Application Notes

PIMBA: a Pipeline for MetaBarcoding Analysis

Renato R. M. Oliveira^{1,2*}, Raíssa Silva¹, Gisele L. Nunes¹ and Guilherme Oliveira^{1*}

¹Instituto Tecnológico Vale, Belém, Pará, Brazil;

²Programa de Pós-Graduação em Bioinformática, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil.

*To whom correspondence should be addressed.

ABSTRACT

Summary: DNA metabarcoding is an emerging monitoring method capable of assessing biodiversity from environmental samples (eDNA). Advances in computational tools have been required due to the increase of Next-Generation Sequencing data. Tools for DNA metabarcoding analysis, such as MOTHUR, QIIME, Obitools, and mBRAVE have been widely used in ecological studies. However, some difficulties are encountered when there is a need to use custom databases. Here we present PIMBA, a Pipeline for MetaBarcoding Analysis, which allows the use of customized databases, as well as other reference databases used by the softwares mentioned here. PIMBA is an open-source and user-friendly pipeline that consolidates all analyses in just three command lines.

Availability and Implementation: <https://github.com/reinator/pimba>

Contact: renato.renison@gmail.com and guilherme.oliveira@pq.itv.org.

Introduction

DNA metabarcoding is a powerful tool that has been widely used for biodiversity monitoring and ecosystem assessment from DNA information. The technique based on the high-throughput sequencing (HTS) allows the multispecies detection with DNA from environmental samples (eDNA) using a specific molecular marker for specific groups (plants, animals, fungi, and bacteria) (Creer *et al.*, 2016). Next-Generation Sequencing (NGS) results in millions of DNA sequences (reads) that are sequenced to decipher the genetic code of species that help to answer taxonomic and functional questions. For DNA metabarcoding analysis, the reads are grouped into operational taxonomic units (OTUs) to differentiate species or taxa based on the similarity of genetic code (Cristescu, 2014; Hering *et al.*, 2018). Similarity of 97% is commonly used as a cutoff, but this value depends on the group to be evaluated (Deiner *et al.*, 2017).

The recent increase of NGS data has caused the development of new tools for DNA metabarcoding analysis, making the metabarcoding method more accessible and user-friendly (Alberdi *et al.*, 2018). Mothur (Schloss *et al.*, 2009), Qiime (Caporaso *et al.*, 2010), Obitools (Boyer *et al.*, 2016), and mBRAVE (Ratnasingham, 2019) are currently the most used tools for metabarcoding analysis. The biggest restriction of these pipelines is to make it difficult to use a customized database for taxonomy assignment. For example, Mothur is useful when analyzing 16S/18S rRNA, Influenza viral, and fungal ITS regions, using Greengenes (DeSantis *et al.*, 2006), Influenza Virus, and SILVA (Quast *et al.*, 2013) databases, respectively. Qiime (and even its updated version, Qiime2) is optimized to analyze metabarcoding data from 16S rRNA, 18S, and fungal ITS marker genes, using Greengenes, SILVA, and UNITE (Abarenkov *et al.*, 2010) databases, respectively. Qiime2 allows the user to train a classifier with a NaïveBayes model, but they report tests by using only a 16S example and with some constraints to the use of this classifier with other marker genes. Qiime does not support adapting its pipeline to use a custom database to analyze sequenced data from different marker genes, such as COI or Plant ITS. Obitools is optimized to analyze data from 16S (SILVA and PR2) and it 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.

We developed PIMBA, a Pipeline for MetaBarcoding Analysis, which adapts the Qiime pipeline for OTUs clustering with additional OTU corrections based on the algorithm LULU (Frøslev *et al.*, 2017). A preliminary abundance and diversity analysis are also automatically delivered. The main innovation of this pipeline is the ease of using both standard and customized databases minimizing errors in taxonomic assignments making the results more reliable.

Implementation

Besides implementing all the features provided by the other metabarcoding tools, PIMBA also allows the user to apply different databases and not only those commonly used by most of the available softwares. PIMBA can be used with single or paired-end reads and is divided into three steps: (1) preprocessing, which promotes the demultiplexing and quality treatment of reads (Figure 1A); (2) clustering, in which reads are clustered into OTUs, along with errors correction and taxonomic assignment (Figure 1B), and (3) plotting, in which alpha and beta diversity plots are built by Phyloseq (McMurdie and Holmes, 2013), including rarefaction curves and Principal Coordinates Analysis (PCoA), respectively. A metadata file is required for this last step.

Fig. 1: Workflow of PIMBA steps for preprocessing (1A), clustering (1B) and plotting (1C).

Results

The table below shows the accuracy of PIMBA when using SILVA 16S (which uses the same steps as QIIME) or the NCBI database.

Database	Number of OTUs	PIMBA's sensibility
SILVA	100	85%
NCBI	100	85%

Results show that whether using SILVA or the NCBI database, PIMBA can obtain the same sensibility when identifying taxa in a simulated database. Since for the SILVA database the steps were the same that

Funding

Conflict of Interest: none declared.

- Abarenkov, K. *et al.* (2010) The UNITE database for molecular identification of fungi – recent updates and future perspectives. *New Phytol.*, **186**, 281–285.
- Alberdi, A. *et al.* (2018) Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods Ecol. Evol.*, **9**, 134–147.
- Almeida, A. *et al.* (2018) Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. *Gigascience*, **7**.
- Boyer, F. *et al.* (2016) obitools : a unix -inspired software package for DNA metabarcoding. *Mol. Ecol. Resour.*, **16**, 176–182.
- Caporaso, J.G. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.
- Creer, S. *et al.* (2016) The ecologist's field guide to sequence-based identification of biodiversity. *Methods Ecol. Evol.*, **7**, 1008–1018.
- Cristescu, M.E. (2014) From barcoding single individuals to metabarcoding biological communities: Towards an integrative approach to the study of global biodiversity. *Trends Ecol. Evol.*, **29**, 566–571.
- Deiner, K. *et al.* (2017) Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Mol. Ecol.*, **26**, 5872–5895.
- DeSantis, T.Z. *et al.* (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, **72**, 5069–5072.
- Frøslev, T.G. *et al.* (2017) Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nat. Commun.*, **8**, 1–11.
- Hering, D. *et al.* (2018) Implementation options for DNA-based identification into ecological status assessment under the European Water Framework Directive. *Water Res.*, **138**, 192–205.
- McMurdie, P.J. and Holmes, S. (2013) phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One*, **8**, e61217.
- Quast, C. *et al.* (2013) The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.
- Ratnasingham, S. (2019) mBRAVE: The Multiplex Barcode Research And Visualization Environment. *Biodivers. Inf. Sci. Stand.*, **3**, e37986.
- Schloss, P.D. *et al.* (2009) Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.