Output Curation User Guide

Authors: Fabricio dos Anjos Santa Rosa
Tiago Ferreira Leão
Renato R. M. Oliveira
Document version: 1.0
Date: 30/06/2025

This section outlines the steps for curating and validating taxonomic outputs derived from the PIMBA 3.0 workflow. The process includes automatic filtering of inconsistent records, flagging for manual validation, and final preparation for downstream analyses.

1. Importing and Merging Datasets

The initial step involves combining the taxonomic assignments, OTU table, and their corresponding DNA sequences into a single dataset. This unified dataset serves as the basis for quality control and manual curation steps.

2. Automated Flag Generation

A series of automated checks are applied to identify taxonomic entries that require manual review. Flags are generated based on the following criteria:

- Species names composed solely of capital letters, or numbers, that do not follow the standard 'sp.' format;
- Entries with a single-character value;
- Complex or malformed taxonomic strings, e.g., those with multiple special character blocks.

These conditions help detect potentially invalid or unreliable taxonomic identifications early in the process.

3. NCBI Taxonomic Validation

For entries not removed during the initial checks, a second validation step is performed using NCBI taxonomic data. If the full species name (Genus + Species) is not found as a valid taxon in NCBI, the species name is removed and only the genus is retained as the most reliable identification for that record.

4. Excel Output Structure

The output Excel file is automatically generated and includes the following tabs:

- Raw data with sequences: Includes the original input data for each OTU, its associated DNA sequence, and sequence length;

- Filtered data with flags: Contains only the entries that passed basic quality checks, with flags indicating entries that require manual validation;
- Manual review (user-filled): A dedicated tab for the user to perform and record manual corrections and validations.

OBS: After completing the automated filtering and flagging steps, manual review should be performed on the flagged taxonomic records. To do this safely, it is strongly recommended that you make a copy of the "Filtered data with flags" tab and save your manual edits in a separate Excel file. This precaution is important because if the pipeline is re-executed, the original output file (including its tabs) will be overwritten, and any manual work done directly on the file may be lost.

5. Manual Review and Re-import

After completing the manual validation in the Excel file, the reviewed data can be re-imported for downstream analyses and visualization.