feature-table summarize \
  --i-table table-filtered.qza \
  --o-visualization table-filtered.qzv \
  --m-sample-metadata-file WC_qiime_metadata.txt

# download .qzv file and view through view.qiime2.org
```

We can also look at what the specific sequences were for each feature. Typically not needed.

```
# visualize sequences for each feature
qiime feature-table tabulate-seqs \
  --i-data rep-seqs.qza \
  --o-visualization rep-seqs.qzv

# download .qzv file and view through view.qiime2.org
```

Step 7: Create phylogenetic trees and classify taxonomy

In order to do other steps, we need a phylogeny for these samples so that we can then assign taxa to the specific features in this dataset. Here I am using the SILVA reference database, because it will match previous data better and Greengenes2 is super new. This should also be run through a bash script. Note, this script should be run from outside the main folder, not within it.

```
#!/bin/bash
#SBATCH --partition=htc
#SBATCH --job-name=classify
#SBATCH --output=outs/classify.X.out
#SBATCH --error=errs/classify.X.err
#SBATCH --time=0-10:00:00
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=4
#SBATCH --mail-type=END,FAIL
#SBATCH --mail-user=jed213@pitt.edu

module load qiime2/2023.5

# create phylogenetic trees
qiime phylogeny align-to-tree-mafft-fasttree \
  --i-sequences main2/seqs/rep-seqs.qza \
  --o-alignment main2/seqs/aligned-rep-seqs.qza \
  --o-masked-alignment main2/seqs/masked-aligned-rep-seqs.qza \
  --o-tree main2/seqs/unrooted-tree.qza \
  --o-rooted-tree main2/seqs/rooted-tree.qza

# classify features to taxa according to the Greengenes database
```

```
qiime feature-classifier classify-sklearn \
  --i-classifier training-feature-classifiers/silva-138-99-515-806-nb-classifier.qza \
  --i-reads main2/seqs/rep-seqs.qza \
  --o-classification main2/seqs/taxonomy.qza

# visualize each feature associaton with taxa and confidence
qiime metadata tabulate \
  --m-input-file main2/seqs/taxonomy.qza \
  --o-visualization main2/seqs/taxonomy.qzv

# download .qzv file and view through view.qiime2.org
```

Step 8: Filter table based on taxonomy

Now that we know the taxonomy associated with each sequence, we can remove any undesired taxa. Specifically, I will remove an reads that are not mapped to a phyla, are mitochondria, are chloroplasts, or are not Bacteria.

```
# filter taxa with no phylum, not bacteria, or are mitochondria or chloroplasts
qiime taxa filter-table \
  --i-table table-filtered.qza \
  --i-taxonomy taxonomy.qza \
  --p-include p__ \
  --p-exclude mitochondria,chloroplast,archaea,eukaryota \
  --o-filtered-table table-filt-bytaxa.qza
```

We can also visualize taxa in each sample (besides being informative, this can help confirm whether this filtering was done correctly).

```
# create taxa visualization
qiime taxa barplot \
  --i-table table-filt-bytaxa.qza \
  --i-taxonomy taxonomy.qza \
  --m-metadata-file WC_qiime_metadata.txt \
  --o-visualization taxa-bar-plots.qzv

# download .qzv file and view through view.qiime2.org
```

We should also look at read counts again, now that many sequences have been removed through this taxonomic filter.

```
# visualize number of reads per sample after removing these taxa
qiime feature-table summarize \
  --i-table table-filt-bytaxa.qza \
  --o-visualization table-filt-bytaxa.qzv \
  --m-sample-metadata-file WC_qiime_metadata.txt
```

Now the files can be imported into R for analysis. I will need the taxonomy.qza, table-filt-bytaxa.qza, rooted-tree.qza, and WC_qiime_metadata.txt.

Before proceeding, I will generate quick beta div plots in qiime to see how the fecal and tracheal samples group together.

```
# generate core metrics
qiime diversity core-metrics-phylogenetic \
  --i-phylogeny rooted-tree.qza \
  --i-table table-filt-bytaxa.qza \
  --p-sampling-depth 1500
  --m-metadata-file WC_qiime_metadata.txt \
  --output-dir core-metrics-results
```