SortMeRNA User Manual

Evguenia Kopylova
jenya.kopylov@gmail.com

Oct 2014, version 2.0

Contents

1 Introduction 3

2 Installation 3
- 2.1 Install from tarball release . . . 3
- 2.2 Install development version from git . . . 4
- 2.3 Install from precompiled code . . . 4
- 2.4 Uninstall . . . 5

3 Databases 5

4 How to run SortMeRNA 5
- 4.1 Index the rRNA database: command '`indexdb_rna`' . . . 5
  - 4.1.1 Example 1: indexdb_rna using one database . . . 6
  - 4.1.2 Example 2: indexdb_rna using multiple databases . . . 7
- 4.2 A guide to choosing '**sortmerna**' parameters for filtering and read mapping . . . 8
- 4.3 Filter rRNA reads . . . 9
  - 4.3.1 Example 3: multiple databases and the fastest alignment option . . . 11
  - 4.3.2 Filtering paired-end reads . . . 14
  - 4.3.3 Example 4: forward-reverse paired-end reads (2 input files) . . . 15
- 4.4 Read mapping . . . 17
  - 4.4.1 Mapping reads for classification . . . 17
  - 4.4.2 Example 5: mapping reads against the 16S Greengenes 97% id database with multithreading . . . 17
- 4.5 OTU-picking . . . 19

5 SortMeRNA advanced options 19

6 Help 20

7 Citation 20

1 Introduction

`http://bioinfo.lifl.fr/RNA/sortmerna/`
OTU-picking extensions and continuous support developed in the Knight Lab,
BioFrontiers Institute, University of Colorado at Boulder, CO
`https://knightlab.colorado.edu`

SortMeRNA is a local sequence alignment tool for filtering, mapping and OTU-picking. The core algorithm is based on approximate seeds and allows for fast and sensitive analyses of NGS reads. The main application of SortMeRNA is filtering rRNA from metatranscriptomic data. Additional applications include OTU-picking and taxonomy assignation available through QIIME v1.9+ (`http://qiime.org`, currently the development version to be released in early December). SortMeRNA takes as input a file of reads (fasta or fastq format) and one or multiple rRNA database file(s), and sorts apart aligned and rejected reads into two files specified by the user. SortMeRNA works with Illumina, 454, Ion Torrent and PacBio data, and can produce SAM and BLAST-like alignments.

For questions & help, please contact:

1.	Evguenia Kopylova	evguenia.kopylova@lifl.fr
2.	Laurent Noe	laurent.noe@lifl.fr
3.	Helene Touzet	helene.touzet@lifl.fr

Important: This user manual is strictly for SortMeRNA version 2.0.

2 Installation

2.1 Install from tarball release

1. Download `sortmerna-2.0.tar.gz` from `https://github.com/biocore/sortmerna/releases`
2. Extract the source code package into a directory of your choice, enter `sortmerna-2.0` directory and type,

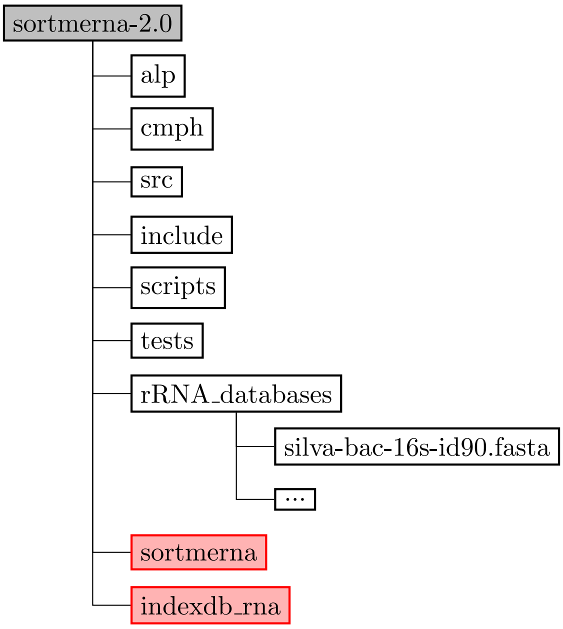
   ```
   > bash ./build.sh
   ```

3. At this point, two executables `indexdb_rna` and `sortmerna` will be located in the `sortmerna-2.0` directory. If the user would like to install the executables into their default installation directory (`/usr/local/bin` for Linux or `/opt/local/bin` for Mac) then type,

   ```
   > make install (with root permissions)
   ```

4. To begin using SortMeRNA, type '`indexdb_rna -h`' or '`sortmerna -h`'. Databases must first be indexed using `indexdb_rna`.

Figure 1: `sortmerna-2.0` directory tree



2.2 Install development version from git

1. Clone the sortmerna directory to your local system

```
> git clone https://github.com/biocore/sortmerna.git
```

2. Build sortmerna

```
> cd sortmerna
> bash ./build.sh
```

2.3 Install from precompiled code

1. Download the latest binary distribution of SortMeRNA from `http://bioinfo.lifl.fr/RNA/sortmerna`
2. Extract the source code package into a directory of your choice,

```
> tar -xvf sortmerna-2.0.tar.gz
> cd sortmerna-2.0
```

3. To begin using SortMeRNA, type '`indexdb_rna -h`' or '`sortmerna -h`'. The user must firstly index the databases with the command `indexdb_rna` before they can run the command `sortmerna`.

2.4 Uninstall

If the user installed SortMeRNA using the command `'make install'`, then they can use the command `'make uninstall'` to uninstall SortMeRNA (with root permissions).

3 Databases

SortMeRNA comes prepackaged with 8 databases,

representative database	**%id**	**# seq (clustered)**	**origin**	**# seq (original)**
silva-bac-16s-id90	90	12798	SILVA SSU Ref NR v.119	464618
silva-arc-16s-id95	95	3193	SILVA SSU Ref NR v.119	18797
silva-euk-18s-id95	95	7348	SILVA SSU Ref NR v.119	51553
silva-bac-23s-id98	98	4488	SILVA LSU Ref v.119	43822
silva-arc-23s-id98	98	251	SILVA LSU Ref v.119	629
silva-euk-28s-id98	98	4935	SILVA LSU Ref v.119	13095
rfam-5s-id98	98	59513	RFAM	116760
rfam-5.8s-id98	98	13034	RFAM	225185

HMMER 3.1b1 and SumaClust v1.0.00 were used to reduce the size of the original databases to the similarity listed in column 2 (%id) of the table above (see `/sortmerna/rRNA_databases/README.txt` for a list of complete steps).

These representative databases were specifically made for fast filtering of rRNA. Approximately the same number of rRNA will be filtered using silva-bac-16s-id90 (12802 rRNA) as using Greengenes 97% (99322 rRNA), but the former will run significantly faster.

id %: members of the cluster must have identity at least this % id with the representative sequence

Remark: The user must first index the fasta database by using the command `indexdb_rna` and then filter/map reads against the database using the command `sortmerna`.

4 How to run SortMeRNA

4.1 Index the rRNA database: command 'indexdb_rna'

The executable `indexdb_rna` indexes an rRNA database.

To see the man page for `indexdb_rna`,

```
>> indexdb_rna -h

  Program:     SortMeRNA version 2.0, 29/11/2014
  Copyright:   2012-2015 Bonsai Bioinformatics Research Group:
               LIFL, University Lille 1, CNRS UMR 8022, INRIA Nord-Europe
               OTU-picking extensions and continuing support developed in the Knight Lab,
               BioFrontiers Institute, University of Colorado at Boulder
  Disclaimer:  SortMeRNA comes with ABSOLUTELY NO WARRANTY; without even the
               implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
               See the GNU Lesser General Public License for more details.
  Contact:     Evguenia Kopylova, jenya.kopylov@gmail.com
               Laurent Noe, laurent.noe@lifl.fr
               Helene Touzet, helene.touzet@lifl.fr


  usage:   ./indexdb_rna --ref db.fasta,db.idx [OPTIONS]:

  --------------------------------------------------------------------------------------------------------
  | parameter        value          description                                                 default |
  --------------------------------------------------------------------------------------------------------
     --ref           STRING,STRING  FASTA reference file, index file                            mandatory
                                     (ex. --ref /path/to/file1.fasta,/path/to/index1)
                                      If passing multiple reference sequence files, separate
                                      them by ':',
                                     (ex. --ref /path/to/file1.fasta,/path/to/index1:/path/to/file2.fasta,path/to/index2)
   [OPTIONS]:
     --fast          BOOL           suggested option for aligning ~99% related species          off
     --sensitive     BOOL           suggested option for aligning ~75-98% related species       on
     --tmpdir        STRING         directory where to write temporary files
     -m              INT            the amount of memory (in Mbytes) for building the index     3072
     -L              INT            seed length                                                 18
     --max_pos       INT            maximum number of positions to store for each unique L-mer  10000
                                     (setting --max_pos 0 will store all positions)
     -v              BOOL           verbose
     -h              BOOL           help
```

There are eight rRNA representative databases provided in the 'sortmerna-2.0/rRNA_databases' folder. All databases were derived from the SILVA SSU and LSU databases (release 119) and the RFAM databases using HMMER 3.1b1 and SumaClust v1.0.00. Additionally, the user can index their own database.

4.1.1 Example 1: indexdb_rna using one database

```
>> ./indexdb_rna --ref ./rRNA_databases/silva-bac-16s-id90.fasta,./index/silva-bac-16s-db -v

  Program:     SortMeRNA version 2.0, 29/11/2014
  Copyright:   2012-2015 Bonsai Bioinformatics Research Group:
               LIFL, University Lille 1, CNRS UMR 8022, INRIA Nord-Europe
               OTU-picking extensions and continuing support developed in the Knight Lab,
               BioFrontiers Institute, University of Colorado at Boulder
  Disclaimer:  SortMeRNA comes with ABSOLUTELY NO WARRANTY; without even the
               implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
               See the GNU Lesser General Public License for more details.
  Contact:     Evguenia Kopylova, jenya.kopylov@gmail.com
               Laurent Noe, laurent.noe@lifl.fr
               Helene Touzet, helene.touzet@lifl.fr


  Parameters summary:
    K-mer size: 19
    K-mer interval: 1
    Maximum positions to store per unique K-mer: 10000
  Total number of databases to index: 1

  Begin indexing file ./rRNA_databases/silva-bac-16s-id90.fasta under index name ./index/silva-bac-16s-db:
  Collecting sequence distribution statistics ..  done  [1.133206 sec]

  start index part # 0:
    (1/3) building burst tries .. done  [23.643256 sec]
    (2/3) building CMPH hash .. done  [22.306709 sec]
    (3/3) building position lookup tables .. done [54.958680 sec]
    total number of sequences in this part = 12798
      writing kmer data to ./index/silva-bac-16s-db.kmer_0.dat
      writing burst tries to ./index/silva-bac-16s-db.bursttrie_0.dat
      writing position lookup table to ./index/silva-bac-16s-db.pos_0.dat
      writing nucleotide distribution statistics to ./index/silva-bac-16s-db.stats
    done.
```

4.1.2 Example 2: indexdb_rna using multiple databases

Multiple databases can be indexed simultaneously by passing them as a ':' separated list to `--ref` (no spaces allowed).

```
>> ./indexdb_rna --ref ./rRNA_databases/silva-bac-16s-id90.fasta,./index/silva-bac-16s-db:\
./rRNA_databases/silva-bac-23s-id98.fasta,./index/silva-bac-23s-db:\
./rRNA_databases/silva-arc-16s-id95.fasta,./index/silva-arc-16s-db:\
./rRNA_databases/silva-arc-23s-id98.fasta,./index/silva-arc-23s-db:\
./rRNA_databases/silva-euk-18s-id95.fasta,./index/silva-euk-18s-db:\
./rRNA_databases/silva-euk-28s-id98.fasta,./index/silva-euk-28s:\
./rRNA_databases/rfam-5s-database-id98.fasta,./index/rfam-5s-db:\
./rRNA_databases/rfam-5.8s-database-id98.fasta,./index/rfam-5.8s-db
```

4.2 A guide to choosing 'sortmerna' parameters for filtering and read mapping

In SortMeRNA version 1.99 beta and up, users have the option to output sequence alignments for their matching rRNA reads in the SAM or BLAST-like formats. Depending on the desired quality of alignments, different parameters choices must be set. Table 1 presents a guide to setting parameters choices for most use cases. In all cases, output alignments are always guaranteed to reach the threshold E-value score (default E-value=1). An E-value of 1 signifies that one random alignment is expected for aligning **all** reads against the reference database. The E-value in SortMeRNA is computed for the entire search space, not per read.

Table 1: SortMeRNA alignment parameter guide

option	speed	description
`--num-alignments INT`	Very fast for `INT = 1`	Output the first alignment passing E-value threshold (**best choice if only filtering is needed**)
	Speed decreases for higher value `INT`	Higher `INT` signifies more alignments will be made & output
	Very slow for `INT = 0`	All alignments reaching the E-value threshold are reported (this option is not suggested for high similarity rRNA databases, due to many possible alignments per read causing a very large file output)
`--best INT`	Fast for `INT = 1`	Only one high-candidate reference sequence will be searched for alignments (determined heuristically using a Longest Increasing Subsequence of seed matches). The single best alignment of those will be reported
	Speed decreases for higher value `INT`	Higher `INT` signifies more alignments will be made, though only the best one will be reported
	Very slow for `INT = 0`	All high-candidate reference sequences will be searched for alignments, though only the best one will be reported

4.3 Filter rRNA reads

The executable `sortmerna` can filter rRNA reads against an indexed rRNA database.

To see the man page for `sortmerna`,

```
>> ./sortmerna -h

  Program:     SortMeRNA version 2.0, 29/11/2014
  Copyright:   2012-2015 Bonsai Bioinformatics Research Group:
               LIFL, University Lille 1, CNRS UMR 8022, INRIA Nord-Europe
               OTU-picking extensions and continuing support developed in the Knight Lab,
               BioFrontiers Institute, University of Colorado at Boulder
  Disclaimer:  SortMeRNA comes with ABSOLUTELY NO WARRANTY; without even the
               implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
               See the GNU Lesser General Public License for more details.
  Contact:     Evguenia Kopylova, jenya.kopylov@gmail.com
               Laurent Noe, laurent.noe@lifl.fr
               Helene Touzet, helene.touzet@lifl.fr


  usage:   ./sortmerna --ref db.fasta,db.idx --reads file.fa --aligned base_name_output [OPTIONS]:

  -------------------------------------------------------------------------------------------------------------
  | parameter         value           description                                                    default |
  -------------------------------------------------------------------------------------------------------------
     --ref             STRING,STRING   FASTA reference file, index file                               mandatory
                                         (ex. --ref /path/to/file1.fasta,/path/to/index1)
                                         If passing multiple reference files, separate
                                         them using the delimiter ':',
                                         (ex. --ref /path/to/file1.fasta,/path/to/index1:/path/to/file2.fasta,path/to/index
     --reads           STRING          FASTA/FASTQ reads file                                         mandatory
     --aligned         STRING          aligned reads filepath + base file name                        mandatory
                                         (appropriate extension will be added)

   [COMMON OPTIONS]:
     --other           STRING          rejected reads filepath + base file name
                                         (appropriate extension will be added)
     --fastx           BOOL            output FASTA/FASTQ file                                        off
                                         (for aligned and/or rejected reads)
     --sam             BOOL            output SAM alignment                                           off
                                         (for aligned reads only)
     --SQ              BOOL            add SQ tags to the SAM file                                    off
     --blast           INT             output alignments in various Blast-like formats
                                        0 - pairwise
                                        1 - tabular (Blast -m 8 format)
                                        2 - tabular + column for CIGAR
                                        3 - tabular + columns for CIGAR and query coverage
     --log             BOOL            output overall statistics                                      off
     --num_alignments  INT             report first INT alignments per read reaching E-value          -1
                                        (--num_alignments 0 signifies all alignments will be output)
       or (default)
     --best            INT             report INT best alignments per read reaching E-value           1
                                         by searching --min_lis INT candidate alignments
                                        (--best 0 signifies all candidate alignments will be searched)
     --min_lis         INT             search all alignments having the first INT longest LIS         2
                                         LIS stands for Longest Increasing Subsequence, it is
                                         computed using seeds' positions to expand hits into
                                         longer matches prior to Smith-Waterman alignment.
     --print_all_reads BOOL            output null alignment strings for non-aligned reads            off
                                         to SAM and/or BLAST tabular files
     --paired_in       BOOL            both paired-end reads go in --aligned fasta/q file             off
                                         (interleaved reads only, see Section 4.2.4 of User Manual)
     --paired_out      BOOL            both paired-end reads go in --other fasta/q file               off
                                         (interleaved reads only, see Section 4.2.4 of User Manual)
     --match           INT             SW score (positive integer) for a match                        2
     --mismatch        INT             SW penalty (negative integer) for a mismatch                   -3
     --gap_open        INT             SW penalty (positive integer) for introducing a gap            5
     --gap_ext         INT             SW penalty (positive integer) for extending a gap              2
     -N                INT             SW penalty for ambiguous letters (N's)                         scored as --mismatch
     -F                BOOL            search only the forward strand                                 off
     -R                BOOL            search only the reverse-complementary strand                   off
     -a                INT             number of threads to use                                       1
     -e                DOUBLE          E-value threshold                                              1
     -m                INT             INT Mbytes for loading the reads into memory                   1024
                                        (maximum -m INT is 4096)
     -v                BOOL            verbose                                                        off


   [OTU PICKING OPTIONS]:
     --id              DOUBLE          %id similarity threshold (the alignment must                   0.97
                                         still pass the E-value threshold)
     --coverage        DOUBLE          %query coverage threshold (the alignment must                  0.97
                                         still pass the E-value threshold)
     --de_novo_otu     BOOL            FASTA/FASTQ file for reads matching database < %id             off
                                         (set using --id) and < %cov (set using --coverage)
                                         (alignment must still pass the E-value threshold)
     --otu_map         BOOL            output OTU map (input to QIIME's make_otu_table.py)            off


   [ADVANCED OPTIONS] (see SortMeRNA user manual for more details):
    --passes           INT,INT,INT     three intervals at which to place the seed on the read         L,L/2,3
                                         (L is the seed length set in ./indexdb_rna)
    --edges            INT             number (or percent if INT followed by % sign) of               4
                                         nucleotides to add to each edge of the read
                                         prior to SW local alignment
    --num_seeds        INT             number of seeds matched before searching                       2
                                         for candidate LIS
    --full_search      BOOL            search for all 0-error and 1-error seed                        off
                                         matches in the index rather than stopping
                                         after finding a 0-error match (<1% gain in
                                         sensitivity with up four-fold decrease in speed)
    --pid              BOOL            add pid to output file names                                   off


   [HELP]:
     -h                BOOL            help
     --version         BOOL            SortMeRNA version number
```

The user can adjust the amount of memory allocated for loading the reads through the command option -m. By default, -m is set to be high enough for 1GB. If the reads file is larger than 1GB, then sortmerna internally divides the file into partial sections of 1GB and executes one section at a time. Hence, if a user has an input file of 15GB and only 1GB of RAM to store it, the file will be processed in partial sections using mmap without having to physically split it prior to execution. Otherwise, the user can increase -m to map larger portions of the file. The limit for -m is given by typing sortmerna -h.

4.3.1 Example 3: multiple databases and the fastest alignment option

```
>> time ./sortmerna --ref ./rRNA_databases/silva-bac-16s-id90.fasta,./index/silva-bac-16s-db:\
./rRNA_databases/silva-bac-23s-id98.fasta,./index/silva-bac-23s-db:\
./rRNA_databases/silva-arc-16s-id95.fasta,./index/silva-arc-16s-db:\
./rRNA_databases/silva-arc-23s-id98.fasta,./index/silva-arc-23s-db:\
./rRNA_databases/silva-euk-18s-id95.fasta,./index/silva-euk-18s-db:\
./rRNA_databases/silva-euk-28s-id98.fasta,./index/silva-euk-28s:\
./rRNA_databases/rfam-5s-database-id98.fasta,./index/rfam-5s-db:\
./rRNA_databases/rfam-5.8s-database-id98.fasta,./index/rfam-5.8s-db\
 --reads SRR106861.fasta --sam --num_alignments 1 --fastx --aligned SRR105861_rRNA\
 --other SRR105861_non_rRNA --log -v


  Program:     SortMeRNA version 2.0, 29/11/2014
  Copyright:   2012-2015 Bonsai Bioinformatics Research Group:
               LIFL, University Lille 1, CNRS UMR 8022, INRIA Nord-Europe
               OTU-picking extensions and continuing support developed in the Knight Lab,
               BioFrontiers Institute, University of Colorado at Boulder
  Disclaimer:  SortMeRNA comes with ABSOLUTELY NO WARRANTY; without even the
               implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
               See the GNU Lesser General Public License for more details.
  Contact:     Evguenia Kopylova, jenya.kopylov@gmail.com
               Laurent Noe, laurent.noe@lifl.fr
               Helene Touzet, helene.touzet@lifl.fr


  Computing read file statistics ... done [2.16 sec]
  size of reads file: 35238748 bytes
  partial section(s) to be executed: 1 of size 35238748 bytes
  Parameters summary:
    Number of seeds = 2
    Edges = 4 (as integer)
    SW match = 2
    SW mismatch = -3
    SW gap open penalty = 5
    SW gap extend penalty = 2
    SW ambiguous nucleotide = -3
    SQ tags are not output
    Number of threads = 1

  Begin mmap reads section # 1:
  Time to mmap reads and set up pointers [0.11 sec]

  Begin analysis of: ./rRNA_databases/silva-bac-16s-id90.fasta
    Seed length = 18
    Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
    Gumbel lambda = 0.602397
    Gumbel K = 0.328927
    Minimal SW score based on E-value = 54
    Loading index part 1/1 ...  done [4.67 sec]
    Begin index search ...  done [83.53 sec]
    Freeing index ...  done [0.87 sec]

  Begin analysis of: ./rRNA_databases/silva-bac-23s-id98.fasta
    Seed length = 18
    Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
    Gumbel lambda = 0.603075
    Gumbel K = 0.330488
    Minimal SW score based on E-value = 53
    Loading index part 1/1 ...  done [3.63 sec]
    Begin index search ...  done [94.76 sec]
    Freeing index ...  done [0.41 sec]

  Begin analysis of: ./rRNA_databases/silva-arc-16s-id95.fasta
    Seed length = 18
    Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
    Gumbel lambda = 0.596230
    Gumbel K = 0.322143
    Minimal SW score based on E-value = 52
    Loading index part 1/1 ...  done [1.14 sec]
    Begin index search ...  done [22.63 sec]
    Freeing index ...  done [0.14 sec]

  Begin analysis of: ./rRNA_databases/silva-arc-23s-id98.fasta
    Seed length = 18
    Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
    Gumbel lambda = 0.597749
    Gumbel K = 0.325630
    Minimal SW score based on E-value = 49
    Loading index part 1/1 ...  done [0.50 sec]
    Begin index search ...  done [13.27 sec]
    Freeing index ...  done [0.06 sec]

  Begin analysis of: ./rRNA_databases/silva-euk-18s-id95.fasta
    Seed length = 18
    Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
    Gumbel lambda = 0.612228
    Gumbel K = 0.334926
    Minimal SW score based on E-value = 52
    Loading index part 1/1 ...  done [3.23 sec]
    Begin index search ...  done [30.28 sec]
    Freeing index ...  done [0.45 sec]

  Begin analysis of: ./rRNA_databases/silva-euk-28s-id98.fasta
    Seed length = 18
    Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
    Gumbel lambda = 0.612068
    Gumbel K = 0.344763
    Minimal SW score based on E-value = 53
    Loading index part 1/1 ...  done [3.43 sec]
    Begin index search ...  done [35.69 sec]
    Freeing index ...  done [0.48 sec]

  Begin analysis of: ./rRNA_databases/rfam-5s-database-id98.fasta
    Seed length = 18
    Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
    Gumbel lambda = 0.616617
    Gumbel K = 0.341306
    Minimal SW score based on E-value = 51
    Loading index part 1/1 ...  done [1.77 sec]
    Begin index search ...  done [13.50 sec]
    Freeing index ...  done [0.22 sec]

  Begin analysis of: ./rRNA_databases/rfam-5.8s-database-id98.fasta
    Seed length = 18
    Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
    Gumbel lambda = 0.617817
    Gumbel K = 0.340589
    Minimal SW score based on E-value = 49
    Loading index part 1/1 ...  done [0.60 sec]
    Begin index search ...  done [8.78 sec]
    Freeing index ...  done [0.07 sec]
    Total number of reads mapped (incl. all reads file sections searched): 104243
```

```
    Writing aligned FASTA/FASTQ ...  done [1.13 sec]
    Writing not-aligned FASTA/FASTQ ...  done [0.10 sec]
```

The option '`--log`' will create an overall statistics file,

```
>> cat SRR105861_rRNA.log
 Time and date

 Command: sortmerna --ref ./rRNA_databases/silva-bac-16s-id90.fasta,./index/silva-bac-16s-db:\
 ./rRNA_databases/silva-bac-23s-id98.fasta,./index/silva-bac-23s-db:\
 ./rRNA_databases/silva-arc-16s-id95.fasta,./index/silva-arc-16s-db:\
 ./rRNA_databases/silva-arc-23s-id98.fasta,./index/silva-arc-23s-db:\
 ./rRNA_databases/silva-euk-18s-id95.fasta,./index/silva-euk-18s-db:\
 ./rRNA_databases/silva-euk-28s-id98.fasta,./index/silva-euk-28s:\
 ./rRNA_databases/rfam-5s-database-id98.fasta,./index/rfam-5s-db:\
 ./rRNA_databases/rfam-5.8s-database-id98.fasta,./index/rfam-5.8s-db\
  --reads /Users/jenya/Downloads/SRR106861.fasta --sam --num_alignments 1\
   --fastx --aligned SRR105861_rRNA --other SRR105861_non_rRNA.fasta fasta -v
 Process pid = 1957
 Parameters summary:
    Index: ./index/silva-bac-16s-db
     Seed length = 18
     Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
     Gumbel lambda = 0.602397
     Gumbel K = 0.328927
     Minimal SW score based on E-value = 54
    Index: ./index/silva-bac-23s-db
     Seed length = 18
     Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
     Gumbel lambda = 0.603075
     Gumbel K = 0.330488
     Minimal SW score based on E-value = 53
    Index: ./index/silva-arc-16s-db
     Seed length = 18
     Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
     Gumbel lambda = 0.596230
     Gumbel K = 0.322143
     Minimal SW score based on E-value = 52
    Index: ./index/silva-arc-23s-db
     Seed length = 18
     Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
     Gumbel lambda = 0.597749
     Gumbel K = 0.325630
     Minimal SW score based on E-value = 49
    Index: ./index/silva-euk-18s-db
     Seed length = 18
     Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
     Gumbel lambda = 0.612228
     Gumbel K = 0.334926
     Minimal SW score based on E-value = 52
    Index: ./index/silva-euk-28s
     Seed length = 18
     Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
     Gumbel lambda = 0.612068
     Gumbel K = 0.344763
     Minimal SW score based on E-value = 53
    Index: ./index/rfam-5s-db
     Seed length = 18
     Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
     Gumbel lambda = 0.616617
     Gumbel K = 0.341306
     Minimal SW score based on E-value = 51
    Index: ./index/rfam-5.8s-db
     Seed length = 18
     Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
     Gumbel lambda = 0.617817
     Gumbel K = 0.340589
     Minimal SW score based on E-value = 49
    Number of seeds = 2
    Edges = 4 (as integer)
    SW match = 2
    SW mismatch = -3
    SW gap open penalty = 5
    SW gap extend penalty = 2
    SW ambiguous nucleotide = -3
    SQ tags are not output
    Number of threads = 1
    Reads file = SRR106861.fasta

 Results:
    Total reads = 113128
    Total reads passing E-value threshold = 104243 (92.15%)
    Total reads failing E-value threshold = 8885 (7.85%)
    Minimum read length = 59
    Maximum read length = 1253
    Mean read length = 267
 By database:
    ./rRNA_databases/silva-bac-16s-id90.fasta               25.73%
    ./rRNA_databases/silva-bac-23s-id98.fasta               64.37%
    ./rRNA_databases/silva-arc-16s-id95.fasta               0.00%
    ./rRNA_databases/silva-arc-23s-id98.fasta               0.00%
    ./rRNA_databases/silva-euk-18s-id95.fasta               0.00%
    ./rRNA_databases/silva-euk-28s-id98.fasta               0.00%
    ./rRNA_databases/rfam-5s-database-id98.fasta                2.04%
    ./rRNA_databases/rfam-5.8s-database-id98.fasta                0.00%
```

4.3.2 Filtering paired-end reads

When writing aligned and non-aligned reads to FASTA/Q files, sometimes the situation arises where one of the paired-end reads aligns and the other one doesn't. Since SortMeRNA looks at each read individually, by default the reads will be split into two separate files. That is, the read that aligned will go into the `--aligned` FASTA/Q file and the pair that didn't align will go into the `--other` FASTA/Q file.

This situation would result in the splitting of some paired reads in the output files and not optimal for users who require paired order of the reads for downstream analyses.

For users who wish to keep the order of their paired-ended reads, two options are available. If one read aligns and the other one not then,

(1) `--paired-in` will put both reads into the file specified by `--aligned`

(2) `--paired-out` will put both reads into the file specified by `--other`

The first option, `--paired-in` is optimal for users that want all reads in the `--other` file to be non-rRNA. However, there are small chances that reads which are non-rRNA will also be put into the `--aligned` file.

The second option, `--paired-out` is optimal for users that want only rRNA reads in the `--aligned` file. However, there are small chances that reads which are rRNA will also be put into the `--other` file.

If neither of these two options is added to the `sortmerna` command, then aligned and non-aligned reads will be properly output to the `--aligned` and `--other` files, possibly breaking the order for a set of paired reads between two output files.

It's important to note that regardless of the options used, the `--log` file will always report the true number of reads classified as rRNA (not the number of reads in the `--aligned` file).

4.3.3 Example 4: forward-reverse paired-end reads (2 input files)

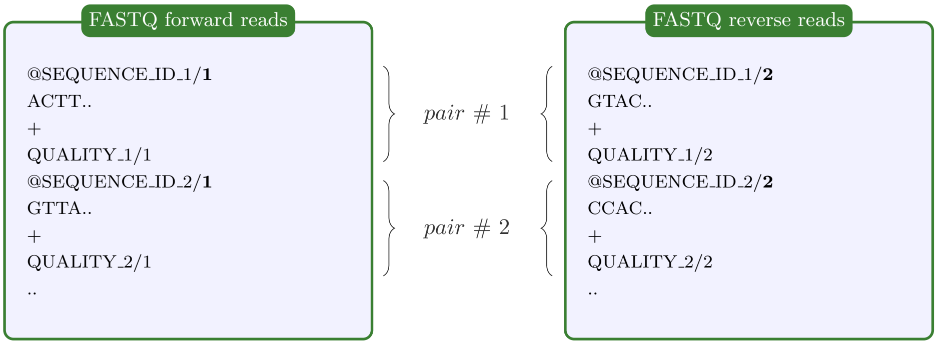


Figure 2: Forward and reverse reads in paired-end sequencing format

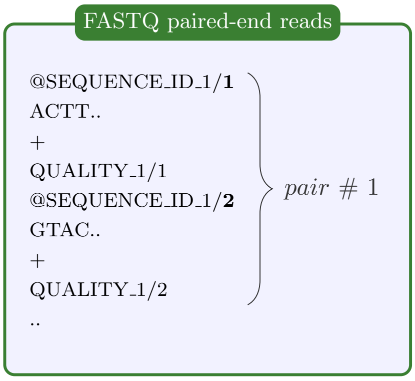


Figure 3: Paired-end read format accepted by SortMeRNA

SortMeRNA accepts only 1 file as input for the reads. If a user has two input files, in the case for the foward and reverse paired-end reads (see Figure 2), they may use the `merge-paired-reads.sh` script found in '`sortmerna/scripts`' folder to interleave the paired reads into the format of Figure 3.

The command for `merge-paired-reads.sh` is the following,

```
 > bash ./merge-paired-reads.sh forward-reads.fastq reverse-reads.fastq outfile.fastq
```

Now, the user may input `outfile.fastq` to SortMeRNA for analysis.

Similarly, for unmerging the paired reads back into two separate files, use the command,

```
> bash ./unmerge-paired-reads.sh merged-reads.fastq forward-reads.fastq reverse-reads.fastq
```

Important: unmerge-paired-reads.sh should only be used if one of the options `--paired_in` or `--paired_out` was used during filtering. Otherwise it may give incorrect results if a paired-read was split during alignment (one read aligned and the other one not).

4.4 Read mapping

4.4.1 Mapping reads for classification

Although SortMeRNA is very sensitive with the small rRNA databases distributed with the source code, these databases are not optimal for classification since often alignments with 75-90% identity will be returned (there are only several thousand rRNA in most of the databases, compared to the original SILVA or Greengenes databases containing millions of rRNA). Classification at the species level generally considers alignments at 97% and above, so it is suggested to use a larger database is species classification is the main goal.

Moreover, SortMeRNA is a local alignment tool, so it's also important to look at the query coverage % for each alignment. In the SAM output format, neither % id or query coverage are reported. If the user wishes for these values, then the Blast tabular format with CIGAR + query coverage option (`--blast 3`) is the way to go.

4.4.2 Example 5: mapping reads against the 16S Greengenes 97% id database with multithreading

This example will generate SAM and BLAST tabular output files. Alignments are classified as significant based on the E-value cutoff (default 1). SortMeRNA's E-value takes into consideration the full size of the reference database as well as the query file, thus the E-value is higher than BLAST's (ex. equivalent to BLAST's 1e-5).

```
>> sortmerna --ref 97_otus_gg_13_8.fasta,./index/97_otus_gg_13_8\
 --reads SRR106861.fasta --blast 3 --sam --log --aligned SRR106861_gg_rRNA -a 20 -v


  Program:     SortMeRNA version 2.0, 29/11/2014
  Copyright:   2012-2015 Bonsai Bioinformatics Research Group:
               LIFL, University Lille 1, CNRS UMR 8022, INRIA Nord-Europe
               OTU-picking extensions and continuing support developed in the Knight Lab,
               BioFrontiers Institute, University of Colorado at Boulder
  Disclaimer:  SortMeRNA comes with ABSOLUTELY NO WARRANTY; without even the
               implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.
               See the GNU Lesser General Public License for more details.
  Contact:     Evguenia Kopylova, jenya.kopylov@gmail.com
               Laurent Noe, laurent.noe@lifl.fr
               Helene Touzet, helene.touzet@lifl.fr


  Computing read file statistics ... done [0.44 sec]
  size of reads file: 35238748 bytes
  partial section(s) to be executed: 1 of size 35238748 bytes
  Parameters summary:
    Number of seeds = 2
    Edges = 4 (as integer)
    SW match = 2
    SW mismatch = -3
    SW gap open penalty = 5
    SW gap extend penalty = 2
    SW ambiguous nucleotide = -3
    SQ tags are not output
    Number of threads = 20

  Begin mmap reads section # 1:
  Time to mmap reads and set up pointers [0.10 sec]
  Begin analysis of: 97_otus_gg_13_8.fasta
    Seed length = 18
    Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
    Gumbel lambda = 0.600470
    Gumbel K = 0.327880
    Minimal SW score based on E-value = 57
    Loading index part 1/1 ...  done [10.76 sec]
    Begin index search ...  done [23.75 sec]
    Freeing index ...  done [1.44 sec]
    Total number of reads mapped (incl. all reads file sections searched): 29089
    Writing alignments ...  done [7.71 sec]
```

This is almost the same number of 16S rRNA as identified by SortMeRNA using the smaller provided database,

```
>> cat SRR106861_gg_rRNA.log
 Date and time

 Command: sortmerna --ref 97_otus_gg_13_8.fasta,./index/97_otus_gg_13_8\
  --reads SRR106861.fasta --blast 3 --sam --log --aligned SRR106861_gg_rRNA -a 20 -v
 Process pid = 44246
 Parameters summary:
    Index: ./index/97_otus_gg_13_8
     Seed length = 18
     Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
     Gumbel lambda = 0.600470
     Gumbel K = 0.327880
     Minimal SW score based on E-value = 57
    Number of seeds = 2
    Edges = 4 (as integer)
    SW match = 2
    SW mismatch = -3
    SW gap open penalty = 5
    SW gap extend penalty = 2
    SW ambiguous nucleotide = -3
    SQ tags are not output
    Number of threads = 20
    Reads file = SRR106861.fasta

 Results:
    Total reads = 113128
    Total reads passing E-value threshold = 29089 (25.71%)
    Total reads failing E-value threshold = 84039 (74.29%)
    Minimum read length = 59
    Maximum read length = 1253
    Mean read length = 267
 By database:
    97_otus_gg_13_8.fasta                25.71%
```

4.5 OTU-picking

SortMeRNA is implemented in QIIME's closed-reference and open-reference OTU-picking workflows. The readers are referred to QIIME's tutorials for an in-depth discussion of these methods `http://qiime.org/tutorials/otu_picking.html`.

5 SortMeRNA advanced options

`--num_seeds INT`

The threshold number of seeds required to match in the primary seed-search filter before moving on to the secondary seed-cluster filter. More specifically, the threshold number of seeds required before searching for a longest increasing subsequence (LIS) of the seeds' positions between the read and the closest matching reference sequence. By default, this is set to 2 seeds.

`--passes INT,INT,INT`

In the primary seed-search filter, SortMeRNA moves a seed of length L (parameter of `indexdb_rna`) across the read using three passes. If at the end of each pass a threshold number of seeds (defined by `--num_seeds`) did not match to the reference database, SortMeRNA attempts to find more seeds by decreasing the interval at which the seed is placed along the read by using another pass. In default mode, these intervals are set to $L, L/2, 3$ for Pass 1, 2 and 3, respectively. Usually, if the read is highly similar to the reference database, a threshold number of seeds will be found in the first pass.

`--edges INT(%)`

The number (or percentage if followed by %) of nucleotides to add to each edge of the alignment region on the reference sequence before performing Smith-Waterman alignment. By default, this is set to 4 nucleotides.

`--full_search FLAG`

During the index traversal, if a seed match is found with 0-errors, SortMeRNA will stop searching for further 1-error matches. This heuristic is based upon the assumption that 0-error matches are more significant than 1-error matches. By turning it off using the `--full_search` flag, the sensitivity may increase (often by less than 1%) but with up to four-fold decrease in speed.

`--pid FLAG`

The pid of the running `sortmerna` process will be added to the output files in order to avoid over-writing output if the same `--aligned STRING` base name is provided for different runs.

6 Help

Any issues or bug reports should be reported to `https://github.com/biocore/sortmerna/issues` or by e-mail to the authors (see list of e-mails in Section 1 of this document). Comments and suggestions are also always appreciated!

7 Citation

If you use SortMeRNA please cite,

Kopylova E., Noé L. and Touzet H., "SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data", *Bioinformatics* (2012), doi: 10.1093/bioinformatics/bts611.