MetaGen User Manual

Xin Xing

Contents

0.1 Overview

MetaGen is a reference-free metagenomic tool for simultaneously binning microbial species and quantifying their distributions using multiple metagenomic samples.

Since MetaGen solely uses the cross-sample abundance patterns for binning, we recommend that the number of samples is larger than 10 samples or 2% of total number of species as a practical guide.

0.2 Installation

0.2.1 Dependencies

- Genome Assembly:
  - Ray Assembler($\geq v2.3.1$)
    - *User can also use MegaHIT Assembler($\geq v1.1$), if a preference is given to MegaHIT.*
- Extracting the reads count mapping matrix:
  - Bowtie2($\geq v2.2.4$)
  - Samtools($\geq v1.3$)
- Binning and distribution estimation:
  - R ($\geq v3.2.1$)

To initialize MetaGen, run the following code in R:

```
packages <- c("Rcpp", "MASS", "mixtools", "doParallel", "foreach",
"seqinr","getopt")
for(i in 1:length(packages)){
        if(packages[i] %in% installed.packages()[, "Package"]){
    next
    }else{
        install.packages(packages[i], dependencies=TRUE)
    }
}
```

0.2.2 Installation

The installation is tested on linux system with Ubuntu Server 16.04/Red Hat Enterprise release 6.7. First, download MetaGen using

```
git clone https://github.com/BioAlgs/MetaGen
```

Add the MetaGen's installation directory, the path of data set and the working directory to bash variables:

```
metagen=/home/username/metagen_v1.0.1
metagen_data=/home/username/test_data
metagen_work_dir=/home/username/example
```

0.3 Genome assembly to get contigs

The first step to begin the analysis is to assemble the reads of multiple samples using

```
cd $metagen_data
mpiexec -n #of_cores Ray -k length_of_kmer -detect-sequence-files ./ -o \
$metagen_work_dir/ray
mkdir $metagen_work_dir/contigs
cp $metagen_work_dir/ray/Contigs.fasta $metagen_work_dir/contigs
```

where "-n" in mpiexec specifies number of cores, "-k" in mpiexec specifies the length of k-mers(default is 31), "-detect-sequence-files" specifies automatically detect of both paired-end reads and single-end reads, "-o" specifies the output directory.
⋆ *If you are prefer to use MegaHIT, see https://github.com/voutcn/megahit for detailed instruction.*

0.3.1 Extracting the read counts mapping matrix

(1) Build bowtie2 index for the assembled contigs using the following code.

```
cd $metagen_work_dir/contigs
bowtie2-build Contigs.fasta ./contigs-ref
```

(2) Extract read counts use the following:

```
bash $metagen/scripts/bowtie2-align.sh \
[options] <ref> <outdir> <reads-dir> <sample-name>
```

where

```
[options]:
-h To show help documentation
-s single-end reads
-p paired-end reads
-a fasta files
-q fastq files(It is recommended to convert fastq to fasta and use -a option)

[input arguments]:
<ref>: The reference name of the bowtie2 index.
<outdir>: The output directory for the alignment result.
<reads-dir>: The directory of original reads.
<sample-name>: The sample name, for paired-ends reads
               sample-name_1.fastq sample-name_2.fastq,
               for single end reads sample-name.fastq.
```

(3) Build read counts mapping matrix (RCMM) using

```
bash MetaGen/scripts/combine-counts.sh \
[options] <ref> <outdir> <reads-dir> <sample-name>
```

where

```
[options]:
-h To show help documentation
-s single-end reads
-p paired-end reads

[input argument]:
<ref>: Specify the bowtie index obtained in (1)
<out-dir>: Specify the output directory.
<reads-dir>: Specify the working directory.
<sample-name>: The sample name, for paired-ends reads
               sample-name_1.fastq sample-name_2.fastq,
               for single end reads sample-name.fastq.

[output]:
$metagen_work_dir/output/count-map.tsv:
  The extracted read counts mapping matrix.
```

0.3.2 Binning using MetaGen

```
Rscript MetaGen/R/metagen.R -m <metagen_path> -w <work_dir>
```

```
[options]:
--metagen_path -m Specify the MetaGen's installation
  directory.
--work_dir -w Specify the working directory of current
  data set.

[optional options:]
--help -h Show help documentation.
--num_threads -n Specify the number of CPU cores used
  for parallel computing. When there is a large number of
  contigs, it is recommended to set multiple CPU cores to
  accelerate the computation. The default number is 1.
--bic_min -i Specify the minimum number of clusters. The
  default is 2.
--bic_max -a Specify the maximum number of clusters. The
  default is 0, which will let the algorithm sets the maximum
  number of cluster automatically.
--bic_step -s Specify the increment of the number of
  clusters from bic_min to bic_max. The default is 1.
--thred -t Specify the threshold for setting the initial
  value. It is recommended to set this number smaller(0.01
  -0.1) when the number of samples is less than $10$ and larger
  (0.1-0.2) when the number of samples is larger than $10$. The
  default value is set to 0.1.
--initial_per -p Specify how many percent contigs are used
  to set the initial value of the algorithm. The default
  value is 1, which means that all the contigs is used to
  find initial value of the algorithm. The number can be
  set to a smaller one, when there are a very large number
  of contigs.
--ctg_len_trim -l Specify the minimum contig length,
  contigs shorter than this value will not be included.
  Default is 500.
--plot-bic -p If the value is "T", output the plot of BIC scores.
```

```
  The default is "F".
-o The value is "1" for the simple metagenomic community.
   The value is "2" for the complex metagenomic community.

[output]:
$metagen_work_dir/output/segs.txt:
  The binning results for each contigs in a table with
  two columns. The first column lists the names of contigs.
  The second column lists the cluster ID for each contig.
$metagen_work_dir/output/scaled_relative_abundance:
  The scaled relative abundance matrix with column sum equals to 1.
  The first row specifies the sample names.
$metagen_work_dir/output/relative_abundance.txt
  The relative abundance matrix with first row specifies
  the sample names.
```

0.4 Example

In this section, a small example is presented to illustrate how to use MetaGen to analyze metagenomic data set. A test data is available to download through the repository: Test-Data. It is recommended to run the following examples on a computer with large memory. If you are running the example on a local computer, please make sure that the RAM of computer is larger than 8Gb.

Set MetaGen's installation directory, the working directory and the path of test data set to bash variables:

```
metagen=/home/username/metagen_v1.0.1
metagen_data=/home/username/Test-Data
metagen_work_dir=/home/username/example
```

Run Ray assembler for pooled assembly:

```
cd $metagen_data
mpiexec -n 10 Ray -k 31 -detect-sequence-files ./ -o \
$metagen_work_dir/ray
```

Build the bowtie2 index for the assembled contigs:

```
mkdir $metagen_work_dir/contigs
```

```
cp $metagen_work_dir/ray/Contigs.fasta $metagen_work_dir/contigs
cd $metagen_work_dir/contigs
bowtie2-build Contigs.fasta ./contigs-ref
```

Align the reads of each sample to the bowtie2 index. Here we use the "xargs" to run alignment in parallel. You can also set the "-P" option of "xargs" to 1, if you do not prefer the parallel computation. For paired-end reads, you need to change the "-s" option of "$metagen/bowtie2-align.sh" to "-p".

```
cd ../
chmod +x $metagen/script/bowtie2-align.sh
ls $metagen_data/*.fasta | \
gawk '{gsub(/.*[/]|.fasta/, "", $0)} 1' | \
xargs -P 6 -n 1 $metagen/script/bowtie2-align.sh -s -a \
$metagen_work_dir/contigs/contigs-ref \
$metagen_work_dir/map $metagen_data
```

Extract the read counts mapping matrix from the alignment results:

```
bash $metagen/script/combine-counts.sh -s $metagen_work_dir
```

Extract the number of reads for each sample using "-s" for single-end reads and "-p" for paired-end reads.

```
bash $metagen/script/sum-reads.sh -s $metagen_data $metagen_work_dir
```

Run the main statistical algorithm of MetaGen for binning and simultaneously estimating the relative abundance:

```
Rscript $metagen/R/metagen.R -m $metagen -w $metagen_work_dir
```

Check the binning result and the estimated scaled relative abundance matrix and relative abundance matrix in:

```
[output]:
$metagen_work_dir/output/segs.txt:
  The binning results for each contigs in a table with
  two columns. The first column lists the names of contigs.
  The second column lists the cluster ID for each contig.
$metagen_work_dir/output/scaled_relative_abundance:
  The scaled relative abundance matrix with column sum equals to 1.
  The first row specifies the sample names.
$metagen_work_dir/output/relative_abundance.txt
  The relative abundance matrix with first row specifies
  the sample names.
\end{verbatim}
```

```
% \subsection{For paired-end reads}
% Set MetaGen's installation directory, the working directory and the path of test data set to b
% \begin{lstlisting}[language=bash]
% metagen=/home/username/metagen_v1.0.1
% metagen_data=/home/username/paired
% metagen_work_dir=/home/username/example
%
```