Reproducible Analysis of RNA-Seq Data – SRP033351

Mengyuan Kan

July 30, 2018

Informatics Tools

Raw reads QC

trimmomatic

- usage: trim raw reads
- version: trimmomatic-0.32
- location: /opt/software/Trimmomatic/0.32/trimmomatic-0.32.jar

fastqc

Use a local version because we will use a custom adapter and primer file for contaminant_list.txt

- usage: report reads quality
- version: FastQC v0.11.27
- location: /home/mengykan/.local/bin/
- local installation:

```bash
cd ~/softwares
wget https://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.11.7.zip
unzip fastqc_v0.11.7.zip
cd FastQC/
chmod 775 fastqc
ln -s /home/mengykan/softwares/FastQC/fastqc /home/mengykan/.local/bin/
```

Include frequently used Illumina index infromation in a modified Configuration file.

Align and mapping

STAR

- usage: align and mapping RNA-Seq reads
- version: STAR_2.5.2b, comment out module load STAR-2.5.2a in .bashrc
- location: /project/bhimeslab/STAR-2.5.2b/bin/Linux_x86_64/STAR. Create symbolic link to local bin

samtools

- usage: sort, index, statistics
- version: 0.1.19-44428cd
- location: /opt/software/samtools/samtools-0.1.19/samtools

Bam File QC Metrics

bamtools

- usage: manipulate bam files
- version: 2.3.0
- location: /opt/software/bamtools/2.3.0/bin/bamtools. Add bamtools library path /opt/software/bamtools/2.3.0/lib to LD_LIBRARY_PATH in .bashrc if it is not there, otherwise will get the error *error while loading shared libraries: libbamtools.so.2.3.0: cannot open shared object file: No such file or directory*

picard

- usage: stats of RNA metrics and insertsize metrics in RNA-Seq data
- version: picard-tools-1.96
- location: /opt/software/picard/picard-tools-1.96

Gene and transcript-level quantification

HTseq

- usage: count mapping reads for DESeq2 DE identification
- version: HTSeq-0.6.1
- location: /home/mengykan/.local/bin/htseq-count
- local installation:

BASH

```
cd ~/softwares
wget https://pypi.python.org/packages/72/0f/566afae6c149762af301a19686cd5fd1876deb2b48d09546dbd5caebbb78/HTSeq-0.6.1.tar.gz#md5=b7f4\
f38a9f4278b9b7f948d1efbc1f05
python setup.py build
python setup.py install --install-platlib=$HOME/.local/python_lib --install-scripts=$HOME/.local/bin
```

kallisto

- kallisto[1] usage: pseudoalign and quantify transcript
- version: 0.44.0
- location: /project/bhimeslab/kallisto_linux-v0.42.3/kallisto (symbolic to ~/.local/bin/kallisto)

BASH

```
cd /project/bhimeslab/softwares
wget https://github.com/pachterlab/kallisto/releases/download/v0.44.0/kallisto_linux-v0.44.0.tar.gz
tar zxvf kallisto_linux-v0.44.0.tar.gz
ln -s /project/bhimeslab/softwares/kallisto_linux-v0.44.0/kallisto /home/mengykan/.local/bin/kallisto
```

DE analysis

sleuth

- sleuth R package usage: identify DE and visualize results
- version: 0.30.0, dependency rhdf5 2.22.0
- location: /home/mengykan/.local/R-3.4/libs

local installation

1. set R local enviroment

add R enviromental path in .bashrc

```
export R_LIBS=/home/mengykan/.local/R-3.4/libs
```

2. install rhdf5

R

```
source("http://bioconductor.org/biocLite.R")
biocLite("rhdf5")
```

3. install devtools

R

```
install.packages("devtools")
```

4. install sleuth

R

```
library(devtools)
withr::with_libpaths(new="~/.local/R/libs", install_github('pachterlab/sleuth'))
```

DESeq2

- DESeq2 R Package [2] usage: identify DE based on the negative binomial distribution and visualize results
- version: 1.10.1
- location: $HOME/.local/R/libs

1. install r package RcppArmadillo [optional]

```
wget https://cran.r-project.org/src/contrib/Archive/RcppArmadillo/RcppArmadillo_0.5.600.2.0.tar.gz
R CMD INSTALL -l ~/.local/R/libs RcppArmadillo_0.5.600.2.0.tar.gz
```

> **Note**
> Run step 2 first to see if this package has been installed. Fails to compile the latest version 0.7.500 using R command `install.packages(`*RcppArmadillo*`,lib=`∼/.*local/R/libs*`)`. It gives the error *Only g++ version 4.6 or greater can be used with RcppArmadillo. error: Please use a different compiler*. Use older version *0.5.600.2.0*.

2. local installation

```
source("http://bioconductor.org/biocLite.R")
biocLite(pkgs="DESeq2",lib.loc="~/.local/R/libs",lib="~/.local/R/libs")
```

Visualization

R packages

- genefilter
- gplots: heatmap2 plots

Reference Genome

Human genome reference files

- Human reference: /project/bhimeslab/Reference/
  - indexed reference genome with ERCC spike-in: hg38/genome.ERCC.fa and hg38/genome.ERCC.fa.fai
  - indexed transcripts for kallisto: hg38/hg38_new.idx
  - known gene/transcript annotations: hg38/genes.gtf (human genes), ERCC92.gtf (ERCC spike-in), hg38/genes.ERCC.g (human genes with ERCC spike-in)
  - rRNA annotations: hg38/rRNA_hg38.gtf
  - refFlat format [3] position file used through Picard command to generate RNA-Seq metrics: hg38/refFlat.txt

STAR index creation

Create reference genome index for STAR if it does not exist. STAR tutorial [4] recommended to remove all files from the genome directory before running the genome generation step. Create a new directory STAR_index under previous reference folder.

```bash
mkdir /project/bhimeslab/Reference/hg38/STAR_index
cd mkdir /project/bhimeslab/Reference/hg38/STAR_index
STAR --runThreadN 12 --runMode genomeGenerate --genomeDir /project/bhimeslab/Reference/hg38/STAR_index --genomeFastaFiles /project/b\
himeslab/Reference/hg38/genome.ERCC.fa --sjdbGTFfile /project/bhimeslab/Reference/hg38/genes.ERCC.gtf
```

- 15 files generated: chrLength.txt, chrNameLength.txt, chrName.txt, chrStart.txt, exonGeTrInfo.tab, exonInfo.tab, geneInfo.tab, Genome, genomeParameters.txt, SA, SAindex, sjdbInfo.txt, sjdbList.fromGTF.out.tab, sjdbList.out.tab, transcriptInfo.tab
- `--sjdbOverhang` defalt is 100

RNA-Seq Pipeline

SRA download and fastqc

```bash
rnaseq_sra_download.py --geo_id GSE52778 --path_start /home/mengykan/Projects/SRP033351 --project_name SRP033351 --template_dir /hom\
e/mengykan/Projects/shared_files/RNASeq --fastqc
for i in *_download.lsf; do bsub < $i; done
```

Note

> Check the error (.screen) files to see if the ftp address is available. For example, SRR1039513_3.fastq.gz is in sqlite database but not in the ftp ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR103/003/SRR1039513

User tailored phenotype file preparation

```
The sample info file used in the following steps should be provided by users.
```

Required columns:

- ’Sample’ column containing sample ID
- ’Status’ column containing variables of comparison state
- ’R1’ and/or ’R2’ columns containing full paths of .fastq files

Other columns:

’Treatment’, ’Disease’, ’Donor’ (i.e. cell line ID if <i>in vitro</i> treatment is used), ’Tissue’, ’ERCC_Mix’ (i.e. ERCC mix ID if ERCC spike-in sample is used), ’protocol’ designating sample preparation kit information.

’Index’ column:

contains index sequence for each sample. If provided, trim raw .fastq files based on corresponding adapter sequences.

If use data from GEO, most GEO phenotype data do not have index information. However, FastQC is able to detect them as *Overrepresented sequences*. Users can tailor the ’Index’ column based on FastQC results. We provide a file with most updated adapter and primer sequences for FastQC detection.

An example phenotype file can be found here: example_files/sample_info_file.txt.

Note

> Column naming is rigid for the following columns: ’Sample’, ’Status’, ’Index’, ’R1’, ’R2’, ’ERCC_Mix’, ’Treatment’, ’Disease’, ’Donor’, because pipeline scripts will recognize these name strings, but the column order can be changed.

Align and QC

alignment and quantification

Run rnaseq_align_and_qc.py. User-defined variable script.

Output files in

BASH

```bash
rnaseq_align_and_qc.py --project_name SRP033351 --samples_in /home/mengykan/Projects/SRP033351/files/SRP033351_Info_Sheet_index.txt \
--aligner star --ref_genome hg38 --library_type PE --index_type truseq_single_index --strand nonstrand --path_start /home/mengykan/P\
rojects/SRP033351 --template_dir /home/mengykan/Projects/shared_files/RNASeq
for i in *_align.lsf; do bsub < $i; done
```

QC report

BASH

```bash
rnaseq_align_and_qc_report.py --project_name SRP033351 --samples_in /home/mengykan/Projects/SRP033351/files/SRP033351_Info_Sheet_ind\
ex.txt --aligner star --ref_genome hg38 --library_type PE --path_start /home/mengykan/Projects/SRP033351 --template_dir /home/mengyk\
an/Projects/shared_files/RNASeq
bsub < SRP033351_qc.lsf
```

Differential expression analysis

DESeq2

BASH

```bash
rnaseq_de_report.py --project_name SRP033351 --samples_in /home/mengykan/Projects/SRP033351/files/SRP033351_Info_Sheet_index_withQC.\
txt --comp /home/mengykan/Projects/SRP033351/files/SRP033351_comp_file.txt --de_package deseq2 --ref_genome hg38 --path_start /home/\
mengykan/Projects/SRP033351 --template_dir /home/mengykan/Projects/shared_files/RNASeq
bsub < SRP033351_deseq2.lsf
```

References

1. [kallisto]
   https://pachterlab.github.io/kallisto/starting

2. [DESeq2 R Package]
   https://www.bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf

3. [refFlat format]
   https://genome.ucsc.edu/FAQ/FAQformat.html

4. [STAR tutorial]
   https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf