RNA-seq示例分析

生物信息学
助教-刘柯
助教-方明昊

目录

- ❖上游处理(STAR, featureCounts)
- ❖下游分析(Deseq2)

❖ 上游分析(STAR， TopHat， HISAT)

File format	Procedure
RNA-seq Sequencer	
↓	Base calling
FASTQ: raw NGS reads	
↓	Aligning
SAM: aligned NGS reads	
↓	SAM to BAM, sorting etc.
BAM: compressed binary SAM	
↓	Quantification
txt/csv: sample gene expression	

和DNA的主要差别在于 RNA Alternative Splicing.

STAR manual 2.7.11a

Alexander Dobin
dobin@cshl.edu

August 15, 2023

Contents

❖ 上游分析(HTSeq, featureCounts, RSEM)

File format

RNA-seq Sequencer

FASTQ: raw NGS reads

SAM: aligned NGS reads

BAM: compressed binary SAM

txt/csv: sample gene expression

Procedure

Base calling

Aligning

SAM to BAM, sorting etc.

Quantification

Software

Rsubread/Subread Users Guide

Rsubread v2.12.3/Subread v2.0.4

20 February 2023

Wei Shi and Yang Liao

Olivia Newton-John Cancer Research Institute
Melbourne, Australia

❖ 上游分析(STAR)

module avail 命令查看集群安装的包

上游比对和定量所需要的
STAR和featurecounts已经安装

```
liuke@mgt:~ — -ssh liuke@nebula.ustc.edu.cn — 80×34
  9. qselect: select PBS batch jobs

-----------------------------------------------------------------
Please do not run jobs directly on the node mgt !!!
Please use qsub to submit your jobs !!!
-----------------------------------------------------------------
(base)[liuke@mgt ~]$module avail
--------------------------- /public/MODULES/COMPILER ---------------------------
cmake/3.19.0  cuDNN/v7.1              INTEL/parallel_studio_xe_2016.2.181
CUDA/8.0      cuDNN/v7.4.2            INTEL/parallel_studio_xe_2017_update4
CUDA/9.0      gcc/7.2.0               openmpi/4.1.2
CUDA/10.2.89  gcc/10.2.0              oracle-jdk
CUDA/11.4.4   INTEL/icc_2017_update4  R/4.1.2
----------------------------- /public/MODULES/APPS -----------------------------
Gaussian/G16          MATLAB/R2019a                vasp/5.4.4/intel2017update4
MaterialsStudio/18.1  singularity/3.1.0
MATLAB/R2017a         vasp/5.4.4/intel2016withGPU
----------------------------- /public/MODULES/BIO ------------------------------
afnl                Encode/Phantompeakqualtools          Relion_3.0beta
amber               Encode/PIQ                           RepeatMasker
amber22             Encode/sample                        RMBlast
Anaconda2           Encode/TophatBAMRepair               samtools
Anaconda3           Encode/WASP                          scipion
bcftools            fastqc                               sratoolkit
bedops              flexbar                              STAR
blast               GATK                                 subread-1.6.4
bowtie              gemtools                             tantan
bowtie2             gromacs/4.5.5                        tophat-1.4.1
bwa                 gromacs/2016.3                       tophat-2.1.1
cdhit               HiC-Pro                              TRF
ChIA-PET2           hisat2                               Trinity
chilin              hmmer                                vsearch
chimera             HOMER
cryolo              hotspot2
```

❖ 上游分析(STAR)

构建索引

liuke@mgt:~/rna/pbs — -ssh liuke@nebula.ustc.edu.cn — 110×24

```
#DO NOT RUN THIS SCRIPT DIRECTLY,
#PLEASE RUN THIS SCRIPT WITH qsub: qsub serial_job.pbs
#
#PBS -N star
#PBS -o log/star.log
#PBS -e log/star.err
#PBS -q batch
#PBS -l walltime=48:00:00
#PBS -l nodes=1:ppn=28

echo Start time: `date`

module load STAR

cd /home/qukun/liuke/Reference/STAR/hg38

STAR --runThreadN 28\
     --runMode genomeGenerate\
     --genomeDir /home/qukun/liuke/Reference/STAR/hg38 \
     --genomeFastaFiles /home/qukun/liuke/Reference/GENCODE/hg38/GRCh38.p13.genome.fa \
     --sjdbGTFfile /home/qukun/liuke/Reference/GENCODE/hg38/gencode.v38.annotation.gtf

echo End time: `date`
                                                                    25,21         Bot
```

❖ 上游分析(STAR)

比对

liuke@mgt:~/rna/pbs — -ssh liuke@nebula.ustc.edu.cn — 92×27

```
#!/bin/sh
#An example for serial job.
#DO NOT RUN THIS SCRIPT DIRECTLY,
#PLEASE RUN THIS SCRIPT WITH qsub: qsub serial_job.pbs
#PBS -N mapping
#PBS -o log/mapping.log
#PBS -e log/mapping.err
#PBS -q batch
#PBS -l walltime=96:00:00
#PBS -l nodes=1:ppn=20

echo Start time: `date`
cd /home/qukun/liuke/rna/data
module load STAR

STAR --runThreadN 20 \
     --genomeDir  /home/qukun/liuke/reference/STAR/hg38 \
     --readFilesIn test.R1.fastq.gz test.R2.fastq.gz \
     --readFilesCommand zcat \
     --outFilterMultimapNmax 1 \
     --clip3pNbases=0 \
     --clip5pNbases=0 \
     --outFileNamePrefix /home/qukun/liuke/rna/result/test \
     --outSAMtype BAM SortedByCoordinate

echo End time: `date`
"mapping.pbs" 26L, 701C
```

❖ 上游分析(STAR)

比对结束后的结果文件包括了比对后的bam文件，以及比对的一些信息。

Prefix.final.out里面包括了比对上的reads的具体数目和比例等

```
(base)[liuke@mgt result]$ls
testAligned.sortedByCoord.out.bam  testLog.out            testSJ.out.tab
testLog.final.out                  testLog.progress.out   test_STARtmp
```

```
(base)[liuke@mgt result]$cat testLog.final.out
                                 Started job on |       Oct 09 16:51:22
                             Started mapping on |       Oct 09 16:54:20
                                    Finished on |       Oct 09 16:58:02
       Mapping speed, Million of reads per hour |       148.77

                          Number of input reads |       9174385
                      Average input read length |       300
                                    UNIQUE READS:
                   Uniquely mapped reads number |       830656
                        Uniquely mapped reads % |       9.05%
                          Average mapped length |       283.93
                       Number of splices: Total |       578650
            Number of splices: Annotated (sjdb) |       567991
                       Number of splices: GT/AG |       568618
                       Number of splices: GC/AG |       2326
                       Number of splices: AT/AC |       0
               Number of splices: Non-canonical |       7706
                      Mismatch rate per base, % |       0.33%
                         Deletion rate per base |       0.02%
                        Deletion average length |       1.42
                        Insertion rate per base |       0.01%
                       Insertion average length |       1.62
                             MULTI-MAPPING READS:
        Number of reads mapped to multiple loci |       0
             % of reads mapped to multiple loci |       0.00%
        Number of reads mapped to too many loci |       4828814
             % of reads mapped to too many loci |       52.63%
                                  UNMAPPED READS:
       % of reads unmapped: too many mismatches |       0.00%
                 % of reads unmapped: too short |       38.30%
                     % of reads unmapped: other |       0.01%
                                  CHIMERIC READS:
```

❖ 上游分析(featureCounts)

计数

liuke@mgt:~/rna/pbs — -ssh liuke@nebula.ustc.edu.cn — 97×23

```
#!/bin/sh
#An example for serial job.
#DO NOT RUN THIS SCRIPT DIRECTLY,
#PLEASE RUN THIS SCRIPT WITH qsub: qsub serial_job.pbs
#PBS -N counting
#PBS -o log/counting.log
#PBS -e log/counting.err
#PBS -q batch
#PBS -l walltime=96:00:00
#PBS -l nodes=1:ppn=20

cd /home/qukun/liuke/rna/data
module load subread-1.6.4

featureCounts -a /home/qukun/liuke/reference/GENCODE/hg38/gencode.v38.annotation.gtf \
 -o /home/qukun/liuke/rna/result/test.count \
 -R BAM /home/qukun/liuke/rna/result/testAligned.sortedByCoord.out.bam \
 -p \
 -T 20 \
 -g gene_name \
 -t gene
~
                                                    10,1          All
```

❖ 上游分析(featureCounts)

结果文件

test.count中会包含具体的gene，分布位置以及计数情况。

Test.count.summary中会包含计数信息，同时被计数的reads会被提取到新后缀为featurecounts.bam文件中

```
liuke@mgt:~/rna/result — -ssh liuke@nebula.ustc.edu.cn — 153×13
(base)[liuke@mgt result]$head test.count
# Program:featureCounts v1.6.4; Command:"featureCounts" "-a" "/home/qukun/liuke/reference/GENCODE/hg38/gencode.v38.annotation.gtf" "-o" "/home/qukun/liuk
e/rna/result/test.count" "-R" "BAM" "/home/qukun/liuke/rna/result/testAligned.sortedByCoord.out.bam" "-p" "-T" "20" "-g" "gene_name" "-t" "gene"
Geneid  Chr     Start   End     Strand  Length  /home/qukun/liuke/rna/result/testAligned.sortedByCoord.out.bam
DDX11L1 chr1    11869   14409   +       2541    0
WASH7P  chr1    14404   29570   -       15167   0
MIR6859-1       chr1    17369   17436   -       68      0
MIR1302-2HG     chr1    29554   31109   +       1556    0
MIR1302-2       chr1    30366   30503   +       138     0
FAM138A chr1    34554   36081   -       1528    0
OR4G4P  chr1    52473   53312   +       840     0
OR4G11P chr1    57598   64116   +       6519    0
(base)[liuke@mgt result]$
```

```
liuke@mgt:~/rna/result — -ssh liuke@nebula.ustc.edu.cn — 76×15
(base)[liuke@mgt result]$cat  test.count.summary
Status  /home/qukun/liuke/rna/result/testAligned.sortedByCoord.out.bam
Assigned        560703
Unassigned_Unmapped     0
Unassigned_MappingQuality       0
Unassigned_Chimera      0
Unassigned_FragmentLength       0
Unassigned_Duplicate    0
Unassigned_MultiMapping 0
Unassigned_Secondary    0
Unassigned_NonSplit     0
Unassigned_NoFeatures   10661
Unassigned_Overlapping_Length   0
Unassigned_Ambiguity    259292
(base)[liuke@mgt result]$
```

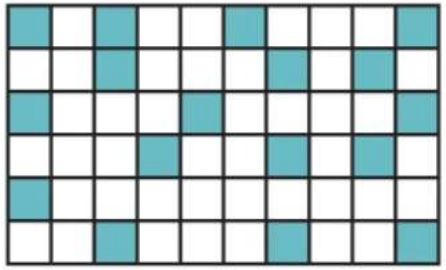
目录

- ❖上游处理(STAR, featureCounts)
- ❖下游分析(Deseq2)

❖ 下游分析(Deseq2)

Sample

Gene



测序reads总数 1 million 2 million 3 million

Gene Name	Sample1 Counts	Sample2 Counts	Sample3 Counts
A (2kb)	10	12	30
B (4kb)	20	25	60

FPKM

	sample1	sample2	sample3
A	10/1/2	12/2/2	30/3/2
B	20/1/4	25/2/4	60/3/4

TPM

	sample1	sample2	sample3
A	10/2/1*	12/2/2*	30/2/3*
B	20/4/1*	25/4/2*	60/4/3*

FPKM

FPKM: FPKM的全称为Fragments Per Kilobase Million，Fragments Per Kilobase of exon model per Million mapped fragments(每千个碱基的转录每百万映射读取的fragments)。通俗讲，把比对到的某个基因的Fragment数目，除以基因的长度，其比值再除以所有基因的总长度。**注意**，这里的基因长度是指基因外显子的总长度。

$$公式(2): FPKM = \frac{ExonMappedFragments * 10^9}{TotalMappedFragments * ExonLength}$$

红色 测序深度值应变为标准化之后的数值

TPM

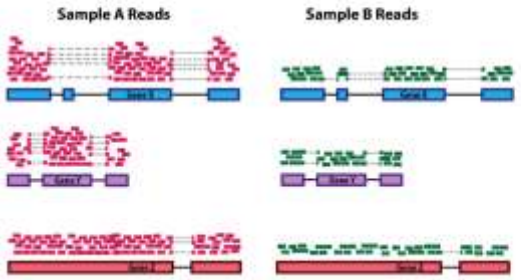
定义：TPM的全称为Transcripts per million，Transcripts Per Kilobase of exon model per Million mapped reads (每千个碱基的转录每百万映射读取的Transcripts)

$$TPM = \frac{Ni/Li * 10^6}{sum(N1/L1 + N2/L2 + ... + Nn/Ln)}$$

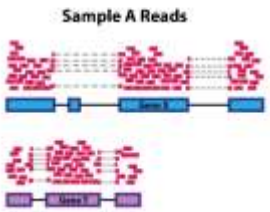
解释：Ni为比对到第i个exon的reads数； Li为第i个exon的长度；sum(N1/L1+N2/L2 + ... + Nn/Ln)为所有 (n个)exon按长度进行标准化之后数值的和。

❖ 下游分析(Deseq2)

- **Sequencing depth:** Accounting for sequencing depth is necessary for comparison of gene expression between samples. In the example below, each gene appears to have doubled in expression in *Sample A* relative to *Sample B*, however this is a consequence of *Sample A* having double the sequencing depth.

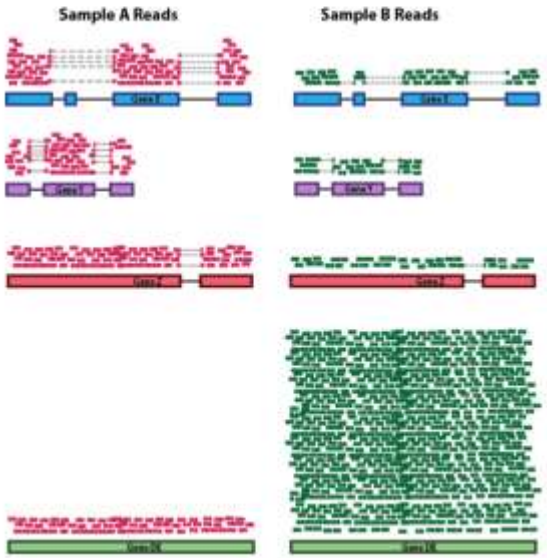


- **Gene length:** Accounting for gene length is necessary for comparing expression between different genes within the same sample. In the example, *Gene X* and *Gene Y* have similar levels of expression, but the number of reads mapped to *Gene X* would be many more than the number mapped to *Gene Y* because *Gene X* is longer.



- **RNA composition:** A few highly differentially expressed genes between samples, differences in the number of genes expressed between samples, or presence of contamination can skew some types of normalization methods. Accounting for RNA composition is recommended for accurate comparison of expression between samples, and is particularly important when performing differential expression analyses [1].

In the example, imagine the sequencing depths are similar between Sample A and Sample B, and every gene except for gene DE presents similar expression level between samples. The counts in Sample B would be greatly skewed by the DE gene, which takes up most of the counts. Other genes for Sample B would therefore appear to be less expressed than those same genes in Sample A.



https://hbctraining.github.io/DGE_workshop_salmon/lessons/02_DGE_count_normalization.html

❖ 下游分析(Deseq2)

TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**
DESeq2's **median of ratios** [1]	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**
EdgeR's **trimmed mean of M values (TMM)** [2]	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**

https://hbctraining.github.io/DGE_workshop_salmon/lessons/02_DGE_count_normalization.html

❖ 下游分析(Deseq2)

	Sample #1	Sample #2	Sample #3
Gene1	0	10	4
Gene2	2	6	12
Gene3	33	55	200

对每个基因取对数ln(默认底数e)　负无穷暂时去掉该基因

Step 1: Take the log of all the values

	log(Sample #1)	log(Sample #2)	log(Sample #3)
Gene1	-Inf	2.3	1.4
Gene2	0.7	1.8	2.5
Gene3	3.5	4.0	5.3

对每个基因的对数 减去 样本间对数均值(本质为比值)

	Sample #1	Sample #2	Sample #3
Gene2	-1.0	0.1	0.5
Gene3	-0.8	-0.3	1.3

$$\log\left(\frac{\text{reads for gene X}}{\text{average for gene X}}\right)$$

	Sample #1	Sample #2	Sample #3
Gene2	-1.0	0.1	0.5
Gene3	-0.8	-0.3	1.3
median =	-0.9	-0.1	0.9

House 基因集

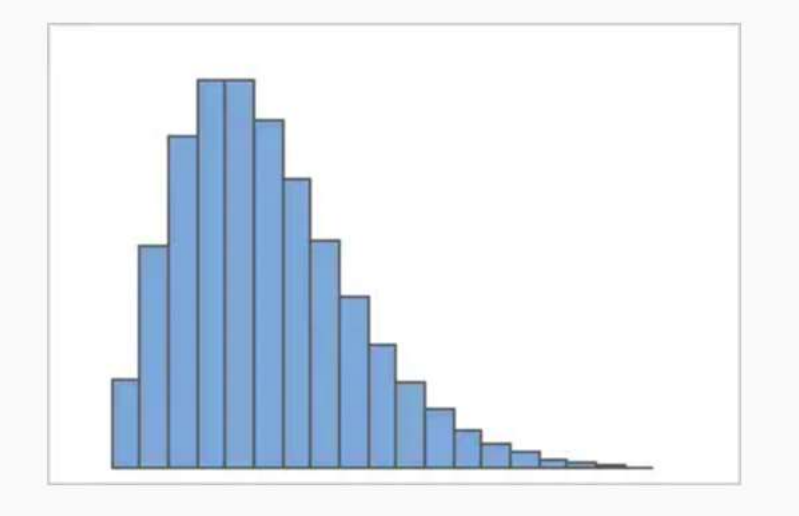
Size factor
(对数变换后用
于标准化矫正)

以对数比值中值数 代表该样本测序深度影响

https://cloud.tencent.com/developer/article/1553526

❖ 下游分析(Deseq2)

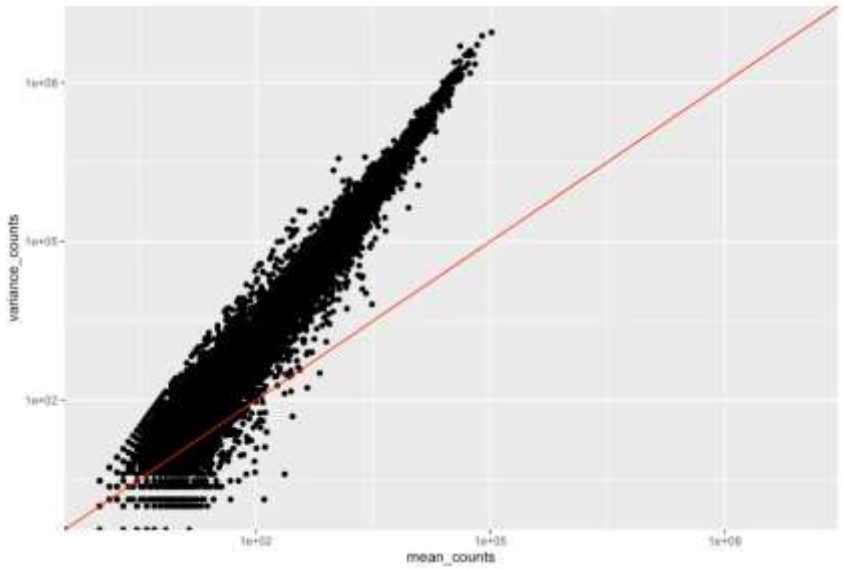
负二项分布



$$\mu = \frac{pr}{1-p}$$

$$\sigma^2 = \frac{pr}{(1-p)^2}$$

方差>均值



DESeq2使用离散度(dispersion)作为方差的度量方式，离散度既可以解释基因表达值的方差也可以解释基因的平均表达值。其具体公式为：`Var = μ + α*μ^2`。其中Var表示方差，μ表示均值，α表示离散度。因此我们可以得到这么一个关系

	离散度
方差增加	离散度增加
平均值增加	离散度降低

GLM 广义线性建模

https://www.jianshu.com/p/bf3fdd21153e

❖ 下游分析(Deseq2)

Analyzing RNA-seq data with DESeq2

Michael I. Love, Simon Anders, and Wolfgang Huber

06/23/2023

Abstract

A basic task in the analysis of count data from RNA-seq is the detection of differentially expressed genes. The count data are presented as a table which reports, for each sample, the number of sequence fragments that have been assigned to each gene. Analogous data also arise for other assay types, including comparative ChIP-Seq, HiC, shRNA screening, and mass spectrometry. An important analysis question is the quantification and statistical inference of systematic changes between conditions, as compared to within-condition variability. The package DESeq2 provides methods to test for differential expression by use of negative binomial generalized linear models; the estimates of dispersion and logarithmic fold changes incorporate data-driven prior distributions. This vignette explains the use of the package and demonstrates typical workflows. An RNA-seq workflow on the Bioconductor website covers similar material to this vignette but at a slower pace, including the generation of count matrices from FASTQ files. DESeq2 package version: 1.40.2

Installation

To install this package, start R (version "4.3") and enter:

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("DESeq2")
```

For older versions of R, please refer to the appropriate Bioconductor release.

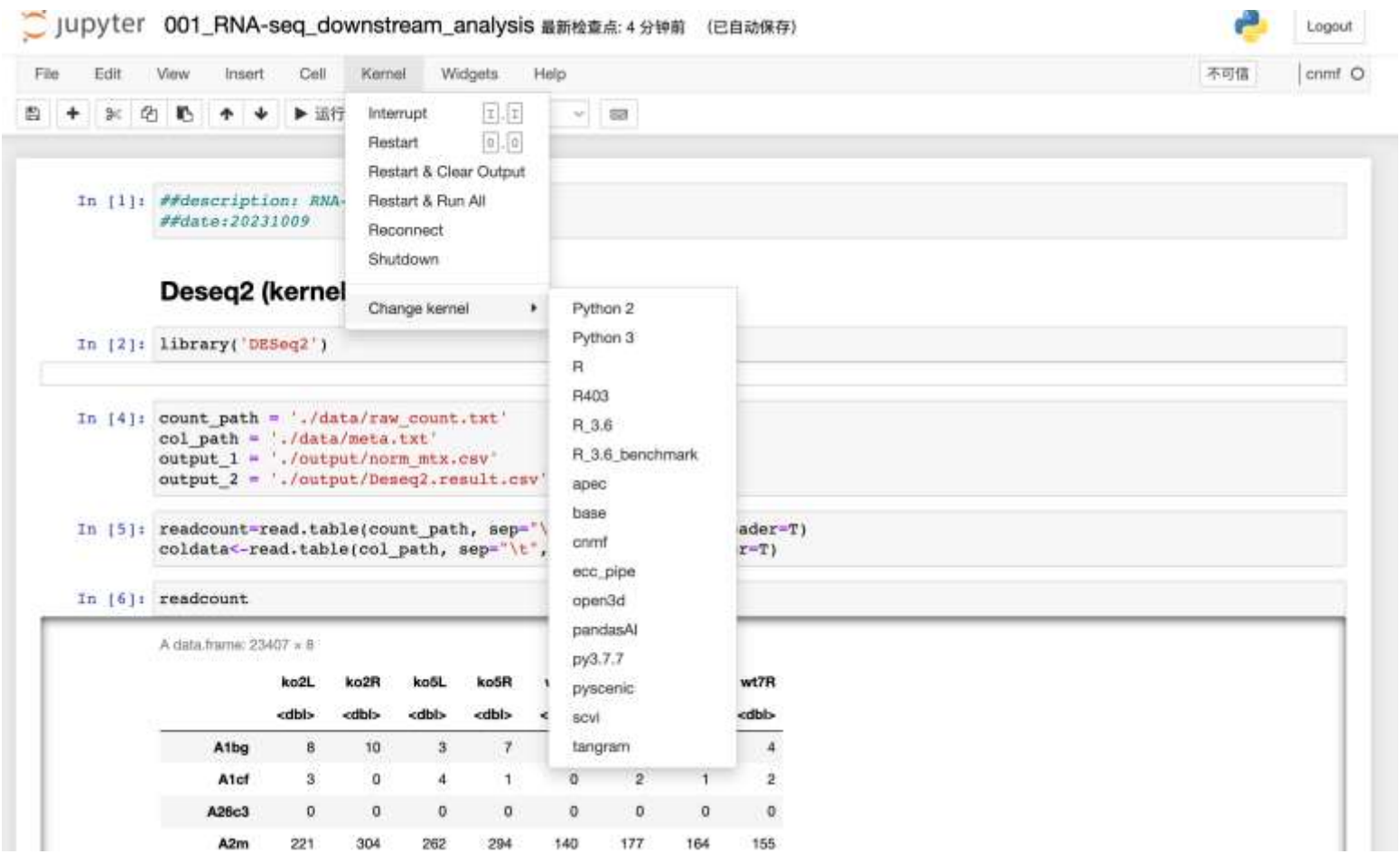
conda install

To install this package run one of the following:

```
conda install -c bioconda bioconductor-deseq2
conda install -c "bioconda/label/broken" bioconductor-deseq2
conda install -c "bioconda/label/cf201901" bioconductor-deseq2
conda install -c "bioconda/label/gcc7" bioconductor-deseq2
```

https://www.bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

❖ Deseq2 使用实例



1. 激活环境请更改成自己的环境

2. 请提交PBS脚本运行,禁止在登录节点运行任务

3.Jupyter禁止直接在登录节点运行,需提交pbs脚本后通过log文件获取地址然后ssh转本机端口浏览器访问[如服务器配置存在困难, 可以本机自行安装anaconda+jupyter]