Qiime2_Parkinson's_Mouse

February 12, 2020

1 Project "Parkinson's Mouse in Qiime 2"

1.1 Germs everywhere! An introduction to microbiome bioinformatics!

1.2 Bioinformatics, Utah Valley University, Feb 2020

This tutorial is an edited version of a Qiime2 tutorial, this version was create to be used in Cyverse Discovery Environment using this Jupyter Notebook and Qiime2 2018-11.

This tutorial will demonstrate a "typical" QIIME 2 analysis of 16S rRNA gene amplicon data, using a set of fecal samples from humanized mice. The original study, Sampson et al, 2016, was designed to determine whether the fecal microbiome contributed to the development of Parkinson's Disease (PD). Several observation studies showed a difference in the microbiome between PD patients and controls, although the organisms identified across studies were not consistent. However, this was sufficient evidence to suggest that there might be a relationship between PD and the fecal microbiome.

To determine whether that relationship was incidental or actually disease associated, a second study was needed. A human cohort study was not feasible; the disease only affects about 1% of the population over 60 years old, PD takes a long time to develop and to be diagnosed, and it would be difficult to determine when to collect the samples. Therefore, a gnotobiotic mouse study was utilized to evaluate the role of the microbiome in the development of PD symptoms. Feces were collected from six donors with Parkinson's disease and six age- and sex-matched neurologically health controls, and then transplanted into mice who were either predisposed to developing Parkinson's disease due to a mutation ("aSyn") or resistant wild type mice ("BDF1"). Mice from different donors were kept in separate cages, but mice from different genetic backgrounds were co-housed. The mice were followed for 7 weeks to see if they developed symptoms of Parkinson's disease.

We'll look a subset of data from two human donors (one healthy and one with PD) whose samples were each transplanted into three separate cages of mice from the susceptible genotype. For this tutorial, a subset of the metadata has been prepared, and the sequences have been subsampled to approximately 5000 sequences per sample to allow the tutorial to run in a short time. The sequences for the full study are accessible at EBI with accession PRJEB17694; processed tables from the full study can be downloaded from the Qiita database from study 10483.

1.3 THIS PART OF THE CODE ENABLES THE PROPER RUN OF QIIME2

To be able to generate visualization files we need to enable matplotlib:

[14]:

```
echo "backend: Agg" > ~/.config/matplotlib/matplotlibrc
```

Create temp directory for R to be able to run DADA2:

[15]:

```
mkdir -p $HOME/tmp
export TMP="$HOME/tmp"
export TMPDIR="$HOME/tmp"
export TEMP="$HOME/tmp"
```

1.4 UNIX COMMANDS

Let's start, by running the coding line below and create a new directory `mkdir` to store your outputs

[ ]:

```
mkdir PD_Mouse_output
```

Let's use a common Unix command `ls` to list the content of the current folder

[ ]:

```
ls
```

1.5 IMPORT AND TRAIN REFERENCE DATABASE

In this project we are going to use the 99_otus (99% identity) GreenGenes database v13. The code line below will import the 99% **Sequences** from the gg_13_8 folder

[ ]:

```
qiime tools import \
--type 'FeatureData[Sequence]' \
--input-path gg_13_8_otus/rep_set/99_otus.fasta \
--output-path PD_Mouse_output/99_otus.qza
```

The code line below will import the 99% **Taxonomy** from the gg_13_8 folder

[ ]:

```
qiime tools import \
--type 'FeatureData[Taxonomy]' \
--input-format HeaderlessTSVTaxonomyFormat \
--input-path gg_13_8_otus/taxonomy/99_otu_taxonomy.txt \
--output-path PD_Mouse_output/99_otu_taxonomy.qza
```

We need to use Qiime2 to train using naive Bayes machine-learning classifier set trimmed to 250 bp of the V4 hypervariable region. Fill the code line below to **extract** the reads based in primers: 515-F -> GTGCCAGCMGCCGCGGTAA and 806R -> GGACTACHVHHHTWTCTAAT. Be patient, this step takes time.

[29]:

```
qiime feature-classifier extract-reads \
--i-sequences 99_otus.qza \
--p-f-primer CCTACGGGNGGCWGCAG \
--p-r-primer GACTACHVGGGTATCTAATCC \
--o-reads 99_otus_trimmed.qza
```

```
Saved FeatureData[Sequence] to: 99_otus_trimmed.qza
```

The code line below will **train** the classifier using Naive Bayes. This step takes 10 minutes

[ ]:

```
qiime feature-classifier fit-classifier-naive-bayes \
--i-reference-reads PD_Mouse_output/99_otus_trimmed.qza \
--i-reference-taxonomy PD_Mouse_output/99_otu_taxonomy.qza \
--o-classifier PD_Mouse_output/99_otus_classifier.qza
```

1.6 IMPORT THE SEQUENCING DATA

Our samples were amplified using the 515f-806r primers and sequenced on an Illumina MiSeq with a 2x150bp kit. The hypervariable region covered by the primers we used is 290bp long, so with 150bp reads our sequences will be slightly too short to be able to do paired-end analysis downstream. Therefore, we're going to work with single-end sequences. We will work with a version of the samples which have already been demultiplexed.

The first step to import your data is to have your manifest file ready.
Within the working directory, you can see a manifest file, double click on it and you will see a new tab.
1. As you can see the absolute path to the fastqs folder that contains the data is not correct. So please replace "ABSOLUTE-PATH-TO-FASTQS-FOLDER" to the actual absolute path to the folde (Hint: it should start with /home).
2. This manifest file is also missing the orientation of the sequences. What reads are we importing, Read 1 or Read 2? Please add after the absolute path a comma and the orientation (forward or reverse) of the reads.
3. Once the file is updated, please click file, and save file.

Use the coding line below to import your Single End Data FASTQ data.
Remember, you can type any of the following to view formats and types:
qiime tools import –show-importable-types
or
qiime tools import –show-importable-formats

[ ]:

```
qiime tools import \
--type 'SampleData[SequencesWithQuality]' \
--input-path manifest \
--output-path PD_Mouse_output/demux_seqs.qza \
```

The code line below will create a **visualization** .qzv file of your data

[ ]:

```
qiime demux summarize \
--i-data PD_Mouse_output/demux_seqs.qza \
--output-dir PD_Mouse_output/demux_seqs.qzv
```

When this **qzv** file is imported into: https://view.qiime2.org/
You should see something like this:

1.7 *QUIZ TIME*

Quick Questions for you: 1. After demultiplexing, which sample has the lowest sequencing depth? 2. What is the median sequence length? 3. What is the median quality score at position 125?

1.8 DENOISE/DEREPLICATE THE DATA

In this project, we'll denoise with DADA2 (using single-end sequences).
The qiime dada2 denoise-single method requires us to set the –p-trunc-len parameter.
This controls the length of the sequences and should be selected based on a drop in quality scores.
In our dataset, the quality scores are relatively evenly distributed along the sequencing run, so we'll use the full **150** bp sequences. However, the selection of the trim length is a relatively subjective measurement and relies on the decision making capacity of the analyst.

[ ]:

```
qiime dada2 denoise-single \
--i-demultiplexed-seqs PD_Mouse_output/demux_seqs.qza \
--p-trim-left 0 \
--p-trunc-len 0 \
--output-dir PD_Mouse_output/dada2/ \
--o-denoising-stats PD_Mouse_output/denoising-stats.qza
```

1.9 UNDERSTANDING OUR DATA

Let's give a look to our metadata file to understand a little bit better our data.
The code line below will display the first 10 lines of the metadata_map_file file:

[1]:

```
head metadata_map_file.tsv
```

```
      File "<ipython-input-1-101ce5afb9f8>", line 1
    head metadata_map_file.tsv
                         ^
SyntaxError: invalid syntax
```

Run the following code, and download the .qzv.
Visit https://view.qiime2.org/ and drag your output

[21]:

```
qiime metadata tabulate \
  --m-input-file metadata_map_file.tsv \
  --o-visualization PD_Mouse_output/metadata.qzv
```

```
Saved Visualization to: metadata.qzv
```

When this **qzv** file is imported into: https://view.qiime2.org/
You should see something like this:

1.10 *QUIZ TIME*

Quick Questions for you: 4. What does the **metadata** plugin do when using the option **tabulate**? 5. What is the shortest and longest period of days post transplant?

Run the following code, and download the **.qzv**

[39]:
```
qiime metadata tabulate \
  --m-input-file PD_Mouse_output/dada2/denoising-stats.qza \
  --o-visualization PD_Mouse_output/dada2/denoising-stats.qzv
```

```
Saved Visualization to: dada2/denoising-stats.qzv
```

Visit https://view.qiime2.org/ and drag your output.
You should see something like this:

The code below will create a feature table (table.qzv) from our DADA output using the metadata file provided.

[ ]:
```
qiime feature-table summarize \
  --i-table PD_Mouse_output/dada2/table.qza \
  --m-sample-metadata-file metadata.tsv \
  --o-visualization PD_Mouse_output/dada2/table.qzv
```

Visit https://view.qiime2.org/ and drag your output.
It shoud look like this:

1.11 *QUIZ TIME*

Quick Questions for you: 6. How many total features remain after denoising? 7. Which sample has the highest total count of features? How many sequences did that sample have prior to DADA2 denoising? 8. How many samples have fewer than 4250 total features? 9. Which features are observed in at least 46 samples? 10. Which sample has the fewest features? How many does it have? 11. If we set a sampling depth of 4,250 sequences, how many samples reach this depth?

1.12 TAXONOMY CLASSIFY THE DATA

We are ready to match our sequences with taxonomic classification using the **feature-classifier** and **sklearn** and the classifier trained in the step 3.

[ ]:
```
qiime feature-classifier classify-sklearn \
--i-classifier PD_Mouse_output/99_otus_classifier.qza \
--i-reads PD_Mouse_output/dada2/representative_sequences.qza \
--o-classification PD_Mouse_output/mouse_classified.qza
```

Now, let's review the taxonomy associated with the sequences using the qiime metadata tabulate method. Please replace asteriscks with the name of your classified artifact (.qza) as name of your output in the line below

[34]:
```
qiime metadata tabulate \
  --m-input-file PD_Mouse_output/mouse_classified.qza \
  --o-visualization PD_Mouse_output/mouse_classified.qzv
```

```
Saved Visualization to: mouse_classsified.qzv
```

Visit https://view.qiime2.org/ and drag your output.
You should look at something like this:

1.13 *QUIZ TIME*

Quick Questions for you: 12. Find the feature, 07f183edd4e4d8aef1dcb2ab24dd7745. What is the taxonomic classification of this sequence? What's the confidence for the assignment? 13. How many features are classified as g___Akkermansia?

1.14 FILTERING AND VISUALIZING THE DATA

We are going to use the **barplot** plugin and create a visualization *artifacts* (.qzv)
Before doing this, we will first filter out any samples with fewer features than our rarefaction threshold (2000). We can filter samples using the q2-feature-table plugin with the filter-samples method. This lets us filter our table based on a variety of criteria such as the number of counts (frequency, –p-min-frequency and –p-max-frequency), number of features (–p-min-features and –p-max-features), or sample metadata (–p-where). Make sure the path and name of the table from dada2 is correct in the code below.

[32]:
```
qiime feature-table filter-samples \
  --i-table PD_Mouse_output/dada2/table.qza \
  --p-min-frequency 2000 \
  --o-filtered-table PD_Mouse_output/dada2/table_2k.qza
```

```
Saved FeatureTable[Frequency] to: dada2/table_2k.qza
```

Now, let's create a barplot from the **new filtered** taxonomic classified file.

[ ]:
```
qiime taxa barplot \
--i-table PD_Mouse_output/dada2/table_2k.qza \
--i-taxonomy PD_Mouse_output/dada2/mouse_classified.qza \
--m-metadata-file metadata_map_file.tsv \
--output-dir PD_Mouse_output/barplot
```

Download the **.qzv** file
Visit https://view.qiime2.org/ and drag your output. Should look like this:

1.15 *QUIZ TIME*

Quick Questions for you:
14. Visualize the data at level 2 (phylum level) and sort the samples by donor. Can you observe a consistent difference in phyla present between the donors?

1.16 PHYLOGENETIC TREE

In the class we reviewed how to build phylogenetic trees based on alignments. In the Atacama Tutorial we checked the code to run a MAFFT alignment in Qiime2. Here,we're going to create a **fragment insertion** tree using the *q2-fragment-insertion* plugin. The authors of the fragment insertion plugin suggest that it can outperform traditional alignment based methods based on short Illumina reads by alignment against a reference tree built out of larger sequences. Our command, qiime fragment-insertion sepp will use the representative sequences (a FeatureData[Sequence] artifact) we generated during denoising to create a phylogenetic tree where the sequences have been inserted into the greengenes 13_8 99% identity reference tree backbone.

In this tutorial, again, I am presenting to you the code that was run to make the tree (output is in the main folder as tree.qza file), but we are not running here. Also, this code was running Qiime2 version 2019 and we are using version 2018, so even if we try the old version cannot run it.

> qiime fragment-insertion sepp –i-representative-sequences dadа2/representative_sequences.qza –i-reference-database sepp-refs-gg-13-8.qza –o-tree tree.qza –o-placements tree_placements.qza –p-threads 1

Let's **export** the tree.qza file to a directory containing a .nwk file

[ ]:

```
qiime tools export \
--input-path tree.qza \
--output-path PD_Mouse_output/exported-tree
```

We can use an interactive visualization tool as: https://itol.embl.de/upload.cgi to open the tree. Should look something like this:

1.17 ALPHA DIVERSITY RAREFACTION

We now have a feature table (observation matrix) of sequence variants in each sample, and a phylogenetic tree representing those variants, so are almost ready to perform various analyses of microbial diversity. However, first we must normalize our data to account for uneven sequencing depth between samples.

Although sequencing depth in a microbiome sample does not directly relate to the original biomass in a community, the relative sequencing depth has a large impact on observed communities (Weiss et al, 2017). Therefore, for most diversity metrics, a normalization approach is needed.

Current best practices suggest the use of rarefaction, a normalization via sub-sampling without replacement. Rarefaction occurs in two steps: first, samples which are below the rarefaction depth are filtered out of the feature table. Then, all remaining samples are subsampled without replacement to get to the specified sequencing depth. It's both important and sometimes challenging to select a rarefaction depth for diversity analyses. Several strategies exist to figure out an appropriate rarefaction depth - we will primarily consider alpha rarefaction in this tutorial, because it is a data-driven way to approach the problem.

We'll use qiime diversity alpha-rarefaction to subsample the table at different depths (between –p-min-depth and –p-max-depth) and calculate the alpha diversity using one or more metrics (–p-metrics). We want to set a maximum depth close to the maximum number of sequences. We also know from the quiz that if we look at a sequencing depth around 4250 sequences per sample, we'll be looking at information from 34 samples. So, let's set this as our maximum sequencing depth.

At each sampling depth, 10 rarefied tables are usually calculated to provide an error estimate, although this can be adjusted using the –p-iterations parameter. We can check and see if there is a relationship between the alpha diversity and metadata by specifying the metadata file for the –m-metadata-file parameter.ASVASVAD

[41]:

```
qiime diversity alpha-rarefaction \
  --i-table PD_Mouse_output/dada2/table_2k.qza \
  --m-metadata-file metadata_map_file.tsv \
  --o-visualization PD_Mouse_output/alpha_rarefaction_curves_2k.qzv \
  --p-min-depth 10 \
  --p-max-depth 4250
```

```
Saved Visualization to: alpha_rarefaction_curves_2k.qzv
```

Visit https://view.qiime2.org/ and drag your output.
You should see something like this:

1.18 *QUIZ TIME*

Quick Questions for you: 15. Are all metadata columns represented in the visualization? If not, which columns were excluded?
16. Which metric shows saturation and stabilization of the diversity?
17. Which mouse genetic background (genotype) has higher diversity, based on the curve? Which has shallower sampling depth?

1.18.1 Alpha Diversity

Alpha diversity asks whether the distribution of features within a sample (or groups of samples) differs between different conditions. The comparison makes no assumptions about the features that are shared between samples; two samples can have the same alpha diversity and not share any features. The rarefied alpha diversity produced by q2-diversity is a univariate, continuous value and can be tested using common non-parametric statistical tests.

We'll start by using the qiime diversity core-metrics-phylogenetic method, which ratifies the input feature table, calculates several commonly used alpha- and beta-diversity metrics, and produces principal coordinate analysis (PCoA) visualizations in Emperor for the beta diversity metrics. By default, the metrics computed are:

Alpha Diversity >Shannon's diversity index
>Observed OTUs
>Faith's phylogenetic diversity
>Pielou's evenness

Beta Diversity >Jaccard distance
>Bray-Curtis distance
>Unweighted UniFrac distance
>Weighted UniFrac distance

The following code takes 10 minutes to run

[42]:

```
qiime diversity core-metrics-phylogenetic \
  --i-table PD_Mouse_output/dada2/table_2k.qza \
  --i-phylogeny tree.qza \
  --m-metadata-file metadata_map_file.tsv \
  --p-sampling-depth 2000 \
  --output-dir core-metrics-results
```

```
Saved FeatureTable[Frequency] to: core-metrics-
results/rarefied_table.qza
Saved SampleData[AlphaDiversity] % Properties(['phylogenetic']) to: core-
metrics-results/faith_pd_vector.qza
Saved SampleData[AlphaDiversity] to: core-metrics-
results/observed_otus_vector.qza
Saved SampleData[AlphaDiversity] to: core-metrics-
results/shannon_vector.qza
Saved SampleData[AlphaDiversity] to: core-metrics-
results/evenness_vector.qza
Saved DistanceMatrix % Properties(['phylogenetic']) to: core-metrics-
results/unweighted_unifrac_distance_matrix.qza
Saved DistanceMatrix % Properties(['phylogenetic']) to: core-metrics-
results/weighted_unifrac_distance_matrix.qza
Saved DistanceMatrix to: core-metrics-
results/jaccard_distance_matrix.qza
Saved DistanceMatrix to: core-metrics-
results/bray_curtis_distance_matrix.qza
Saved PCoAResults to: core-metrics-
results/unweighted_unifrac_pcoa_results.qza
Saved PCoAResults to: core-metrics-
results/weighted_unifrac_pcoa_results.qza
Saved PCoAResults to: core-metrics-results/jaccard_pcoa_results.qza
Saved PCoAResults to: core-metrics-results/bray_curtis_pcoa_results.qza
Saved Visualization to: core-metrics-

results/unweighted_unifrac_emperor.qzv
Saved Visualization to: core-metrics-

results/weighted_unifrac_emperor.qzv
Saved Visualization to: core-metrics-results/jaccard_emperor.qzv
Saved Visualization to: core-metrics-results/bray_curtis_emperor.qzv
```

Let's give a closer look to the eveness_vector

```
[44]: qiime diversity alpha-group-significance \
--i-alpha-diversity core-metrics-results/evenness_vector.qza \
--m-metadata-file metadata_map_file.tsv \
--o-visualization core-metrics-results/evenness_statistics.qzv
```

```
Saved Visualization to: core-metrics-results/evenness_statistics.qzv
```

Download the **.qzv** file
Visit https://view.qiime2.org/ and drag your output. It should look like this:

1.19 *QUIZ TIME*

Quick Questions for you:
18. What is the p-value for the differences of the eveness based in genotype? 19. What is the p-value for the differences of the eveness based in donor? What can you conclude?

1.20 DIFFERENTIAL ABUNDANCE TEST WITH ANCOM

QIIME2 uses ANCOM (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4450248/) to identify differentially abundant taxa

Lets use the plugin **composition** to add one count to every value (pseudocount)
This is done because ANCOM is a log-based method (log0=undefined)

```
[45]: qiime composition add-pseudocount \
--i-table dada2/table_2k.qza \
--o-composition-table dada2/comp-table.qza
```

```
Saved FeatureTable[Composition] to: dada2/comp-table.qza
```

Apply Analysis of Composition of Microbiomes (ANCOM) to identify features that are differentially abundant across groups
In this case we want to test differential abundance between donors.

```
[48]: qiime composition ancom \
--i-table dada2/comp-table.qza \
--m-metadata-file metadata_map_file.tsv \
--m-metadata-column donor \
--o-visualization ancom-donor.qzv
```

```
Saved Visualization to: ancom-donor.qzv
```

Let's repeat the analysis but now to test differential abundance between genotypes. Please fill code line below

[47]:

```
qiime composition ancom \
--i-table dada2/comp-table.qza \
--m-metadata-file metadata_map_file.tsv \
--m-metadata-column genotype \
--o-visualization ancom-genotype.qzv
```

```
Saved Visualization to: ancom-genotype.qzv
```

Download and visualize both, donor and genotype **.qzv** files. They should look like this:

1.21 ***QUIZ TIME***

Quick Questions for you:
20. What is the number of features with a statistical W for each, genotypes and donors?
21. What is the main variable that explains the difference on composition and diversity in the microbial samples of mice with transplanted microbiomes form healthy and PD donors? 22. Use the visualization of the taxonomy classified table and search sequence identifiers for the significantly different features by genotype. What genera do they belong to?

1.22 THIS IS THE END

Thank you very much!
Natalia J. Bayona-Vásquez, Ph.D.