RNA-QC-Chain user's manual

Version 1.0

Introduction

RNA-QC-Chain is a comprehensive and fast RNA-Seq data quality control (QC) method. It contains three steps to provide a highly efficient QC solution for RNA-Seq data. Firstly, it assesses and trims low sequencing-quality reads using the software Parallel-QC. Secondly, by aligning read to SILVA database (a comprehensive rRNA database), rRNA reads could be identified and filtered. The extracted rRNA reads are also used to screen of contaminating species (to determine their existence and to identify the species). Finally, multiple alignment metrics are provided to evaluate the quality of the RNA-Seq data by a tool named SAM-stats. Parallel computation was implemented in RNA-QC-Chain, which could significantly accelerate its processing speed.

Download

The latest version can be downloaded at:

http://bioinfo.single-cell.cn/rna-qc-chain.html

Package Dependency

GCC 4.2 or higher

GNUplot

Install

Extract the package:

tar-xzvf rrna-qc-chain.tar.gz

Configure the environment variables:

export RNAQCChain=Path to RNA-QC-Chain

export PATH=$PATH:$RNAQCChain/bin

Build the executable file:

cd RNA-QC-Chain

make

Tools in toolkit

RQC-parallel-qc

The quality control tool for reads quality assessment and trimming

RQC-rRNA-filter

rRNA fragment prediction, filtration and contamination detection

RQC-SAM-stats

Statistics assessment of read alignment results

Usage

RQC-parallel-qc

The **parallel-qc** accepts paired-end or single-end reads in FASTA or FASTQ format.

RQC-parallel-qc [Options] Value

[Options]:

-i Input file name(s) **[Required]**

Input file must be in FASTQ or FASTA format, supporting 1 (single-end sequences, or paired-end sequences in a single file) or 2 (paired-end sequences in two separated files) names.

-Q Quality file name(s)

Must have the same IDs as the input sequences. Supporting 1 (single-end sequences, or paired-end sequences in single file) or 2 (paired-end sequences in separated files, in the same order as the input file names) names. Applicable only when the input is in FASTA format.

-o Output file directory **[Required]**

-f Output format. Q or F

Q for FASTQ and F for FASTA. Default is FASTQ

-p T or F

If the input is paired-end reads. Default is T.

-k T or F

If keep the pairs for output. Default is T.

-d T or F

If drop the duplicated reads. Default is F

-b begin(int) end(int)

Base trim begin and end. Keep 0 0 to turn off. Default is turn off.

-q Quality (int) ratio (float, 0~1), default in Sanger format. Add '-P' if in Phred format.

Trim quality value and threshold ratio that minimum percentage of bases must have the trim quality value. Keep 0 0 to turn off. Default is turn off.

-m Tag sequences file, must be in FASTA format. Maximum sequence number is 100. Default is turn off. The recommended tag sequences are located at Parallel-QC_directory/Default_tag_sequence/primer-combined.fa.

-g GC min_ratio (float, 0~1) max_ratio (float, 0~1).
GC proportion trim, must between min_ratio ~ max_ratio. Default is turn off.

-t Thread number. Default thread number is 1.

-h Print help.

RQC-rRNA-filter

The rRNA-filter accepts paired-end or single-end reads in FASTA or FASTQ format.

RQC-rRNA-filter [Options] Value

[Options]

-i Input file name(s) [Required]
Input file must be in FASTQ or FASTA format, supporting 1 or 2 names.

-Q Quality file name(s)
Available only when the input is in FASTA format, supporting 1 or 2 names.

-o Output directory [Required]

-f Q or F
Output format. Q for FASTQ and F for FASTA. Default is FASTQ

-p T or F
If the input is pair-end reads. Default is T.

-b T or F
If enable the 16S rRNA filter. Default is T.

-u T or F
If enable the 18S rRNA filter. Default is T.

-B T or F
If enable the 23S rRNA filter. Default is T.

-U T or F
If enable the 28S rRNA filter. Default is T.

#Embedded parameters for Parallel-META

-m T or F
If enable the rRNA classification by Parallel-META. Default is T.

-e Float value
Expectation value of rRNA mapping, default is 1e-30.

-n Integer value
To assign the core number of CPU. Default is to automatically detect the system hardware configuration.

-h Print help.

RQC-SAM-stats

The SAM-stats accepts the GTF file and alignment results in SAM format.

RQC-SAM-stats [Option] Value

[Options]

-i or s Input file name [Conflict with -b]
Input file in SAM format.

-b Input file name [Conflict with -i & -s]
Input file in BAM format.

-r Reference file name **[Required]**
Reference file must be in GTF/GFF3 format.

-o Output directory **[Required]**

-p T or F
If input file is paired-end. Default is T

-k Input the statistical key

-h Print help.

Results

RQC-parallel-qc

For parallel-qc, all analysis results will be in the directory assigned by parameter '**-o'**. If the output reads are not kept into pairs, the suffix of '-1' and '-2' will be removed for each pair of output files. In the output directory, files include:

Analysis_report.txt: the overall information of the quality control analysis.
read-1.fq & read-2.fq: the output sequence file after quality control analysis.
trim-qual-1.fq & trim-qual-2.fq: trimmed reads by quality trimming step.
trim-primer-1.fq & trim-primer-2.fq: trimmed reads by tag sequence trimming.
trim-gc-1.fq & trim-gc-2.fq: trimmed reads by GC proportion trimming.
trim-dup-1.fq & trim-dup-2.fq: trimmed reads by duplication trimming.

RQC-rRNA-filter

For rRNA-filter, all analysis results will be in the directory assigned by parameter '**-o**'. In the output directory, files include:

Analysis_report.txt: the overall information of the rRNA filtering.
read.fasta / read.fastq: the purified input sequences data after rRNA filtering.
16S_rRNA.fasta / 18S_rRNA.fasta / 23S_rRNA.fasta / 28S_rRNA.fasta: the extracted rRNA sequences from the input data.

16S_rRNA / 18S rRNA: the classification results of extracted rRNA sequences by Parallel-META software.

RQC-SAM-stats

For SAM-stats, all statistical results will be in the directory assigned by parameter '**-o**'. In the output directory, files include:

Analysis_report.txt: the overall information of the alignment results

Gene_report.txt: the mapping information of the genes in the GTF file.

Mapping_region_distribution.png / txt: the mapping information of statistical keys.

Genebody_coverage_bias.png / txt: the read mapping coverage along the genes from 5' to 3'.

Coverage_distribution.png / txt: the coverage distribution of mapped genes.

Insert_size_distribution.png / txt: the distribution of insertion length.

Notice

1. Please set the environmental variables following the instruction in "Install" section.
2. The output path will be cleared initially, and please ensure that parallel-qc has the write permission of the output path.
3. To enable the 16S rRNA and 18S rRNA classification in "rRNA-filter", please install Parallel-META software and set its environment variables correctly.
4. Make sure the input is in FASTA or FASTAQ format, and select the correct option.
5. Parameter for quality file '-Q' is available only when the input file(s) are in FASTA format.
6. Please assign '-p F' if the input file is not pair-ended.

Contact

Any problems please feel free to contact:

Dr. Kang Ning: ningkang@hust.edu.cn

Dr. Qian Zhou: zhouqian@ysfri.ac.cn

Dr. Xiaoquan Su: suxq@qibebt.ac.cn