# 生物信息学-2023秋-大作业

# 服务器上预装软件和原始数据路径

1.module load bowtie2, fastqc, picard, HOMER, samtools, STAR

2.Bowtie2 使用 reference 地址:

mm9: /home/bioinfo2023/bioclass2023/software/bowtie2_ref/indexes/mm9
hg19: /home/bioinfo2023/bioclass2023/software/bowtie2_ref/indexes/hg19

3.STAR 所用 reference 地址:

mm10: /home/bioinfo2023/bioclass2023/software/STAR_ref/mm10_star_index
hg38: /home/bioinfo2023/bioclass2023/software/STAR_ref/hg38_star_index

3.Macs3:

/home/bioinfo2023/bioclass2023/miniconda3/bin/macs3

4.ATAC-pipe 地址:

/home/bioinfo2023/bioclass2023/software/ATAC-pipe-master

5.作业原始fq.gz数据 地址:

/home/bioinfo2023/bioclass2023/homework/RNA-seq/01.raw_data
/home/bioinfo2023/bioclass2023/homework/CHIP-seq/01.raw_data
/home/bioinfo2023/bioclass2023/homework/ATAC-seq/01.raw_data

- RNA-seq
- CHIP-seq
- ATAC-seq

# RNA-seq作业

**目标：**利用公开的数据，完成一项RNA-seq的分析

**数据：**https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA516223 (Group A: SRR8467686, SRR8467687, SRR8467688, SRR8467689; Group B:SRR8467690, SRR8467691, SRR8467692, SRR8467693) **nebula地址：** /home/bioinfo2023/bioclass2023/homework/RNA-seq/01.raw_data

**要求：**

1.从原始测序数据开始（fastq）

2.至少包含以下分析内容：

- 数据下载（1分）
- 数据质量控制（1分）
- 数据比对和基因表达定量(得到Gene x Cell的矩阵)（3分）
- 差异表达分析(Heatmap或火山图) （3分）
- 差异基因功能富集分析(展示差异基因GO Term)（2分）

3.在2023年xx月xx日之前，将PPT和代码发送到邮箱：组长为周一班的 请发送邮件至 liuk0617@mail.ustc.edu.cn 周三班发至fang0426@mail.ustc.edu.cn，邮件名和作业压缩包命名一致。

4.邮件主题：生物信息学RNA-seq数据分析实践作业+组长学号+组长姓名

5.PPT第一页务必说明组员姓名学号和分工

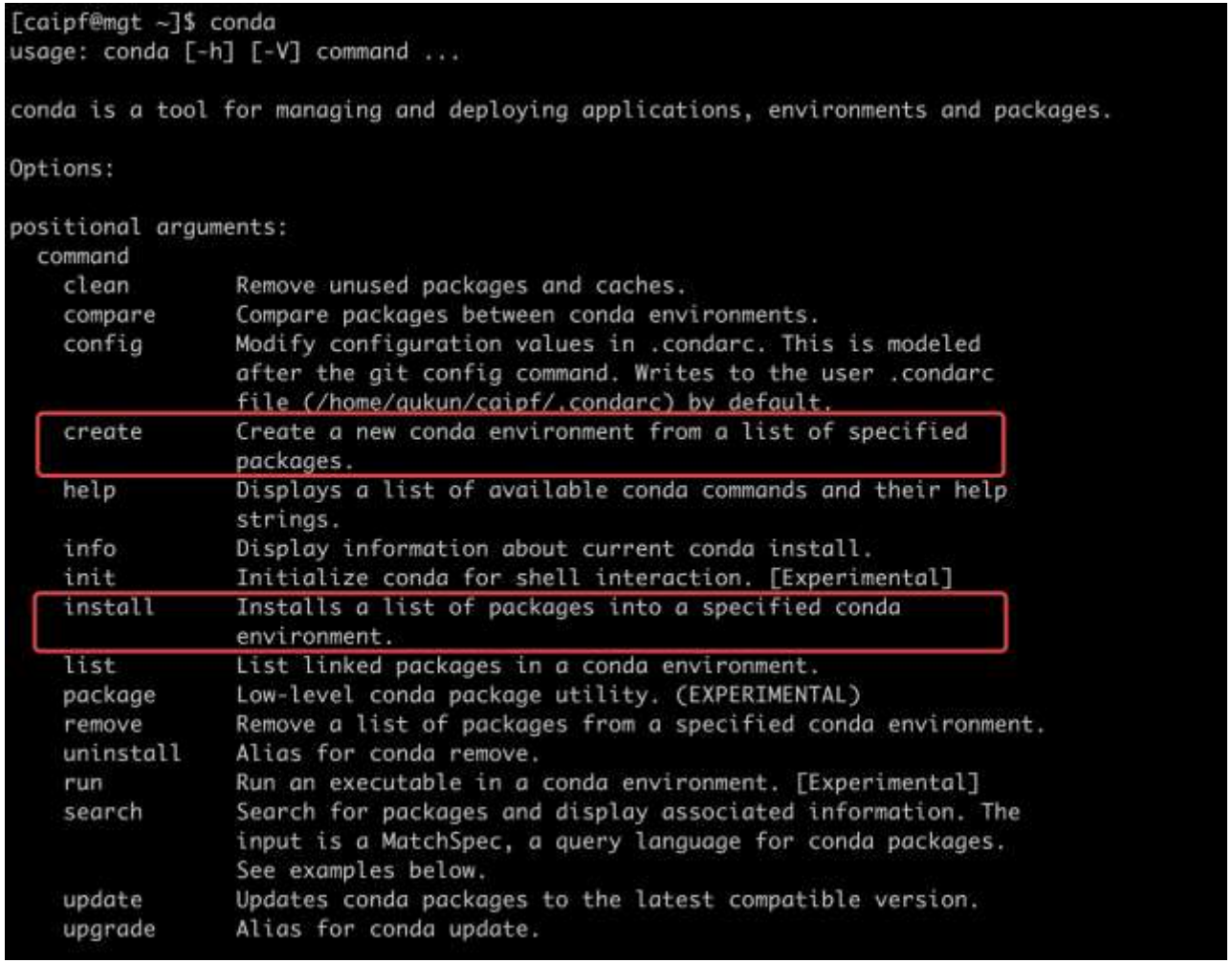
# Softwares on nebula server

```
[caipf@mgt ~]$ module avail
------------------------------------------------------------ /public/MODULES/COMPILER ------------------------------------------------------------
CUDA/8.0  CUDA/10.2.89  cuDNN/v7.4.2  INTEL/icc_2017_update4              INTEL/parallel_studio_xe_2017_update4
CUDA/9.0  cuDNN/v7.1    gcc/7.2.0     INTEL/parallel_studio_xe_2016.2.181  oracle-jdk
-------------------------------------------------------------- /public/MODULES/APPS --------------------------------------------------------------
Gaussian/G16  MaterialsStudio/18.1  MATLAB/R2017a  MATLAB/R2019a  singularity/3.1.0  vasp/5.4.4/intel2016withGPU  vasp/5.4.4/intel2017update4
--------------------------------------------------------------- /public/MODULES/BIO ---------------------------------------------------------------
afnl       chilin            Encode/Flux                fastqc          IGVTools      Relion3                                 tophat-1.4.1
amber      chimera           Encode/fseq                flexbar         juicer        Relion3.1_beta_SinglePrecisionOnGPU     tophat-2.1.1
Anaconda2  cryolo            Encode/gerp                GATK            kentUtils     Relion3_SinglePrecisionOnGPU            TRF
Anaconda3  cufflinks         Encode/kentUtils           gemtools        lammps        Relion_3.0beta                          Trinity
bcftools   demuxlet          Encode/mfinder             gromacs/4.5.5   macs2         RepeatMasker                            vsearch
bedops     Dynamo            Encode/npIDR               gromacs/2016.3  meme          RMBlast
blast      ea-utils          Encode/PeakSeq             HiC-Pro         modwt         samtools
bowtie     Encode/AlleleSeq  Encode/Phantompeakqualtools hisat2         nuc_dynamics  scipion
bowtie2    Encode/bismark    Encode/PIQ                 hmmer           picard        sratoolkit
bwa        Encode/ChromHMM   Encode/sample              HOMER           preseq        STAR
cdhit      Encode/Cluster    Encode/TophatBAMRepair     hotspot2        prinseq-lite  subread-1.6.4
ChIA-PET2  Encode/cxrepo-bed Encode/WASP                IGV             Relion        tantan
------------------------------------------------------- /public/MODULES/to_be_deleted -------------------------------------------------------
```

Sratoolkit (NCBI下载数据)
Bowtie, tophat, STAR, hisat2 （将测序reads比对到基因组数据上）
Fastqc, picard, IGVTools (测序数据进行质量评估和结果查看）
Samtools, deeptools（对比对后的sam, bam文件进行操作）
Tophat+cufflinks （这个组合现在基本已经过时）
RSEM, HTSeq, featurecounts, self-made scripts（计算基因表达量, reads counts/RPKM/TPM）
DESeq2/EdgeR（差异分析）
Enrichr, Metascape, String, GSEA (基因集功能分析)

# RNA-seq环境配置

服务器上首先下载conda (https://docs.conda.io/projects/conda/en/latest/user-guide/install/linux.html)




E.g.:
# create a environment
conda create –n RNA python=3.7
conda info –env
conda activate RNA

# install packages
conda install
pip install HTSeq
pip install deeptools

# 数据集下载

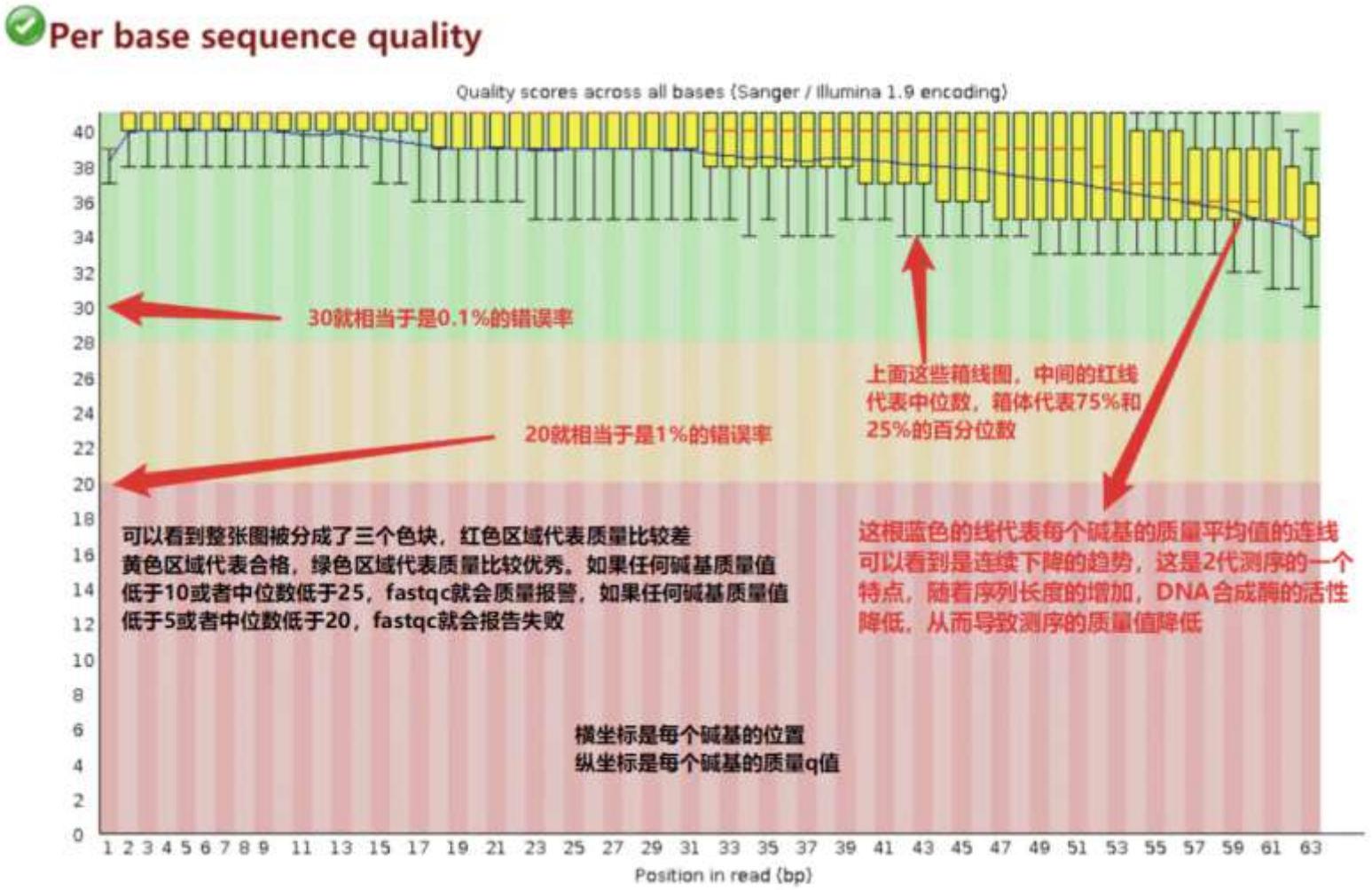
You selected 8 Items

| | Run | BioSample | Assay Type | AvgSpotLen | Bases | Bytes | Experiment | GEO_Accession | LibrarySelection | Sample Name |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | SRR8467686 | SAMN10785301 | RNA-Seq | 50 | 2.09 G | 1.57 Gb | SRX5274092 | GSM3573385 | cDNA | GSM3573385 |
| 2 | SRR8467687 | SAMN10785299 | RNA-Seq | 50 | 1.78 G | 1.35 Gb | SRX5274093 | GSM3573386 | cDNA | GSM3573386 |
| 3 | SRR8467688 | SAMN10785298 | RNA-Seq | 50 | 1.41 G | 1.09 Gb | SRX5274094 | GSM3573387 | cDNA | GSM3573387 |
| 4 | SRR8467689 | SAMN10785297 | RNA-Seq | 50 | 1.40 G | 1.08 Gb | SRX5274095 | GSM3573388 | cDNA | GSM3573388 |
| 5 | SRR8467690 | SAMN10785296 | RNA-Seq | 50 | 1.95 G | 1.46 Gb | SRX5274096 | GSM3573389 | cDNA | GSM3573389 |
| 6 | SRR8467691 | SAMN10785294 | RNA-Seq | 50 | 2.10 G | 1.58 Gb | SRX5274097 | GSM3573390 | cDNA | GSM3573390 |
| 7 | SRR8467692 | SAMN10785293 | RNA-Seq | 50 | 2.51 G | 1.88 Gb | SRX5274098 | GSM3573391 | cDNA | GSM3573391 |
| 8 | SRR8467693 | SAMN10785292 | RNA-Seq | 50 | 1.78 G | 1.39 Gb | SRX5274099 | GSM3573392 | cDNA | GSM3573392 |

**GEO:GSE125400** https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA516223

**参考工具和代码**:
**Sratoolkit**下的prefetch命令下载数据
**Fast-dump**命令转换sra文件为fastq文件， 例如：
fastq-dump –gzip /home/qukun/caipf/ncbi/public/sra/SRR8467682.sra, 如果是双端测序（paired-end）还需添加–split-3参数

# 数据质量控制




**FastQC**旨在提供一种简单的方法，对来自高通量测序管道的原始序列数据做一些质量控制检查

过滤标准：

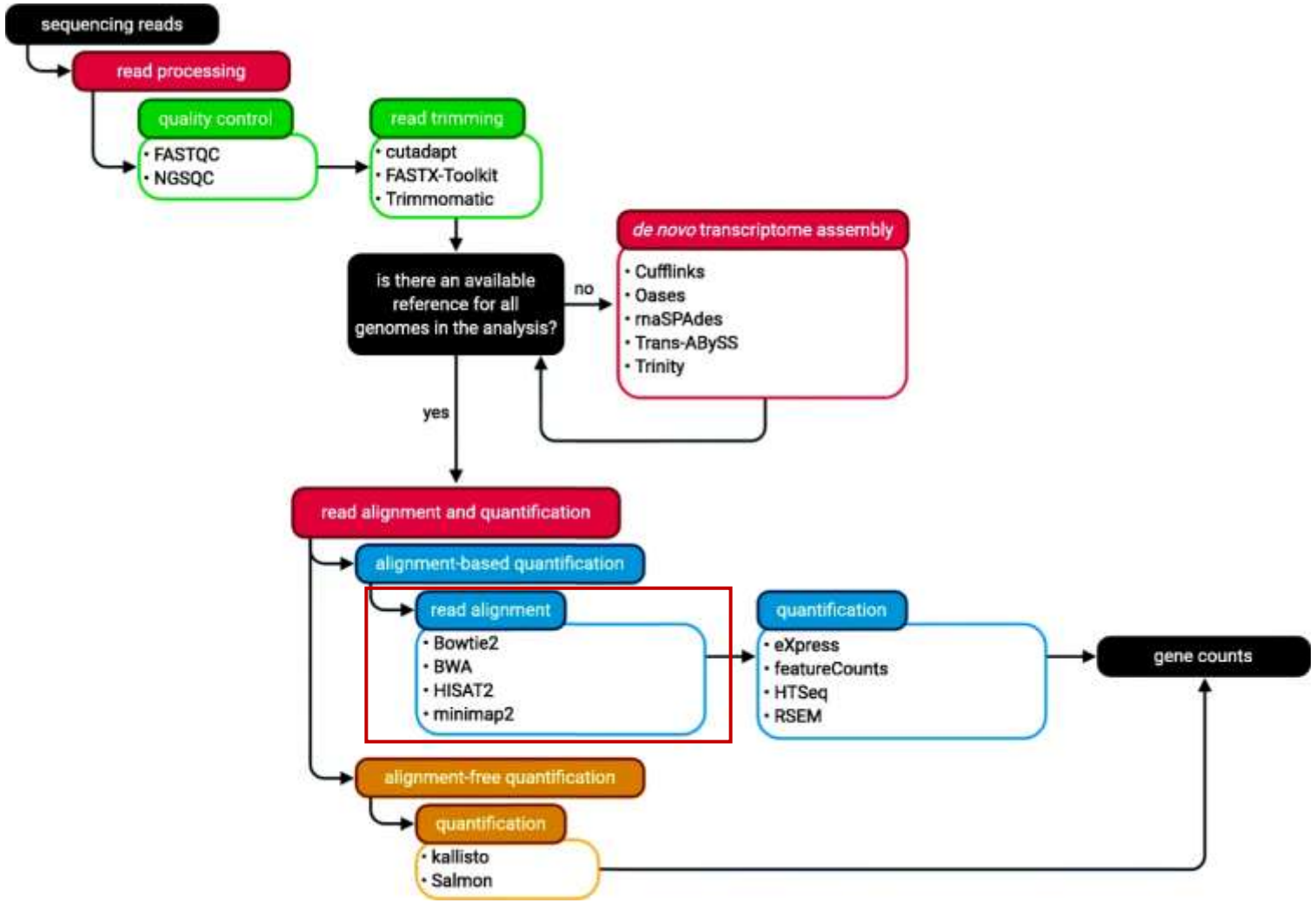
- 去除碱基质量值低于20的reads
- 去除N比例高于百分之5的reads
- 去除Index或接头
- 去除一些reads的head或tail

数据过滤的软件：**cutadapter**，**trimmomatic**，**trim_galore**，**fastp**。

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

https://www.jianshu.com/p/f223206b3378

# 比对基因组



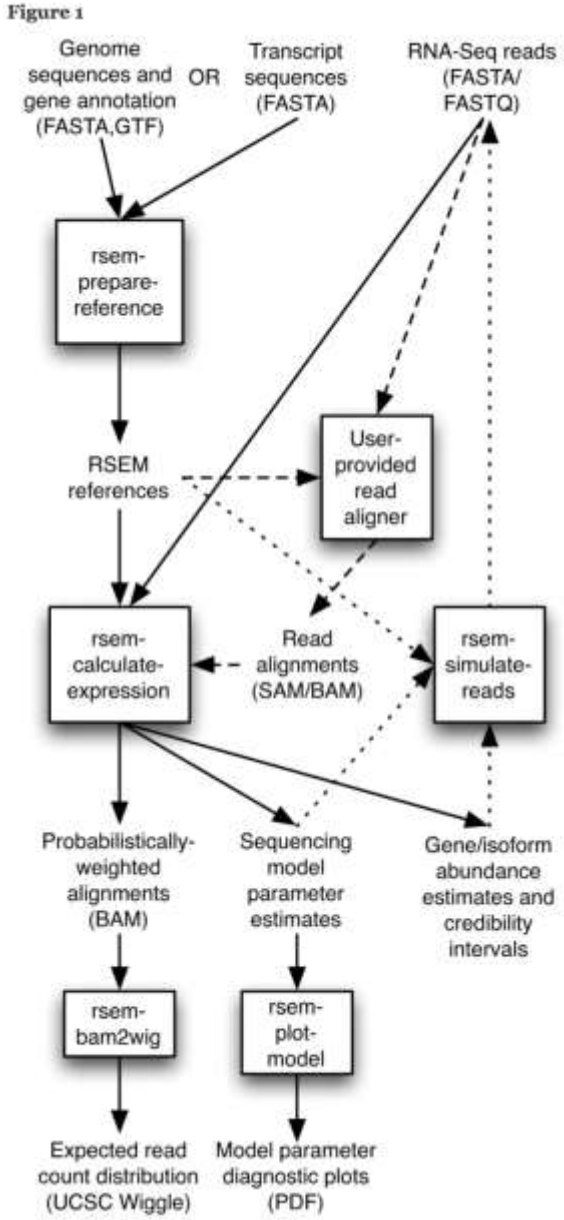

常用软件

- **STAR**
https://github.com/alexdobin/STAR/

- **TopHat2**
http://ccb.jhu.edu/software/tophat/index.shtml

- **HISAT2**
https://daehwankimlab.github.io/hisat2/

# 基因表达定量




- **RSEM**计算TPM、FPKM/RPKM的值
- **HTSeq**计算reads counts

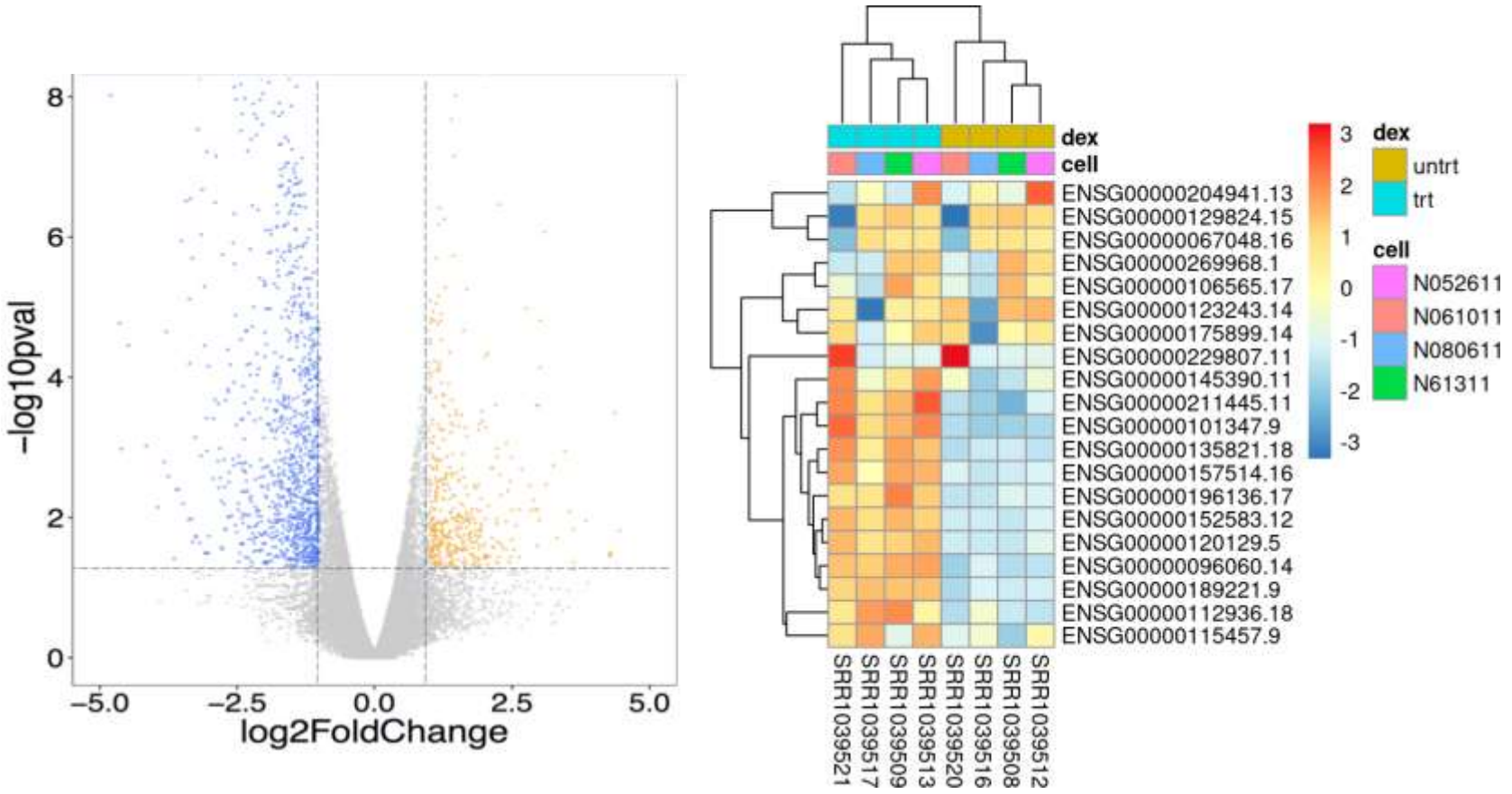
RSEM:
https://github.com/deweylab/RSEM
Htseq:
https://htseq.readthedocs.io/en/master/index.html

参考文献:

Li et al., RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinformatics.

Anders et al., HTSeq—a Python framework to work with high-throughput sequencing data;
Bioinformatics.

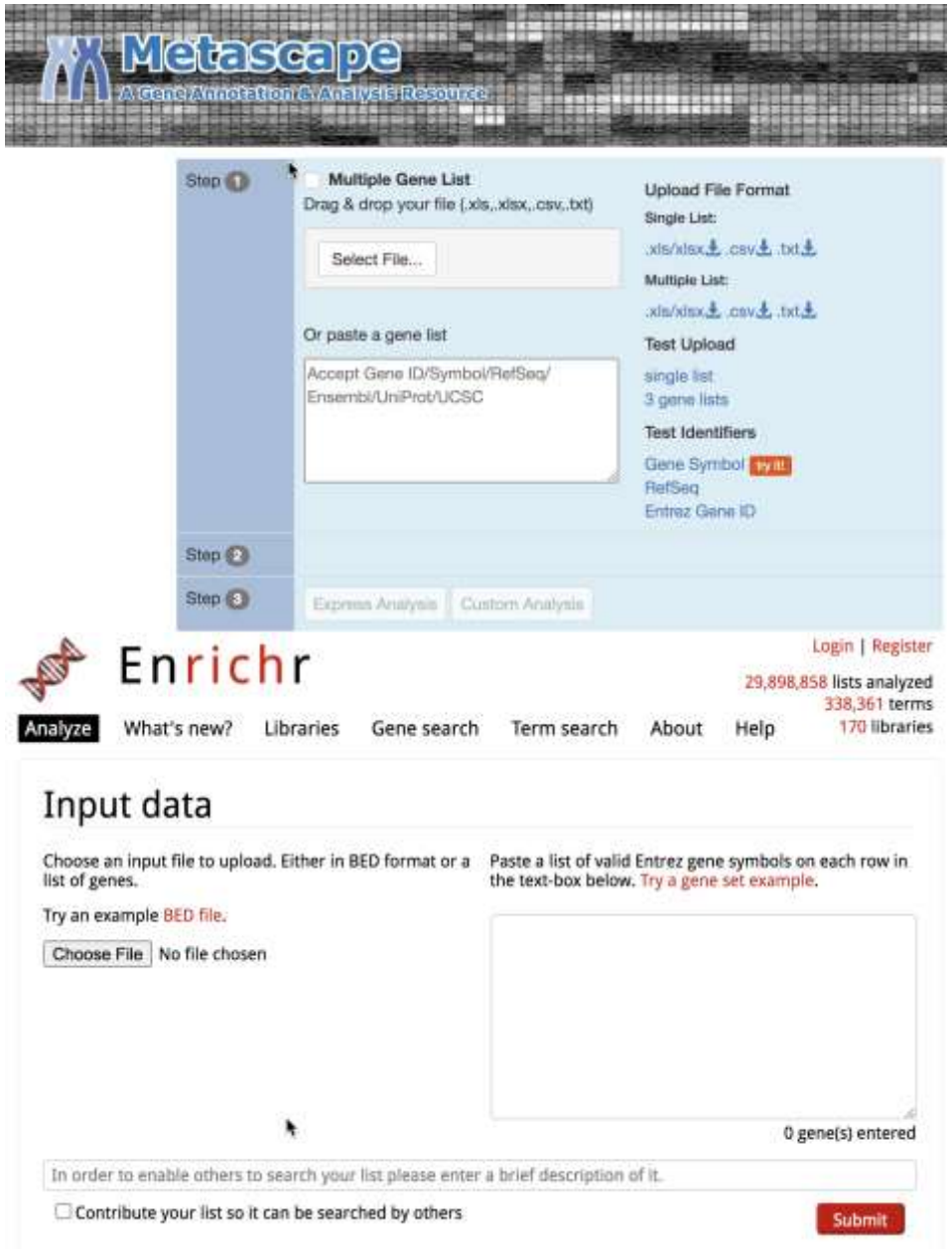
# 差异分析

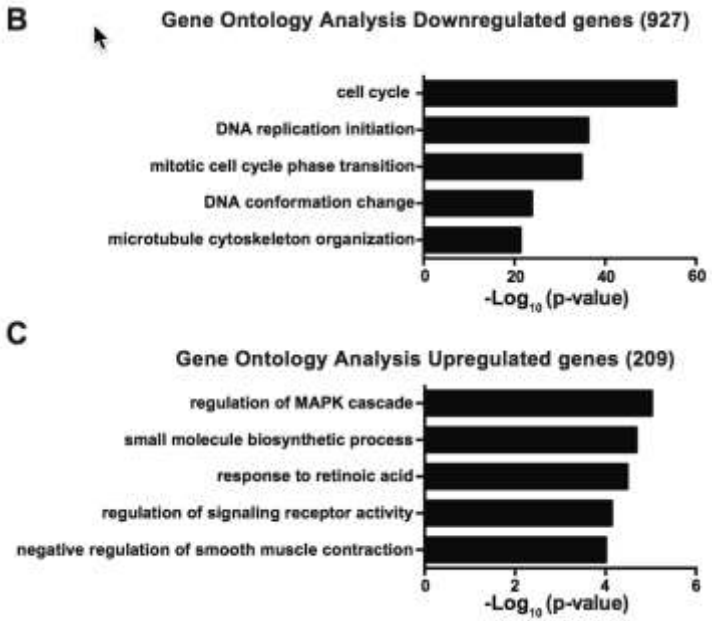



常用工具：

- **DESeq2** https://genepattern.github.io/DESeq2/v1/index.html
- **Limma** https://kasperdanielhansen.github.io/genbioconductor/html/limma.html
- **edgeR** https://bioconductor.org/packages/release/bioc/html/edgeR.html

# 基因功能分析




常用工具

- **Metascape**: https://metascape.org/gp/index.html#/main/step1
- **Enrichr**: https://maayanlab.cloud/Enrichr/

# 参考资料

## RNA-seq workflow: rnaseqGene

RNA-seq workflow: gene-level exploratory analysis and differential expression

*Michael I. Love*[1,2], *Simon Anders*[3], *Vladislav Kim*[4] *and Wolfgang Huber*[4]

[1]Department of Biostatistics, UNC-Chapel Hill, Chapel Hill, NC, US
[2]Department of Genetics, UNC-Chapel Hill, Chapel Hill, NC, US
[3]Zentrum für Molekulare Biologie der Universität Heidelberg, Heidelberg, Germany
[4]European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

16 October, 2019

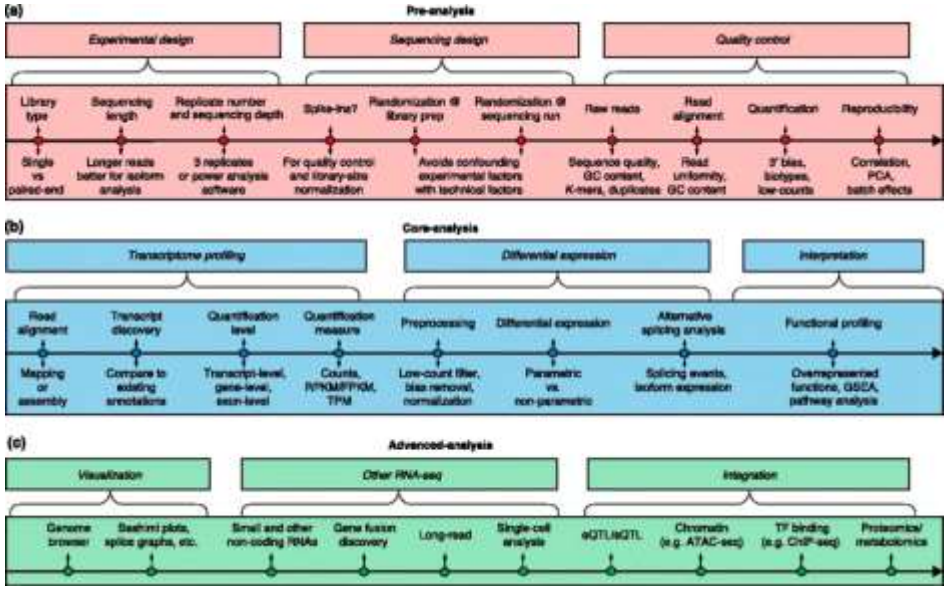
**Abstract**

Here we walk through an end-to-end gene-level RNA-seq differential expression workflow using Bioconductor packages. We will start from the FASTQ files, show how these were quantified to the reference transcripts, and prepare gene-level count datasets for downstream analysis. We will perform exploratory data analysis (EDA) for quality assessment and to explore the relationship between samples, perform differential gene expression analysis, and visually explore the results.

**参考**

- https://bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html
- http://master.bioconductor.org/packages/release/workflows/html/rnaseqGene.html

## A survey of best practices for RNA-seq data analysis




**参考**

- https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

❖RNA-seq

❖CHIP-seq

❖ATAC-seq

# ChIP-seq作业

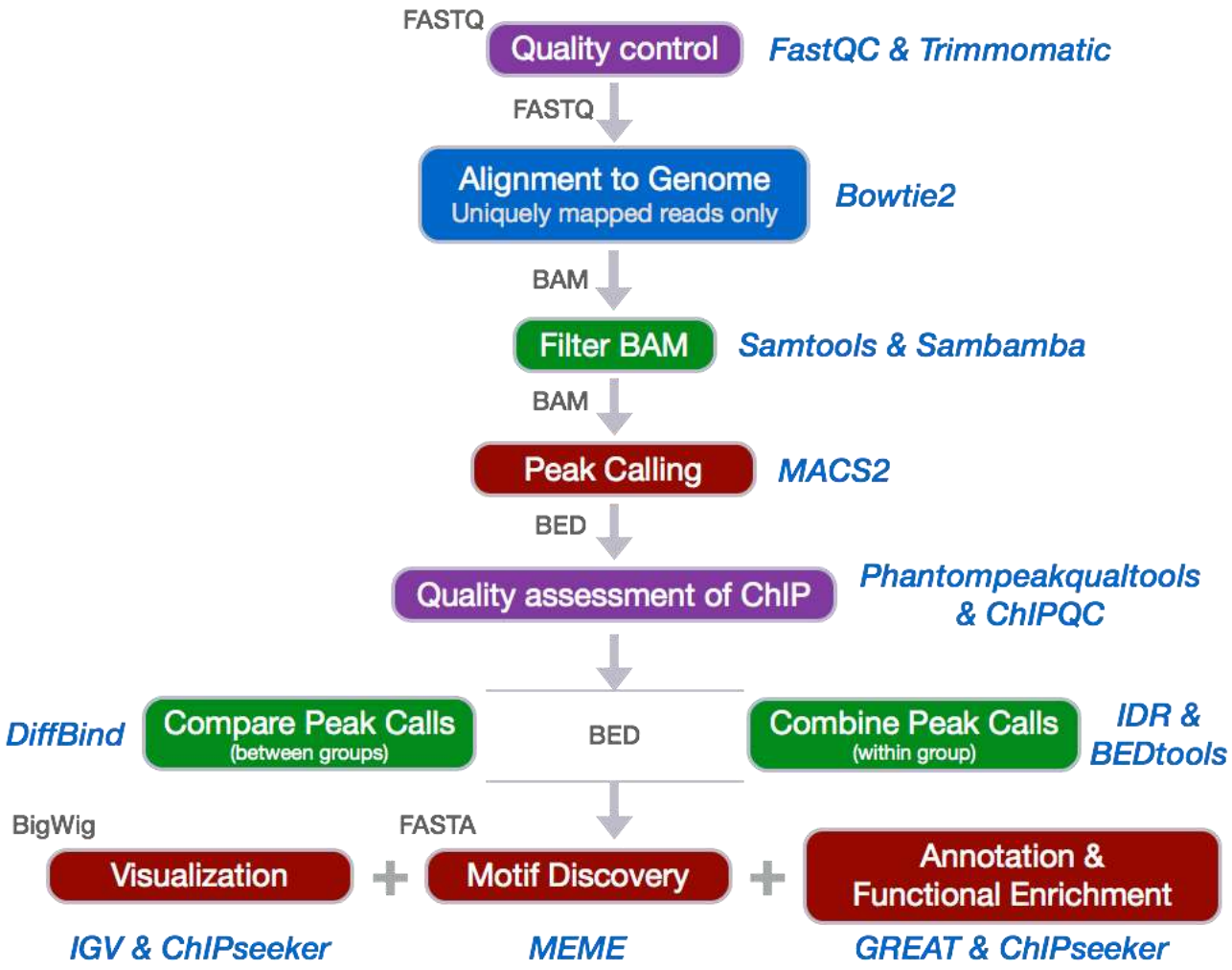
**目标**：利用公开的数据，完成一项ChIP-seq的分析
**数据**：https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE39241. SRR号：FOXA1（SRR520342, SRR520343, SRR520344）, Input（SRR520348）

**数据nebula地址**： /home/bioinfo2023/bioclass2023/homework/CHIP-seq/01.raw_data

**要求**：
1. 从原始测序数据开始（fastq），包含以下分析内容（总分10分）：
    - 质量控制（1分）
    - 数据比对（1分）
    - 去除PCR重复（1分）
    - Peak Calling+选择top2000富集的peak（2分）
    - Motif search（2分）
    - IGV可视化（包含质量较好的和较差的peak作为对比）（3分）

3.在2023年xx月xx日之前，将PPT和代码发送到邮箱：组长为周一班的 请发送邮件至liuk0617@mail.ustc.edu.cn 周三班发至fang0426@mail.ustc.edu.cn，邮件名和作业压缩包命名一致。
4.邮件主题：生物信息学ChIP-seq数据分析实践作业+组长学号+组长姓名
5.PPT第一页务必说明组员姓名学号和分工

# 常用工具




- **FastQC**
  https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- **Bowtie2**
  http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
- **MACS2**
  https://github.com/macs3-project/MACS
- **MEME**
  https://meme-suite.org/meme/
- **HOMER**
  http://homer.ucsd.edu/homer/motif/
- **IGV**
  https://igv.org/

# 参考资料

## Introduction to ChIP-seq using high performance computing

- https://github.com/hbctraining/Intro-to-ChIPseq/tree/master/lessons

## ChIP-seq-analysis

- https://github.com/crazyhottommy/ChIP-seq-analysis

- RNA-seq
- CHIP-seq
- ATAC-seq

# ATAC-seq作业

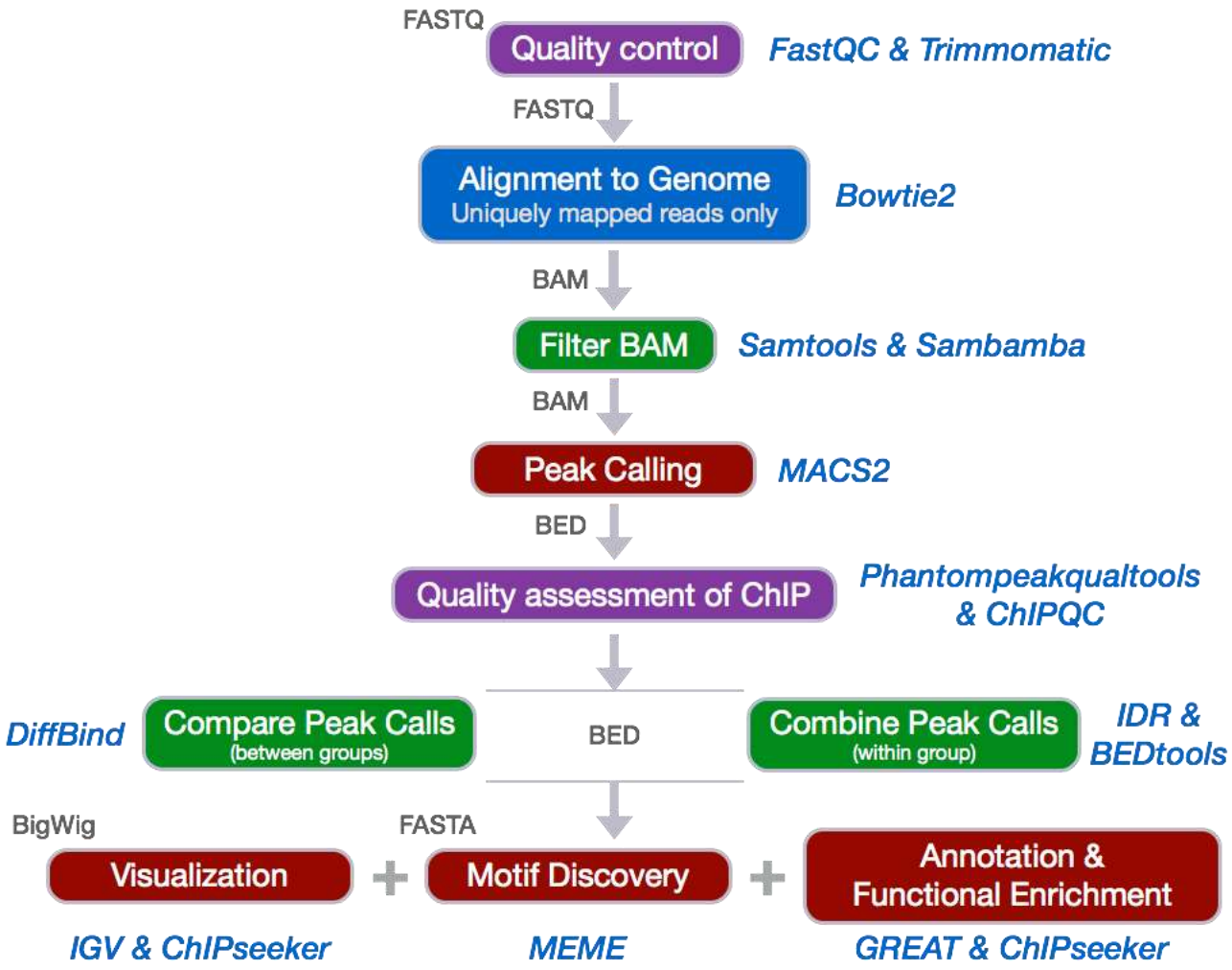
**目标：**利用公开的数据，完成一项ATAC-seq的分析

**数据：** 人类的单核细胞 (RAM007, RAM009)；T细胞 (RATA045, RATA046)
**nebula地址：** /home/bioinfo2023/bioclass2023/homework/ATAC-seq/01.raw_data
**要求：**
1. 从原始测序数据开始（fastq），包含以下分析内容（总分10分）：

- 成功安装ATAC-pipe （1分）
- 质量控制（TSS富集图+片段分布图+QC表格）（2分）
- 差异分析+绘制heatmap图（2分）
- 差异peak的GO分析（2分）
- Motif search（1分）
- IGV可视化（展示组间差异的peak例子）（2分）

3. 在2023年xx月xx日之前，将PPT和代码发送到邮箱：组长为周一班的 请发送邮件至 liuk0617@mail.ustc.edu.cn 周三班发至fang0426@mail.ustc.edu.cn，邮件名和作业压缩包命名一致。
4.邮件主题：生物信息学ATAC-seq数据分析实践作业+组长学号+组长姓名
5.PPT第一页务必说明组员姓名学号和分工

# 常用工具




- **ATAC-pipe**
  https://github.com/QuKunLab/ATAC-pipe
- **HOMER**
  http://homer.ucsd.edu/homer/motif/
- **Cluster3.0**
  http://bonsai.hgc.jp/~mdehoon/software/cluster/
- **TreeView**
  https://sourceforge.net/projects/jtreeview/
- **IGV**
  https://igv.org/

# 参考资料

**ATAC-pipe: general analysis of genome-wide chromatin accessibility**

- https://academic.oup.com/bib/article/20/5/1934/5047123?login=true

# 课程参考资料WEB

### 助教实践PPT

| 日期 | 课件 |
| --- | --- |
| 2023-09-18 | linux_入门 |
| 2023-09-25 | 测序文件与bowtie2 |
| 2023-10-09 | RNA-seq示例　下游代码示例 |
| 2023-10-23 | 数据库介绍(上)　数据库介绍(下) |

### 课程大作业

1.请严格遵守作业说明文档的要求完成作业，包括但不限于作业的提交方式以及文档的命名方式

2.提交作业时，组长为周一班的 请发送邮件至 liuk0617@mail.ustc.edu.cn 周三班发至 fang0426@mail.ustc.edu.cn，邮件名和作业压缩包命名一致。

3.请务必在截止日期之前提交作业，时间以收到邮件的时间为准

4.邮箱有自动回复，请确认自动回复以保证正常提交了作业。

5.禁止任何形式的抄袭。

| 日期 | 作业 | 截止日期 |
| --- | --- | --- |
| / | / | / |
| / | / | / |

### 大作业参考代码教程

| | |
| --- | --- |
| RNA-seq | 参考代码教程 |
| ATAC-seq | 参考代码教程 |

### 考试信息

| 内容 | 说明 |
| --- | --- |
| 考试答疑 | 待定 |
| 考试时间 | 待定 |
| 考试地点 | 待定 |
| 考试形式 | 一页A4纸(可正反) 半开卷考试 |

https://ustc-fmh.github.io/