SortMeRNA User Manual

Evguenia Kopylova
jenya.kopylov@gmail.com

March 2014, version 1.99 beta

Contents

1 Introduction

SortMeRNA is a software designed to filter metatranscriptomic reads data. It takes as input a file of reads (fasta or fastq format) and one or multiple rRNA database file(s), and sorts apart the accepted reads and the 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.

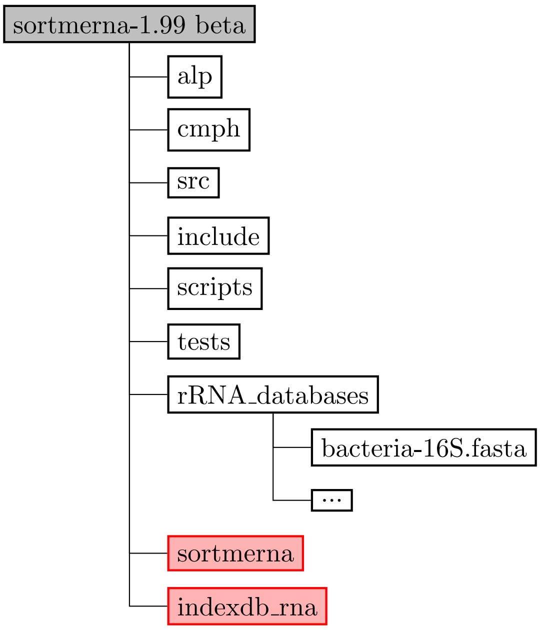
For questions & help, please contact:

```
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 1.99 beta.

2 Installation

Figure 1: `sortmerna-1.99-beta` directory tree



2.1 Install from source code

1. Download `sortmerna-1.99-beta.tar.gz` from `http://bioinfo.lifl.fr/RNA/sortmerna`

2. Extract the source code package into a directory of your choice, enter `sortmerna-1.99-beta` and type,

   ```
   > ./configure
   > make
   ```

3. At this point, two executables `indexdb_rna` and `sortmerna` will be located in the `sortmerna-1.99-beta` 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`.

2.2 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-1.99-beta.tar.gz
   > cd sortmerna-1.99-beta
   ```

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`.

2.3 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 %**	**average id %**	**# seq**	**origin**	**# seq**	**filtered to remove**
silva-bac-16s-database-id85.fasta	85	91.6	8174	SILVA SSU Ref NR v.111	244077	23s
silva-arc-16s-database-id95.fasta	95	96.7	3845	SILVA SSU Ref NR v.111	10919	23s
silva-euk-18s-database-id95.fasta	95	96.7	4512	SILVA SSU Ref NR v.111	31862	26s,28s,23s
silva-bac-23s-database-id95.fasta	98	99.4	3055	SILVA LSU Ref v.111	19580	16s,26s,28s
silva-arc-23s-database-id95.fasta	98	99.5	164	SILVA LSU Ref v.111	405	16s,26s,28s
silva-euk-28s-database-id95.fasta	98	99.1	4578	SILVA LSU Ref v.111	9321	18s
rfam-5s-database-id98.fasta	98	99.2	59513	RFAM	116760	–
rfam-5.8s-database-id98.fasta	98	98.9	13034	RFAM	225185	–

The tool UCLUST was used to reduce the size of the original databases.

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

Remark: The user must first index the fasta database by using the command `indexdb_rna` and then filter 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


  usage:   ./indexdb_rna <input> <output> <options>:

  --------------------------------------------------------------------------------------------------------
  | parameter        value          description                                                 default |
  --------------------------------------------------------------------------------------------------------
   <input>:
     --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          FLAG            suggested option for aligning ~99% related species         off
     --sensitive     FLAG            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 250
                                      (setting --max_pos 0 will store all positions)
     -v              FLAG            verbose
     -h              FLAG            help
```

There are eight rRNA representative databases provided in the '`sortmerna-1.99-beta/rRNA_databases`' folder. All databases were derived from the SILVA SSU and LSU databases (release 111) and the RFAM databases using the tool UCLUST. 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-database-id85.fasta,./index/silva-bac-16s -v

  Program:    SortMeRNA version 1.99 beta, 11/03/2014
```

```
Copyright:  2012-2014 Bonsai Bioinformatics Research Group
            LIFL, University Lille 1, CNRS UMR 8022, INRIA Nord-Europe
            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: 250

Total number of databases to index: 1

Begin indexing file ./rRNA_databases/silva-bac-16s-database-id85.fasta under index name ./index/silva-bac-16s:
Collecting sequence distribution statistics ..  done  [0.781479 sec]

start index part # 0:
  (1/3) building burst tries .. done  [14.726437 sec]
  (2/3) building CMPH hash .. done  [22.519546 sec]
  (3/3) building position lookup tables .. done [21.117368 sec]
  total number of sequences in this part = 8174
    writing kmer data to ./index/silva-bac-16s.kmer_0.dat
    writing burst tries to ./index/silva-bac-16s.bursttrie_0.dat
    writing position lookup table to ./index/silva-bac-16s.pos_0.dat
    writing nucleotide distribution statistics to ./index/silva-bac-16s.stats
  done.
```

4.1.2 Example 2: indexdb_rna using all eight 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-database-id85.fasta,./index/silva-bac-16s:\
./rRNA_databases/silva-bac-23s-database-id98.fasta,./index/silva-bac-23s:\
./rRNA_databases/silva-arc-16s-database-id95.fasta,./index/silva-arc-16s:\
./rRNA_databases/silva-arc-23s-database-id98.fasta,./index/silva-arc-23s:\
./rRNA_databases/silva-euk-18s-database-id95.fasta,./index/silva-euk-18s:\
./rRNA_databases/silva-euk-28s-database-id98.fasta,./index/silva-euk-28s:\
./rRNA_databases/rfam-5.8s-database-id98.fasta,./index/rfam-5.8s:\
./rRNA_databases/rfam-5s-database-id98.fasta,./index/rfam-5s
```

4.2 Filter reads against the indexed rRNA database: command ‘sortmerna’

The executable **sortmerna** filters rRNA reads against an indexed rRNA database.

To see the man page for **sortmerna**,

```
>> ./sortmerna -h

  usage:   ./sortmerna <input> <output> <options>:

  -------------------------------------------------------------------------------------------------------------
  | parameter          value          description                                                    default |
  -------------------------------------------------------------------------------------------------------------
   <input>:
     --reads           STRING         FASTA/FASTQ reads file                                         mandatory
     --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/index2)

   <output>:
     --aligned         STRING         aligned reads base file name
                                        (appropriate extension will be added)
     --other           STRING         rejected reads base file name
                                        (appropriate extension will be added)
     --fastx           FLAG           output FASTA/FASTQ file                                        off
                                        (for aligned and/or rejected reads)
     --sam             FLAG           output SAM alignment                                           off
                                        (for aligned reads only)
     --SQ              FLAG           add SQ tags to the SAM file                                    off
     --blast           FLAG           output BLAST-like alignment                                    off
                                        (for aligned reads only)
     --log             FLAG           output overall statistics                                      off

   For alignments (with --sam or --blast options):

     --feeling-lucky   FLAG           report the first alignment per read reaching E-value           off
       or
     --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 single best alignment per read reaching E-value         2
                                        from alignments of INT best candidate reference sequences
                                        (ex. --best 2: find all alignments for the first 2
                                        best matching reference sequences and report the
                                        the single best alignment; --best 0 signifies
                                        all highest scoring reference sequences will be searched)

   <options>:
     --paired_in       FLAG           both paired-end reads go in --aligned fasta/q file             off
     --paired_out      FLAG           both paired-end reads go in --other fasta/q file               off
     --match           INT            SW score (positive integer) for a match                        2
     --mismatch        INT            SW score (negative integer) for a mismatch                     -3
     --gap_open        INT            SW score (positive integer) for introducing a gap               5
     --gap_ext         INT            SW score (positive integer) for extending a gap                2
     -N                INT            SW score for ambiguous letters (N's)             scored as --mismatch
     -F                FLAG           search only the forward strand                                 off
     -R                FLAG           search only the reverse-complementary strand                   off
     -a                INT            number of threads to use                                       1
     -e                DOUBLE         E-value                                                        1
     -m                INT              INT Mbytes for loading the reads into memory                 1024
                                         (maximum -m INT is 4096)
     -v                FLAG             verbose                                                      off

   advanced <options>: (see SortMeRNA user manual for more details)
    --passes          STRING           values for seed skip lengths for Pass 1, 2 and 3             L,L/2,3
                                         must be in the form 'INT,INT,INT', respectively
                                         (L is the seed length set in ./indexdb)
    --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     FLAG             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             FLAG             add pid to output file names                                 off

   help:
     -h                FLAG             help
     --version        FLAG             SortMeRNA version number
```

The command sortmerna takes as input a list of rRNA databases (must be indexed) and a set of reads (in fasta or fastq format), and filters out the reads matching to at least one of the rRNA databases.

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.

4.2.1 A guide to choosing parameters for filtering and quality of alignments

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
`--feeling-lucky`	Very fast	The first alignment reaching the E-value threshold will be reported (if a high-scoring alignment was found on the forward strand, the reverse-complementary strand will not be searched)
`--num-alignments INT`	Very fast for `INT = 1`	Same behavior as option `--feeling-lucky`
	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

4.2.2 Example 2: sortmerna using multiple databases and the fastest alignment option

```
>> time ./sortmerna --ref ./rRNA_databases/silva-bac-16s-database-id85.fasta,./index/silva-bac-16s:\
./rRNA_databases/silva-bac-23s-database-id98.fasta,./index/silva-bac-23s:\
./rRNA_databases/silva-arc-16s-database-id95.fasta,./index/silva-arc-16s:\
./rRNA_databases/silva-arc-23s-database-id98.fasta,./index/silva-arc-23s:\
./rRNA_databases/silva-euk-18s-database-id95.fasta,./index/silva-euk-18s:\
./rRNA_databases/silva-euk-28s-database-id98.fasta,./index/silva-euk-28s:\
./rRNA_databases/rfam-5.8s-database-id98.fasta,./index/rfam-5.8s:\
./rRNA_databases/rfam-5s-database-id98.fasta,./index/rfam-5s \
--reads SRR106861.fasta --feeling-lucky --sam --fastx --aligned accept --other other --log -v


  Program:     SortMeRNA version 1.99 beta, 11/03/2014
  Copyright:   2012-2014 Bonsai Bioinformatics Research Group
               LIFL, University Lille 1, CNRS UMR 8022, INRIA Nord-Europe
               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 [2.31 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 (OpenMP is not supported with your current C++ compiler).

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

  Begin analysis of: ./rRNA_databases/silva-bac-16s-database-id85.fasta
    Seed length = 18
    Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
    Gumbel lambda = 0.602506
    Gumbel K = 0.328589
    Minimal SW score based on E-value = 53
    Loading index part 1/1 ...  done [3.26 sec]
    Begin index search ...  done [27.78 sec]
    Freeing index ...  done [0.45 sec]

  Begin analysis of: ./rRNA_databases/silva-bac-23s-database-id98.fasta
    Seed length = 18
    Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
    Gumbel lambda = 0.602275
    Gumbel K = 0.333737
    Minimal SW score based on E-value = 53
    Loading index part 1/1 ...  done [2.04 sec]
    Begin index search ...  done [23.04 sec]
    Freeing index ...  done [0.31 sec]
```

```
Begin analysis of: ./rRNA_databases/silva-arc-16s-database-id95.fasta
  Seed length = 18
  Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
  Gumbel lambda = 0.596068
  Gumbel K = 0.321832
  Minimal SW score based on E-value = 52
  Loading index part 1/1 ...  done [1.21 sec]
  Begin index search ...  done [10.90 sec]
  Freeing index ...  done [0.17 sec]

Begin analysis of: ./rRNA_databases/silva-arc-23s-database-id98.fasta
  Seed length = 18
  Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
  Gumbel lambda = 0.596330
  Gumbel K = 0.324091
  Minimal SW score based on E-value = 48
  Loading index part 1/1 ...  done [0.31 sec]
  Begin index search ...  done [8.73 sec]
  Freeing index ...  done [0.06 sec]

Begin analysis of: ./rRNA_databases/silva-euk-18s-database-id95.fasta
  Seed length = 18
  Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
  Gumbel lambda = 0.611988
  Gumbel K = 0.337232
  Minimal SW score based on E-value = 51
  Loading index part 1/1 ...  done [1.76 sec]
  Begin index search ...  done [15.63 sec]
  Freeing index ...  done [0.27 sec]

Begin analysis of: ./rRNA_databases/silva-euk-28s-database-id98.fasta
  Seed length = 18
  Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
  Gumbel lambda = 0.611523
  Gumbel K = 0.335218
  Minimal SW score based on E-value = 53
  Loading index part 1/1 ...  done [2.86 sec]
  Begin index search ...  done [19.54 sec]
  Freeing index ...  done [0.48 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 [0.55 sec]
  Begin index search ...  done [5.71 sec]
  Freeing index ...  done [0.07 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 [1.54 sec]
  Begin index search ...  done [7.62 sec]
  Freeing index ...  done [0.21 sec]
  Total number of reads mapped (incl. all reads file sections searched): 104249
  Writing alignments ...  done [5.14 sec]
  Writing aligned FASTA/FASTQ ...  done [0.93 sec]
```

```
    Writing not-aligned FASTA/FASTQ ...  done [0.08 sec]

real        2m30.574s
user        2m26.740s
sys        0m2.420s
```

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

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

 SortMeRNA command: <command will be here>
 Process pid = 50199
 Parameters summary:
    Index: ./index/silva-bac-16s
     Seed length = 18
     Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
     Gumbel lambda = 0.602506
     Gumbel K = 0.328589
     Minimal SW score based on E-value = 53
    Index: ./index/silva-bac-23s
     Seed length = 18
     Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
     Gumbel lambda = 0.602275
     Gumbel K = 0.333737
     Minimal SW score based on E-value = 53
    Index: ./index/silva-arc-16s
     Seed length = 18
     Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
     Gumbel lambda = 0.596068
     Gumbel K = 0.321832
     Minimal SW score based on E-value = 52
    Index: ./index/silva-arc-23s
     Seed length = 18
     Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
     Gumbel lambda = 0.596330
     Gumbel K = 0.324091
     Minimal SW score based on E-value = 48
    Index: ./index/silva-euk-18s
     Seed length = 18
     Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
     Gumbel lambda = 0.611988
     Gumbel K = 0.337232
     Minimal SW score based on E-value = 51
    Index: ./index/silva-euk-28s
     Seed length = 18
     Pass 1 = 18, Pass 2 = 9, Pass 3 = 3
     Gumbel lambda = 0.611523
     Gumbel K = 0.335218
     Minimal SW score based on E-value = 53
    Index: ./index/rfam-5.8s
     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
    Index: ./index/rfam5s
     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
     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 (OpenMP is not supported with your current C++ compiler).
     Reads file = SRR106861.fasta

 Results:
     Total reads = 113128
 By database:
     aligned reads = 104249 (92.15%)
     non-aligned reads = 8879
     ./rRNA_databases/silva-bac-16s-database-id85.fasta                30.69%
     ./rRNA_databases/silva-bac-23s-database-id98.fasta                55.63%
     ./rRNA_databases/silva-arc-16s-database-id95.fasta                0.26%
     ./rRNA_databases/silva-arc-23s-database-id98.fasta                0.11%
     ./rRNA_databases/silva-euk-18s-database-id95.fasta                0.01%
     ./rRNA_databases/silva-euk-28s-database-id98.fasta                3.14%
     ./rRNA_databases/rfam-5.8s-database-id98.fasta                0.01%
     ./rRNA_databases/rfam-5s-database-id98.fasta                2.31%
```

4.2.3 Filtering paired-ended reads

When outputting matching and non-matching reads into FASTA/Q files, sometimes the situation arises where one of the paired-ended reads matches and the other one doesn't. For users who wish to keep the order of their paired-ended reads, we provide two options:

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

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

And, by default the reads will be split into two `--aligned` and `--other` files.

4.2.4 Example 5: sortmerna on forward-reverse paired-end reads (2 input files)

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
```

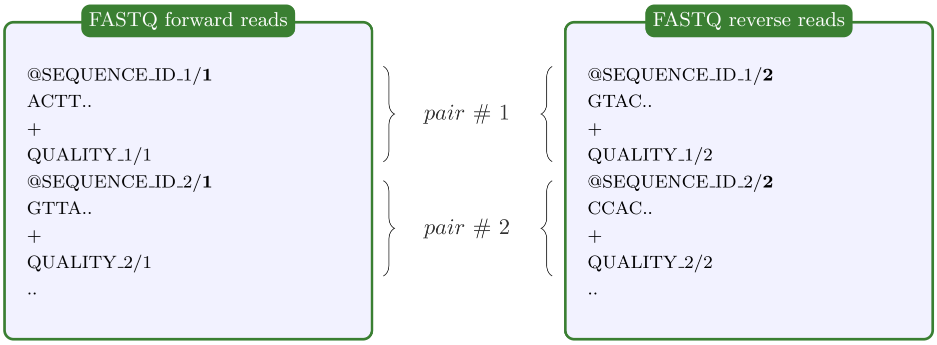


Figure 2: Forward and reverse reads in paired-end sequencing format

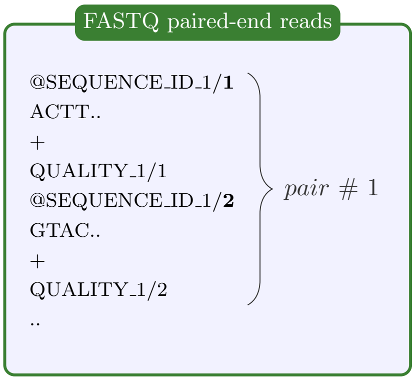


Figure 3: Paired-end read format accepted by SortMeRNA

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.