PIMBA: A PIpeline for MetaBarcoding Analysis that allows the use of a personalized reference database.

Renato Renison Moreira Oliveira, Guilherme Oliveira

Universidade Federal de Minas Gerais

Abstract

DNA metabarcoding is an emerging field in biodiversity analysis. The increase in data generation thanks to Next-Generation Sequencing caused the development of new DNA metabarcoding techniques. Mothur, Qiime, Obitools, and mBRAVE are currently the most used metabarcoding analysis tools. The only problem with those pipelines is that they do not allow the researcher to use a personalized database. For example, Mothur is only useful when analyzing 16S data. Qiime (and even its updated version, Qiime2) is optimized to analyze metabarcoding data from 16S, 18S and fungal ITS marker genes, using Greengenes, SILVA and UNITE databases, respectively. If the researcher is interested in using the Qiime pipeline to analyze data sequenced from different marker genes, such as COI or Plant ITS, Qiime does not give support to adapt its pipeline for such personalized use. Obitools is optimized to analyze data from 16S (SILVA and PR2). Obitools also allows the use of the NCBI database for taxonomic assignment. mBRAVE is optimized to use only the BOLD database as a reference, allowing the researcher to use a personalized database only after BOLD submission. Here we present PIMBA, a pipeline for metabarcoding analysis which adapts the Qiime pipeline in order to allow the researcher to additionally use a personalized or the NCBI databases. PIMBA also implements all the features provided by the other metabarcoding tools such as reference databases for 16S, 18S and Fungal ITS. PIMBA performs the NCBI taxonomic assignment by blasting the resulting OTUs against the NCBI nt database and using taxdump, it retrieves the full taxonomic information, creating all the files that Qiime needs to perform its final analysis. PIMBA also allows the researcher to analyze datasets from marker genes without a published reference database (COI, Plant ITS, rbcL, matK, etc.), where it is only needed a fasta file with the full taxonomy written in the header of the sequences. We used PIMBA to analyze two metabarcoding datasets (Plant ITS and Invertebrates COI). PIMBA was able to generate all the outputs needed to infer results and additional metabarcoding analysis, such as alfa and beta diversity analyses, PCoA and taxonomic bar plots. Next steps include benchmarking PIMBA against the most used metabarcode tools. We will also create a docker image, making PIMBA easily deployable by the user.

Funding: 372439/2019-5, CNPq