COMPASS: a **COM** prehensive **P**latform for sm **A**ll RNA-**S**eq data Analy**S**is (v1.0)

Jiang Li[1], Alvin Kho[2], and Kelan Tantisira[1,3]

[1]The Channing Division of Network Medicine, Department of Medicine, Brigham & Women's Hospital and Harvard Medical School, Boston, MA, USA.
[2]Boston Children's Hospital.
[3]Division of Pulmonary and Critical Care Medicine, Department of Medicine, Brigham and Women's Hospital, and Harvard Medical School, Boston, MA, USA.

November 25, 2018

Contents

1 Introduction

COMPASS was composed of five functional modules: Quality Control, Alignment, Annotation, Microbe and Function. They are integrated into a pipeline and each module can also process independently (Figure 1).

Quality Control To deal with fastq files and filter out the adapter sequences and reads with low quality.

Alignment To align the clean reads to the reference genome.

Annotation To annotate different kinds of circulating RNAs based on the alignment result.

Microbe To predict the possible species of microbes existed in the samples.

Function To perform differential expression analysis and other functional studies to be extended.

2 Installation

2.1 JAVA Virtual Machine

COMPASS was achieved by Java language, so Java Runtime Environment (JRE) version 8 (or up) is required. The JRE can be downloaded in ORACLE website (http://www.oracle.com/technetwork/java/javase/downloads/index.html).

2.2 STAR

COMPASS will take STAR as the default aligner. STAR can be downloaded from Google Code (https://code.google.com/archive/p/rna-star/downloads).

3 Quick Examples

3.1 Run COMPASS

java-jarCOMPASS.jar-inHBRNA_AGTCAA_L001_R1.fastq.gz;S-001570893_CGATGT_L001_R1.fastq.gz-refhg38-qc-raTGGAATTCTCGGGTGCCAAGG-rb4-rh20-rt20-rr20-rlh8;17-aln-mtstar-midx2;3-ann-ac1;2;3;4;5;6-aic-mic-mtoolBlast-mdbarchaea;viruses-fun-fd-fdclass1;2;3;4;5;6-fdcase1;2;1-fdctrl2;2

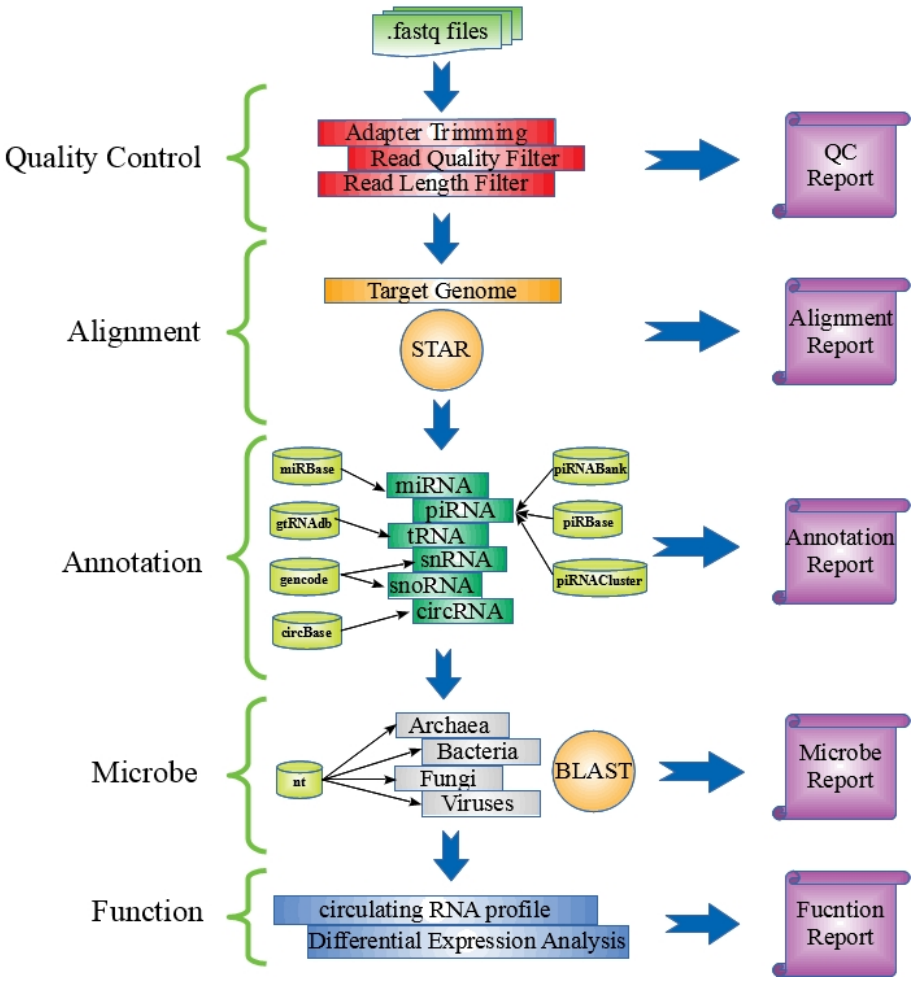

Figure 1: The structure of COMPASS.

4 Options

4.1 General Settings

4.1.1 -h/-help

To display the help information of COMPASS.

4.1.2 -t/-threads *n*

To set the maximum of threads that COMPASS will use when running. The default setting is 1.

4.1.3 -pro/-project_name *ProjectName*

To set the project name. The default setting is COMPASS.

4.1.4 -ref/-ref_genome *hg19/hg38*

To set the reference genome that is used for alignment. Currently, COMPASS supports hg19 (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz) and hg38 (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz) genome version.

4.1.5 -in/-input *file1;file2;...;fileN*

To set the input file. The valid format is fastq file or SAM file.

4.1.6 -inf/-in_file *file.list*

To set the input files through a file list. In the file list, each line should only contain one file without any delimiter.

4.1.7 -out/-output */my/output/path/*

To set the output files. If no setting, COMPASS will create an output directory in the user working path and take the input prefix in default.

4.2 Quality Control

4.2.1 -qc/-quality_control

To open or close the quality control module.

4.2.2 -ra/-rm_adapter *seq*

To remove the adapter sequences at the 3' (3-prime) end. The commonly used adapter sequences from different kits are listed below:

TruSeq Small RNA (Illumina) TGGAATTCTCGGGTGCCAAGG

Small RNA Kits V1 (Illumina) TCGTATGCCGTCTTCTGCTTGT

Small RNA Kits V1.5 (Illumina) ATCTCGTATGCCGTCTTCTGCTTG

NEXTflex Small RNA Sequencing Kit v3 for Illumina Platforms (Bioo Scientific) TGGAATTCTCGGGTGCCAAGG

LEXOGEN Small RNA-Seq Library Prep Kit (Illumina) TGGAATTC TCGGGTGCCAAGGAACTCCAGTCAC

4.2.3 -rb/-rm_bias *n*

To remove *n* random bases in both 5' (5-prime) and 3' (3-prime) ends after removing the adapter sequence.

4.2.4 -rh/-rm_low_quality_head *score*

To remove the low quality bases with the score less than *score* from 5' (5-prime) end.

4.2.5 -rt/-rm_low_quality_tail *score*

To remove the low quality bases with the score less than *score* from 3' (3-prime) end.

4.2.6 -rr/-rm_low_quality_read *score*

To remove the low quality reads with the average score less than *score*.

4.2.7 -rhh/-rm_head_hard *n*

To remove *n* bases from the 5' (5-prime) end.

4.2.8 -rth/-rm_tail_hard *n*

To remove *n* bases from the 3' (3-prime) end.

4.2.9 -rlh/-rm_read_hard *D1;D2;...;Dn*

To divide the reads into several groups according to *[0,D1),[D1,D2),...,[Dn-1,Dn]*.

4.3 Alignment

4.3.1 -aln/-alignment

To open or close the alignment module.

4.3.2 -mt/-mapping_tool *star/bowtie/bowtie2*

To set the aligner used in COMPASS. The default aligner is *star*

4.3.3 -mp/-mapping_param

To set parameters of the aligner. The default settings for *star/bowtie/bowtie2* are listed below:

- *star*

  –**runMode** alignReads

  –**outSAMtype** SAM

  –**outSAMattributes** Standard

  –**readFilesCommand** zcat

  –**outSAMunmapped** Within

  –**outReadsUnmapped** None

  –**alignEndsType** EndToEnd

  –**alignIntroMax** 1

  –**alignIntroMin** 21

  –**outFilterMismatchNmax** 1

  –**outFilterMultimapScoreRange** 0

  –**outFilterScoreMinOverLread** 0

  –**outFilterMatchNminOverLread** 0

  –**outFilterMismatchNoverLmax** 0.3

  –**outFilterMatchNmin** 16

  –**outFilterMultimapNmax** 20

- *bowtie*

- *bowtie2*

4.3.4 -midx/-mapping_index *R1;R2;...;Rn*

To set the read group that will be used for alignment. The default value is "last" , which means the group with the longest reads. Otherwise, the number *Rn* denotes the index of region when setting the parameter -rlh/-rm_read_hard *D1;D2;...;Dn*.

4.3.5 -mref/-mapping_reference *hg19/hg38*

To set the reference genome in alignment. The default value is the same as the parameter -ref/-ref_genome *hg19/hg38*.

4.4 Annotation

4.4.1 -ann/-annotation

To open/close the annotation module.

4.4.2 -ac/-ann_class *A1;A2;...;An*

To set the small RNA categories that will be annotated. The index of small RNA is listed:

1 miRNA

2 piRNA

3 tRNA

4 snoRNA

5 snRNA

6 circRNA

4.4.3 -aol/-ann_overlap *n*

To set the overlap rate between reads and gene regions. The default value is 1.0.

4.4.4 -aic/-ann_inCluster

To show whether or not piRNAs are in the piRNA clusters when annotating piRNAs. The default value is false.

4.4.5 -atd/-ann_threshold *n*

To set the threshold of read counts of small RNAs. If set, only the small RNAs with the read count more than *n* are displayed. The default value is 1.

4.4.6 -armsm/-ann_remove_sam

If added, the original sam file from alignment module will be removed.

4.5 Microbe

4.5.1 -mic/-microbe

To open/close the microbe module.

4.5.2 -mtool/-mic_tool *blast*

To set the tool that will be used for microbe profiling. Currently, only *blast* is supported.

4.5.3 -mdb/mic_database *viruses;bacteria;fungi;archaea*

To set the microbial databases used in blast.

4.6 Function

4.6.1 -fun/-function

To open/close the function module.

4.6.2 -fd/-fun_diff_expr

To open/close the function of differential expression analysis.

4.6.3 -fdclass/-fun_diff_class *A1;A2;...;An*

To set the small RNAs that will be performed the differential expression analysis. The format is the same as the parameter -ac/-ann_class *A1;A2;...;An*.

4.6.4 -fdcase/-fun_diff_case *ID1;ID2;...;IDn*

To set the IDs of case samples.

4.6.5 -fdctrl/-fun_diff_control *ID1;ID2;...;IDn*

To set the IDs of control samples.

4.6.6 -fdtest/-fun_diff_test *mwu*

To set the statistic test between case and control samples. Currently, only Mann-Whitney U test is supported.

4.6.7 -fdmic/-fun_diff_mic

If added, COMPASS will detect the annotation files of microbes. It is valid when running function module separately.

4.6.8 -fmtool/-fun_mtool *blast*

To set the tool that was used for microbe profiling. This parameter can facilitate COMPASS to decide the input files.

4.6.9 -fmdb/-fun_mdb *viruses;bacteria;fungi;archaea*

To set the microbial databases used in blast. This parameter can facilitate COMPASS to decide the input files.

4.6.10 -fdann/-fun_diff_ann

If added, COMPASS will detect the annotation files of all small RNAs. It is valid when running function module separately.

4.6.11 -fm/-fun_merge

To open/close the function of merging.

4.6.12 -fms/-fun_merge_samples *ID1;ID2;...;IDn*

To extract read counts from each sample and merge them in one file by different kinds of small RNAs. The categories are set by the parameter -fdclass/-fun_diff_class *A1;A2;...;An*.

5 FAQ

5.0.1 How much memory does COMPASS need?

COMPASS does not cost lots of memory, but if STAR was taken as aligner, and 30G memory is considered at least for human genome.