##写在前面##

这里是CUT&Tag Tutorial的Tutorial，本教程旨在帮助对计算机知识不太熟悉的同学们理解并掌握CUT&Tag技术。通过详细的步骤说明和实际操作示例，我们希望能够让每一位读者顺利完成这一实验，并且在过程中提高对相关工具和命令行操作的熟悉度。本教程不仅涵盖了基本的实验步骤，还包括了常见问题的解决方法和实际操作中的注意事项。希望通过本教程，大家可以自信地应用CUT&Tag技术进行实验研究。

使用过程中结合tutorial一起看，哪里不懂就找对应的区域

* 首先说一下Rstudio需要用到的library，在console里输入以下几行

library(dplyr)

library(ggplot2)

library(ggpubr)

library(viridis)

library(GenomicRanges)

// library(chromVAR) 这个后面用到的时候会细说

* 再介绍需要用到的conda tools

Bowtie2 (version >= 2.3.4.3)

samtools (version >= 1.10)

bedtools (version >= 2.29.1)

Picard (version >= 2.18.29)// 最新picard识别不了数据,具体下面说

SEACR (version >= 1.3)

deepTools (version >= 2.0)// deepTools内function也有问题, 下面说

* 这里具体说一下$projPath。首先创建一个名为cuttag\_project的directory/目录，后续所有的内容都会存在cuttag\_project这个目录里面，所以我们可以设置一个projPath直接导向cuttag\_project。

**mkdir -p ~/cuttag\_project //在根目录创建(也可以在其他地方，按自己喜好来)**

**(\*)cd cuttag\_project && pwd //进入cuttag\_project并查看路径**

**projPath=”你的cuttag\_project路径” //ex: projPath”/Users/tonyzh/cuttag\_project”**

注：(\*)表示可省略

* Tutorial最开始时介绍了很多packages，很多同学可能一看到包就头疼，不会下，不会装，但是这些都不是问题，因为我们不需要一步到位将全部包都下载下来，**只需要先下载conda**，**其他包**在后续过程中**需要时再下**。
* 首先，CONDA！！

<https://docs.anaconda.com/miniconda/#quick-command-line-install>

此链接通过终端terminal下载，将页面拉到最下面，选择你的操作系统，再将页面内提供的几句代码复制到你的终端里。

（\*）这里要注意，此教程通篇使用linux环境指令，不要用错环境了哦

在完成rm .sh步骤后需要将miniconda初始化，但是链接为两个shell提供了初始化指令，bash与zsh，我们需要确定自己系统使用的是哪个shell

**Echo $SHELL**

后再选择。

初始化后可通过 conda --version来查看conda是否已正确安装，**如果**弹出bash/zsh：conda command not found之类的error：

**echo ‘export PATH=”$HOME/miniconda3/bin:$PATH”’ >> ~/.zshrc**

**source ~/.zshrc**

bash将zshrc替换为bashrc。Ex: ~/.bashrc

* 至此，miniconda3下载好了，接下来下载并配置conda

**conda install conda //下载conda**

**conda config --append channels conda-forge //添加conda-forge channel**

**conda config --append channels bioconda //添加bioconda channel**

* 导入data

可以看到我们一共有六组数据，分别是H3K27me3的rep1rep2，H3K4me3的rep1rep2，和IgG的rep1和rep2. 教程里面只提供了IgG\_rep2的command line下载方法，其他数据需要

<https://www.ebi.ac.uk/ena/browser/home>

进入该网址，在Enter accession处复制粘贴SRA entry：SRX########

往下翻可以看到下载数据的位置，我们有两种选择，通过wget command或者直接浏览器下载。

wget和chrome真不一定谁快，建议两边同时，节省点时间

Wget: **wget -O $projPath/data/#data name#//R#.fastq.gz ftp://#links below#**

**Ex: wget -O $projPath/data/IgG\_rep2/IgG\_rep2\_R1\_001.fastq.gz ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR875/001/SRR8754611/SRR8754611\_1.fastq.gz**

（\*）实际的操作其实是下载TSV格式的download report再通过report内提供的链接wget，为提高效率我将所有链接都贴在下面了，可以直接使用

（\*）将数据下载到data 里，需要提前创建路径，data内数据作为备份是不直接使用的，要确保data内数据纯净 ex: **mkdir -p $projPath/data/K27me3\_rep2**

Chrome：直接下载就好

K27me3\_rep1: SRX8754646

[ftp.sra.ebi.ac.uk/vol1/fastq/SRR122/017/SRR12246717/SRR12246717\_1.fastq.gz](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR122/017/SRR12246717/SRR12246717_1.fastq.gz) R1

[ftp.sra.ebi.ac.uk/vol1/fastq/SRR122/017/SRR12246717/SRR12246717\_2.fastq.gz](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR122/017/SRR12246717/SRR12246717_2.fastq.gz) R2

K27me3\_rep2: SRX7713678 [ftp.sra.ebi.ac.uk/vol1/fastq/SRR110/040/SRR11074240/SRR11074240\_1.fastq.gz](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR110/040/SRR11074240/SRR11074240_1.fastq.gz) R1

[ftp.sra.ebi.ac.uk/vol1/fastq/SRR110/040/SRR11074240/SRR11074240\_2.fastq.gz](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR110/040/SRR11074240/SRR11074240_2.fastq.gz) R2

K4me3\_rep1: SRX7713692

[ftp.sra.ebi.ac.uk/vol1/fastq/SRR110/054/SRR11074254/SRR11074254\_1.fastq.gz](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR110/054/SRR11074254/SRR11074254_1.fastq.gz) R1

[ftp.sra.ebi.ac.uk/vol1/fastq/SRR110/054/SRR11074254/SRR11074254\_2.fastq.gz](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR110/054/SRR11074254/SRR11074254_2.fastq.gz) R2

K4me3\_rep2: SRX7713696

[ftp.sra.ebi.ac.uk/vol1/fastq/SRR110/058/SRR11074258/SRR11074258\_1.fastq.gz](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR110/058/SRR11074258/SRR11074258_1.fastq.gz) R1

[ftp.sra.ebi.ac.uk/vol1/fastq/SRR110/058/SRR11074258/SRR11074258\_2.fastq.gz](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR110/058/SRR11074258/SRR11074258_2.fastq.gz) R2

IgG\_rep1: SRX8468909

[ftp.sra.ebi.ac.uk/vol1/fastq/SRR119/024/SRR11923224/SRR11923224\_1.fastq.gz](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR119/024/SRR11923224/SRR11923224_1.fastq.gz) R1

[ftp.sra.ebi.ac.uk/vol1/fastq/SRR119/024/SRR11923224/SRR11923224\_2.fastq.gz](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR119/024/SRR11923224/SRR11923224_2.fastq.gz) R2

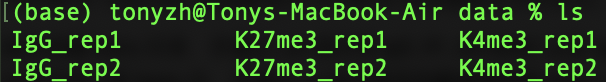
IgG\_rep2: SRX5545346

[ftp.sra.ebi.ac.uk/vol1/fastq/SRR875/001/SRR8754611/SRR8754611\_1.fastq.gz](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR875/001/SRR8754611/SRR8754611_1.fastq.gz) R1\_001

[ftp.sra.ebi.ac.uk/vol1/fastq/SRR875/001/SRR8754611/SRR8754611\_2.fastq.gz](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR875/001/SRR8754611/SRR8754611_2.fastq.gz) R1\_002

[ftp.sra.ebi.ac.uk/vol1/fastq/SRR875/002/SRR8754612/SRR8754612\_1.fastq.gz](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR875/002/SRR8754612/SRR8754612_1.fastq.gz) R2\_001

[ftp.sra.ebi.ac.uk/vol1/fastq/SRR875/002/SRR8754612/SRR8754612\_2.fastq.gz](ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR875/002/SRR8754612/SRR8754612_2.fastq.gz) R2\_002



* 2

首先进行2.2. Merge technical replicates/lanes if needed

histName可以无视, 但需要明白histName在后续指令中的含义, 他只是一个变量

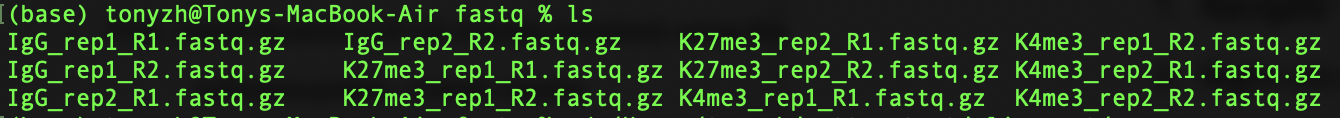
首先创建fastq目录，并通过cat指令将IgG\_R1\_001&002以及IgG\_R2\_001&002合并并导入到fastq内，并通过cp指令将其他数据复制到fastq内

**mkdir -p ${projPath}/fastq**

**下面cat可以用我的也可以用tutorial里给的，一样的**

**cat ${projPath}/data/IgG\_rep1/IgG\_rep1\_R1\_001.fastq.gz ${projpath}/data/IgG\_rep1/IgG\_rep1\_R1\_002.fastq.gz > ${projPath}/fastq/IgG\_rep1\_R1.fastq.gz**

**cp ${projpath}/data/K27me3\_rep1/K27me3\_rep1\_R1.fastq.gz ${projPath}/fastq/K27me3\_rep1\_R1.fastq.gz**



* 然后再回到上面

2.2.1 Obtain FastQC直接照抄

2.2.2就是把histName替换成想用的数据名称，其他照抄

* 3.1.1
* 先下载bowtie2

conda install -c bioconda bowtie2

* 先无视cores和ref，先将四个目录建好，设好projPath的前提下照抄就行
* 在四个路径的基础上再创建一个bowtie2Index的目录

**mkdir -p ${projPath}/bowtie2Index**

* 重点提一下bowtie2-build，bowtie2-build将hg38基因组转换成bowtie2可以识别的6个文件…..具体如下

<https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#the-bowtie2-build-indexer>

“bowtie2-build builds a Bowtie index from a set of DNA sequences. bowtie2-buildoutputs a set of 6 files with suffixes .1.bt2, .2.bt2, .3.bt2, .4.bt2, .rev.1.bt2, and .rev.2.bt2. In the case of a large index these suffixes will have a bt2ltermination. These files together constitute the index: they are all that is needed to align reads to that reference. The original sequence [FASTA](https://en.wikipedia.org/wiki/FASTA) files are no longer used by Bowtie 2 once the index is built.”

* 回到tutorial本身，了解bowtie2-build原理之后可以着手将hg38.fa转成.bt2

**bowtie2-build ${projPath}/hg38.fa ${projPath}/bowtie2Index/hg38**

你问我hg38.fa在哪？

**wget -P $projPath https://hgdownload.cse.ucsc.edu/goldenpath/hg38/bigZips/hg38.fa.gz**

下载好之后hg38.fa就会在cuttag\_project里了，别忘了bowtie2-build哦

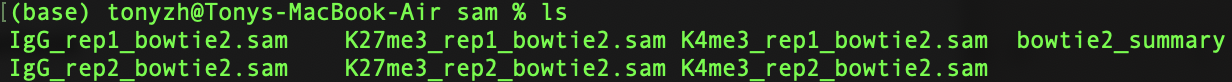
* 现在可以设置cores和ref了

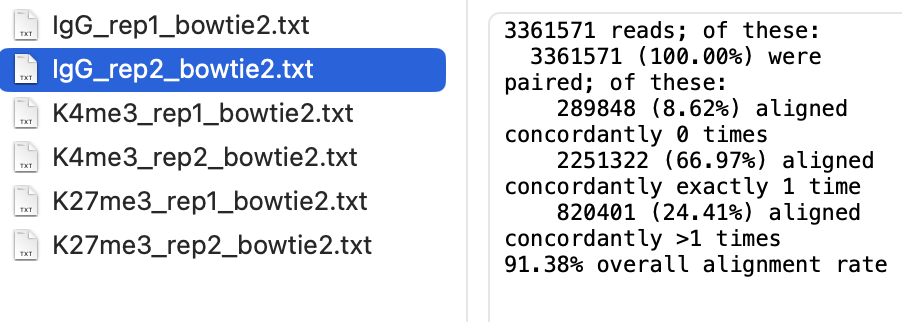
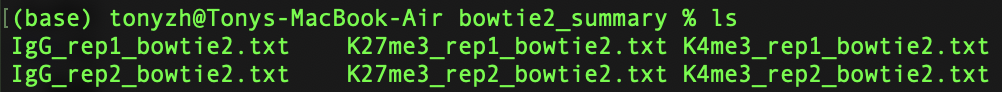
**cores=8**

**ref=“/Users/tonyzh/cuttag\_project/bowtie2Index/hg38”**

（\*）细心的同学会发现bowtie2Index里面没有叫hg38的文件啊，这就是下面那一大串中“-x ${ref}” 的 -x的作用：-x <bt2-idx>，-x后面跟所有.bt2文件的basename，告诉bowtie2应该用哪种指标，比如-x hg38，bowtie2就知道应该用hg38的指标来进行对准，从而他可以读取全部以hg38为basename的文件

* 然后就可以把这一长串输入进去了，注意typo，路径，output如下





* 3.1.2
* 这一步其实和上一步差不多，只不过需要转换的基因组变成了E.Coli，E.Coli下载

<https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_002853715.1/>

* 转好之后设置spikeInRef和chromSize

chromSize还是通过wget来下载

**wget -P $projPath/hg38.chrom.sizes https://hgdownload.cse.ucsc.edu/goldenpath/hg38/bigZips/hg38.chrom.sizes**

* 接着往下，bowtie2一长串
* 接着往下seqDepthDouble那里，注意seqDepthDouble=后面是`(1左边的)

注意这里三行是一起的，一组基因进行完才能更换下一组基因

scale\_factor一定要记录下来，我的数据可能不太一样，这里要记录你的

K27me3\_rep1: 3.69685767097966728280

K27me3\_rep2: 4.42869796279893711248

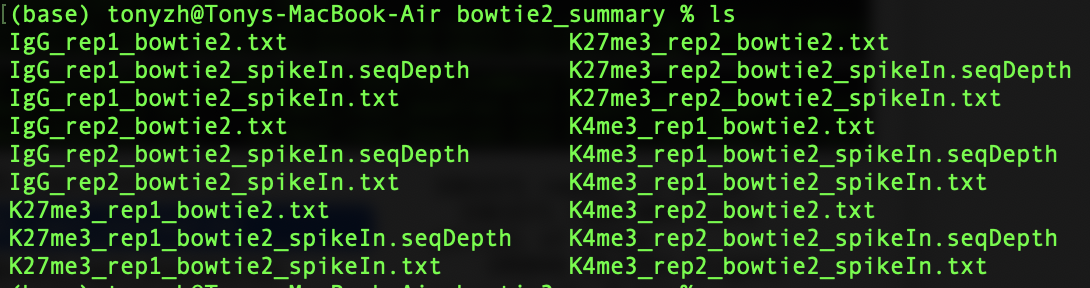
K4me3\_rep1: 4.71698113207547169811

K4me3\_rep2: 2.28675966155957008918

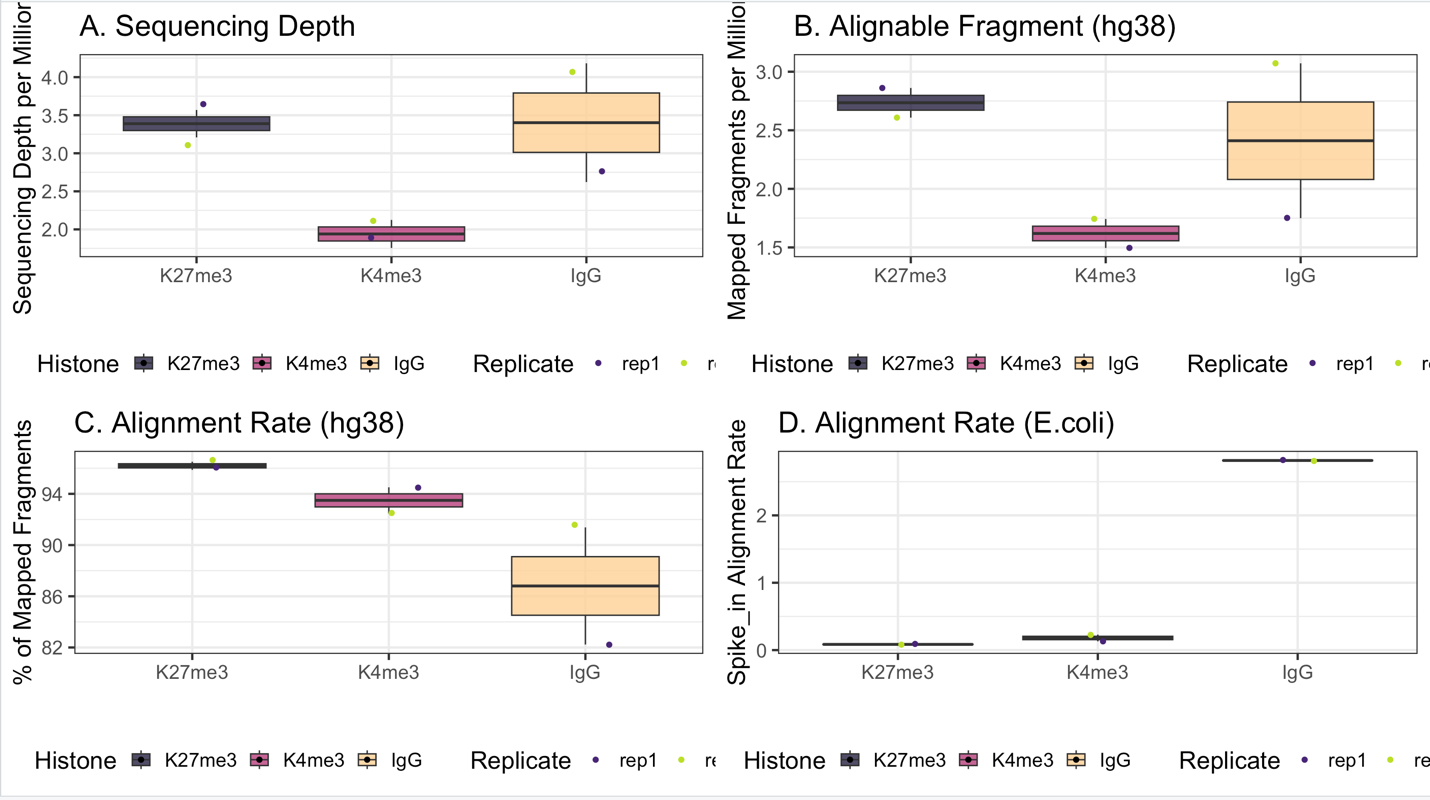
IgG\_rep1: .16649462222370217441

IgG\_rep2: .10593444776372380770

* Output如下



* 3.2.1:
  + projPath将他的路径替换成你的路径
  + sampleList：将如果只是在过tutorial那么可以直接复制，如果在用自己的数据，那要将你的数据的名字放进来
  + histList保持一样的pattern就好
* 将3.2.x全部敲到同一个Rscript里面，其他不用变



* 3.3 picard v3.x有问题，需要下载v2.x，我下的是cuttag最低需求，目前无法确定是否相比新版本有功能上的缺失但是v2.x确实可以运行

**conda install -c bioconda picard=“2.18.29” //这是最低可用版本, 可酌情提高版本**

* 3.3要注意picardCMD需要严格按照我下面说的来
* 查询picard.jar的路径

find $(conda info --base) -name “picard.jar”

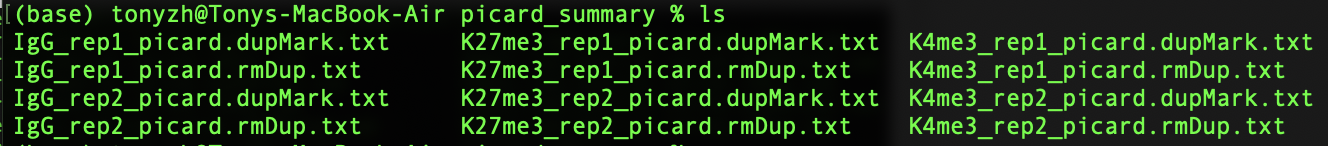
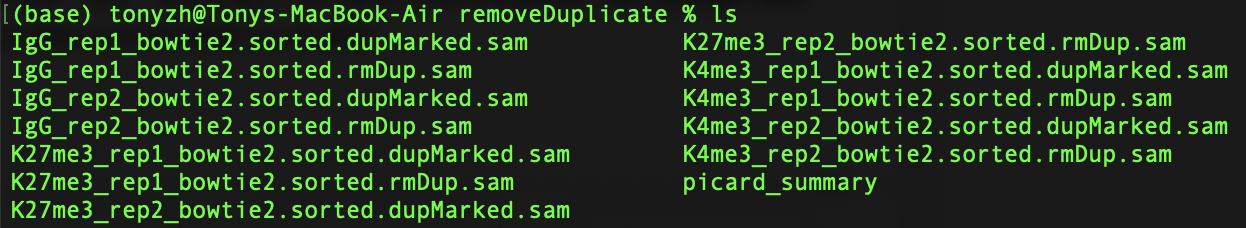
* 将产出的路径复制并附给picardCMD

picardCMD=”/路径/picard.jar”

* 注意tutorial里的picardCMD包含了两个指令java -jar，但我们的picardCMD里并没有这两个指令，所以在进行3.3linux command阶段需要手动在前面输入java -jar 然后再$picardCMD 再继续tutorial

// ex: java -jar $picardCMD SortSam / MarkDuplicate ……

* output如下



A diagram of a variety of colors

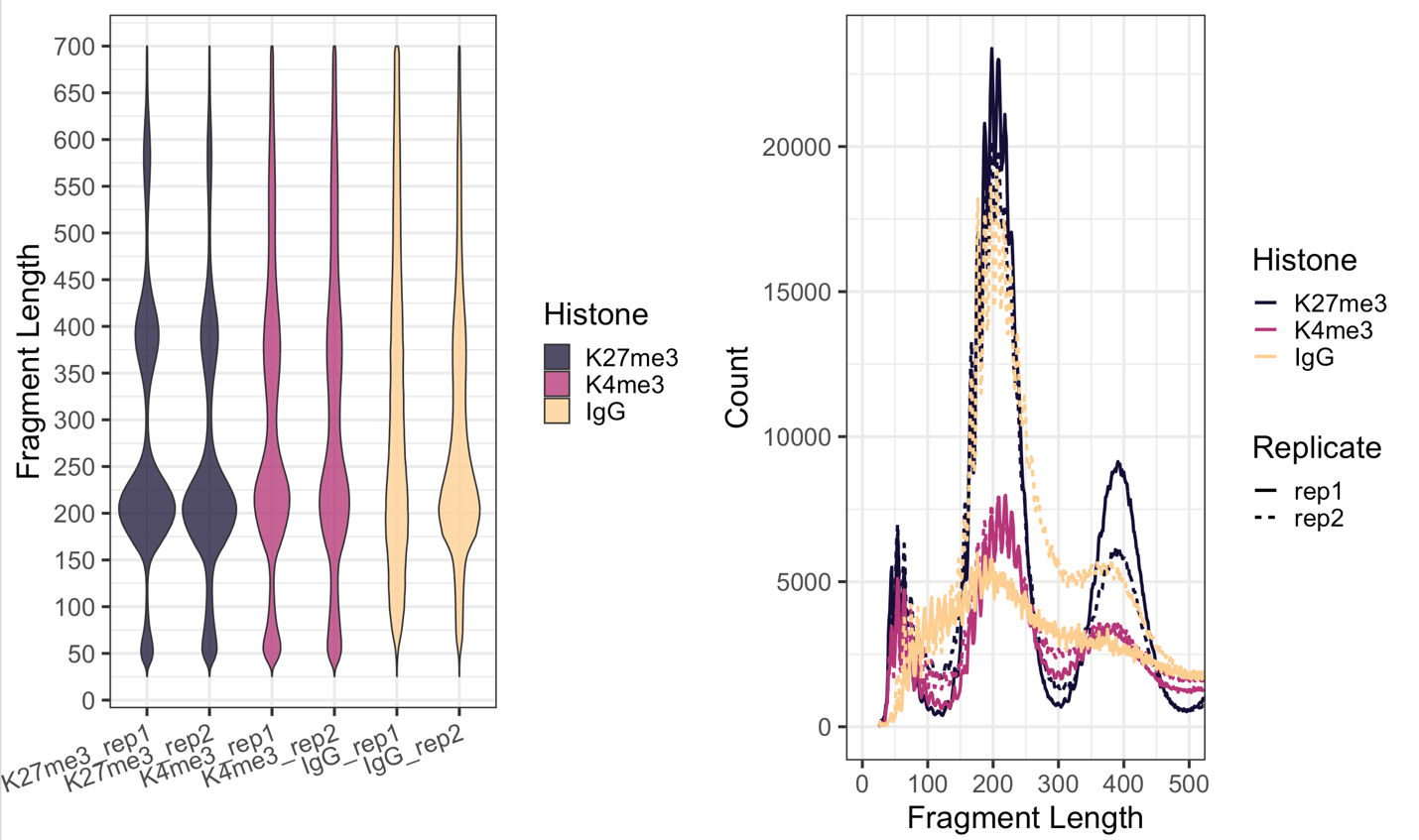
Description automatically generated with medium confidence

* 3.4
* 下载samtools

**conda install -c bioconda samtools**

**samtools --version**

* 照抄，注意typo

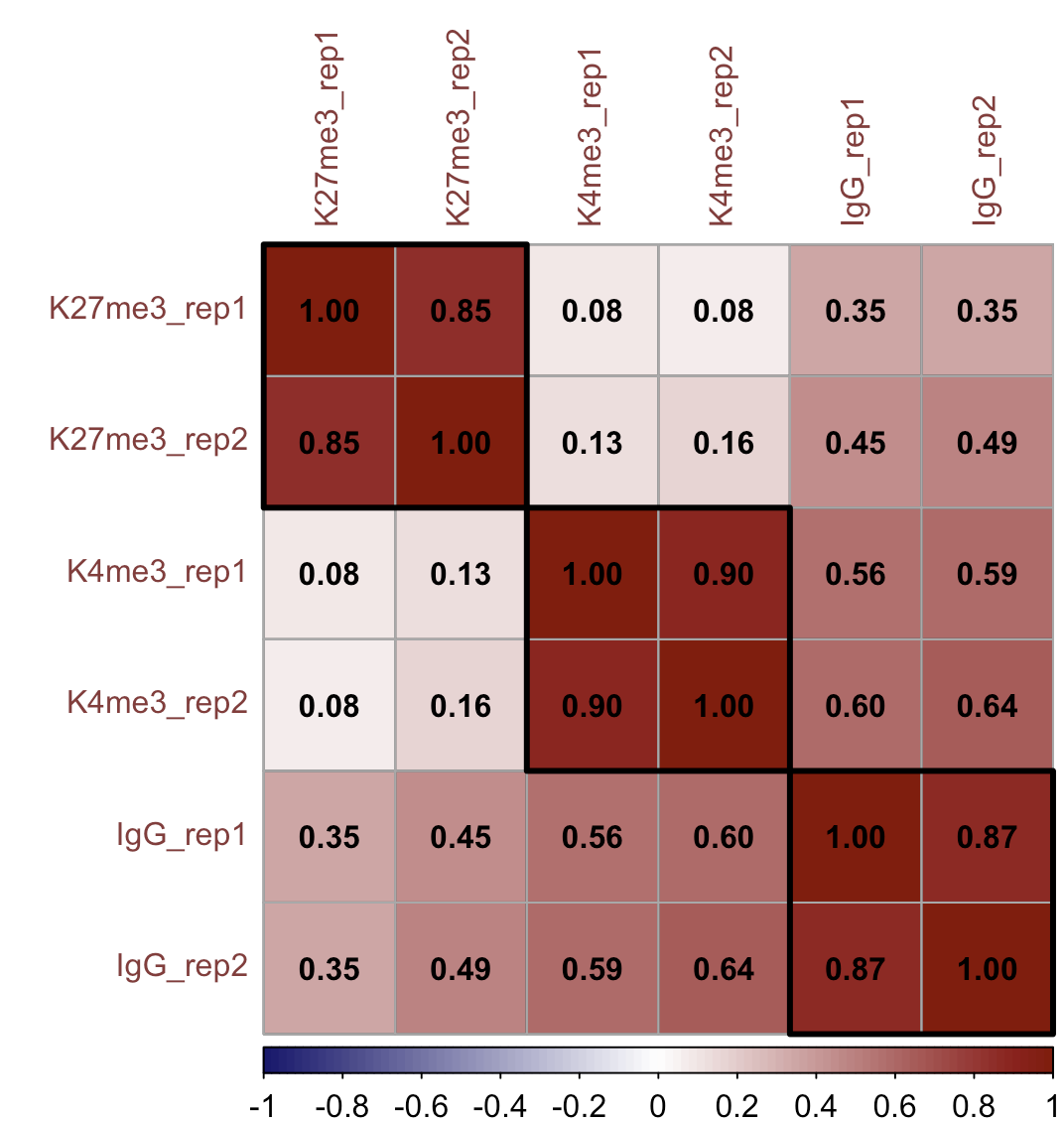


* 4.1
* 照抄，注意typo
* 4.2
* 下载bedtools

**conda install -c bioconda bedtools**

**bedtools --version**

* 4.3



* 5
* 照抄

A graph of a normalization of a fragment count

Description automatically generated with medium confidence

* 6
* 下载seacr

**conda install -c bioconda seacr**

* R下载chromVAR

新建R文档，输入以下内容

if (!require("BiocManager", quietly = TRUE))

install.packages("BiocManager")

BiocManager::install("chromVAR")

再在console里library(chromVAR)

* SEACR\_x.x.sh这个要找一下，但是一般都在miniconda3/lib/SEACR\_x.x.sh，反正找到了之后pwd，把返还的路径/SEACR\_x.x.sh赋给seacr
* histControl

A screenshot of a graph

Description automatically generated

* 7：先下载deeptools这个工具和上面说的picard有很多相似之处--都很不稳定

不多说直接**conda install -c bioconda deeptools**

Conda下不了的话 **pip install deeptools**

* 7.2. 直接照抄就行
* 7.2.1 需要用到hg38\_annotation file，从illumina下载[Illumina Exome Panel v1.2 (CEX) BED File (hg38)](https://support.illumina.com.cn/content/dam/illumina-support/documents/downloads/productfiles/trusight/hg38/Illumina_Exome_TargetedRegions_v1.2.hg38.bed" \t "_blank) 网址：  
  <https://support.illumina.com.cn/downloads/enrichment-bed-files-hg38.html>
* 再下载bedops，可以将GTF/GFF转成bed格式

<https://bedops.readthedocs.io/en/latest/content/installation.html#installation>可以选下package也可以via source code

下载并安装好后可以通过**convert2bed --version** 来确认

* 将下载好的GFF/GTF转成bed

**convert2bed -i gff --do-not-sort < /路径/ncbi\_dataset.gff > /路径/hg38\_gene.bed**

**convert2bed -I gtf --do-not-sort < /路径/ncbi\_dataset.gff > /路径/hg38\_gene.bed**

* computeMatrix一长串

Debug

4.2

## Filter and keep the mapped read pairs

samtools view -bS -F 0x04 $projPath/alignment/sam${histName}\_bowtie2.qualityScore$minQualityScore.sam >$projPath/alignment/bam/${histName}\_bowtie2.mapped.bam

if did 4.1: 只需要改这一步 其他不用动

## Filter and keep the mapped read pairs

samtools view -bS -F 0x04 $projPath/alignment/sam/${histName}\_bowtie2.sam >$projPath/alignment/bam/${histName}\_bowtie2.mapped.bam

## Convert into bed file format

bedtools bamtobed -i $projPath/alignment/bam/${histName}\_bowtie2.mapped.bam -bedpe >$projPath/alignment/bed/${histName}\_bowtie2.bed

## Keep the read pairs that are on the same chromosome and fragment length less than 1000bp.

awk '$1==$4 && $6-$2 < 1000 {print $0}' $projPath/alignment/bed/${histName}\_bowtie2.bed >$projPath/alignment/bed/${histName}\_bowtie2.clean.bed

## Only extract the fragment related columns

cut -f 1,2,6 $projPath/alignment/bed/${histName}\_bowtie2.clean.bed | sort -k1,1 -k2,2n -k3,3n >$projPath/alignment/bed/${histName}\_bowtie2.fragments.bed

3.3 & 5.1

find $(conda info --base) -name "picard.jar" //找picard.jar地址

picardCMD不能有java -jar，picardCMD只存picard.jar地址,手打java -jar -> java -jar $picardCMD …

picard更新了，所以tutorial里面给的代码是过时了的，下面的是更新后的

## sort by coordinate

SortSam -I /Users/tonyzh/cuttag\_project/alignment/sam/K27me3\_bowtie2.sam -O /Users/tonyzh/cuttag\_project/alignment/sam/K27me3\_bowtie2.sorted.sam -SORT\_ORDER coordinate

第一步SortSam没有问题，

## mark duplicates 第二步问题很大，MarkDuplicate识别不了数据，需要AddOrReplaceReadGroups

java -jar $picardCMD MarkDuplicates -I /Users/tonyzh/cuttag\_test/alignment/sam/K27me3\_rep1\_bowtie2.sorted.sam -O /Users/tonyzh/cuttag\_test/alignment/removeDuplicate/K27me3\_rep1\_bowtie2.sorted.dupMarked.sam -METRICS\_FILE /Users/tonyzh/cuttag\_test/alignment/removeDuplicate/picard\_summary/K27me3\_rep1\_picard.dupMark.txt

15:33:40.632 INFO NativeLibraryLoader - Loading libgkl\_compression.dylib from jar:file:/Users/tonyzh/miniconda3/pkgs/picard-3.2.0-hdfd78af\_0/share/picard-3.2.0-0/picard.jar!/com/intel/gkl/native/libgkl\_compression.dylib

15:33:40.658 WARN NativeLibraryLoader - Unable to load libgkl\_compression.dylib from native/libgkl\_compression.dylib (/private/var/folders/jp/t28g7vnx1k161cpfpdty8xnw0000gn/T/libgkl\_compression5281022966999022486.dylib: dlopen(/private/var/folders/jp/t28g7vnx1k161cpfpdty8xnw0000gn/T/libgkl\_compression5281022966999022486.dylib, 0x0001): tried: '/private/var/folders/jp/t28g7vnx1k161cpfpdty8xnw0000gn/T/libgkl\_compression5281022966999022486.dylib' (mach-o file, but is an incompatible architecture (have 'x86\_64', need 'arm64')), '/System/Volumes/Preboot/Cryptexes/OS/private/var/folders/jp/t28g7vnx1k161cpfpdty8xnw0000gn/T/libgkl\_compression5281022966999022486.dylib' (no such file), '/private/var/folders/jp/t28g7vnx1k161cpfpdty8xnw0000gn/T/libgkl\_compression5281022966999022486.dylib' (mach-o file, but is an incompatible architecture (have 'x86\_64', need 'arm64')))

15:33:40.659 INFO NativeLibraryLoader - Loading libgkl\_compression.dylib from jar:file:/Users/tonyzh/miniconda3/pkgs/picard-3.2.0-hdfd78af\_0/share/picard-3.2.0-0/picard.jar!/com/intel/gkl/native/libgkl\_compression.dylib

15:33:40.661 WARN NativeLibraryLoader - Unable to load libgkl\_compression.dylib from native/libgkl\_compression.dylib (/private/var/folders/jp/t28g7vnx1k161cpfpdty8xnw0000gn/T/libgkl\_compression5265586390462787491.dylib: dlopen(/private/var/folders/jp/t28g7vnx1k161cpfpdty8xnw0000gn/T/libgkl\_compression5265586390462787491.dylib, 0x0001): tried: '/private/var/folders/jp/t28g7vnx1k161cpfpdty8xnw0000gn/T/libgkl\_compression5265586390462787491.dylib' (mach-o file, but is an incompatible architecture (have 'x86\_64', need 'arm64')), '/System/Volumes/Preboot/Cryptexes/OS/private/var/folders/jp/t28g7vnx1k161cpfpdty8xnw0000gn/T/libgkl\_compression5265586390462787491.dylib' (no such file), '/private/var/folders/jp/t28g7vnx1k161cpfpdty8xnw0000gn/T/libgkl\_compression5265586390462787491.dylib' (mach-o file, but is an incompatible architecture (have 'x86\_64', need 'arm64')))

[Wed Aug 28 15:33:40 CST 2024] MarkDuplicates --INPUT /Users/tonyzh/cuttag\_test/alignment/sam/K27me3\_rep1\_bowtie2.sorted.sam --OUTPUT /Users/tonyzh/cuttag\_test/alignment/removeDuplicate/K27me3\_rep1\_bowtie2.sorted.dupMarked.sam --METRICS\_FILE /Users/tonyzh/cuttag\_test/alignment/removeDuplicate/picard\_summary/K27me3\_rep1\_picard.dupMark.txt --MAX\_SEQUENCES\_FOR\_DISK\_READ\_ENDS\_MAP 50000 --MAX\_FILE\_HANDLES\_FOR\_READ\_ENDS\_MAP 8000 --SORTING\_COLLECTION\_SIZE\_RATIO 0.25 --TAG\_DUPLICATE\_SET\_MEMBERS false --REMOVE\_SEQUENCING\_DUPLICATES false --TAGGING\_POLICY DontTag --CLEAR\_DT true --DUPLEX\_UMI false --FLOW\_MODE false --FLOW\_DUP\_STRATEGY FLOW\_QUALITY\_SUM\_STRATEGY --USE\_END\_IN\_UNPAIRED\_READS false --USE\_UNPAIRED\_CLIPPED\_END false --UNPAIRED\_END\_UNCERTAINTY 0 --UNPAIRED\_START\_UNCERTAINTY 0 --FLOW\_SKIP\_FIRST\_N\_FLOWS 0 --FLOW\_Q\_IS\_KNOWN\_END false --FLOW\_EFFECTIVE\_QUALITY\_THRESHOLD 15 --ADD\_PG\_TAG\_TO\_READS true --REMOVE\_DUPLICATES false --ASSUME\_SORTED false --DUPLICATE\_SCORING\_STRATEGY SUM\_OF\_BASE\_QUALITIES --PROGRAM\_RECORD\_ID MarkDuplicates --PROGRAM\_GROUP\_NAME MarkDuplicates --READ\_NAME\_REGEX <optimized capture of last three ':' separated fields as numeric values> --OPTICAL\_DUPLICATE\_PIXEL\_DISTANCE 100 --MAX\_OPTICAL\_DUPLICATE\_SET\_SIZE 300000 --VERBOSITY INFO --QUIET false --VALIDATION\_STRINGENCY STRICT --COMPRESSION\_LEVEL 5 --MAX\_RECORDS\_IN\_RAM 500000 --CREATE\_INDEX false --CREATE\_MD5\_FILE false --help false --version false --showHidden false --USE\_JDK\_DEFLATER false --USE\_JDK\_INFLATER false

[Wed Aug 28 15:33:40 CST 2024] Executing as tonyzh@Tonys-MacBook-Air.local on Mac OS X 13.4 aarch64; OpenJDK 64-Bit Server VM 22.0.1+8; Deflater: Jdk; Inflater: Jdk; Provider GCS is available; Picard version: Version:3.2.0-1-g3948afb6b

INFO 2024-08-28 15:33:40 MarkDuplicates Start of doWork freeMemory: 25798224; totalMemory: 35651584; maxMemory: 2147483648

INFO 2024-08-28 15:33:40 MarkDuplicates Reading input file and constructing read end information.

INFO 2024-08-28 15:33:40 MarkDuplicates Will retain up to 7780737 data points before spilling to disk.

WARNING 2024-08-28 15:33:40 AbstractOpticalDuplicateFinderCommandLineProgram A field field parsed out of a read name was expected to contain an integer and did not. Read name: SRR12246717.2375893. Cause: String 'SRR12246717.2375893' did not start with a parsable number.

[Wed Aug 28 15:33:40 CST 2024] picard.sam.markduplicates.MarkDuplicates done. Elapsed time: 0.00 minutes.

Runtime.totalMemory()=124780544

To get help, see http://broadinstitute.github.io/picard/index.html#GettingHelp

Exception in thread "main" java.lang.NullPointerException: Cannot invoke "htsjdk.samtools.SAMReadGroupRecord.getReadGroupId()" because the return value of "htsjdk.samtools.SAMRecord.getReadGroup()" is null

at picard.sam.markduplicates.MarkDuplicates.buildSortedReadEndLists(MarkDuplicates.java:558)

at picard.sam.markduplicates.MarkDuplicates.doWork(MarkDuplicates.java:270)

at picard.cmdline.CommandLineProgram.instanceMain(CommandLineProgram.java:281)

at picard.cmdline.PicardCommandLine.instanceMain(PicardCommandLine.java:105)

at picard.cmdline.PicardCommandLine.main(PicardCommandLine.java:115)

MarkDuplicates -I /Users/tonyzh/cuttag\_project/alignment/sam/K27me3\_bowite2.sorted.sam -O /Users/tonyzh/cuttag\_project/alignment/removeDuplicate/K27me3\_bowtie2.sorted.dupMarked.sam -METRICS\_FILE /Users/tonyzh/cuttag\_project/alignment/removeDuplicate/picard\_summary/K27me3\_picard.dupMark.txt

Pi card最终解决办法：下载2.x.x版本picard，完美解决

6

seacr=”/Users/tonyzh/miniconda3/bin/SEACR\_1.3.sh” //慎重，如果报错no such file or directory, 就手打地址

histName=“K27me3”

histControl=“IgG\_rep2”

% bash $seacr $projPath/alignment/bedgraph/K27me3\_bowtie2.fragments.normalized.bedgraph $projPath/alignment/bedgraph/IgG\_rep2\_bowtie2.fragments.normalized.bedgraph non stringent $projPath/peakCalling/SEACR/K27me3\_seacr\_control.peaks

Calling enriched regions with control file

Proceeding without normalization of control to experimental bedgraph

Using stringent threshold

Creating experimental AUC file: Mon Aug 26 09:52:35 CST 2024

/Users/tonyzh/miniconda3/bin/SEACR\_1.3.sh: line 102: ?W?gn??C??Z?.auc.bed: Illegal byte sequence

* password=`head /dev/urandom | base64 | tr -d '/+-=' | head -c 13; echo ''`
* password2=`head /dev/urandom | base64 | tr -d '/+-=' | head -c 13; echo ''`
* 将SEACR\_x.x.sh中56 57两行改成这样可以解决上述问题
* 我是macOS，seacr v1.3，因为他password生成是通过dev/urandom，所以可能其他OS就不会遇到这个问题

7

7.2.1:

cores=8

computeMatrix scale-regions -S $projPath/alignment/bigwig/K27me3\_rep1\_raw.bw \

$projPath/alignment/bigwig/K27me3\_rep2\_raw.bw \

$projPath/alignment/bigwig/K4me3\_rep1\_raw.bw \

$projPath/alignment/bigwig/K4me3\_rep2\_raw.bw \

-R $projPath/data/hg38\_gene/hg38\_gene.tsv \

--beforeRegionStartLength 3000 \

--regionBodyLength 5000 \

--afterRegionStartLength 3000 \

--skipZeros -o $projPath/data/hg38\_gene/matrix\_gene.mat.gz -p $cores

/\* Error:

computeMatrix scale-regions -S $projPath/alignment/bigwig/K27me3\_raw.bw -R $projPath/data/hg38\_gene/hg38\_gene.tsv --beforeRegionStartLength 3000 --regionBodyLength 5000 --afterRegionStartLength 3000 --skipZeros -o $projPath/data/hg38\_gene/matrix\_gene.mat.gz -p $cores

Traceback (most recent call last):

File "/Users/tonyzh/miniconda3/bin/computeMatrix", line 8, in <module>

sys.exit(main())

^^^^^^

File "/Users/tonyzh/miniconda3/lib/python3.12/site-packages/deeptools/computeMatrix.py", line 398, in main

hm.computeMatrix(scores\_file\_list, args.regionsFileName, parameters, blackListFileName=args.blackListFileName, verbose=args.verbose, allArgs=args)

File "/Users/tonyzh/miniconda3/lib/python3.12/site-packages/deeptools/heatmapper.py", line 252, in computeMatrix

res, labels = mapReduce.mapReduce([score\_file\_list, parameters],

^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

File "/Users/tonyzh/miniconda3/lib/python3.12/site-packages/deeptools/mapReduce.py", line 85, in mapReduce

bed\_interval\_tree = GTF(bedFile, defaultGroup=defaultGroup, transcriptID=transcriptID, exonID=exonID, transcript\_id\_designator=transcript\_id\_designator, keepExons=keepExons)

^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

File "/Users/tonyzh/miniconda3/lib/python3.12/site-packages/deeptoolsintervals/parse.py", line 591, in \_\_init\_\_

ftype = self.inferType(fp, line, labelColumn)

^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

File "/Users/tonyzh/miniconda3/lib/python3.12/site-packages/deeptoolsintervals/parse.py", line 166, in inferType

raise RuntimeError('{0} does not seem to be a recognized file type!'.format(self.filename))

RuntimeError: /Users/tonyzh/cuttag\_project/data/hg38\_gene/hg38\_gene.tsv does not seem to be a recognized file type! \*/

**Solutions:**

**Use Conda with osx-64 Emulation**: You can create a conda environment that emulates the osx-64 architecture, which is compatible with most bioinformatics tools, including deeptools. This might involve running conda in "Rosetta 2" emulation mode.

* + First, install Rosetta 2 (if not already installed) using the command:

softwareupdate --install-rosetta

* + Create a new conda environment for osx-64:

CONDA\_SUBDIR=osx-64 conda create -n bio-env python=3.9

* + Activate the environment and install deeptools:

conda activate bio-env // conda deactivate

conda config --env --set subdir osx-64

conda install -c bioconda deeptools

//这也是系统问题，win有可能遇不到这个问题，mac如果遇到就这样搞

但重新下载完deeptools还是报错

% computeMatrix scale-regions -S $projPath/alignment/bigwig/K27me3\_raw.bw -R /Users/tonyzh/Downloads/hg38.ncbiRefSeq.gtf --beforeRegionStartLength 3000 --regionBodyLength 5000 --afterRegionStartLength 3000 --skipZeros -o $projPath/data/hg38\_gene/matrix\_gene.mat.gz -p 8

multiprocessing.pool.RemoteTraceback: """

Traceback (most recent call last):

File "/Users/tonyzh/miniconda3/lib/python3.12/multiprocessing/pool.py", line 125, in worker

result = (True, func(\*args, \*\*kwds) ^^^^^^^^^^^^^^^^^^^

File "/Users/tonyzh/miniconda3/lib/python3.12/multiprocessing/pool.py", line 48, in mapstar

return list(map(\*args)) ^^^^^^^^^^^^^^^^

File "/Users/tonyzh/miniconda3/lib/python3.12/site-packages/deeptools/heatmapper.py", line 174, in compute\_sub\_matrix\_wrapper

return heatmapper.compute\_sub\_matrix\_worker(\*args) ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

File "/Users/tonyzh/miniconda3/lib/python3.12/site-packages/deeptools/heatmapper.py", line 379, in compute\_sub\_matrix\_worker

sub\_matrix[:] = np.NAN ^^^^^^

File "/Users/tonyzh/miniconda3/lib/python3.12/site-packages/numpy/\_\_init\_\_.py", line 414, in \_\_getattr\_\_

raise AttributeError("module {!r} has no attribute "

AttributeError: module 'numpy' has no attribute 'NAN'"""

The above exception was the direct cause of the following exception:

Traceback (most recent call last):

File "/Users/tonyzh/miniconda3/bin/computeMatrix", line 8, in <module>

sys.exit(main()) ^^^^^^

File "/Users/tonyzh/miniconda3/lib/python3.12/site-packages/deeptools/computeMatrix.py", line 398, in main

hm.computeMatrix(scores\_file\_list, args.regionsFileName, parameters, blackListFileName=args.blackListFileName, verbose=args.verbose, allArgs=args)

File "/Users/tonyzh/miniconda3/lib/python3.12/site-packages/deeptools/heatmapper.py", line 252, in computeMatrix

res, labels = mapReduce.mapReduce([score\_file\_list, parameters], ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

File "/Users/tonyzh/miniconda3/lib/python3.12/site-packages/deeptools/mapReduce.py", line 142, in mapReduce

res = pool.map\_async(func, TASKS).get(9999999) ^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^

File "/Users/tonyzh/miniconda3/lib/python3.12/multiprocessing/pool.py", line 774, in get

raise self.\_value

AttributeError: module 'numpy' has no attribute 'NAN'

* 把heatmapper.py 379行的np.NAN -> np.nan就好了

