MetaShot User Guide

Requirements

MetaShot scripts use freely available Python packages and third party tools both requiring a previous installation performed by the users.

It requires a working Python (2.6 or 2.7) environment and the following modules:

- NumPy: release 1.7.1 or superior (http://www.numpy.org)

- BioPython: release 1.61 (http://biopython.org)

- Pysam: 0.7.4 (http://pysam.readthedocs.io/en/stable/). Please note that

- Psutil: release 3.3.0 or superior (https://github.com/giampaolo/psutil)

- ETE2: release 2.3.10 (http://etetoolkit.org/download/)

The following tools need to be installed on your machine:

- Bowtie2: release2.2.3 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml)

- STAR: release STAR_2.4.2a (https://github.com/alexdobin/STAR)

- Krona Graph: release 2.6 (https://github.com/marbl/Krona/wiki)

- FaQCs: release 2.08 (https://github.com/chienchi/FaQCs) (please install it by using conda: *conda install faqcs -c bioconda*)

- Samtools: release 0.1.18 or superior (http://samtools.sourceforge.net)

MetaShot requires at least 30 Gb of RAM to perform the entire analysis. Its reference collections require about 420Gb of free space on your storage system. To complete the analysis of 500 million PE reads it requires about 1.3 Tb.

MetaShot setting up
Download the reference collections

All the reference collections, taxonomies and the other files needed for the MetaShot computation are stored in a compressed folder "MetaShot_reference_data.tar.gz", freely available at https://recascloud.ba.infn.it/index.php/s/9dFFTJUr0bOU2mN (the file is about 49 GB. Md5sum: 2def0475c8c3b49c15181de0c226b11a).

To decompress the folder, type the following command in your terminal:

    tar xvfz MetaShot_reference_data.tar.gz

The "MetaShot_reference_data" folder contains 9 subfolders:

- Bacteria (contains the data for all the Prokaryotes), Fungi, Protist and Virus folders contain the FASTA files to produce the bowtie indexes for each reference collection and the specific input data for taxonomic assignment. In the "bowtie_index" subfolder a bash script is supplied to correctly format and name the bowtie-index. For Prokaryotes, the reference collection is divided in 15 section, called "split", each containing the bash script for bowtie2 index building; To build the bowtie2 index type the following commands:

  chmod 755 build_index.sh

  ./build_index.sh

- Homo_sapiens: contains the hg19 release of human genome. Please follow the instructions available in the STAR guide;

- Phix: contains the Phix genome. To build the bowtie2 index type the following commands:

  cd PhiX

  chmod 755 build_index.sh

  ./build_index.sh

- krona_tax: contains the NCBI taxonomy pre-formatted needed for Krona graph production;
- New_TANGO_perl_version. Contains the executable TANGO Perl scripts,
- Metashot_reference_taxonomy: contains the taxonomy data.

Configure the Parameters File

In the MetaShot package, a textual file called "parameters_file.txt" is available that contains all the required information for the pipeline execution.

The user must complete this file by adding the following information to the "General data" section:

1. Adding the complete path to the MetaShot package:

  cd Metashot

  pwd

  Copy the result in correspondence of "script_path", as in the following example:

  script_path : /home/path/to/the/script_folder/

2. Adding the complete path to the MetaShot_reference_data folder:

cd MetaShot_reference_data

pwd

Copy the result in correspondence of "reference_path", as in the following example:

reference_path : /home/path/to/the/reference_folder/

Usage
Taxonomic assignment

The whole MetaShot analysis is performed by executing the "MetaShot_Master_script.py" script. This script requires the following mandatory and optional parameters to launch the analysis process:

- **-m** A text file containing the R1 and R2 PE reads file names, tab separated [MANDATORY]. If a sample has been split in more flowcell lines, please insert in the file one line per each PE reads file. Example: The sample1 has been sequenced in 3 flowcell lines. The read file content will be the following:

    sample1_L001_R1_1.fastq   sample1_L001_R2_1.fastq

    sample1_L002_R1_2.fastq   sample1_L002_R2_2.fastq

    sample1_L003_R1_3.fastq   sample1_L003_R2_3.fastq

- **-p** Parameters files: a file containing all the information required for the MetaShot application [MANDATORY]
- **-h** print this help

Paired End (PE) reads extraction

Following the end of MetaShot analysis, the user could be interested in extracting specific PE reads. The PE_extraction.py script can be used to extract the PE reads belonging to specific taxa in the NCBI taxonomy or to extract all the unassigned or ambiguous PE reads. This script requires as input a single NCBI taxonomy identifier or a list of NCBI taxonomy identifiers.
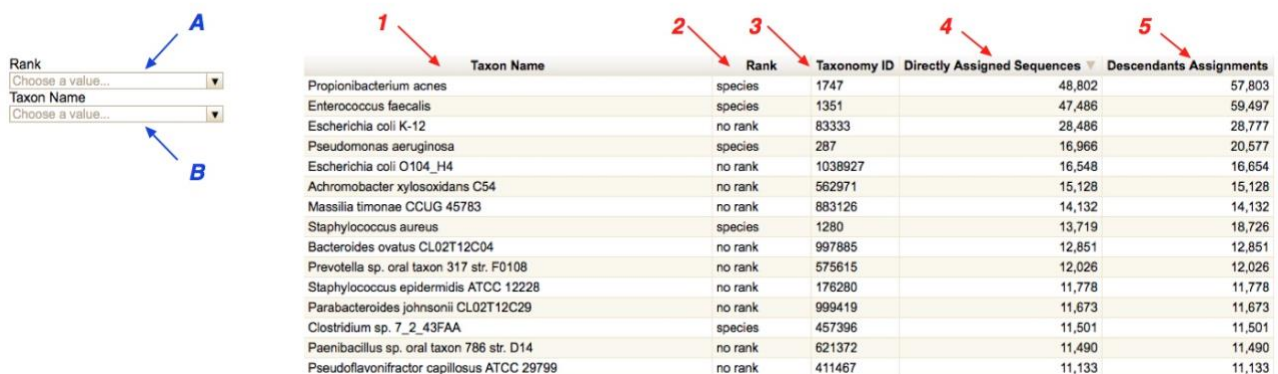
Options:

- **-t** NCBI taxonomy ID. To extract human sequences please use 9606

- **-l** A text file containing a list of NCBI taxonomy ID, one per line

- **-u** extract unassigned PE reads

- **-a** extract ambiguous PE reads

- **-h** print this help

Result files interpretation

MetaShot produces the following results:

1. an HTML interactive table reporting for each node of the inferred taxonomy i) the taxon name; ii) the NCBI taxonomy ID; and iii) the number of PE reads assigned (Figure 1); This interactive table are based on the Google API Code and requires a working internet connection.

2. a CSV file containing the same information reported in the interactive table;

Taxon Name	Rank	Taxonomy ID	Directly Assigned Sequences	Descendants Assignments
Propionibacterium acnes	species	1747	48,802	57,803
Enterococcus faecalis	species	1351	47,486	59,497
Escherichia coli K-12	no rank	83333	28,486	28,777
Pseudomonas aeruginosa	species	287	16,966	20,577
Escherichia coli O104_H4	no rank	1038927	16,548	16,654
Achromobacter xylosoxidans C54	no rank	562971	15,128	15,128
Massilia timonae CCUG 45783	no rank	883126	14,132	14,132
Staphylococcus aureus	species	1280	13,719	18,726
Bacteroides ovatus CL02T12C04	no rank	997885	12,851	12,851
Prevotella sp. oral taxon 317 str. F0108	no rank	575615	12,026	12,026
Staphylococcus epidermidis ATCC 12228	no rank	176280	11,778	11,778
Parabacteroides johnsonii CL02T12C29	no rank	999419	11,673	11,673
Clostridium sp. 7_2_43FAA	species	457396	11,501	11,501
Paenibacillus sp. oral taxon 786 str. D14	no rank	621372	11,490	11,490
Pseudoflavonifractor capillosus ATCC 29799	no rank	411467	11,133	11,133

Figure 1: an example of the interactive tables produced by MetaShot. In red are enumerated the field in common with the CSV file. In particular: (1) Taxon Name: the NCBI scientific name of the node; (2) Rank: the taxonomic rank of the node; (3) Taxonomy ID: the NCBI taxonomy identifier of the node; (4) Directly Assigned Sequences: number of PE reads directly assigned to the node; (5) Descendants Assignments: the number of PE reads assigned to the node and to its descendants (i.e. the first node is the species *Propionibacterium acnes* and MetaShot assigns directly to this node 48,802 and 9,001 to its descendants, so the fifth field in the line is equal to 57,803). In blue are listed the interactive fields to show the results on the basis of the NCBI taxonomic rank (A) or on a specific taxon name (B).

3. a Krona graph to graphically inspect the inferred microbiome. In particular, a Krona graph is produced for each division and a total one is produced to summarize the entire assignments. Information about the Krona graph is available following this link.