May 21, 2021

Sequence processing and assembly workflow using CLC workbench, SortMeRNA, and MegaHit.

Helena Pound[1], Steven W Wilhelm[1]

[1]The University of Tennessee, Knoxville

 dx.doi.org/10.17504/protocols.io.buvdnw26

The Aquatic Microbial Ecology Research Group - AMERG (The Buchan, Zinser and Wilhelm labs)
Great Lakes Center for Fresh Waters and Human Health

Helena Pound
University of Tennessee, Knoxville

ABSTRACT

The protocol details one of many methods available to process and assemble sequence data using CLC workbench, SortMeRNA, and MegaHit.

DOI

dx.doi.org/10.17504/protocols.io.buvdnw26

PROTOCOL CITATION

Helena Pound, Steven W Wilhelm 2021. Sequence processing and assembly workflow using CLC workbench, SortMeRNA, and MegaHit. . **protocols.io** https://dx.doi.org/10.17504/protocols.io.buvdnw26

KEYWORDS

null, metatranscriptome, metagenome, microbial ecology, assembly, SortMeRNA, MegaHit

CREATED

May 10, 2021

LAST MODIFIED

May 21, 2021

PROTOCOL INTEGER ID

49797

MATERIALS TEXT

Raw sequence files (fastq), CLC Workbench

SAFETY WARNINGS

Sitting for long periods of time can be hard on one physically. Remember to get up and stretch / move.

BEFORE STARTING

Please download all raw sequence files in fastq format. User will also need to download SortMeRNA version 4 along with its 8 databases and MegaHit version 1.2.9.

1 Upload sequence files to CLC workbench, indicating whether the reads are paired-end or single-end. Choose quality

control parameters.

1.1 We recommend removing failed reads and not demultiplexing.

2 If multiple lanes were run for a single sample and they have not yet been interleaved, now is the best time to create a new sequence list that contains all sequencing data from multiple lanes/runs for a single sample.

3 Trim sequences and remove adapters in CLC.

3.1 We recommend using a quality score of 0.02 (the lower the score, the more stringent) and using ambiguous trimming with a limit of 2. Read length parameters varies based on the length of reads you requested from the sequencer. Adapter removal can either performed automatically or by uploading adapter list provided by sequencer.

4 Trimmed reads should then be exported as fastq files, maintaining 2 files for pair-end reads.

5 Remove any residual rRNA from sequences using SortMeRNA version 4. Note, this is not necessary if you have sequenced DNA. We recommend using all 8 databases provided. See example code below.

5.1

SortMERNA example

```
sortmerna -ref rfam-5.8s-database-id98.fasta -refrfam-5s-database-id98.fasta -ref silva-arc-16s-id95.fasta -ref silva-arc-23s-id98.fasta -ref silva-bac-16s-id90.fasta -ref silva-bac-23s-id98.fasta -ref silva-euk-18s-id95.fasta -ref silva-euk-28s-id98.fasta -reads sample_R1.fastq -reads sample_R2.fastq -workdir sample_folder -fastx -paired_in -other sample_notaligned.fasta
```

SortMeRNA example for paired-end reads

Linux

The example is for paired reads, denoted by the R1 and R2 read files and the -paired_in function. Read files can be compressed with .gz or uncompressed. Note that each run must have a new working directory or it will overwrite existing files. This is particularly important if you are running multiple terminals of SortMeRNA at once. The output file denoted as sample_notaligned.fasta is the remaining interleaved, pair-end "clean" sequences you will use to assemble.

5.2 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.

6 Assemble "clean" sequence files using MegaHit version 1.2.9. The clean files should be named sample_notaligned.fasta.

Citation: Helena Pound, Steven W Wilhelm (05/21/2021). Sequence processing and assembly workflow usingÃÂ CLC workbench, SortMeRNA, and MegaHit.ÃÂ . https://dx.doi.org/10.17504/protocols.io.buvdnw26

6.1

MegaHit Example

megahit -12 sample1_notaligned.fasta, sample2_notaligned.fasta, sample3_notaligned.fasta -o assembly_folder

MegaHit example using interleaved paired-end reads from 3 files after SortMeRNA processing.

Linux

This example uses the interleaved paired-end "clean" sequence files generated by SortMeRNA. Additional assembly options for other sequence types and stringency parameters are available. The final assembly will be labeled final.contigs.fa in the assembly_folder indicated.

6.2 Li, D., Luo, R., Liu, C.M., Leung, C.M., Ting, H.F., Sadakane, K., Yamashita, H. and Lam, T.W. MEGAHIT v1.0: A Fast and Scalable Metagenome Assembler driven by Advanced Methodologies and Community Practices. Methods (2016).

Citation: Helena Pound, Steven W Wilhelm (05/21/2021). Sequence processing and assembly workflow usingÃÂ CLC workbench, SortMeRNA, and MegaHit.ÃÂ . https://dx.doi.org/10.17504/protocols.io.buvdnw26