A quick and simple RNA–seq preprocessing

Huanglab

Overview

1. read alignment using STAR
2. prepare count table using GenomicAlignments
3. rpkm/cpms using edgeR and log2
4. Add annotation based on org.Mm.eg.db(EntrezID, Gene Symbol)
5. Calling differential expression using DEseq

Packages and User–guides

```
$ module load anaconda/201903
$ module load star/2.5.2b
$ module load python/2.7.15
```

```
#load the script from the internet that is used in install bioconductor;
#these packages have been installed.
#source("http://bioconductor.org/biocLite.R")
#biocLite("GenomicRanges")
#biocLite("GenomicFeatures")
#biocLite("Rsamtools")
#biocLite("GenomicAlignments")
#biocLite("BiocParallel")
#biocLite("DESeq")
#biocLite("edgeR")
#biocLite("org.Mm.eg.db")
```

Package guides could be found **/cluster/huanglab/Trainning/RNA–seq/package_guide** or search on the http://www.bioconductor.org

main codes

Code could be downloaded from：**/cluster/huanglab/Trainning/RNA–seq/Code/**

my_STAR_alignReads_bat.py

Set Paths and Parameters

```
STAR='/cluster/apps/star/2.5.2b/STAR '
genomeDir = '/cluster/database/Index_old/STAR-Index/mm10/'
path_in = 'Fastq/'
path_out = 'STAR/'
file_ins=sorted(glob.glob(path_in+'.fastq.gz'))
# For fq.gz:file_ins=sorted(glob.glob(path_in+'*.fq.gz'))

## parameters
runMode = 'alignReads'
runThreadN = '8'
outSAMtype = 'BAM SortedByCoordinate'
readFilesCommand = 'zcat'
```

- ## my_RNAseq_pipelines_Rsamtools.R

Set Paths and Parameters

```
genome = 'mm10'  # mm9 mm10 hg18 hg19
setwd("/cluster/huanglab/metong/RNAseq/")

file_exbygene =
paste("/cluster/huanglab/metong/RNAseq/",genome,"_refGene_exbygene.RData",sep="")

file_txbygene =
paste("/cluster/huanglab/metong/RNAseq/",genome,"_refGene_txbygene.RData",sep="")

path_in = 'STAR/Bam/' #Path for Bam file
path_out = 'DESeq/'
sampleInfo_file = 'sampleInfo.csv' #Label for samples
contrasts = c('Primitive-definitive')        # change it to conditions for comparisons

se = summarizeOverlaps(features=bygene, reads=bamfiles,
                        mode="Union",
                        singleEnd=FALSE,
                        ignore.strand=TRUE,
                        fragments=TRUE ) ## For singleEnd, singleEnd=TRUE

#singleEnd (Default TRUE) A logical indicating if reads are single or paired-end. In
Bioconductor > 2.12 it is not necessary to sort paired-end BAM files by qname. When counting
with summarizeOverlaps, setting singleEnd=FALSE will trigger paired-end reading and counting.
It is fine to also set asMates=TRUE in the BamFile but is not necessary when singleEnd=FALSE.
   More details find GenomicAlignments.pdf
pvalue_cutoff<-1.0E-300 #set pvalue_cutoff
```

Demo data

Step 0. load packages

```
$ module load anaconda/201903
$ module load python/2.7.15
$ module load star/2.5.2b
```

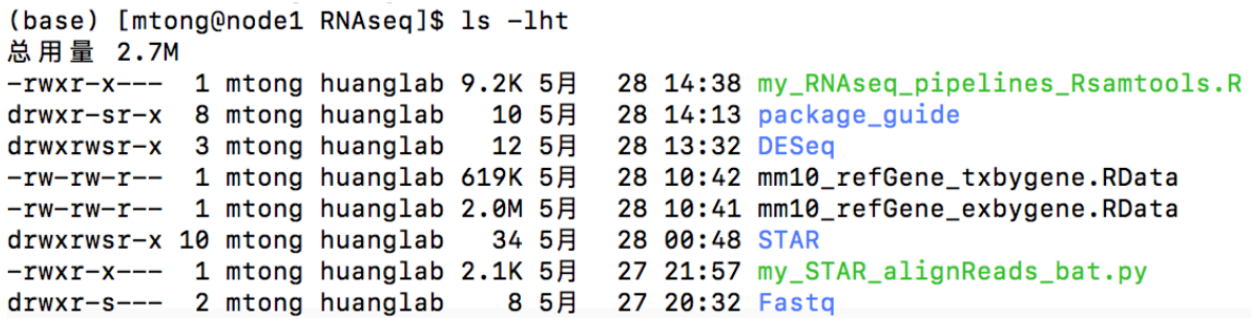
Step 1. Check input files and folders

demo Input files could be downloaded from:

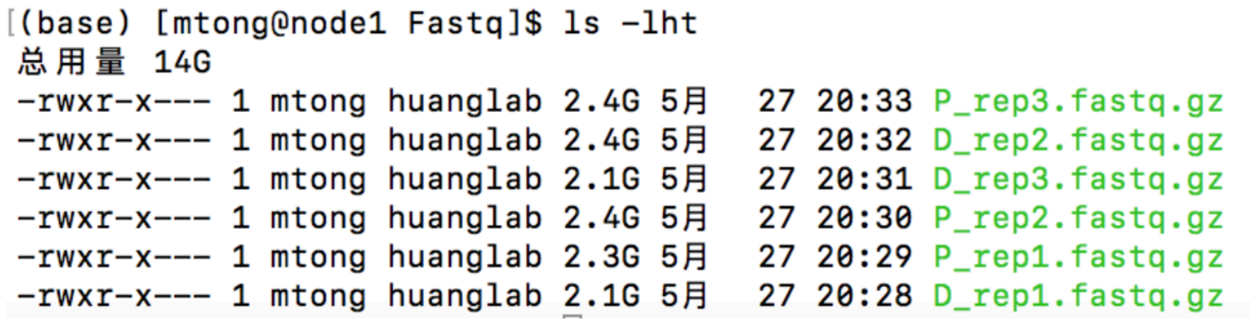
/cluster/huanglab/Trainning/RNA-seq/Rawdata/Fastq/

Data: Primitive (P) and definitive (D) erythroblast, 3 replicates

sampleInfo.csv (saved in the /cluster/huanglab/Trainning/RNA-seq/Rawdata/Fastq/)



```
(base) [mtong@node1 RNAseq]$ ls -lht
总用量 2.7M
-rwxr-x---  1 mtong huanglab 9.2K 5月  28 14:38 my_RNAseq_pipelines_Rsamtools.R
drwxr-sr-x  8 mtong huanglab   10 5月  28 14:13 package_guide
drwxrwsr-x  3 mtong huanglab   12 5月  28 13:32 DESeq
-rw-rw-r--  1 mtong huanglab 619K 5月  28 10:42 mm10_refGene_txbygene.RData
-rw-rw-r--  1 mtong huanglab 2.0M 5月  28 10:41 mm10_refGene_exbygene.RData
drwxrwsr-x 10 mtong huanglab   34 5月  28 00:48 STAR
-rwxr-x---  1 mtong huanglab 2.1K 5月  27 21:57 my_STAR_alignReads_bat.py
drwxr-s---  2 mtong huanglab    8 5月  27 20:32 Fastq
```



```
(base) [mtong@node1 Fastq]$ ls -lht
总用量 14G
-rwxr-x--- 1 mtong huanglab 2.4G 5月  27 20:33 P_rep3.fastq.gz
-rwxr-x--- 1 mtong huanglab 2.4G 5月  27 20:32 D_rep2.fastq.gz
-rwxr-x--- 1 mtong huanglab 2.1G 5月  27 20:31 D_rep3.fastq.gz
-rwxr-x--- 1 mtong huanglab 2.4G 5月  27 20:30 P_rep2.fastq.gz
-rwxr-x--- 1 mtong huanglab 2.3G 5月  27 20:29 P_rep1.fastq.gz
-rwxr-x--- 1 mtong huanglab 2.1G 5月  27 20:28 D_rep1.fastq.gz
```

Step 2. alignReads by STAR

```
$ nohup python my_STAR_alignReads_bat.py > STAR/logs/my_STAR_alignReads_bat_Fastq.log &
```

Output:

D_rep1_star_sorted.bam D_rep3_star_sorted.bam P_rep2_star_sorted.bam

D_rep2_star_sorted.bam P_rep1_star_sorted.bam P_rep3_star_sorted.bam

Step 3. count table using GenomicAlignments/Calling differential expression using DEseq2/Add annotation

prepare count table using GenomicAlignments.

rpkm/cpms using edgeR and log2

Calling differential expression using DEseq2

Add annotation (EntrezID, Gene Symbol)

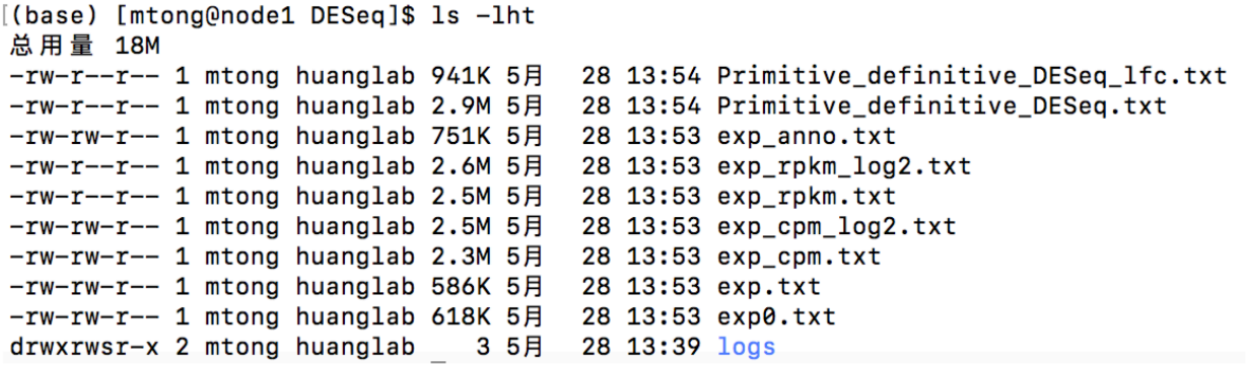
```
$ nohup R CMD BATCH my_RNAseq_pipelines_Rsamtools.R > DESeq/logs/my_RNAseq_pipelines_
Rsamtools.log &
```

Output:

Primitive_definitive_DESeq.txt ; exp0.txt ; exp_cpm.txt

exp_rpkm_log2.txt; exp_rpkm.txt; exp.txt

exp_cpm_log2.txt; exp_anno.txt; Primitive_definitive_DESeq_lfc.txt



```
[(base) [mtong@node1 DESeq]$ ls -lht
总用量 18M
-rw-r--r-- 1 mtong huanglab 941K 5月   28 13:54 Primitive_definitive_DESeq_lfc.txt
-rw-r--r-- 1 mtong huanglab 2.9M 5月   28 13:54 Primitive_definitive_DESeq.txt
-rw-rw-r-- 1 mtong huanglab 751K 5月   28 13:53 exp_anno.txt
-rw-r--r-- 1 mtong huanglab 2.6M 5月   28 13:53 exp_rpkm_log2.txt
-rw-r--r-- 1 mtong huanglab 2.5M 5月   28 13:53 exp_rpkm.txt
-rw-rw-r-- 1 mtong huanglab 2.5M 5月   28 13:53 exp_cpm_log2.txt
-rw-rw-r-- 1 mtong huanglab 2.3M 5月   28 13:53 exp_cpm.txt
-rw-rw-r-- 1 mtong huanglab 586K 5月   28 13:53 exp.txt
-rw-rw-r-- 1 mtong huanglab 618K 5月   28 13:53 exp0.txt
drwxrwsr-x 2 mtong huanglab _  3 5月   28 13:39 logs
```