A logo with a letter and a logo

AI-generated content may be incorrect.

**Fast and accurate sample inference from amplicon data with single-nucleotide resolution**

**Current version:**

v.1.26 (as of 3-9-2025)

**A short introduction to DADA2**

[DADA2](https://benjjneb.github.io/dada2/) is an open-source software package designed to correct errors in Illumina-sequenced DNA data. It is primarily used to analyze amplicon data, which consists of DNA sequences obtained through PCR amplification.

The amplicon sequencing process causes errors to sequencing data and these error severely complicate the interpretation of the results.

To solve this problem, DADA2 implements a novel algorithm that models the errors, which are introduced during amplicon sequencing. This error model will be used to infer the true sample composition.

The DADA2 method is more sensitive and specific than the [OTU](https://en.wikipedia.org/wiki/Operational_taxonomic_unit) (Operational Taxonomic Unit) method. DADA2 detects real biological variation which are missed by OUT methods while outputting fewer spurious sequences (wrong/unwanted sequences).

**DADA2 pipeline**

1. **Input**

The input for DADA2 is a set of demultiplexed FASTQ files, meaning each sample has its own individual FASTQ file (or two files for paired-end sequencing) from an amplicon sequencing study. If demultiplexing has not been performed yet, tools like **QIIME’s** [split\_libraries\_fastq.py](http://qiime.org/scripts/split_libraries_fastq.html) and [split\_sequence\_file\_on\_sample\_ids.py](http://qiime.org/scripts/split_sequence_file_on_sample_ids.html), or the **idemp** utility can be used.

**NOTE: Non-biological bases such as PCR primers should be removed before proceeding**

1. **Quality control (plotQualityProfile())**

DADA2 has a built-in plotQualityProfile() function, which is an alternative to FastQC. It allows the user to visually inspect the quality of the reads without using an external tool.

Since DADA2 natively handles compressed (.gz) files, it can perform quality control directly without the need to unpack them first.

1. **Filter and Trim (filterAndTrim())**

With the filterAndTrim() function, DADA2 allows the user to filter and trim the sequencing reads to remove low-quality sequences and reduce errors. This step is crucial to ensure that only high-quality data is used in the analysis. The filtered sequences are saved as gzipped FASTQ files to save disk space.

1. **Dereplicate (derepFastq())**

This function is used to group identical sequences together (dereplication) by finding exact matches in the dataset and counting the occurrences. This step reduces computational load and improves analysis efficiency.

1. **Learn Error Rates (learnErrors())**

The DADA2 algorithm is used to estimate the error rates in the sequencing data, as sequencing technologies can introduce errors that affect the identification of true biological sequences.

The estimation is done by modeling errors, such as substitutions, insertions, and deletions, based on the actual data. This function uses the quality scores associated with each base to identify where errors are most likely to occur.

By estimating error rates directly from the data, this method can adapt to different sequencing runs and protocols, rather than relying on fixed error models.

1. **Infer sample composition (dada(…))**

The core method of DADA2 focuses on sample composition inference, which is the process of accurately identifying the true biological sequences present in a sample.

The DADA2 algorithm is more sensitive and specific than traditional OTU (Operational Taxonomic Unit) methods. It can detect subtle sequence differences and distinguish true biological variants from sequencing errors.

Unlike OTU-based methods, which cluster sequences, DADA2 uses **denoising** to correct sequencing errors and infer the true biological composition at a higher resolution. The output consists of **sequence variants (SVs)**, representing the true biological sequences, rather than OTUs grouped by similarity.

1. **Merge forward/reverse reads (mergePairs…())**

The mergePairs() function combines the forward and reverse reads into an single, complete sequence. This is done by matching the overlapping parts (more than 20 bp) of the forward and reverse reads. If the reads don’t overlap enough or don’t match well, they will be discarded.

1. **Remove chimeras (removeBimeraDenovo())**

The removeBimeraDenovo() function is used to identify and remove chimeric sequences, which are spurious sequences formed during PCR amplification when an incomplete amplicon primes the next amplification cycle. These chimeras consist of fragments from different sequences, leading to false results.

1. **Make a sequence table (makeSequenceTable())**

The makeSequenceTable() function combines the inferred sequence variants from all of the samples into a unified table, where rows represent samples and columns represent unique, non-chimeric sequences.

After the table is created, the function removeBimeraDenovo() is applied again to remove any remaining chimeric sequences.

The result is a clean, non-chimeric sequence table, similar to an OTU table, which can be used for downstream analysis, often with the phyloseq package for further exploration of sample composition.

**Advantages of DADA2**

* **High accuracy**: Identifies exact DNA sequences instead of approximating groups
* **Better taxonomic resolution**: Distinguishes closely related species more precisely
* **Integration QIIME2**: DADA2 works well with QIIME2, which is a widely used microbiome analysis tool

**Extra**

* [**A detailed tutorial about DADA2**](https://www.bioconductor.org/packages/devel/bioc/vignettes/dada2/inst/doc/dada2-intro.html)
* **Installation:**

Binaries for the current version of DADA2 are available from Bioconductor version 3.16. To install the package, start R (version 4.2.0 or newer).

Enter the following in R:

A close up of a text

AI-generated content may be incorrect.

**Sources**

1. <https://benjjneb.github.io/dada2/>
2. <https://www.nature.com/articles/nmeth.3869>
3. <https://callahanlab.cvm.ncsu.edu/software/>
4. <https://www.bioconductor.org/packages/devel/bioc/vignettes/dada2/inst/doc/dada2-intro.html>