Summary eDNA report

This report summarises the output from a eDNA pipeline based on QIIME2 for project Example eDNA project.

Raw sequence data

The maximum sequence length for reads in the `fastq_data` folder was: 76bp.

Primers

The following primer sequences were listed in the `config.yaml` file used to setup the pipeline. These primers are removed from the reads before downstream processing.

forward primer: 5'-ACACCGCCCGTCACTCT-3'

reverse primer: 5'-CTTCCGGTACACTTACCATG-3'

Metadata

The following metadata was supplied in `metadata.csv`. This file maps samples to the sequencing data located in the `fastq_data` folder. Note only for the R1 file name is shown.

Sample	R1 filename
sample1	sample1_S1_L001_R1_001.fastq.gz
sample2	sample2_S1_L001_R1_001.fastq.gz
sample3	sample3_S1_L001_R1_001.fastq.gz
sample4	sample4_S1_L001_R1_001.fastq.gz
sample5	sample5_S1_L001_R1_001.fastq.gz
sample6	sample6_S1_L001_R1_001.fastq.gz

Based on the metadata supplied the following raw sequence counts were observed for each sample (this represents the number of sequenced read pairs).

Table 2. Raw sequence counts for each sample.

Sample ID	Raw sequence count
sample1	2000
sample2	2000
sample3	2000
sample4	2000
sample5	2000
sample6	2000

The QC plot for the R1 forward read is shown in the figure below. Note a quality score (y-axis) of >20 corresponds to base call accuracy of >99%.
The box plot shows the 25th, 50th (median), and 75th percentiles. The whiskers indicate the min and max values. The median quality score for each base is shown by the orange bar.

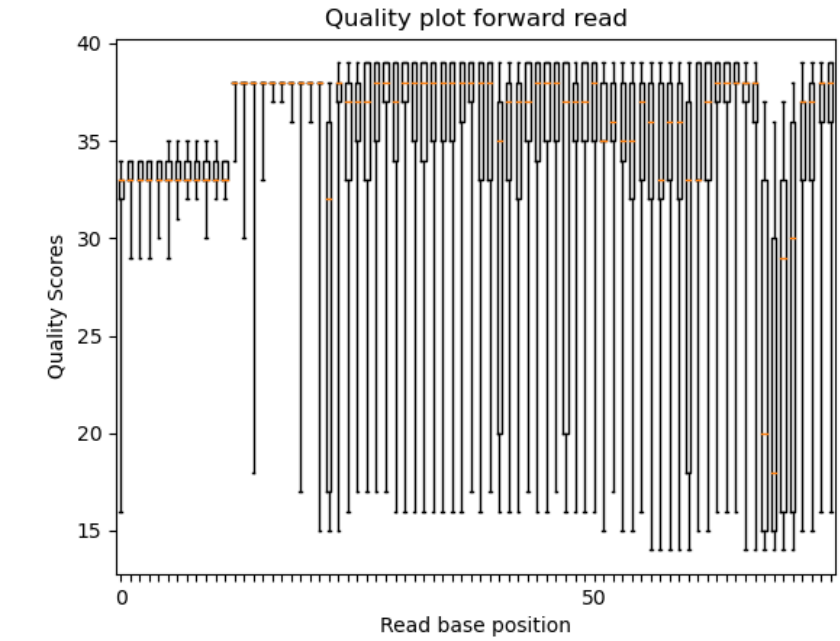

Figure 1: QC plot of foward R1 read

And the reverse R2 read.

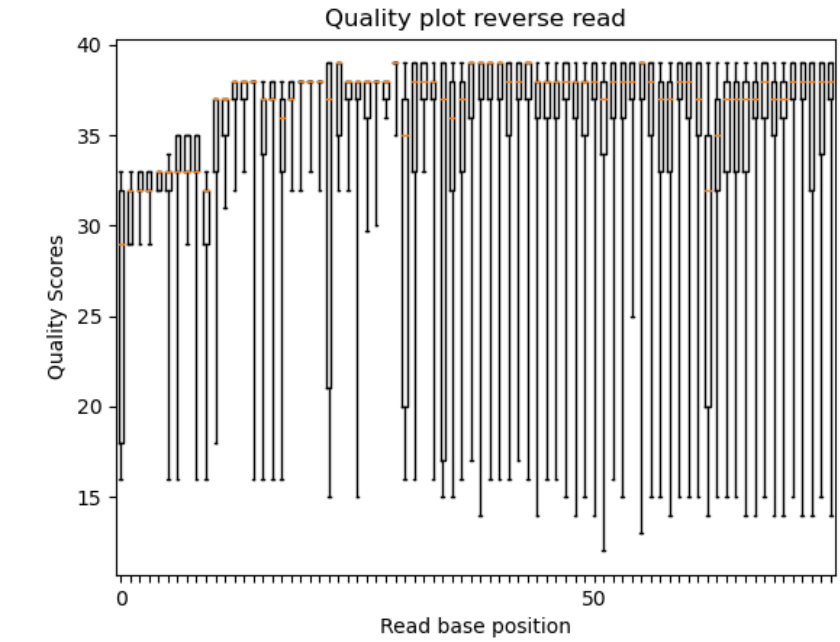

Figure 2: QC plot of reverse R2 read

Checking for primer sequences

An important sanity check is to confirm that primer sequences are being correctly identified and removed from the reads before downstream processing. If the primers are not correctly removed this will result in taxonomic errors (so it is important to check that this is working correctly). The panel below shows the first 3 reads before the primers sequences are removed by cutadapt. The forward primer (5-ACACCGCCCGTCACTCT-3) should be located at the 5' end of these sequences (note that sequence errors often occur at the very 5' end of the read). The first line shows the file these sequences were obtained from. A full description of the fastq format can be found here.

```
sample1_S1_L001_R1_001.fastq.gz
@M07867:16:000000000-DKGKT:1:1101:3869:21829 1:N:0:1
ACACCGCCCGTCATTCTCCCCAAGTTCAACCTGTCCTTCTAACTAAGAATTTAACCTAACAAAGGGGAGTCAAGT
+
BBBBBDDBBDBBGGGGGGGGGG22FGHHHHHHFHHHHHHHDHHHFHCDD3FFFGHH53BF3GAGFGG?01FEGDH
@M07867:16:000000000-DKGKT:1:1101:12655:3099 1:N:0:1
ACACCGCCCGTCACTCTCCTCGAAAAAACAACAATTTATAAATAAACTGCACCCCAACACAGAGATGAGGCAAGT
+
CCCCCCCCCCCCCGGGGGGGGGGBGEAGFGHHHHHHGHHHHGHFGGHHBFFHHHGEEFEFFBBBGHHHF3BCAE1G
@M07867:16:000000000-DKGKT:1:1101:13088:27401 1:N:0:1
TCACCGCCCGTCACTCTCCCCATGTTCTTATACATTAATAACTAATACCCCCCCAGAACAAAGGGGAGGCATGTCG
+
1A11A?DDDDDAAEGGGFGGGB0FF2FFFDFGG2BDADFE21B1EGHGEGEEEEAB/0B00B>C?/B>EA/1BBB?
```

The next panel is after cutadapt has removed the primers.

```
sample1_0_L001_R1_001.fastq.gz
@M07867:16:000000000-DKGKT:1:1101:3869:21829 1:N:0:1
CCCCAAGTTCAACCTGTCCTTCTAACTAAGAATTTAACCTAACAAAGGGGAGTCAAGT
+
GGGGG22FGHHHHHHFHHHHHHHDHHHFHCDD3FFFGHH53BF3GAGFGG?01FEGDH
@M07867:16:000000000-DKGKT:1:1101:12655:3099 1:N:0:1
CCTCGAAAAAACAACAATTTATAAATAAACTGCACCCCAACACAGAGATGAGGCAAGT
+
GGGGGBGEAGFGHHHHHHGHHHHGHFGGHHBFFHHHGEEFEFFBBBGHHHF3BCAE1G
@M07867:16:000000000-DKGKT:1:1101:13088:27401 1:N:0:1
CCCCATGTTCTTATACATTAATAACTAATACCCCCCCAGAACAAAGGGGAGGCATGTCG
```

```
+
FGGGB0FF2FFFDFGG2BDADFE21B1EGHGEGEEEEAB/0B00B>C?/B>EA/1BBB?
```

A summary of the primer trimming log file is shown below, the full file can be found in `logs/qimme2.trim.log`. We would expect a very high percentage (>90%) of the reads to contain primers otherwise it could indicate PCR issues. If any of these outputs show a low percentage please check the `logs/qimme2.trim.log` file to identify the problematic file.

```
  Read 1 with adapter:                  1,474 (73.7%)
  Read 2 with adapter:                  1,743 (87.2%)
--
  Read 1 with adapter:                  1,469 (73.5%)
  Read 2 with adapter:                  1,764 (88.2%)
--
  Read 1 with adapter:                  1,437 (71.9%)
  Read 2 with adapter:                  1,759 (87.9%)
--
  Read 1 with adapter:                  1,458 (72.9%)
  Read 2 with adapter:                  1,740 (87.0%)
--
  Read 1 with adapter:                  1,507 (75.3%)
  Read 2 with adapter:                  1,752 (87.6%)
--
  Read 1 with adapter:                  1,469 (73.5%)
  Read 2 with adapter:                  1,758 (87.9%)
```

Denoising with DADA2

This step filters out noisy sequences, correct errors in marginal sequences, removes chimeric sequences, removes singletons, joins denoised paired-end reads, and then de-replicates these sequences. The de-replicated sequences are used to assign taxonomy and generate taxonomic counts.

For target loci with variable length it is important to limit sequence truncation (in this case `p-trunc-len-f` and `p-trunc-len-r` should be set to `0`), otherwise this could create a sequence length bias. For fixed length loci the trimming parameters need to be

adjusted based the sequence quality profile (Figures 1 and 2) to ensure that the read overlaps includes high quality calls, whilst maximising the sequence overlap. The R2 read is normally trimmed more aggressively.

The settings used in this run were (from the `config.yaml` file):

```
--p-trunc-len-f 0
--p-trunc-len-r  0
--p-max-ee-f 2
--p-max-ee-r 4
--p-trunc-q 2
--consensus-method: consensus
```

A full explanation of the above settings is available here.

The tables below summarise the results of the DADA2 filtering and de-noising steps. The de-noising and chimera removal steps should only remove a small proportion (<30%) of reads otherwise the above DADA2 settings may need to be adjusted. The actual number of sequences retained depends on several factors, such as sequence error profile and paired-end read overlap length.

Table 3. De-noising results from the *dada2* algorithm.

Sample	**Input**	**Filtered**	**Passed filter (%)**	**de-noised**
sample1	1283	1279	99.69	1270
sample2	1309	1307	99.85	1298
sample3	1274	1271	99.76	1265
sample4	1269	1267	99.84	1260
sample5	1316	1311	99.62	1304
sample6	1293	1292	99.92	1288

The read merging and de-noising summary is below. Note the final non-chimeric sequences are used for assigning taxonomy and generating the taxonomy counts for each sample.

Table 4. Final non-chimeric sequence counts.

Sample	Merged	Input merged (%)	Non-chimeric	Non-chimeric (%)
sample1	1213	94.54	1213	94.54
sample2	1255	95.87	1255	95.87
sample3	1228	96.39	1228	96.39
sample4	1223	96.38	1223	96.38
sample5	1252	95.14	1252	95.14
sample6	1233	95.36	1233	95.36

Representative eDNA sequences identified

The following summary statics describe the representative sequences used to assign taxonomy and taxonomic counts that make up the eDNA profile.

Table 5. Summary metrics for the representative sequences used for taxonomic assignment (min, max, mean, range and standard deviation are base-pairs).

Metric	value
count	8
min	61
max	65
mean	63.125
range	4
std	1.24642

Final eDNA taxonomic count table

The final spreadsheet (CSV file) of taxonomic counts across the samples can be found at the following folder paths: `final_results/asv_count_tax_seqs_summary.csv`.
Each sample is represented by a column in the spreadsheet showing

the number of sequences for each species (rows). Note these are raw sample counts (they have not been normalised by the number of reads obtained for each sample).

The columns descriptions in the final spreadsheet are shown below in Table 6.

Table 6. Column descriptions for `asv_count_tax_seqs.csv`.

Column	Description
Feature_ID	Hash identification for reference sequence
Taxon	Taxonomic breakdown of reference sequence
Taxa_confidence	Confidence of taxonomic assignment
Reference_variants	The number of ASVs variants for this species
reference sequence	The reference DNA sequence use to make the assignment

Software versions used

A complete list of packages and versions is available in: `./env/qiime2-2022.2-py38-linux-conda.yml`

This pipeline was written by Dave Wheeler (DPI's Chief Scientist Unit). For any issues please make contact: dave.wheeler@dpi.nsw.gov.au