by João F. Matias Rodrigues, Thomas S. B. Schmidt,
Janko Tackmann, and Christian von Mering
Institute of Molecular Life Sciences
University of Zürich
Switzerland

MAPseq v1.0 (October 2016)

bringing speed, accuracy and consistency to metagenomic ribosomal RNA analysis

Contents

1 Introduction

MAPseq is a set of fast and accurate sequence read classification tools designed to classify ribosomal RNA sequences in terms of their taxonomy and OTU classification. This is done by using a reference set of full-length ribosomal RNA sequences for which known taxonomies are known, and for which a set of high quality OTU clusters has been previously generated. For each read, the best guess and correspoding confidence in the assignment is shown at each taxonomic and OTU level.

2 INSTALLATION

2.1 Linux/Unix/MacOSX

To install MAPseq on a linux, unix, or MacOSX simply type:

./configure make make install

In the directory where you unpacked the package contents. Alternatively, if you want the program to be installed to another location instead of the default system wide /usr/local/ directory, you can change the ./configure command to:

./configure –prefix=$HOME/usr

This would install the program binaries to a directory usr/bin inside your home directory (i.e.: $HOME/usr/bin/mapseq), after you type the command "make install".

3 USING MAPseq

MAPseq takes as input a fasta file with raw sequence data which should have been previously demultiplexed and quality filtered usually from a fastq file produced by the sequencing machine.

If the input sequences can be found in the file "rawseqs.fa". Then to classify the reads one simply has to run the following command:

mapseq rawseqs.fa >rawseqs.fa.mseq

This will classify all the sequences found in rawseqs against the standard reference dataset provided with MAPseq.

You can change the number of threads that MAPseq uses with the -nthreads <no_threads>argument.

4 EXAMPLE OUTPUT

In the results output, each line indicates a classification of the read.

For example: SRR044946.347 GQ156763:1..1446 548 0.91985428 505 22 22 1 540 263 800 0.99072355 20 Bacteria 1 1 Firmicutes 0.55452305 1 Clostridia 0.55452305 1 Clostridiales 0.55452305 1 Ruminococcaceae 0.31190208 0.3119020760059357 Ruminococcus 0 0.2104288786649704 Ruminococcus gnavus 0 0.0604640431702137 Bacteria 0.58272612 1 F6159 0.22964814 1 G35588 0 1 S61033 0 0.7381679934055649 SS52094 0 0.2980887881680916

Each field is tab separated and indicates the following:

1 Query sequence id

2 Reference sequence id (highest alignment score)

3 Alignment bitscore

4 Pairwise identity

5 Matches

6 Mismatches

7 Gaps

8 Query start pos

9 Query end pos

10 Reference start pos

11 Reference end pos

12 (empty)

After the first empty field the taxonomy classifications and confidences are shown, every taxonomy classification is separated by an empty field. Although different fasta reference and taxonomy databases can be specified by the user, by default mapseq maps reads to the NCBI taxonomy and to OTU taxonomies

NCBI taxonomy fields:

13,14,15 kingdom, combined confidence (score+cutoff), score confidence

16,17,18 phylum, combined confidence, score confidence

19,20,21 class, combined confidence, score confidence

22,23,24 order, combined confidence, score confidence

25,26,27 family, combined confidence, score confidence

28,29,30 genus, combined confidence, score confidence

31,32,33 species, combined confidence, score confidence

34 (empty)

OTU taxonomy fields:

35,36,37 kingdom, combined confidence (score+cutoff), score confidence

38,39,40 90% otus, combined confidence, score confidence

41,42,43 96% otus, combined confidence, score confidence

44,45,46 98% otus, combined confidence, score confidence

47,48,49 99% otus, combined confidence, score confidence

The combined confidence is computed based on a score confidence, used to control misclassification errors, and a identity cutoff confidence, used to ensure that the query isnt misclassified due to the inexistence of a sequence representative in the database of the true classification. The score confidence is calculated by comparing the identity of the assigned taxonomy to the identity of the first sequence not matching the assigned taxonomy. The identity cutoff confidence uses preoptimized cutoffs at each taxonomic level to calculate the confidence that the query is not too divergent from the assigned taxonomy.

For further information contact: João F. Matias Rodrigues <jfmrod@gmail.com>