

MEGAnnotator

Multi-threaded Enhanced prokaryotic Genome Annotator

Lugli Gabriele Andrea

July 09, 2015

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

Table of contents

1. What's MEGAnnotator?	3
2. What could MEGAnnotator do?	3
a. Genomic Assembly	3
b. Metagenomic Assembly.....	3
c. Genes Annotation only.....	3
3. System requirements.....	3
4. Installation	3
5. Software requirements and dependencies.....	4
6. Databases	4
a. Pfam-A database	4
b. NCBI (nr) database	4
7. Input data	5
a. Raw Data.....	5
b. Reference Genome (optional)	5
8. Output data.....	5
a. Assembly results	5
b. Alignment results (optional)	5
c. Improvement quality results.....	5
d. Annotation results	5
9. Usage	5
10. Tutorial.....	6
a. Genomic Assembly	6
b. Metagenomic Assembly.....	10
c. Genes Annotation only.....	12

1. What's MEGAnnotator?

MEGAnnotator is a **M**ulti-threaded **E**nhanced prokaryotic **G**enome **A**nnotator. This pipeline allows the generation of an annotated GenBank file fulfilling the NCBI guidelines for assembled microbial genomes submission, based on DNA shotgun sequencing reads, and minimizes manual intervention, removes waiting times between software program executions, while also improving the final quality of both assembly and annotation outputs.

2. What could MEGAnnotator do?

MEGAnnotator has three program sections:

a. Genomic Assembly

Starting from genomic raw reads, MEGAnnotator performs the assembly followed by contigs selection, quality controls, ORFs prediction and genes annotation concluding with the generation of a GenBank file.

b. Metagenomic Assembly

Starting from metagenomics raw reads, MEGAnnotator performs the assembly followed by the ORFs prediction and genes annotation of the resulting contigs.

c. Genes Annotation only

Starting from a pre-assembled genome, MEGAnnotator performs the ORFs prediction and the genes annotation.

3. System requirements

MEGAnnotator should run on all Unix platforms, although it has not tested in all platforms.

4. Installation

First, place the distribution tarball to your work directory. Then, uncompress the distribution tarball typing:

```
gzip -d MEGAnnotator-master.zip
```

MEGAnnotator is a bash script, so it is unnecessary to compile. However, to do a complete analysis, several extra programs are invoked by MEGAnnotator. Therefore, before running MEGAnnotator, users should install the programs listed in the next paragraph.

5. Software requirements and dependencies

MEGAnnotator requires the following programs or package for full functionality:

- Java version 1.7 or superior
- readseq (type “**sudo apt-get install readseq**” to install)
- bwa (type “**sudo apt-get install bwa**” to install)
- samtools (type “**sudo apt-get install samtools**” to install)
- tabix (type “**sudo apt-get install tabix**” to install)
- hmmscan (type “**sudo apt-get install hmmer**” to install)
- emboss software suit (type “**sudo apt-get install emboss**” to install)
- RNAmmer (visit <http://www.cbs.dtu.dk/services/RNAmmer/> to install)
Must be included in the PATH.
- tRNAscan-SE (visit <http://selab.janelia.org/tRNAscan-SE/> to install)
Must be included in the PATH.
- GATK (visit <https://www.broadinstitute.org/gatk/download/> to install)
Must be placed in the bin folder of MEGAnnotator. Simply copy the jar file and rename it as “GenomeAnalysisTK.jar” (if different).

6. Databases

To perform the genes annotation is essential to have available the Pfam-A and NCBI (nr) databases. While the Pfam database is pre-formatted available online, the latter need to be formatted using prerapsearch. User database folder path will be requested by MEGAnnotator at each run.

a. Pfam-A database

Visit the Pfam website ftp for the Pfam-A.hmm.gz download:

ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release

Place the compressed file to your own databases folder and decompress the database:

```
gzip -d Pfam-A.hmm.gz
```

b. NCBI (nr) database

Visit the NCBI website ftp for the nr.gz file download:

<ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/>

Place the compressed file to your own database folder and decompress the sequences:

```
gzip -d nr.gz
```

From console, move to the MEGAnnotator main directory to build the database with:

```
bin/./prerapsearch -d /folder/rapsearch_nr -n /folder/nr
```

N.B. place the complete path of the decompressed nr file location in the above command (where *folder* is displayed). The amount of disk space needed for the database building is about 150 Gigabyte. It will require several hours.

7. Input data

a. Raw Data

Raw data should be supplied as fastq file, furthermore MEGAnnotator is capable to manage paired-end illumina data.

b. Reference Genome (optional)

Reference Genome should be supplied as fasta file. The reference genome is optional; its usage is limited in case users would order the obtained contigs. However, the assembly program does not use the reference genome to perform the contigs creation.

8. Output data

a. Assembly results

Within the assembly output, you can find the multifasta file containing the contigs, the info file with the assembly output information and the assembly log file.

b. Alignment results (optional)

Output file regarding the final alignment performed against the reference genome.

c. Improvement quality results

Multifasta file containing the improved contigs, tabular file with the nucleotide substitutions and vcf files.

d. Annotation results

The result consists in a GenBank file. Within the GenBank file, the contigs are represented by `fasta_records`. For a correct visualization of the GenBank file, we suggest the utilization of the free software Artemis (<https://www.sanger.ac.uk/resources/software/artemis/>)

9. Usage

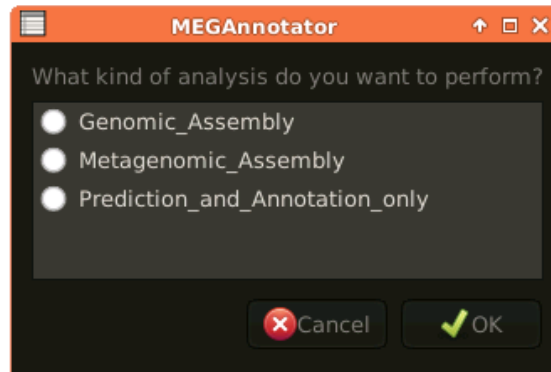
To correctly run MEGAnnotator.sh, the bin and lib folders must be located in the MEGAnnotator script directory, as well as the script annotation.sh and metagenomics.sh. Please do not change any file within these directories, otherwise the pipeline may be compromised.

Simply run the script typing:

```
./MEGAnnotator.sh
```

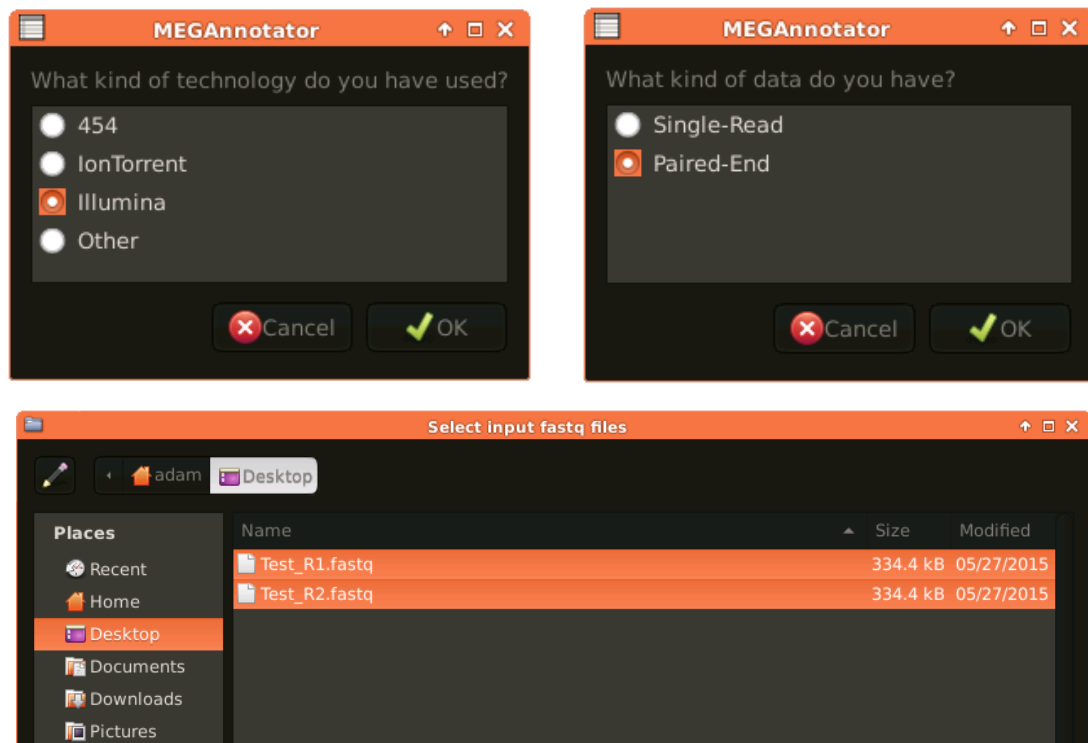
10. Tutorial

MEGAnnotator starts with a list dialog for the pipeline selection. The first choice will define the typology of analysis you want to perform.



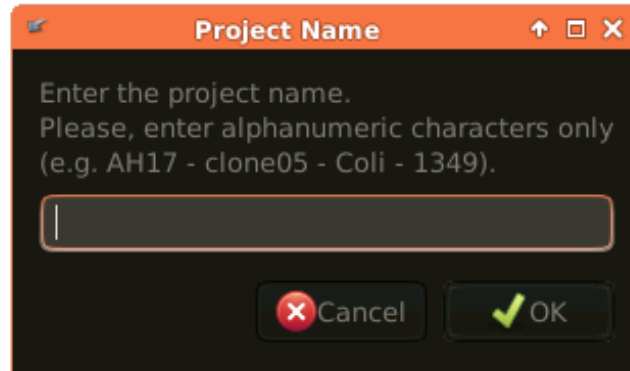
a. Genomic Assembly

The genomic pipeline starts with the definition and selection of the raw reads that will be used from the assembler as input. In case the input is represented by Illumina data, a second list dialog allows the user to select paired- or single-end sequenced reads.



It is essential to provide the raw reads in fastq files, otherwise MEGAnnotator cannot manage the input. In the example, two illumina paired-end fastq files were selected.

Consequently, a text entry dialog awaits the project name to be insert. Please, enter alphanumeric characters only (e.g. AH17 – clone05 – Coli – 1349).

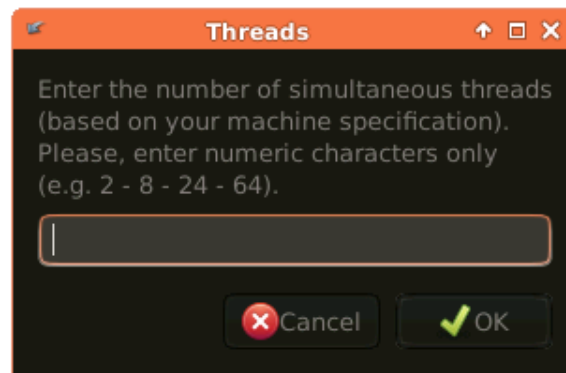
A dialog box titled "Project Name" with a dark background and orange header. It contains the text: "Enter the project name. Please, enter alphanumeric characters only (e.g. AH17 - clone05 - Coli - 1349)." Below the text is a text input field. At the bottom are two buttons: "Cancel" with a red 'X' icon and "OK" with a green checkmark icon.

Project Name

Enter the project name.
Please, enter alphanumeric characters only
(e.g. AH17 - clone05 - Coli - 1349).

Cancel OK

Then, MEGAnnotator needs the number of threads you want allocate for the analyses. Please, enter numeric characters only (e.g. 2 – 8 – 24 – 64).

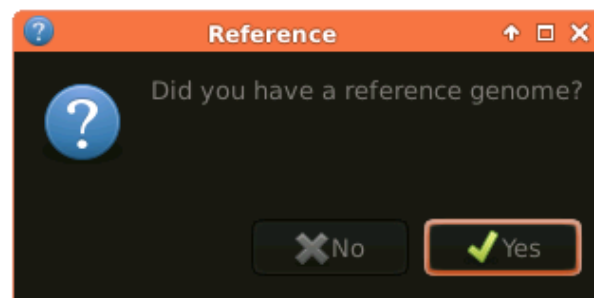
A dialog box titled "Threads" with a dark background and orange header. It contains the text: "Enter the number of simultaneous threads (based on your machine specification). Please, enter numeric characters only (e.g. 2 - 8 - 24 - 64)." Below the text is a text input field. At the bottom are two buttons: "Cancel" with a red 'X' icon and "OK" with a green checkmark icon.

Threads

Enter the number of simultaneous threads
(based on your machine specification).
Please, enter numeric characters only
(e.g. 2 - 8 - 24 - 64).

Cancel OK

Following, you can give as input the genome reference you want to use for the contigs reordering after the assembly (optional). It is essential to provide the nucleotide sequence of the reference genome in fasta file.

A dialog box titled "Reference" with a dark background and orange header. It contains a question mark icon and the text: "Did you have a reference genome?". At the bottom are two buttons: "No" with a grey 'X' icon and "Yes" with a green checkmark icon and an orange border.

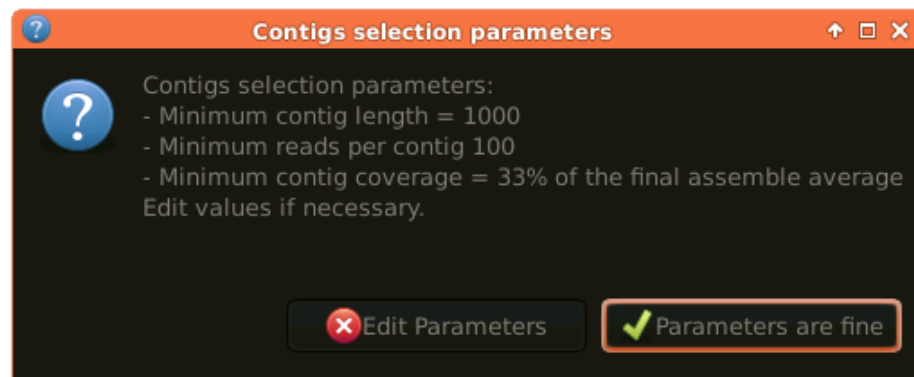
Reference

Did you have a reference genome?

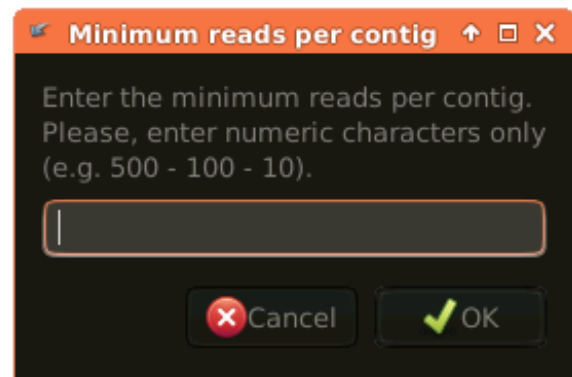
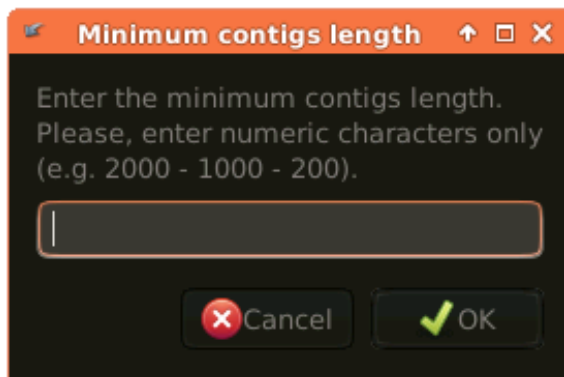
No Yes



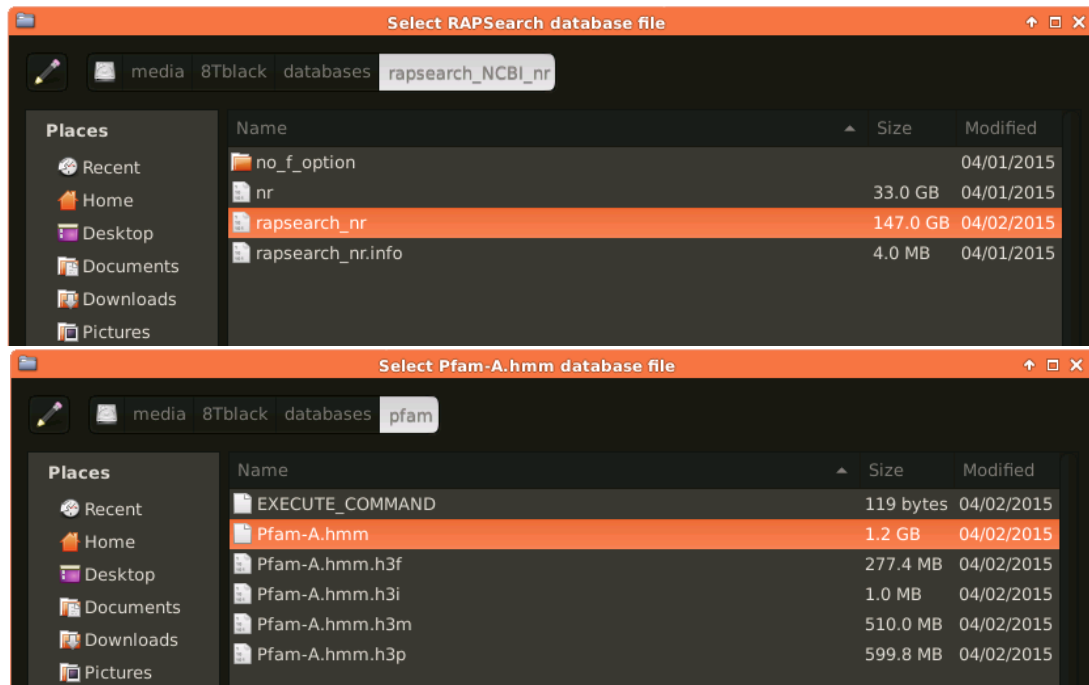
Thereafter, the users can change the parameters set up for the contigs selection or continue with the parameter listed below.



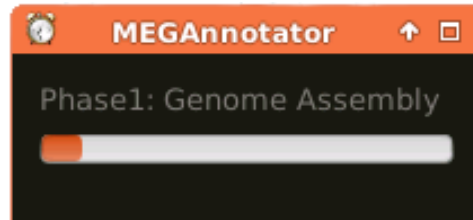
Whether the user chose to edit the parameters, two text entry dialog gives the opportunity to set the minimum contig length and reads per contig. Please, enter numeric characters only.



In the two following steps, the user have to select the databases needed for the genome annotation. It is important to have already build the databases (see chapter 6).



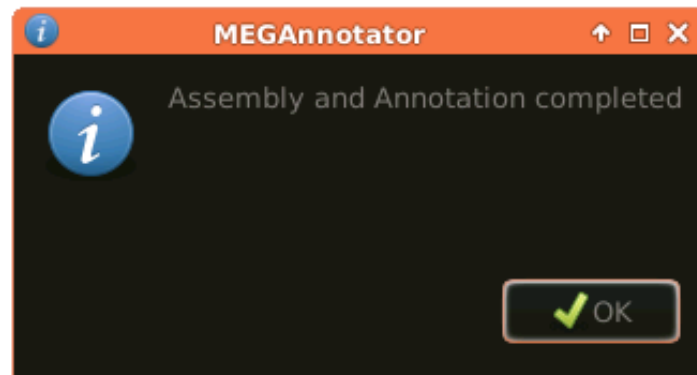
After that, a progress dialog shows the progress status of the analysis.



Here, the list of the ten phases:

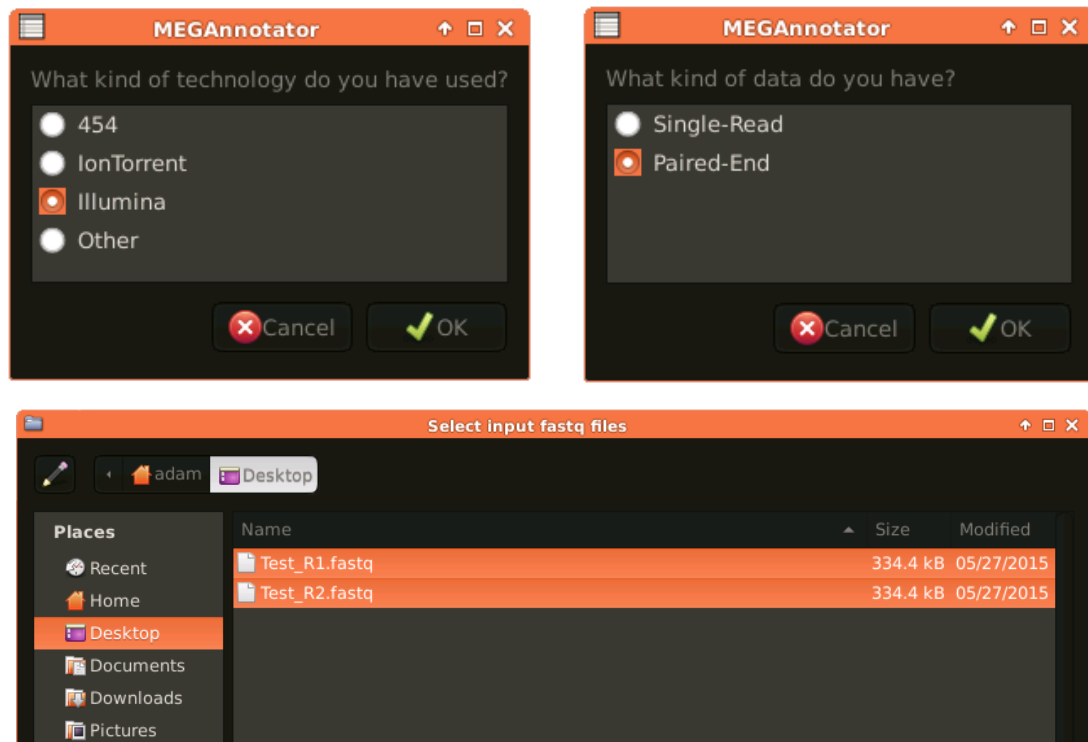
- Phase 1: Genome Assembly
- Phase 2: Contigs selection
- Phase 3: Alignment vs. reference genome
- Phase 4: Improvement of quality output
- Phase 5: Genes prediction
- Phase 6: RapSearch annotation
- Phase 7: pfam prediction
- Phase 8: merging annotation
- Phase 9: gbk generation and rRNA prediction
- Phase 10: genbank finalization and tRNA prediction

At the end, when all the phases ends correctly, the following message will show.



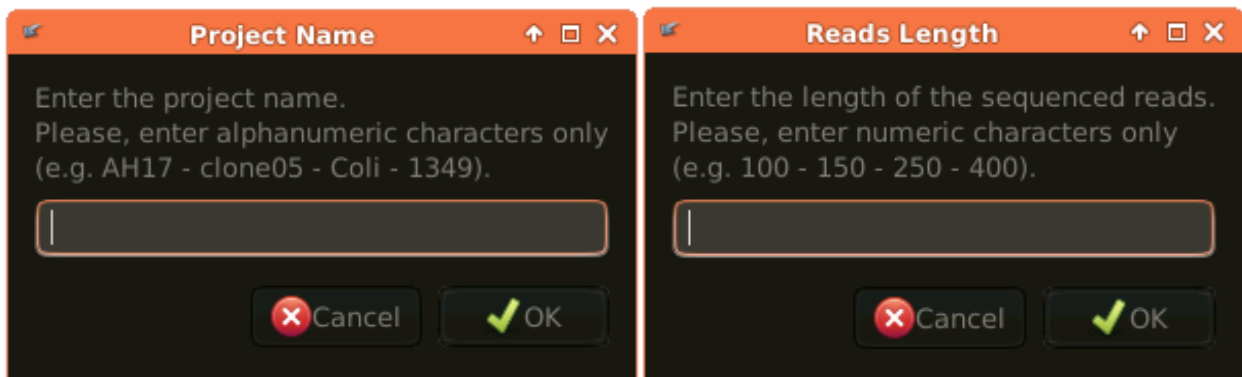
b. Metagenomic Assembly

The metagenomic pipeline starts with the definition and selection of the raw reads that will be used from the assembler as input as well as the genomic pipeline presented above. In case the input is represented by Illumina data, a second list dialog allows the user to select paired or single end sequenced reads.



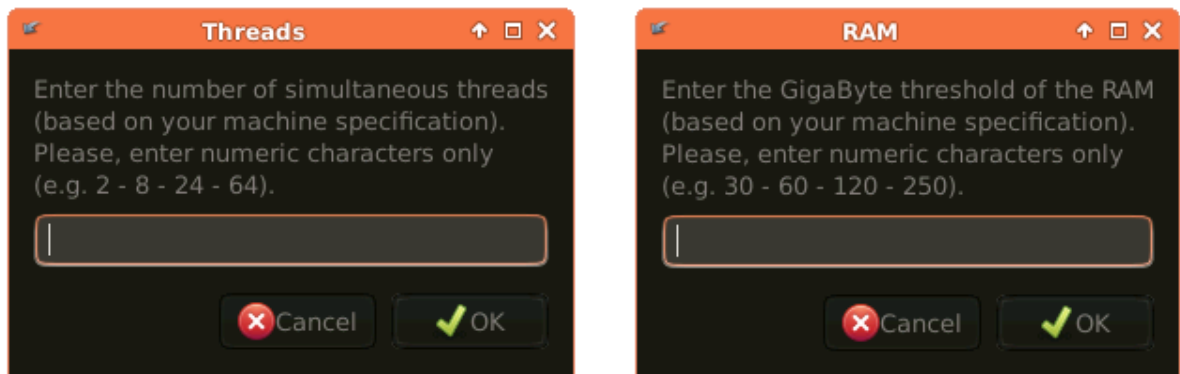
It is essential to provide the raw reads in fastq files, otherwise MEGAnnotator cannot manage the input. In the example, two illumina paired-end fastq files were selected.

Consequentially, two text entry dialog awaits the project name and the reads length to be insert. Please, enter alphanumeric characters only for the project name (e.g. AH17 – clone05 – Coli – 1349) and numeric characters only for the reads length (e.g. 100 – 150 – 250 – 400).



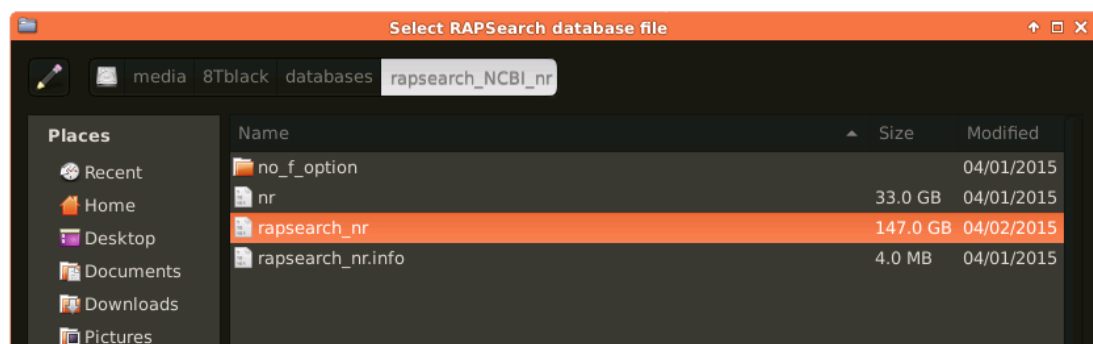
The image shows two side-by-side dialog boxes. The left dialog is titled 'Project Name' and contains the text: 'Enter the project name. Please, enter alphanumeric characters only (e.g. AH17 - clone05 - Coli - 1349).' Below the text is a text input field. The right dialog is titled 'Reads Length' and contains the text: 'Enter the length of the sequenced reads. Please, enter numeric characters only (e.g. 100 - 150 - 250 - 400).' Below the text is a text input field. Both dialogs have 'Cancel' and 'OK' buttons at the bottom.

Then, MEGAnnotator needs the number of threads you want allocate for the analyses and the number of GigaByte of RAM the user wants to allocate. Please, enter numeric characters only.

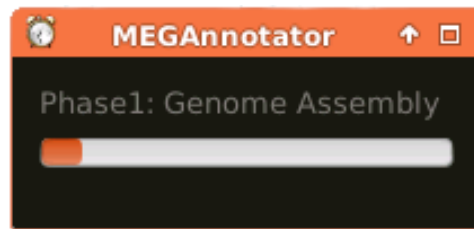


The image shows two side-by-side dialog boxes. The left dialog is titled 'Threads' and contains the text: 'Enter the number of simultaneous threads (based on your machine specification). Please, enter numeric characters only (e.g. 2 - 8 - 24 - 64).' Below the text is a text input field. The right dialog is titled 'RAM' and contains the text: 'Enter the GigaByte threshold of the RAM (based on your machine specification). Please, enter numeric characters only (e.g. 30 - 60 - 120 - 250).' Below the text is a text input field. Both dialogs have 'Cancel' and 'OK' buttons at the bottom.

In the following step, the user have to select the databases needed for the genome annotation. It is important to have already build the NCBI nr database (see chapter 6).



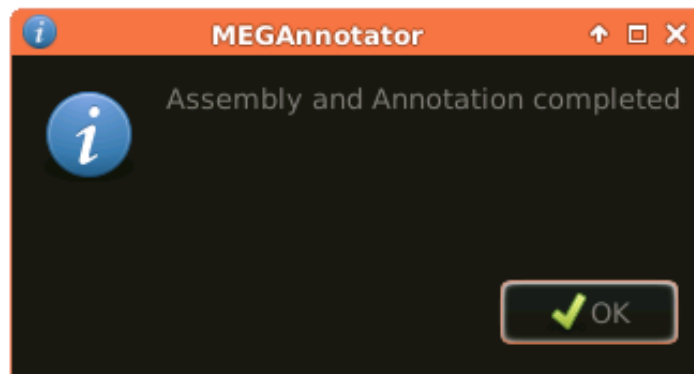
After that, a progress dialog shows the progress status of the analysis.



Here, the list of the six phases:

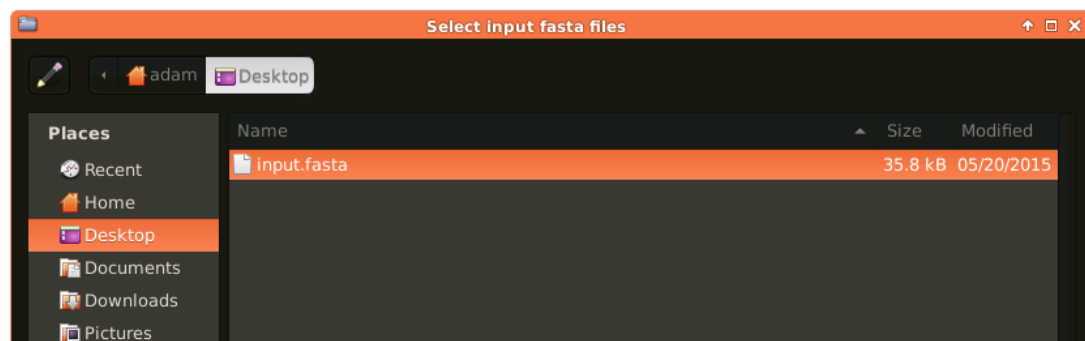
- Phase 1: Metagenome Assembly
- Phase 2: Genes prediction
- Phase 3: RapSearch annotation
- Phase 4: merging annotation
- Phase 5: gbk generation and rRNA prediction
- Phase 6: genbank finalization and tRNA prediction

In the end, where all the phases ends correctly, the following message will show.

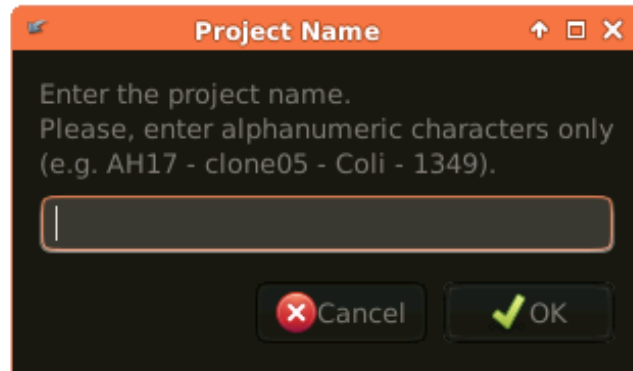


c. Genes Annotation only

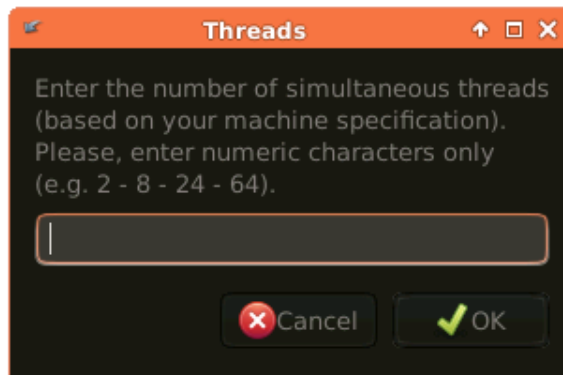
The genes annotation pipeline starts with the selection of the multifasta file that the user wants to use as input.



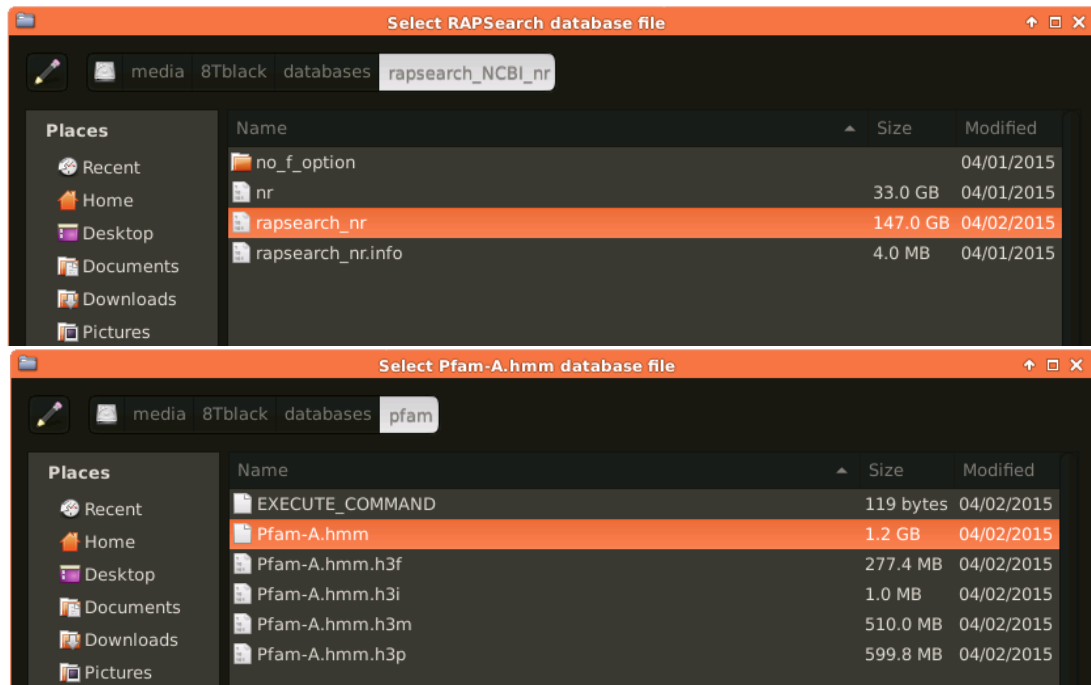
Consequently, a text entry dialog awaits the project name to be inserted. Please, enter alphanumeric characters only (e.g. AH17 – clone05 – Coli – 1349).

A screenshot of a 'Project Name' dialog box. The title bar is orange with a small icon on the left and standard window controls (maximize, close) on the right. The main area has a dark background with white text: 'Enter the project name. Please, enter alphanumeric characters only (e.g. AH17 - clone05 - Coli - 1349)'. Below the text is a white text input field. At the bottom are two buttons: 'Cancel' with a red 'X' icon and 'OK' with a green checkmark icon.

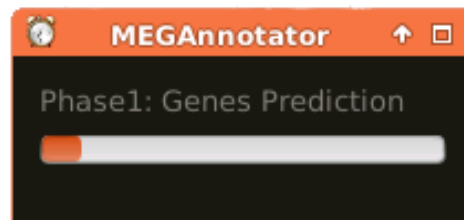
Then, MEGAnnotator needs the number of threads you want allocate for the analyses. Please, enter numeric characters only (e.g. 2 – 8 – 24 – 64).

A screenshot of a 'Threads' dialog box. The title bar is orange with a small icon on the left and standard window controls (maximize, close) on the right. The main area has a dark background with white text: 'Enter the number of simultaneous threads (based on your machine specification). Please, enter numeric characters only (e.g. 2 - 8 - 24 - 64)'. Below the text is a white text input field. At the bottom are two buttons: 'Cancel' with a red 'X' icon and 'OK' with a green checkmark icon.

In the two following steps, the user have to select the databases needed for the genome annotation. It is important to have already build the databases (see chapter 6).



After that, the script starts showing a progress dialog.



Here, the list of the six phases:

- Phase 1: Genes prediction
- Phase 2: RapSearch annotation
- Phase 3: pfam prediction
- Phase 4: merging annotation
- Phase 5: gbk generation and rRNA prediction
- Phase 6: genbank finalization and tRNA prediction

In the end, where all the phases ends correctly, the following message will show.

