RaPDTool: Rapid Profiling and Deconvolution Tool for metagenomes

(Herramienta rápida de creación de perfiles y deconvolución a partir de metagenomas)

Authors: Ayixon Sánchez-Reyes & Luz Bretón-Deval

Affiliations

Cátedras Conacyt-Instituto de Biotecnología, Universidad Nacional Autónoma de México, Avenida Universidad 2001, Chamilpa, 62210, Cuernavaca, Morelos, México

Correspondence to:

Ayixon Sánchez-Reyes: ayixon.sanchez@mail.ibt.unam.mx

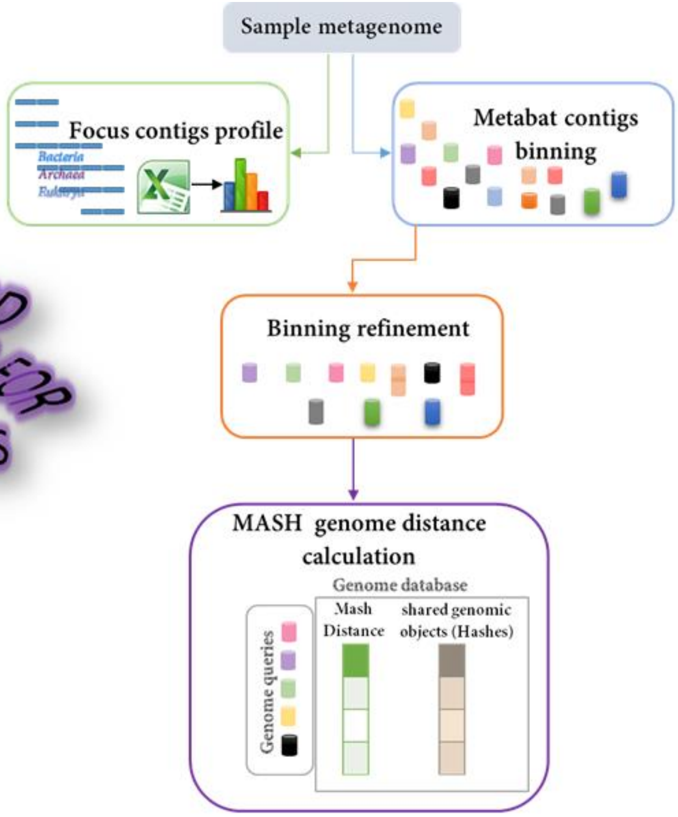


Figure 1. RaPDTool acyclic workflow

Abstract: RaPDTool offer a simple and easy-to-use tool for microbial communities profiling, contigs binning and "genomic-distance" exploration by connecting a series of bioinformatic tools in a single workflow. First, generate a taxonomic profile from massive sequencing data (fasta short reads, metagenome assemblies). Second, aggregate a metagenome into individual genomes or bins, and refine the set of MAGs. Finally, evaluate the probable "taxonomic neighborhoods" of each resulting genome bin by calculation of genomic distance against a custom database. The tool is available on: https://github.com/ayixon/RaPDTool.

Development:

RaPDTool connect the following four functions:

1. Generate a taxonomic profile from massive sequencing data (fasta short reads, metagenome assemblies). RaPDTool can use raw reads or metagenomic assemblies and call FOCUS profiler to report the organisms/abundance present in the metagenome.

2. Deconvolve a metagenome into individual genomes or bins. If the input consists on a metagenome assembly, RaPDTool automatically call Metabat2 to aggregate individual genome bins.

3. Refine the set of Metagenome Assembled Genomes (MAGs or bins). The bins are subsequently refined with Binning_refiner (https://github.com/songweizhi/Binning_refiner) to produce a non-redundant set.

4. Evaluate the probable "taxonomic neighborhoods" of each resulting genome bin. RaPDTool compare each bin against curated taxonomic mash databases like type material genome database (https://figshare.com/ndownloader/files/30851626). Alternatively, it can be compared against the database GTDB-r202 (https://figshare.com/ndownloader/files/30863182). Both databases are offered as representations or sketches that reduce storage space and computing time.

Dependencies:

FOCUS (https://github.com/metageni/FOCUS)

Metabat2 (https://bitbucket.org/berkeleylab/metabat/src/master/)

Binning_refiner (https://github.com/songweizhi/Binning_refiner)

Mash (https://github.com/marbl/Mash)

Before running the application, make sure all dependencies are installed in your PATH directory. For convenience we recommend installing the dependencies from the Conda package manager (https://bioconda.github.io/).

How to install:

RaPDTool it is written in python an C and runs natively by calling the script: rapdtool.py.
Also you will need the accompanying C scripts from the GitHub repository (https://github.com/ayixon/RaPDTool).

You can clone the repository by calling the following command:

```
git clone https://github.com/ayixon/RaPDTool.git
```

Subsequently, you can change the directory and access the main script

Usage:

```
rapdtool.py [-h] [-i INPUT] [-d DATABASE] [-r ROOT] [-c COMMENT]


Focus/Metabat/Binning_refiner/Mash (fmbm) script


optional arguments: -h, --help show this help message and exit


-i INPUT, --input INPUT
                        process this file
```

```
-d DATABASE, --database DATABASE

                        use this database


-r ROOT, --root ROOT  fmbm root subdirectory (default: user home)


-c COMMENT, --comment COMMENT

                        "comment for this execution"
```

example:

rapdtool.py -i INPUT.fasta -d DATABASE.msh -r OUTPUT_FOLDER

Database currently available:

NCBI Prokaryotic type material genomes (https://figshare.com/ndownloader/files/30851626)

GTDB-r202 Genome Taxonomy Database

(https://figshare.com/ndownloader/files/30863182)

Output files:

The RaPDTool output is stored in the $HOME directory. The -r option allows to assign a name to the output folder. The pipeline results are stored in subdirectories easily identifiable by the user:

genomadb: user database

input: input metagenome

profiles: Focus profiling results

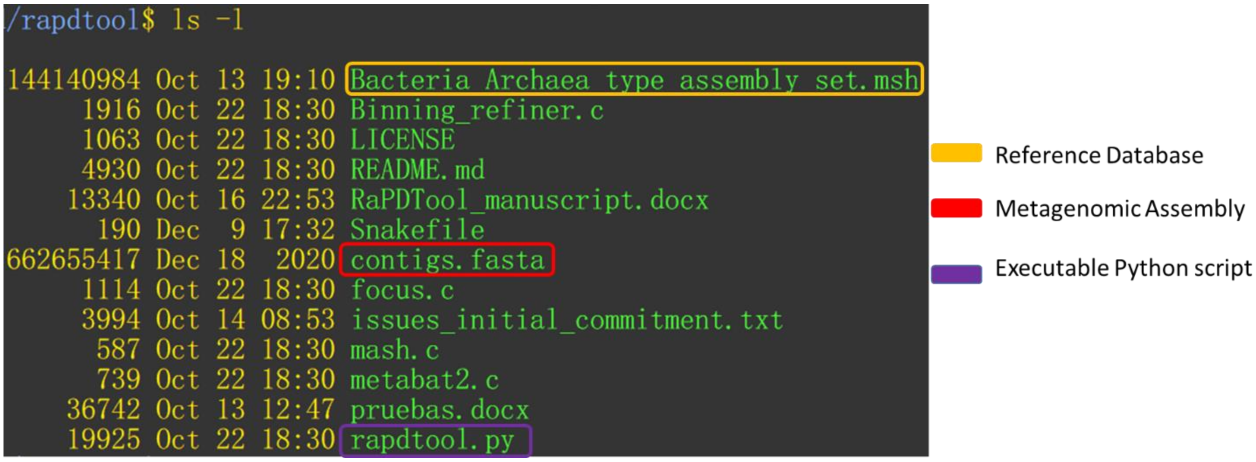
result: Metabat and Binning_refiner result

workf: Summary mash distance calculation

RaPDTool produces individual mash comparisons for every genome bin obtained against the user database (If you select prokaryotic NCBI Type Material DB there will be near to 17,000 records, GTDB contains many more). For this reason, the subdirectory "allresults" contain the ten closest hits from the mash paired comparison for each genome. This simplifies the interpretation of the results by limiting the Mash comparison to the ten closest neighbors to the query, which can be useful in phylogenetics and taxonomy. The user can take this list as the basis for a finer comparison by estimating the Overall genome relatedness index (OGRI) like Average Nucleotide Identity.

Proof of concept and validation

The RaPDTool app requires two mandatory elements, a group of contigs representing a metagenomic community and a reference database. These elements are identified with the –i and *–d* options when executing the program. Figure 2 shows how the application repository looks like for local execution.



```
/rapdtool$ ls -l

144140984 Oct 13 19:10 Bacteria Archaea type assembly set.msh
     1916 Oct 22 18:30 Binning_refiner.c
     1063 Oct 22 18:30 LICENSE
     4930 Oct 22 18:30 README.md
    13340 Oct 16 22:53 RaPDTool_manuscript.docx
      190 Dec  9 17:32 Snakefile
662655417 Dec 18  2020 contigs.fasta
     1114 Oct 22 18:30 focus.c
     3994 Oct 14 08:53 issues_initial_commitment.txt
      587 Oct 22 18:30 mash.c
      739 Oct 22 18:30 metabat2.c
    36742 Oct 13 12:47 pruebas.docx
    19925 Oct 22 18:30 rapdtool.py
```

Figure 2. Repository content for execution. The essential elements of the application are highlighted in color.

The file **contigs.fasta** represents a metagenome from the Apatlaco river micro-basin, Morelos, México (Accessible in:https://figshare.com/ndownloader/files/28075893). The Bacteria_Archaea_type_assembly_set.msh database is a representation of the NCBI Prokaryotic type material genomes, it contains 17,442 prokaryotic genomes (Accessible at: https://figshare.com/ndownloader/files/30851626).

A correct execution with this test data would be with the command:

```
./rapdtool.py -i contigs.fasta -d Bacteria_Archaea_type_assembly_set.msh -r proof_out
```

The results of this initial test can be consulted on the web: https://figshare.com/ndownloader/files/31735784. The execution file in **proof_out\fmbm** named **logfmbm.txt** contains the details of the execution.

In the case of the proof of concept we use a computer equipment LENOVO workstation (MT_11D2_BU_Think_FM_ThinkCentre M90s) with Intel (R) Core (TM) i3-10300 CPU @ 3.70GHz, 3696 Mhz, with 8 logic processors, and total physical memory of 128 GB. Execution takes ~ 4 minutes total for taxonomic profiling, metagenomic binning, refinement, and genomic distance calculation operations. Although the input dataset is a regular sized Metagenome with ~ 600 Mbp. 31 individual genomes were resolved after filtering (Folder: ***contigs_refined_bins***), The taxonomic profile of the community was obtained based on the information of the contigs (Folder: **profilesfmbm**), as well as the estimation of genomic distances against the entire database (17,442 records; Folder: **workfmbm\outmash\contigs**) as well as filtering the 10 phylogenetic neighbors with closest distances to each resolved genome (Folder: **proof_out\allresultsfmbm**).

References:

- Sánchez-Reyes, A.; Fernández-López, M.G. Mash Sketched Reference Dataset for Genome-Based Taxonomy and Comparative Genomics. Preprints 2021, 2021060368 (doi: http://dx.doi.org/10.20944/preprints202106.0368.v1).

- Mash: fast genome and metagenome distance estimation using MinHash. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. Genome Biol. 2016 Jun 20;17(1):132. doi: 10.1186/s13059-016-0997-x.

- Silva, G. G. Z., D. A. Cuevas, B. E. Dutilh, and R. A. Edwards, 2014: FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. PeerJ, 2, e425, doi:10.7717/peerj.425.

- Song WZ, Thomas T (2017) Binning_refiner: Improving genome bins through the combination of different binning programs. Bioinformatics, 33(12), 1873-1875.

- Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ, 7, e7359. https://doi.org/10.7717/peerj.7359.