NCBI SRA database search

Search parameters/queries:

- Virome & viral metagenomics studies
- Related to Homo sapiens or human organism
- RNA-seg strategy only
- Single-end & Paired-end layout
- Illumina & Ion torrent platform

Filter output of initial search

Filtering conditions:

- Number of reads (>=100000 & <=15000000)
- Average read size/length (>=100 & <=500)
- Context-based search of virus-related samples
- Exclusion of DNA and overrepresented viruses in the examined samples

Fetch data for analysis

Requirements:

- SILVA SSU + LSU rRNA database (removed rRNAs with >1% ambiguous N's) (SILVA rRNA ref FASTA)
- RNA-seq data per sample (RNA-seq FASTQ)
- Studied viral genomes per sample (virus ref FASTA)
- QC of RNA-seq data by fastp

rRNA contamination / chimera analysis

Analysis steps:

- 1) BWA align each RNA-seq FASTQ on the SILVA rRNA ref FASTA
- 2) BWA align each RNA-seq FASTQ on the corresponding virus ref FASTA
- 3) Detect rRNA-virus chimeric reads as found in common between the alignments of step 1 & 2
- 4) Calculate rRNA contamination & rRNA-virus chimera statistics

YES

rRNA-virus chimeras present?

NO

Implementation of the compared methods

Compared methods:

- RAW method i.e. the raw unprocessed reads
- BWA method i.e. the unmapped reads after BWA alignment on the SILVA rRNA ref FASTA
- SORTMERNA method i.e. the unmapped reads after SortMeRNA alignment on the SILVA rRNA ref FASTA
- ZWA2 method i.e. the clean reads after ZWA2 mapping and trimming based on the SILVA rRNA ref FASTA

Analysis steps:

- 1) Implement the compared methods separately on each RNA-seq FASTQ
- 2) Measure performance of each method

Virus mapping statistics / analysis

Analysis steps:

- 1) BWA align the treated RNA-seq FASTQ by each compared method on the virus ref FASTA
- Calculate virus mapping / alignment statistics by samtools

De novo assembly statistics / analysis

Analysis steps:

- 1) MEGAHIT *de novo* assembly on the RNA-seq FASTQ files treated by each compared method
- Construction of a local BLASTn database based on the corresponding virus ref FASTA
- 3) BLASTn of the generated MEGAHIT contigs on the local viral database
- Calculate de novo assembly statistics based on BLASTn output report & virus-specific assembly metrics

Legend: Individual file operation Per RNA-seq sample operation Per compared method operation Decision node Process Process notes