MEGAnnotator2

Multi-threaded Enhanced prokaryotic Genome Annotator

Lugli Gabriele Andrea

November 25, 2022

For any suggestion or problem related to MEGAnnotator2: gabrieleandrea.lugli@unipr.it

Table of contents

1. What's MEGAnnotator2?

MEGAnnotator2 is a pipeline able to manage all next-generation sequencing methodologies producing short- and long-read DNA sequences. Starting from raw sequencing data, the pipeline can manage multiple analyses leading to the assembly of high-quality genome sequences and the functional classification of their genetic repertoire, providing the user with a useful report constituting features and statistics related to the microbial genome. The pipeline is fully automated from the installation to the delivery of the output, thus requiring minimal bioinformatics knowledge to be executed.

2. What could MEGAnnotator2 do?

a. Quality filtering of the data

Starting from genomic raw reads, MEGAnnotator2 performs a quality filtering step aiming at removing DNA sequences that are too short or that display low quality. Based on the input file typology, the pipeline will perform a short read filtering (single or paired-end based on the technology) or a long read filtering of the data.

b. Genomic Assembly of the filtered reads

Assemblies of DNA sequences can be performed using a combination of short and long sequences obtained by any NGS platform as well as modern third-generation sequencers such as PacBio and Nanopore.

c. Genome quality check (optional)

Assembled data is assessed with multiple validation methods. The first screening is represented by the identification of the assembled genomes of the microbial species using the 16S/18S rRNA gene sequence and ANI values. Additionally, the quality of the assembled genome is evaluated.

d. Gene prediction and functional annotation

MEGAnnotator2 performs gene prediction among the assembled genome sequence. Then, functional annotation of each gene sequence is attributed using a pre-processed database installed together with MEGAnnotator2. Additionally, non-coding genes are predicted as well.

e. Metabolic profiling (optional)

Predicted gene sequences are investigated to retrieve each attributable enzymatic reaction.

3. System requirements

MEGAnnotator2 should run on all Unix platforms, although it has been tested only in different Ubuntu LTS versions, such as v. 18.04, v. 20.04, and v. 22.04. Notably, if MEGAnnotator2 will be executed in a different Unix platform than Ubuntu, the source code of several programs integrated into the pipeline should be downloaded and compiled by the user.

The user needs to have at least 100GB of free disk space to download all the requested software and databases.

4. Installation

MEGAnnotator2 is a bash script, so it doesn't need to be compiled. However, to perform a complete analysis, several extra programs are invoked by MEGAnnotator2. Most of the requested programs are included in the MEGAnnotator2 package, while additional dependencies, listed in the next paragraph, will be installed into the system together with MEGAnnotator2.

To install MEGAnnotator2, download the installer from the following link and follow the instructions:

http://probiogenomics.unipr.it/sw/MEGAnnotator2/MEGAnnotator2_installer.sh

Then, execute the program as superuser typing "sudo ./MEGAnnotator2_installer.sh"

5. Software requirements and dependencies

In the MEGAnnotator2 package are included the following programs:

- Barrnap (https://github.com/tseemann/barrnap)
- blast software suite (Morgulis et al., 2008, Bioinformatics)
- Bwa (Li et al., 2009, Bioinformatics)
- CANU v2.0 (Koren et al., 2017, Genome Res.)
- CheckM (Parks et al., 2015, Genome Res.)
- DIAMOND (Buchfink et al., 2014, Nat Methods.)
- fastANI (Jain et al., 2018, Nat Commun.)
- FastQC (https://github.com/s-andrews/FastQC)
- fastq-mcf (https://github.com/ExpressionAnalysis/ea-utils)
- Filtlong (https://github.com/rrwick/Filtlong)
- HMMscan (Potter et al., 2018, Nucleic. Acids Res.)
- InterProScan (Jones et al., 2014, Bioinformatics)
- Mauve v 2.3.1 (Rissman et al., 2009, Bioinformatics)
- Polypolish (Wick et al., 2022, PLoS Comput Biol.)
- prodigal (Hyatt et al., 2010, BMC Bioinformatics.)

- SamToFastq.jar (https://github.com/broadinstitute/picard/blob/master/src/main/java/picard/)
- Seqkit (https://github.com/shenwei356/seqkit/releases)
- SPAdes v3.15.4 (Bankevich et al., 2012, J Comput Biol.)
- tRNAscan-SE 2.0 (Chan et al., 2021, Nucleic Acids Res.)

MEGAnnotator2 requires the following programs or package for full functionality:

- Java version 1.7 or superior (type "**sudo apt install default-jdk**" to install)
- python-is-python3 (type "**sudo apt install python-is-python3**" to install)
- artemis (type "**sudo apt install artemis**" to install)
- bedtools (type "**sudo apt install bedtools**" to install)
- checkm-genome (type "**sudo pip3 install checkm-genome**" to install)
- emboss software suit (type "**sudo apt-get install emboss**" to install)
- gawk (type "**sudo apt install gawk**" to install)
- hmmer (type "**sudo apt install hmmer**" to install)
- matplotlib (type "**sudo pip3 install matplotlib**" to install)
- numpy (type "**sudo pip3 install numpy**" to install)
- pysam (type "**sudo pip3 install pysam**" to install)
- readseq (type "**sudo apt-get install readseq**" to install)
- samtools (type "**sudo apt-get install samtools**" to install)
- tree (type "**sudo apt install tree**" to install)
- trnascan-se (type "**sudo apt install trnascan-se**" to install)

Above listed software will be installed during the installation of MEGAnnotator2 using the script "MEGAnnotator2_installer.sh". Thus, if you have corrected installed MEGAnnotator2, you can ignore the software dependencies above listed.

6. Databases

To perform the assembly-based analyses, MEGAnnotator2 requires a range of databases to be downloaded: a manually curated RefSeq NCBI (amino acid) database for functional predictions, a reference genome (nucleotide) database for taxonomic classification of microorganisms, a 16S and 18S rRNA gene sequence (nucleotide) database, and a custom metabolites (amino acid) database for identification of enzymatic reactions.

All these databases will be downloaded during the installation of the MEGAnnotator2 package. Furthermore, databases can be updated by typing "MEGAnnotator2 -u" and following the instructions.

The amount of disk space needed for the database is about 50 Gb.

7. Input data

MEGAnnotator2 accepts both single-end, paired-end, and long reads data in .fastq format. Then, for the correct execution of the pipeline, refer to the file “parameters” containing a detailed list of analyses and settings.

Here you can find the complete list of the editable parameters implemented in the MEGAnnotator2 pipeline (the reported list reflects the settings order in the parameters file):

Databases:

- Custom database for reads filtering: the path of a Custom Database for filtering the input files.
- Annotation database: the path of the RefSeq NCBI database (formatted with DIAMOND).
- Silva database: the path of the silva database for species classification (formatted with makeblastdb).
- K-mer database: the path of the K-mer database for species classification.
- Reference genomes: the path of the folder containing the microorganisms’ chromosomal sequences.
- Metabolic reactions Prokaryotes database: the path of the custom pathways database (formatted with DIAMOND).

System:

- Number of threads: number of multiple threads for multithreading programs included in the pipeline (based on the CPU technology used for the analyses) (numerical variable).

Additional analyses:

- Species prediction (rRNA): set “YES” to perform species classification with the Silva database.
- Species prediction (ANI): set “YES” to perform species classification with the genome database.
- Genome polyshing: set “YES” to perform short-read polishing after a long-read assembly (only usable if both technologies are provided as input).
- Alignment analyses: set “YES” to perform genome alignment of the sequences after assembly using the genome database.
- Annotation analyses: set “YES” to perform functional prediction of genes using the annotation database.
- Metabolic analyses: set “YES” to perform the metabolic prediction using the metabolic database.

Reads filtering:

- Custom filtering: set “YES” to perform a read filtering step using a custom database (the custom database must be indicated in the database section).
- Reads quality report: set “YES” to perform a quality screening of the input files.

- Reads minimum length: smaller reads will be removed during the filtering step (numerical variable).
- Reads minimum quality value: lower quality reads will be removed during the filtering step (numerical variable).
- Long reads minimum length: smaller long-reads will be removed during the filtering step (numerical variable).
- Long reads percentage of the best reads: percentage of long-reads used for the assembly. Those long-reads with the lowest quality are discarded (numerical variable).
- Long reads maximum total bases: maximum number of bases used for the assembly. Those long-reads with the lowest quality are discarded (numerical variable).
- Long reads split: break long-reads if quality is low (the numerical value is used only for hybrid reads filtering).

Contigs analyses:

- Contigs quality report: set "YES" to perform a quality screening of the assembled genome.
- Contigs minimum length (short reads): shorter assembled contigs than the parameter will be discarded (numerical variable).
- Contigs minimum length (long reads): shorter assembled contigs than the parameter will be discarded (numerical variable).
- Long reads estimated genome size: estimated size of the genome (alphanumerical variable). Example: 5m = 5000000.
- Truncated genes minimum length: remove those predicted genes at the edge of contigs that do not have a start or stop codon and are smaller than the variable (numerical variable).
- Kingdom: must be set as "bacteria" or "eukarya"
- Predicted species: name of the predicted species that will be used to perform genome alignment ("Predicted species" must be reported with an undescore between genera and species). Example: Bifidobacterium_breve.
- Annotation e-value cutoff: alignments with higher e-value than the parameter will be discarded during the annotation of the genes (numerical variable). Example: 0.00000001 stands for e^{-8}.
- Annotation minimum percentage query coverage: alignments shorter than the parameter will be discarded during the annotation of the genes (numerical variable).
- Domain e-value cutoff: alignments with higher e-value than the parameter will be discarded during the annotation of domains (numerical variable).

Functional analyses:

- Metabolic reactions e-value cutoff: alignments with higher e-value than the parameter will be discarded during the pathways classification (numerical variable). Example: 0.00000001 stands for e^{-8}.
- Metabolic reactions minimum percentage query coverage: alignments shorter than the parameter will be discarded during the pathways classification (numerical variable).

8. Output data

Results are distributed in the output folder within sub-folders listed below:

- filtered_reads
  Folder containing the filtered fastq files and the report of their quality.
- Project_assembly
  Folder containing the results of the performed assembly.
- Mauve_alignment
  Folder containing the alignment with the reference genome of the species.
- Metabolic_reactions
  Folder containing the results of the metabolic screening.
- checkM_report.txt: report of the quality check performed on the assembled data.
- genome_info.txt: report of all the analyses performed.
- Project.gbk: GenBank file that can be open using the software Artemis.
- Project.blastn: results of the 16S/18S rRNA gene screening.
- Project_aaORFs.fasta: list of amino acid genes predicted on the assembled genome sequence.
- Project_contigs.fasta: list of assembled sequences.
- Project_contics_name.txt: list of contigs names attributed during the analysis.
- Polishing_report: list of modifications made by short-read polishing after long-read assembly.

By using the command "**MEGAnnotator2 -x report**", multiple outputs folder is combined in a single folder named "extracted_output" as follows:

- aaORFs
  Folder containing amino acid genes predicted between all genomes.
- assembly
  Folder containing the results of all assemblies.
- contigs
  Folder containing the list of all assembled sequences.
- Filtered_reads
  Folder containing the report of the filtering.
- Gbks
  Folder containing the GenBank files.
- mauve_alignment.
  Folder containing all alignments with the reference genomes.
- checkM_report.txt: report of the quality check performed on the assembled genomes.
- genome_info.txt: report of all the analyses performed.

9. Usage

MEGAnnotator2 has been placed in the $PATH to be executed as the bash command "**MEGAnnotator2**".

Using the help arguments, the bash version of MEGAnnotator2 (**MEGAnnotator2 -h**) will display the following message to guide the user in the program usage:

```
        MEGAnnotator2 v2.0.0 (12 December 2022)

usage: MEGAnnotator2 -n [name] -s -i [file]                        single-end reads mode
   or: MEGAnnotator2 -n [name] -p -f [file] -r [file]              paired-end reads mode
   or: MEGAnnotator2 -n [name] -l -i [file]                        long reads mode
   or: MEGAnnotator2 -n [name] -y -i [file] -f [file] -r [file]    hybrid reads mode
   or: MEGAnnotator2 -n [name] -o -i [file] -f [file] -r [file]    longpolished reads mode

Arguments:
  -n          Project name (alphanumerical variable)
  -t          Threads number (numerical variable)
  -s          Single reads technology (assembly with short reads)
  -p          Paired reads technology (assembly with short reads)
  -l          Long reads technology (assembly with long reads)
  -y          Hybrid reads technology (assembly with long and short reads)
  -o          Hybrid reads technology (assembly with long reads and polishing with short)
  -i          Input single reads path (to be used in -s, -l, -y, or -o mode)
  -f          Forward reads path (to be used in -p, -y, or -o mode)
  -r          Reverse reads path (to be used in -p, -y, or -o mode)
  -k          Kingdon (bacteria by default, or eukarya)
  -h          Print Help (this message) and exit
  -u          Update software and databases
  -x          Additional eXtra features (prepare_miseq, prepare_nextseq, prepare_nanopore,
report)
```

The bash command can be integrated into a bash script to queue multiple MEGAnnotator2 executions easily. Follow the instruction to use the program with paired-end reads (-p option), single-end reads (-s option), long reads (-l option), or hybrid reads (-y option) as input.

The program can be executed in any location across the machine in which MEGAnnotator2 has been installed. MEGAnnotator2 will generate hyperlinks of each input file and the software folder. The parameters file will then be copied from the MEGAnnotator2 folder if no parameters file has been provided in the location.