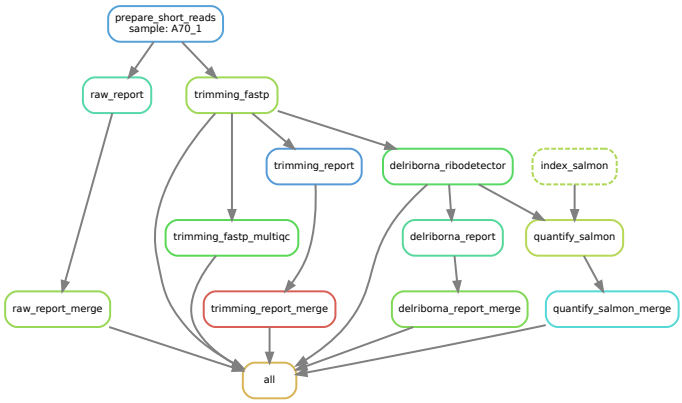
Transcriptome analysis pipeline from RNA-seq data



Feature

- Transcript quantification using STAR or salmon
- CDR3 sequence assembly using TRUST4
- HLA-typing using arcasHLA

Installation

```
➤ git clone https://github.com/ohmeta/rnapi
➤ echo "export PYTHONPATH=/path/to/rnapi:$PYTHONPATH" >> ~/.bashrc
# relogin
```

Run

Overview

```
➤ python toolkit/rnapi/run_rnapi.py --help

usage: rnapi [-h] [-v]  ...

Omics for All, Open Source for All

  RNA sequence analysis pipeline

options:
```

```
  -h, --help     show this help message and exit
  -v, --version  print software version and exit

available subcommands:

    init         init project
    rnaseq_wf    RNA seq analysis pipeline
    scrnaseq_wf  scRNA seq analysis pipeline
```

rnaseq_wf

```
➤ python toolkit/rnapi/run_rnapi.py rnaseq_wf --help
usage: rnapi rnaseq_wf [-h] [-d WORKDIR] [--check-samples] [--config
CONFIG] [--profile PROFILE] [--cores CORES] [--local-cores LOCAL_CORES] [-
-jobs JOBS] [--list] [--debug] [--dry-run] [--run-local] [--run-remote]
                       [--cluster-engine {slurm,sge,lsf,pbs-torque}] [--
wait WAIT] [--use-conda] [--conda-prefix CONDA_PREFIX] [--conda-create-
envs-only]
                       [TASK]

positional arguments:
  TASK                  pipeline end point. Allowed values are
prepare_short_reads_all, raw_fastqc_all, raw_report_all, raw_all,
trimming_fastp_all, trimming_report_all, trimming_all,
delriborna_ribodetector_all, delriborna_report_all, delriborna_all,
align_reads_star_all, align_genome_star_all, align_transcriptome_star_all,
align_star_all, align_hisat2_all, align_all, quantify_gene_star_all,
quantify_transcript_star_all, quantify_all, pseudo_align_salmon_all,
pseudo_align_kallisto_all, quantification_salmon_all,
quantification_sleuth_all, assembly_xcr_trust4_all, assembly_all,
hlatyping_arcashla_all, hlatyping_all, all (default: all)

optional arguments:
  -h, --help            show this help message and exit
  -d, --workdir WORKDIR
                        project workdir (default: ./)
  --check-samples       check samples, default: False
  --config CONFIG       config.yaml (default: ./config.yaml)
  --profile PROFILE     cluster profile name (default: ./profiles/slurm)
  --cores CORES         all job cores, available on '--run-local'
(default: 32)
  --local-cores LOCAL_CORES
                        local job cores, available on '--run-remote'
(default: 8)
  --jobs JOBS           cluster job numbers, available on '--run-remote'
(default: 80)
  --list                list pipeline rules
  --debug               debug pipeline
  --dry-run             dry run pipeline
  --run-local           run pipeline on local computer
  --run-remote          run pipeline on remote cluster
```

```
  --cluster-engine {slurm,sge,lsf,pbs-torque}
                        cluster workflow manager engine, support
slurm(sbatch) and sge(qsub) (default: slurm)
  --wait WAIT           wait given seconds (default: 60)
  --use-conda           use conda environment
  --conda-prefix CONDA_PREFIX
                        conda environment prefix (default: ~/.conda/envs)
  --conda-create-envs-only
                        conda create environments only
```

Real world

Step 1: Prepare samples.tsv like below format

id	fq1	fq2
s1	s1.1.fq.gz	s1.2.fq.gz
s2	s2.1.fq.gz	s2.2.fq.gz
s3	s3.1.fq.gz	s3.2.fq.gz

Step 2: Init

```
➤ mkdir -p rnapi_test
➤ cd rnapi_test
➤ python /path/to/rnapi/run_rnapi.py init -d . -s samples.tsv
```

Step 4: Update config

```
# edit config.yaml
# for example

➤ cat config.yaml
reference:
  # dna:
/home/jiezhu/databases/ensembl/release_104/fasta/mus_musculus/dna/Mus_musc
ulus.GRCm39.dna.primary_assembly.fa.gz
  # dna:
/home/jiezhu/databases/ensembl/release_104/fasta/homo_sapiens/dna/Homo_sap
iens.GRCh38.dna.primary_assembly.fa.gz
  # dna:
/home/jiezhu/databases/ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/r
elease_M28/GRCm39.primary_assembly.genome.fa.gz
  dna:
/home/jiezhu/databases/ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/r
elease_39/GRCh38.primary_assembly.genome.fa.gz
```

```
  # cdna:
/home/jiezhu/databases/ensembl/release_104/fasta/mus_musculus/cdna/Mus_mus
culus.GRCm39.cdna.all.fa.gz
  # cdna:
/home/jiezhu/databases/ensembl/release_104/fasta/homo_sapiens/cdna/Homo_sa
piens.GRCh38.cdna.all.fa.gz
  # cdna:
/home/jiezhu/databases/ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/r
elease_M28/gencode.vM28.transcripts.fa.gz
  cdna:
/home/jiezhu/databases/ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/r
elease_39/gencode.v39.transcripts.fa.gz

  # gtf:
/home/jiezhu/databases/ensembl/release_104/gtf/mus_musculus/Mus_musculus.G
RCm39.104.gtf
  # gtf:
/home/jiezhu/databases/ensembl/release_104/gtf/homo_sapiens/Homo_sapiens.G
RCh38.104.gtf
  # gtf:
/home/jiezhu/databases/ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/r
elease_M28/gencode.vM28.primary_assembly.annotation.gtf
  gtf:
/home/jiezhu/databases/ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/r
elease_39/gencode.v39.primary_assembly.annotation.gtf

  # index_rsem:
/home/jiezhu/databases/ensembl/release_104/fasta/mus_musculus/dna_index/in
dex_rsem/mus_musculus
  # index_rsem:
/home/jiezhu/databases/ensembl/release_104/fasta/homo_sapiens/dna_index/in
dex_rsem/homo_sapiens
  # index_rsem:
/home/jiezhu/databases/ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/r
elease_M28/index_rsem/mus_musculus
  index_rsem:
/home/jiezhu/databases/ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/r
elease_39/index_rsem/homo_sapiens

  # index_star:
/home/jiezhu/databases/ensembl/release_104/fasta/mus_musculus/dna_index/in
dex_star
  # index_star:
/home/jiezhu/databases/ensembl/release_104/fasta/homo_sapiens/dna_index/in
dex_star
  # index_star:
/home/jiezhu/databases/ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/r
elease_M28/index_star
  index_star:
/home/jiezhu/databases/ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/r
elease_39/index_star

  # index_salmon:
/home/jiezhu/databases/ensembl/release_104/fasta/mus_musculus/cdna_index/i
```

```
ndex_salmon
  # index_salmon:
/home/jiezhu/databases/ensembl/release_104/fasta/homo_sapiens/cdna_index/i
ndex_salmon
  # index_salmon:
/home/jiezhu/databases/ftp.ebi.ac.uk/pub/databases/gencode/Gencode_mouse/r
elease_M28/index_salmon
  index_salmon:
/home/jiezhu/databases/ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/r
elease_39/index_salmon


params:
  samples: samples.tsv
  fq_encoding: sanger   # fastq quality encoding. available values:
'sanger', 'solexa', 'illumina-1.3+', 'illumina-1.5+', 'illumina-1.8+'.
(default "sanger")
  reads_layout: pe
  interleaved: false
  strandedness: reverse   # "", "forward", "reverse"

  raw:
    threads: 8
    save_reads: true
    fastqc:
      do: false

  trimming:
    save_reads: true
    fastp:
      do: true
      threads: 4
      use_slide_window: false # strict when using slide window
      disable_adapter_trimming: false
      detect_adapter_for_se: true # If activated, adapter_sequence will
not used
      detect_adapter_for_pe: true # If activated, adapter_sequence and
adapter_sequence_r2 will not used
      adapter_sequence: AAGTCGGAGGCCAAGCGGTCTTAGGAAGACAA   # MGI adapter 3
      adapter_sequence_r2: AAGTCGGATCGTAGCCATGTCGTTCTGTGAGCCAAGGAGTTG   #
MGI adapter 5
      # "AGATCGGAAGAGCACACGTCTGAACTCCAGTCA"         # eg: Illumina TruSeq
adapter 3
      # "AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT"         # eg: Illumina TruSeq
adapter 5
      compression: 6
      cut_front_window_size: 4
      cut_front_mean_quality: 20
      cut_tail_window_size: 4
      cut_tail_mean_quality: 20
      cut_right_window_size: 4
      cut_right_mean_quality: 20
      length_required: 51
      n_base_limit: 5
```

```
      dedup: false
      dup_calc_accuracy: 3 # [1, 2, 3, 4, 5, 6] # only used when dedup:
True

  delriborna:
    threads: 8
    ribodetector:
      do: true
      GPU: true
      reads_len: 100
      chunk_size: 256 # control memory usage when using CPU
      extra: --memory 12   # only work for GPU

  qcreport:
    do: true
    seqkit:
      threads: 4

  align:
    threads: 8
    star:
      do: true
      sjdboverhang: 99 # reads_len - 1
      quant_mode:
        TranscriptomeSAM: true
        # output SAM/BAM alignments to transcriptome into a separate file
        GeneCounts: false
        # count reads per gene

  quantify:
    threads: 8
    salmon:
      do: true
      index_add_genome: true
      kmer_len: 31
      lib_type: A   # To allow Salmon to automatically infer the library
type, simply provide -l A or --libType A to Salmon
      extra: --gcBias

  assembly:
    threads: 8
    trust4:
      do: true
      coordinate_fasta:
/home/jiezhu/databases/funcgenomics/IMGT/TRUST4/Homo_sapien/human_IMGT+C.f
a
      reference_fasta:
/home/jiezhu/databases/funcgenomics/IMGT/TRUST4/Homo_sapien/human_IMGT+C.f
a

  hlatyping:
    threads: 8
    arcashla:
      do: true
```

```
      IMGTHLA_version: latest # 3.46.0 # latest
      unmapped: false
      genes: [DPB1, DRB1, DRA, L, K, B, DOB, DRB3, DMA, G, DMB, C, DQA1,
DOA, F, E,
        DPA1, DRB5, DQB1, H, A, J]
      # genes: ["A", "B", "C", "DPB1", "DQA1", "DQB1", "DRB1"]
      # population: ["prior", "native_american", "asian_pacific_islander",
"caucasian", "black", "hispanic"]

output:
  raw: results/00.raw
  trimming: results/01.trimming
  delriborna: results/02.delriborna
  qcreport: results/02.qcreport
  align: results/03.align
  quantify: results/04.quantify
  assembly: results/05.assembly
  hlatyping: results/06.hlatyping


envs:
  fastp: /home/jiezhu/toolkit/rnapi/test/envs/fastp.yaml
  multiqc: /home/jiezhu/toolkit/rnapi/test/envs/multiqc.yaml
  delriborna: /home/jiezhu/toolkit/rnapi/test/envs/delriborna.yaml
  align: /home/jiezhu/toolkit/rnapi/test/envs/align.yaml
  trust4: /home/jiezhu/toolkit/rnapi/test/envs/trust4.yaml
  arcashla: /home/jiezhu/toolkit/rnapi/test/envs/arcashla.yaml
```

Step 5: dry-run rnaseq_wf

```
➤ python /path/to/rnapi/run_rnapi.py rnaseq_wf all --dry-run

........

Job stats:
job                             count    min threads    max threads
------------------------------  -------  -------------  -------------
align_reads_star                    6              8              8
align_transcriptome_star            6              8              8
all                                 1              1              1
assembly_xcr_trust4                 6              8              8
delriborna_report                   6              4              4
delriborna_report_merge             1              4              4
delriborna_ribodetector             6              8              8
hlatyping_arcashla_extract          6              8              8
hlatyping_arcashla_genotype         6              8              8
hlatyping_arcashla_reference        1              1              1
index_rsem                          1              8              8
prepare_short_reads                 6              8              8
quantify_salmon                     6              8              8
quantify_salmon_merge               1              8              8
```

```
quantify_transcript_star              6               8               8
quantify_transcript_star_merge        1               8               8
raw_report                            6               4               4
raw_report_merge                      1               4               4
trimming_fastp                        6               4               4
trimming_fastp_multiqc                1               1               1
trimming_report                       6               4               4
trimming_report_merge                 1               4               4
total                                87               1               8

Reasons:
    (check individual jobs above for details)
    input files updated by another job:
        align_reads_star, align_transcriptome_star, all,
assembly_xcr_trust4, delriborna_report, delriborna_report_merge,
delriborna_ribodetector, hlatyping_arcashla_extract,
hlatyping_arcashla_genotype, quantify_salmon, quantify_salmon_merge,
quantify_transcript_star, quantify_transcript_star_merge, raw_report,
raw_report_merge, trimming_fastp, trimming_fastp_multiqc, trimming_report,
trimming_report_merge
    missing output files:
        align_reads_star, align_transcriptome_star, assembly_xcr_trust4,
delriborna_report, delriborna_report_merge, delriborna_ribodetector,
hlatyping_arcashla_extract, hlatyping_arcashla_genotype,
hlatyping_arcashla_reference, index_rsem, prepare_short_reads,
quantify_salmon, quantify_salmon_merge, quantify_transcript_star,
quantify_transcript_star_merge, raw_report, raw_report_merge,
trimming_fastp, trimming_fastp_multiqc, trimming_report,
trimming_report_merge

This was a dry-run (flag -n). The order of jobs does not reflect the order
of execution.

Real running cmd:
snakemake --snakefile
/home/jiezhu/toolkit/rnapi/rnapi/snakefiles/rnaseq_wf.smk --configfile
./config.yaml --cores 32 --until all --rerun-incomplete --keep-going --
printshellcmds --reason --dry-run
```

Step 6: run rnaseq_wf local or remote

```
➤ python /path/to/rnapi/run_rnapi.py rnaseq_wf all \
  --run-local \
  --use-conda \
  --local-cores 42 \
  --jobs 5

# or
➤ python /path/to/rnapi/run_rnapi.py rnaseq_wf all \
  --run-remote \
  --use-conda \
```

```
    --local-cores 8 \
    --cores 320 \
    --jobs 40
```

Note

- If you run rnapi at SLURM/SGE system, you may need to edit *profiles/slurm/cluster.yaml* or *profiles/sge/cluster.yaml* at your working folder to update the resources requirement
- rnapi reply snakemake to use conda/mamba to create enviroments automatically, so basically you only need snakemake installed at your working environment, then when run pipeline, just specific **--use-conda** parameter, then the softwares required by rnapi will be installed by conda/mamba accorading to the *envs/*.yaml* files. If you want to used different software version, just edit *envs/*.yaml* and update it