



## Environmental Analysis Program

Pomona College  
Biogeochemistry Laboratory

---

SOP-L705

# Bioinformatics Pipeline (QIIME 2 / DADA2)

Standard Operating Procedure

|                     |                            |
|---------------------|----------------------------|
| <b>SOP Number</b>   | SOP-L705                   |
| <b>Revision</b>     | 0.1                        |
| <b>Author(s)</b>    | Marc Los Huertos           |
| <b>Date</b>         | 2026-06-26                 |
| <b>Organization</b> | EA Program, Pomona College |

FOR INTERNAL USE ONLY

This document is the property of the Environmental Analysis Program at Pomona College.

Unauthorized distribution is not permitted.

---

Table of contents

1 Purpose 1

2 Scope 2

3 Responsibilities 2

4 Definitions 2

5 Software Requirements 3

6 Procedure 3

6.1 Data Organization . . . . . 3

6.2 Import Raw Reads . . . . . 4

6.3 Denoising with DADA2 . . . . . 4

6.4 Taxonomic Classification . . . . . 5

6.5 Phylogenetic Tree Construction . . . . . 5

6.6 Filtering . . . . . 6

6.7 Export for R Analysis . . . . . 6

7 Quality Control 6

8 References 7

9 Related SOPs 7

10 Training Record 8

1. Purpose

This SOP describes the bioinformatics workflow for processing amplicon sequencing data (16S rRNA, ITS, 18S rRNA) from raw demultiplexed FASTQ files through quality filtering, denoising, taxonomic classification, and generation of analysis-ready feature tables. The pipeline uses QIIME 2 as the primary framework with DADA2 as the denoising algorithm (Bolyen et al., 2019; Callahan et al., 2016). DADA2 resolves amplicon sequence variants (ASVs) at single-nucleotide resolution, replacing the older OTU-clustering approach with a more reproducible and sensitive method (Callahan et al., 2017).

## 2. Scope

- **Applies to:** Processing of Illumina paired-end amplicon sequencing data (MiSeq, NextSeq) for 16S, ITS, and 18S marker genes generated from EA Program environmental samples
- **Does not apply to:** Shotgun metagenomics; long-read sequencing (PacBio, Nanopore); transcriptomics (RNA-seq); ddPCR data analysis (see [SOP-L702](#))
- **Location:** EA Program lab computers or personal workstations with QIIME 2 installed; high-performance computing cluster if available

## 3. Responsibilities

| Role              | Responsibility                                         |
|-------------------|--------------------------------------------------------|
| Lab PI            | Approve pipeline parameters and taxonomy database      |
| Trained Personnel | Run the pipeline per this SOP; document all parameters |
| New Students      | Complete training before independent analysis          |

## 4. Definitions

- **ASV (Amplicon Sequence Variant):** An exact DNA sequence resolved from amplicon sequencing data by denoising algorithms (DADA2, Deblur); a higher-resolution alternative to OTUs that retains single-nucleotide differences between sequences.
- **DADA2:** A denoising algorithm that models sequencing error rates to distinguish true biological sequences from sequencing errors; produces ASVs directly from paired-end reads without clustering.
- **Demultiplexing:** The computational step of assigning reads to their source sample based on index sequences; produces per-sample FASTQ files from a multiplexed sequencing run.
- **Feature table (OTU/ASV table):** A matrix of counts where rows are samples, columns are ASVs or OTUs, and each cell contains the number of reads from that sample assigned to that sequence variant; the primary data object in QIIME 2.
- **FASTQ:** A text-based file format storing DNA sequences and their per-base quality scores (Phred scores); the standard output format from Illumina sequencers; paired-end data produces R1 and R2 FASTQ files per sample.
- **OTU (Operational Taxonomic Unit):** A cluster of sequences grouped by a similarity threshold (typically 97% identity); a coarser representation of diversity than ASVs; largely superseded by ASV-based approaches.
- **Phred quality score (Q score):** A logarithmic measure of base-call accuracy:  $Q = -10 \log_{10}(P_{error})$ ; Q20 = 99% accuracy; Q30 = 99.9% accuracy; used to set quality-trimming thresholds in DADA2.
- **QIIME 2:** An open-source bioinformatics platform for microbiome analysis; manages data as typed artifacts (.qza) with embedded provenance records; wraps DADA2, taxonomic classifiers, diversity analyses, and visualization tools.
- **Rarefaction:** Subsampling each sample to the same number of reads to correct for unequal sequencing depth before diversity comparisons; samples below the rarefaction depth are excluded.

- **Taxonomic classifier:** A trained machine-learning model (e.g., naive Bayes classifier trained on SILVA or UNITE) that assigns taxonomy to ASV sequences by comparing them to a reference database.

## 5. Software Requirements

| Software               | Version (minimum)    | Notes                                                                                      |
|------------------------|----------------------|--------------------------------------------------------------------------------------------|
| QIIME 2                | 2024.2 or later      | Install via conda/mamba; see <a href="https://docs.qiime2.org">https://docs.qiime2.org</a> |
| DADA2 (QIIME 2 plugin) | Bundled with QIIME 2 | R package also available for standalone use                                                |
| R and RStudio          | R 4.3+               | For downstream statistical analysis; see <a href="#">SOP-L801</a>                          |
| phyloseq (R package)   | 1.44+                | For importing QIIME 2 artifacts into R (McMurdie & Holmes, 2013)                           |
| FastQC                 | 0.12+                | Optional; for initial read quality inspection                                              |

Install QIIME 2 in a dedicated conda environment to avoid dependency conflicts:

```
# Example installation (check docs.qiime2.org for current commands)
conda env create -n qiime2-2024.2 \
  --file https://data.qiime2.org/distro/amplicon/qiime2-amplicon-2024.2-py39-linux-conda.yml
conda activate qiime2-2024.2
```

## 6. Procedure

### 6.1. Data Organization

1. Create a project directory with the following structure:

```
project_name/
  raw_data/      # Demultiplexed FASTQ files from sequencing facility
  qiime2/        # QIIME 2 artifacts and visualizations
  results/       # Exported tables, trees, taxonomy
  scripts/       # Shell and R scripts
  metadata/      # Sample metadata TSV file
```

2. Prepare a sample metadata file (`metadata.tsv`) in tab-separated format. The first column must be `sample-id` matching the FASTQ filenames. Include all experimental variables (site, treatment, date, depth, replicate). Validate the file at <https://keemei.qiime2.org/> (Rideout et al., 2014).

## 6.2. Import Raw Reads

Import demultiplexed paired-end FASTQ files into a QIIME 2 artifact:

```
qiime tools import \  
  --type 'SampleData[PairedEndSequencesWithQuality]' \  
  --input-path raw_data/ \  
  --input-format CasavaOneEightSingleLanePerSampleDirFmt \  
  --output-path qiime2/demux.qza
```

Visualize read quality:

```
qiime demux summarize \  
  --i-data qiime2/demux.qza \  
  --o-visualization qiime2/demux-summary.qzv
```

Open `demux-summary.qzv` at <https://view.qiime2.org> to inspect quality score profiles. Use these plots to determine truncation parameters for DADA2 (Section 5.3).

## 6.3. Denoising with DADA2

DADA2 performs quality filtering, chimera removal, and denoising in a single step, producing an ASV feature table and representative sequences (Callahan et al., 2016).

```
qiime dada2 denoise-paired \  
  --i-demultiplexed-seqs qiime2/demux.qza \  
  --p-trunc-len-f 240 \  
  --p-trunc-len-r 200 \  
  --p-trim-left-f 0 \  
  --p-trim-left-r 0 \  
  --p-n-threads 4 \  
  --o-table qiime2/table.qza \  
  --o-representative-sequences qiime2/rep-seqs.qza \  
  --o-denoising-stats qiime2/dada2-stats.qza
```

### ! Setting Truncation Parameters

Choose `--p-trunc-len-f` and `--p-trunc-len-r` based on the quality plots from Section 5.2. Truncate where the median quality score drops below 25–30. The forward and reverse reads must still overlap by at least 20 bp after truncation for successful merging. For 16S V4 (515F/806R, ~290 bp amplicon) with 2x250 reads, typical values are 240 (forward) and 200 (reverse).

Review denoising statistics:

```
qiime metadata tabulate \  
  --m-input-file qiime2/dada2-stats.qza \  
  --o-visualization qiime2/dada2-stats.qzv
```

Check the percentage of reads passing each filter step (input, filtered, denoised, merged, non-chimeric). A healthy run retains > 60% of input reads through the full pipeline.

## 6.4. Taxonomic Classification

Assign taxonomy to ASVs using a pre-trained Naive Bayes classifier. Download the appropriate classifier from the QIIME 2 Data Resources page (<https://docs.qiime2.org/2024.2/data-resources/>).

**For 16S (SILVA 138.1):**

```
qiime feature-classifier classify-sklearn \  
  --i-classifier silva-138.1-ssu-nr99-515f-806r-classifier.qza \  
  --i-reads qiime2/rep-seqs.qza \  
  --o-classification qiime2/taxonomy.qza
```

**For ITS (UNITE):**

```
qiime feature-classifier classify-sklearn \  
  --i-classifier unite-ver9-dynamic-all-29.11.2022-classifier.qza \  
  --i-reads qiime2/rep-seqs.qza \  
  --o-classification qiime2/taxonomy.qza
```

Visualize taxonomy:

```
qiime taxa barplot \  
  --i-table qiime2/table.qza \  
  --i-taxonomy qiime2/taxonomy.qza \  
  --m-metadata-file metadata/metadata.tsv \  
  --o-visualization qiime2/taxa-barplot.qzv
```

## 6.5. Phylogenetic Tree Construction

Build a rooted phylogenetic tree for diversity analyses:

```
qiime phylogeny align-to-tree-mafft-fasttree \  
  --i-sequences qiime2/rep-seqs.qza \  
  --o-alignment qiime2/aligned-rep-seqs.qza \  
  --o-masked-alignment qiime2/masked-aligned-rep-seqs.qza \  
  --o-tree qiime2/unrooted-tree.qza \  
  --o-rooted-tree qiime2/rooted-tree.qza
```

## 6.6. Filtering

Remove contaminant ASVs identified in negative controls, mitochondrial and chloroplast sequences, and low-frequency ASVs:

```
# Remove mitochondria and chloroplast
qiime taxa filter-table \
  --i-table qiime2/table.qza \
  --i-taxonomy qiime2/taxonomy.qza \
  --p-exclude mitochondria,chloroplast \
  --o-filtered-table qiime2/table-filtered.qza

# Remove low-frequency ASVs (e.g., singletons)
qiime feature-table filter-features \
  --i-table qiime2/table-filtered.qza \
  --p-min-frequency 2 \
  --o-filtered-table qiime2/table-final.qza
```

For decontamination using negative control data, export the feature table to R and use the `decontam` package (Davis et al., 2018).

## 6.7. Export for R Analysis

Export QIIME 2 artifacts for downstream analysis in R with `phyloseq` (McMurdie & Holmes, 2013):

```
qiime tools export \
  --input-path qiime2/table-final.qza \
  --output-path results/

qiime tools export \
  --input-path qiime2/taxonomy.qza \
  --output-path results/

qiime tools export \
  --input-path qiime2/rooted-tree.qza \
  --output-path results/
```

Alternatively, import QIIME 2 artifacts directly into R using the `qiime2R` package.

## 7. Quality Control

- Inspect the DADA2 denoising stats for every run. Flag samples with fewer than 1,000 reads after denoising.
- Verify that negative controls (NTC, extraction blank) contain fewer than 100 reads. If controls contain substantial reads, use `decontam` to identify and remove contaminant ASVs (Davis et al., 2018; Salter et al., 2014).

- Confirm that the positive control (mock community) recovers the expected taxa at the expected relative abundances.
- Document all QIIME 2 and DADA2 parameters in the project README and in the analysis scripts. Reproducibility requires recording the exact software versions, classifier version, and truncation lengths.
- Archive raw FASTQ files, the QIIME 2 artifacts, and the analysis scripts in the project GitHub repository per [SOP-L801](#).

## 8. References

- Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, 37(8), 852–857. <https://doi.org/10.1038/s41587-019-0209-9>
- Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, 11(12), 2639–2643. <https://doi.org/10.1038/ismej.2017.119>
- Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13(7), 581–583. <https://doi.org/10.1038/nmeth.3869>
- Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., & Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*, 6(1), 226. <https://doi.org/10.1186/s40168-018-0605-2>
- McMurdie, P. J., & Holmes, S. (2013). Phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLOS ONE*, 8(4), e61217. <https://doi.org/10.1371/journal.pone.0061217>
- Rideout, J. R., He, Y., Navas-Molina, J. A., Walters, W. A., Ursell, L. K., Gibbons, S. M., Chase, J., McDonald, D., Gonzalez, A., Robbins-Pianka, A., et al. (2014). Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ*, 2, e545. <https://doi.org/10.7717/peerj.545>
- Salter, S. J., Cox, M. J., Sherrill-Mix, S., Turner, R. M., Sherrill-Mix, S., Sherrill-Mix, S., & Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, 12, 87. <https://doi.org/10.1186/s12915-014-0087-z>

## 9. Related SOPs

| SOP                  | Title                       | Relationship                        |
|----------------------|-----------------------------|-------------------------------------|
| <a href="#">L700</a> | DNA Extraction              | Upstream: sample preparation        |
| <a href="#">L703</a> | PCR and Gel Electrophoresis | Amplicon verification               |
| <a href="#">L704</a> | Amplicon Library Prep       | Upstream: library preparation       |
| <a href="#">L706</a> | eDNA Metabarcoding          | eDNA-specific application           |
| <a href="#">L801</a> | RStudio and GitHub          | Data management and version control |
| <a href="#">L802</a> | Quality Control and QA      | General QA/QC framework             |

## 10. Training Record

Before running the bioinformatics pipeline independently, personnel must:

1. Read this SOP in full
2. Complete the QIIME 2 “Moving Pictures” tutorial (<https://docs.qiime2.org/2024.2/tutorials/moving-pictures/>)
3. Process a test dataset under supervision, from import through taxonomy assignment and export to R
4. Demonstrate correct interpretation of DADA2 denoising statistics and quality plots
5. Be signed off by the PI or lab manager