ATAC/CUT&Tag 分析培训

学习文档

目录

一、基础知识

1、FASTA 格式。

用于表示**核酸序列或多肽序列**的格式。其中核酸或氨基酸均以单个字母来表示，且允许在序列前添加序列名及注释。该格式已成为生物信息学领域的一项标准。

FASTA 格式首先以大于号“>”开头，接着是序列的标识符（ID），然后是序列的描述信息（非必须），最后换行后是序列信息，序列中允许换行，空行，直到下一个大于号，表示该序列的结束。如：

```
>LOC_Os01g01019.1 gene=LOC_Os01g01019
MEEAGERDADETHAWSGTASPAALWKTVASSAAMLKLALAMISAAFRTTPFSMSMQLCPNATMSLHSPSI
FDVVSSITPIMSCIINNRLVAEKAGATMQRWRAHSSPSAMTRPLPNMGMRLSSYDIVCQLAHLHFSHVCC
LV.
>LOC_Os01g01030.1 gene=LOC_Os01g01030
MDSIRRRSAGGILGILFLVLLRWAGAGDPYAYYEWEVSYVWGAPLGGVKKQEAIGINGQLPGPALNVTTN
WNLVVNVRNGLDEPLLLTWHGVQQRKSPWQDGVGGTNCGIPPGWNWTYQFQVKDQVGSFFYAPSTALHRA
AGGYGAITINNRDVIPLPFPLPDGGDITLFLADWYARDHRALRRALDAGDPLGPPDGVLINALGPYRYND
```

2、GFF3 格式

为 general feature format 缩写，目前采用的是 version 3，即我们常说的 gff3 文件。该文件常用来对基因组进行注释，表示基因，外显子，CDS，UTR 等在基因组上的位置。

seqname - 序列的 ID，可以是染色体的 ID 也可以是 Scaffold 或者 Contig 的 ID。

source - 产生此文件的软件，如 Stringtie 产生的则为 Stringtie，Cufflinks 产生的则为 Cufflinks，不知道的使用点 “.” 表示。

feature - 后面 start 和 end 之间区域代表的特征，如果此区域是基因，则此处为 gene，如果是外显子，则为 exon，如果是转录本，则为 transcript，如果是非编码 RNA 则为 lncRNA，如果是重复序列，则为 TE，等等，主要表明这

一块区域的特征。

start -上述 feature 的在序列上的起始位置。

end - 上述 feature 的在序列上的终止位置。

score - 一个浮点数值，也可以为点 “.” 。有值的时候代表上述 feature 的可靠性。因为无论是 gene 还是 mRNA，都是基于预测差生的，因而必然会有一个值来衡量预测准确性。

strand - + (forward)或者 - (reverse)，代表上述 feature 是位于正链还是负链上。

frame - 内含子相位，只能为'0', '1' or '2'，或者为点 “.”。'0' 代表 feature 起始碱基为三联体密码子的第一个碱基，'1' 代表三联体密码子的第 1 个碱基，2 代表第 2 个碱基。

attribute -备注列。主要备注该 feature 的一些信息，常见的是 gene 或者 transcript 等的 ID 信息以及 FPKM 值等，多个备注信息之间通常用分号分隔。

3、FASTQ 格式

FASTQ 格式是常用的生物信息学文件格式，用于存储高通量测序（Next Generation Sequencing, NGS）数据。它包含两个主要部分：序列标识行、序列行、序列质量行和可选的描述行。以下是一个示例：

```
@Sequence_1
ATCGATCGATCGATCG
+
HHHHHIIIIIIIIII
```

在上面的示例中，以 "@" 开头的行表示序列标识符，后面的 "Sequence_1" 是该序列的标识符。接下来的行包含了实际的序列数据和序列质量。

序列行是由碱基字母组成的，表示测序得到的 DNA 或 RNA 序列。

质量行使用 ASCII 码字符来表示测序质量，与序列行一一对应。较高的 ASCII 码值表示较高的测序质量，例如上面示例中的 "H" 表示较高的质量值。

需要注意的是，FASTQ 文件中的每个序列都包含了四行：序列标识行、序列行、"+" 行（可选的描述行，有时也包含其他信息）和序列质量行。

FASTQ 格式不仅存储了测序的序列数据，还提供了相应的质量信息，这对于后续的序列处理、比对、变异检测和基因组组装等分析非常重要。

4、SAM 格式

"SAM"（Sequence Alignment/Map）是一种文本文件格式，用于存储测序数据在参考基因组上的比对结果。SAM 格式广泛用于记录比对算法输出的测序读取与参考基因组的比对位置、比对质量以及其他相关的元数据。

SAM 文件以文本形式表示，并且可以被人类和计算机读取。它由多行组成，每行代表一个测序 read 的比对结果。每一行包含了一系列字段，对应于该比对结果的各个信息，例如读取名称、比对的参考序列名称、比对起始位置、对齐方式、碱基序列、质量分数等。

```
E100045579L1C037R01602585007    147     hg19_chrM       150     60      100M    =       71      -179    CCTATTATTTATCGCACCTACGTTCAATATTATAGGCGAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATAACAATTGAA    =GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGHGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG    MC:Z:100M       MD:Z:32C67      NM:i:1  AS:i:95 XS:i:31 RG:Z:sample2    PG:Z:MarkDuplicates-6B83BE14
E100045579L1C038R01702501650    2227    hg19_chrM       150     60      64H36M  =       219     125     CCTATTATTTATCGCACCTACGTTCAATATTATAGG    >EGGFFFFGGGFECFF?DGFCEFFFGGGGGFGGGFE    SA:Z:hg19_chrM,183,+,33S67M,60,0;       MC:Z:91M9H      MD:Z:32C3       NM:i:1  AS:i:32 XS:i:0  RG:Z:sample2    PG:Z:MarkDuplicates-6B83BE14
E100045579L1C039R01403582319    177     hg19_chrM       150     60      21S79M  =       15611   15483   CTAGTTTGTTAGGGACGGATCCCTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAG    GGGFGGGFGGGGFFFGFFGFFFFGGGGGGGGGFGFFGFFGGGGGGFGGGGGGGFGFFFGGGGGGGFGGGFGGGGFGFGGGGGGGGGGGGGGFFGGFGGF    MC:Z:100M       MD:Z:79NM:i:0   AS:i:79 XS:i:24 RG:Z:sample2    PG:Z:MarkDuplicates-6B83BE14
E100045579L1C040R02601229750    147     hg19_chrM       150     60      100M    =       1       -249    CCTATTATTTATCGCACCTACGTTCAATATTATAGACGAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATAACAATTGAA    5A;GEAE<ECC)5)09-.8=1F?)>FGBG9/>EGF*./FGDFGF7<EE38DGG>GFA?<DFFFFFGCFFFE9FFF4EF=:DFEFFFFFGFGFFGCGDFGE    MC:Z:40S60M     MD:Z:32C2G64    NM:i:2  AS:i:90 XS:i:31 RG:Z:sample2    PG:Z:MarkDuplicates-6B83BE14
```

SAM 文件还包含了一些额外的标志位和标签字段，用于描述特定的比对特性和附加的信息。这使得 SAM 格式非常灵活，可以记录各种类型的比对结果，支持多样的分析需求。

需要注意的是，SAM 格式是一个文本文件格式，相对于二进制的 BAM 格式，其文件大小通常较大。在实际使用中，BAM 格式更常见，因为它将 SAM 格式进行了压缩和二进制化，提供了更高的存储效率和读写速度。

SAM 文件可以通过多种生物信息学工具进行创建、转换和处理，如 Samtools、Picard、GATK 等。这些工具提供了丰富的功能，如排序、索引、筛选、变异检测和可视化等，以支持基因组比对数据的分析和解释。

SAM 格式中的每一列代表着不同的含义。以下是 SAM 格式中列的含义：

QNAME：测序读取（read）的名称或标识符。

FLAG：比对标志位，用于描述比对的各种特性和属性。

RNAME：参考序列的名称或标识符。

POS：比对在参考序列上的起始位置。

MAPQ：比对质量分数，表示比对的可信度。

CIGAR：比对的 CIGAR 字符串，描述了比对的碱基操作、插入和删除情况。M：alignment match，表示 read 比对上或者未比对上（错配）；I：insertion to the reference，表示 read 的碱基序列相对于第三列参考序列有碱基的插入；D：deletion from the reference，表示 read 的碱基序列相对于第三列的参考序列有碱基的删除；N：skipped region from the reference，表示跳过参考序列的碱基。常出现在转录组数据比对基因组的结果中，N 出现的位置表示可能是内含子位置；S：soft clipping (clipped sequences present in SEQ)，软裁剪,指 read 跳过的长度才能比

对上;H:硬裁剪,指参考基因组序列跳过的长度才能比对上;比如 51M1106N30M，前 51 个碱基可以匹配，然后跳过参考基因组上 1106 个碱基后又有 30 个碱基匹配。

RNEXT：下一个比对的参考序列的名称或标识符。

PNEXT：下一个比对的起始位置。

TLEN：模板长度，表示测序读取在参考序列上的覆盖长度。

SEQ：测序读取的碱基序列。

QUAL：碱基序列的质量分数。

除了这些必需的列，SAM 格式还可以包含一些可选的标签（Tag），以提供额外的元数据信息，如比对算法使用的特定参数、分析结果等。这些标签以格式为 "TAG:VALUE" 的形式出现在每行的列中，例如 "AS:i:100" 表示比对得分为 100。部分标签信息如下：

AS:i	匹配的得分，只有当 Align≥1 time 才出现
XS:i	第二好的匹配的得分，当 Align>1 time 出现
YS:i	mate 序列匹配的得分
XN:i	在参考序列上模糊碱基的个数
XM:i	**错配的碱基个数**
XO:i	gap open 的个数，针对于比对中的插入和缺失
XG:i	gap 延伸的个数，针对于比对中的插入和缺失

NM:i	**编辑距离**（插入/缺失/替换）。但是不包含头尾被剪切的序列。一般来说等于序列中 error base 的个数

需要注意的是，SAM 格式中的列是通过制表符（Tab）进行分隔的，因此可以使用文本编辑器或相关的生物信息学工具进行解析和处理。了解每列的含义可以帮助我们理解和分析测序数据的比对结果。

二、数据统计

目前的标准要求实验应该有两个或更多的生物学重复。对于单端测序，每个重复应该有 2500 万条非重复、非线粒体的 read；对于双端测序，则要求 5000 万条 read。具体我们可以使用 FastQC 进行质控！

FastQC 提供了一系列有关测序数据质量的有用信息，包括测序质量分布、GC 含量分布、碱基偏好性、序列重复性等。通过分析这些指标，可以评估测序数据的质量，并在后续的数据分析中做出相应的调整或过滤操作。下面是 FastQC 的详细用法说明：

2.1 安装 FastQC

从 FastQC 官方网站（https://www.bioinformatics.babraham.ac.uk/projects/fastqc/）下载最新版本的 FastQC 压缩包。

```
wget https://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.12.1.zip
```

解压缩压缩包并进入 FastQC 目录。

运行以下命令安装 FastQC：

```
chmod +x fastqc      # 赋予可执行权限
```

2.2 运行 FastQC

在命令行中运行以下命令来运行 FastQC：

```
fastqc [options] input_file(s)
```

常用选项：

-o, --outdir <output directory>：指定输出结果文件夹，默认为当前目录。

-t, --threads <threads>：指定使用的线程数。

--casava：针对 Illumina Casava 1.8 或更高版本的数据进行相应的处理。

--noextract：不提取 fastqc.zip 文件中的原始数据。

--nogroup：在多个输入文件时，禁止将结果分组到单个报告中。

--quiet：静默模式，只输出错误消息。

具体例子：

```
fastqc CK_rep1_1.fq.gz CK_rep1_2.fq.gz CK_rep2_1.fq.gz CK_rep2_2.fq.gz
Treat_rep1_1.fq.gz Treat_rep1_2.fq.gz Treat_rep2_1.fq.gz Treat_rep2_2.fq.gz
```

2.3 输出结果

FastQC 将生成一个 HTML 格式的结果报告，其中包含多个质量指标和图形展示。可以通过 Web 浏览器打开该报告来查看数据质量评估结果。可以通过使用-o 选项指定的目录中找到报告文件。具体每个文件介绍如下：

FastQC 结果报告提供了关于测序数据质量的详细信息，帮助用户评估数据的可靠性和适用性。以下是 FastQC 结果报告中一些常见部分的解读：

1. Basic Statistics（基本统计）：

- Filename：输入文件的名称。

- File Type：输入文件的类型，如 FASTQ。

- Encoding：测序数据的编码方式，常见的有 Sanger、Illumina 1.8+等。

- Total Sequences：总测序序列数。

- Sequence Length：测序序列的长度。

- %GC：测序序列中 G 和 C 碱基的百分比。

2. Per base sequence quality（每个碱基的质量）：

- 柱状图显示了测序序列每个碱基位置的质量得分。理想情况下，质量得分应保持稳定且高于阈值线。

- 质量得分通常使用 Phred 质量分数表示，范围从 0 到 40，高质量的得分通常大于 20。

3. Per sequence quality scores（每个序列的质量得分）：

- 盒形图显示了每个测序序列的平均质量得分分布情况。

- 如果序列质量得分分布不均匀或偏低，可能需要进行质量过滤或修剪操作。

4. Per base sequence content（每个碱基的序列含量）：

- 图表显示了每个碱基位置上 A、T、G、C 等碱基的分布情况。

- 如果某个碱基在特定位置上的分布异常，可能需要检查适配序列或其他污染问题。

5. Per sequence GC content（每个序列的 GC 含量）：

- 直方图显示了测序序列中不同 GC 含量的比例。

- 异常的 GC 含量分布可能意味着样本污染或 PCR 偏好等问题。

6. Sequence Duplication Levels（序列重复水平）：

- 饼图和柱状图展示了重复序列的比例和重复级别分布。

- 高比例的重复序列可能源于 PCR 扩增或方法特异性引起的偏差。

7. Overrepresented sequences（过表示序列）：

- 列出了在测序数据中频繁出现的序列。

- 过多的过表示序列可能来自污染、适配序列残留或其他非特定序列。

8. Adapter Content（接头序列含量）：

- 柱状图显示了测序数据中适配序列的存在情况和含量。

- 高含量的适配序列可能需要进行修剪操作。

通过分析 FastQC 结果报告，用户可以确定测序数据中存在的质量问题，并根据需要采取相应的数据处理方法，如质量过滤、序列修剪或适配序列去除等，以保证后续分析的准确性和可靠性。

三、比对

染色质谱技术，如 ChIP-seq，ATAC-seq 和 Hi-C，被广泛用于研究转录因子结合、染色质可及性和染色质高级结构。标准的分析流程，例如 ENCODE 项目使用的流程，首先通过流行的短读取比对工具 BWA-MEM 或 Bowtie2 进行 read 比对，然后使用 SAMtools 和 Picard 进行比对排序和去重等处理。与下游的峰值调用（如 MACS2）相比，这些步骤是常见的瓶颈，可能需要数小时或数天来完成，而峰值调用通常只需要几分钟。造成这种低效的原因之一是，对于大多数染色质相关的生物学研究来说，并不需要全面的碱基级别比对结果用于变异鉴定。此外，在标准工作流程中，比对过滤、去重和其他预处理步骤是按顺序使用不同的方法处理的，每个步骤都需要从压缩文件中解析。这种重复的输入/输出显著增加了运行时间。Chromap 采用快速基于排序的过程生成映射候选项，并使用快速比对算法选择最佳候选项。为了更好地处理短 read，Chromap 考虑每个最小化器命中，并使用 read 信息来修复由于低频最小化器缺失而导致的剩余缺失比对。利用染色质谱图仅富集在整个基因组子集中的观察结果，Chromap 在这些区域中缓存候选 read 比对位置，以加速包含相同最小化器的未来 read 的比对。除了 read 比对，Chromap 还包括测序接头修剪、重复去除和 scATAC-seq 条形码纠正，进一步提高处理效率。Chromap 显著降低了计算时间，同时保持准确性。典型的使用场景包括：（1）修剪测序接头，将批量 ATAC-seq 或 ChIP-seq 基因组测序数据映射到基因组并去除重复；（2）修剪测序接头，将单细胞 ATAC-seq 测序数据映射到基因组，进行条形码纠正，去除重复并执行 Tn5 位移；（3）将 Hi-C 数据 split-alignment 到参考基因组。在这三种情况下，Chromap 的速度是 BWA 的 10-20 倍，并且准确性更

高。对于染色质相关的数据比对率或映射率（即映射到基因组的 reads 数占总 reads 数的百分比）应该大于 95%，但是大于 80%的数据也可接受。

3.1chromap 安装

要从源代码编译，您需要安装 GCC 编译器（版本 >= 7.3.0）、GNU make 和 zlib 开发文件

git clone https://github.com/haowenz/chromap.git

cd chromap && make

Chromap 还可以在 bioconda 上获得。因此，您可以使用 Conda 轻松安装

conda install -c bioconda -c conda-forge chromap

3.2 chromap 使用

#首先创建一个索引，然后按照不同数据类型进行映射

chromap -i -r test/ref.fa -o ref.index

ATAC-seq 数据

chromap --preset atac -x index -r ref.fa -1 read1.fq -2 read2.fq -o aln.bed

scATAC-seq 数据

chromap --preset atac -x index -r ref.fa -1 read1.fq -2 read2.fq -o aln.bed -b barcode.fq.gz --barcode-whitelist whitelist.txt

ChIP-seq 数据

chromap --preset chip -x index -r ref.fa -1 read1.fq -2 read2.fq -o aln.bed

Hi-C 数据，输出配对格式

chromap --preset hic -x index -r ref.fa -1 read1.fq -2 read2.fq -o aln.pairs

Hi-C 数据输出 SAM 格式

chromap --preset hic -x index -r ref.fa -1 read1.fq -2 read2.fq --SAM -o aln.sam

1、在映射之前，需要创建并保存参考序列的索引

```
chromap -i -r Arabidopsis_thaliana.TAIR10.genome.fa -o Arabidopsis_thaliana.TAIR10.index --min-frag-length 100
```

用户可以通过 --min-frag-length 输入预期测序实验中的最小片段长度，例如 read 长度。然后 Chromap 将选择适当的 k-mer 长度和窗口大小来构建索引。对于人类基因组，构建索引只需要几分钟时间。

2、比对。

在没有任何预设参数的情况下，Chromap 接受一个参考数据库和一个查询序列文件作为输入，并产生近似映射结果，但不进行碱基级别的对齐，输出格式为 BED；您可以要求 Chromap 输出 SAM 格式的对齐结果（使用--SAM），但请注意，对 SAM 文件的处理尚未完全优化，可能会很慢。因此，在可能的情况下，不推荐生成 SAM 格式的输出。同时 Chromap 可以接受多个文库的输入：chromap -x index -r ref.fa -1 query1.fq,query2.fq,query3.fq --SAM -o alignment.sam

Chromap 也支持以 gzip 压缩的 FASTA 和 FASTQ 格式作为输入。您无需在转换为 FASTA 或 FASTQ 格式之前解压缩 gzip 压缩的文件。

重要的是要注意，一旦建立了索引，索引参数（例如 -k、-w 和 --min-frag-length）在映射过程中无法更改。如果您要对不同类型的数据运行 Chromap，则可能需要保留使用不同参数生成的多个索引。这使得 Chromap 与始终使用相同的索引的 BWA 不同。Chromap 可以在几分钟内构建人类基因组索引文件。

为了支持不同的数据类型（例如 ChIP-seq、Hi-C、ATAC-seq），Chromap 需要进行性能和准确性的优化。通常建议使用带有 --preset 选项的预设参数，该选项可以同时设置多个参数。

1）映射 ChIP-seq 短读取

```
chromap --preset chip -x Arabidopsis_thaliana.TAIR10.index -r Arabidopsis_thaliana.TAIR10.genome.fa -1 ../01.Data_access/CK_rep1_1.fq.gz -2 ../01.Data_access/CK_rep1_2.fq.gz -o CK_rep1.aln.bed -t 1
chromap --preset chip -x Arabidopsis_thaliana.TAIR10.index -r Arabidopsis_thaliana.TAIR10.genome.fa -1 ../01.Data_access/CK_rep1_1.fq.gz -2 ../01.Data_access/CK_rep1_2.fq.gz -o CK_rep1.aln.SAM --SAM -t 1
```

这组参数经过调整，用于映射 ChIP-seq 读取。Chromap 将映射配对 read，并设置最大插入大小为 2000 (-l 2000)，然后使用低内存模式（--low-mem）去除重复（--remove-pcr-duplicates)。输出结果以 BED 格式（--BED）输出。在输出的 BED 文件中，每一行表示一个片段（即一个读取对）的映射，列为：

```
染色体 起始位置 终止位置 N mapq 链
```

这里的链是 read1 的链（由 -1 指定）。如果需要每个 read 的映射起始和终止位置，应该使用 --TagAlign 重写预设参数中的 --BED 选项，如下所示：

```
chromap --preset chip -x index -r ref.fa -1 read1.fq.gz -2 read2.fq.gz --TagAlign -o aln.tagAlign
```

对于每个 read 对，输出文件中将有两行，分别表示一对 read 中的映射位置。

2）映射 ATAC-seq/scATAC-seq 短读取

```
chromap --preset atac -x index -r ref.fa -1 read1.fq.gz -2 read2.fq.gz -o aln.bed # ATAC-seq reads
```

chromap --preset atac -x index -r ref.fa -1 read1.fq.gz -2 read2.fq.gz -o aln.bed -b barcode.fq.gz --barcode-whitelist whitelist.txt # scATAC-seq reads

这组参数适用于映射 ATAC-seq/scATAC-seq 读取。Chromap 将修剪 3'端的接头（--trim-adapters），以最大插入大小为 2000bp 进行配对末端读取的映射（-l 2000），然后在细胞水平上去除重复序列（--remove-pcr-duplicates-at-cell-level)。还将应用 Tn5 偏移(--Tn5-shift)到片段上。正向映射的起始位置增加了 4bp，反向映射的终止位置减少了 5bp。处理过程在低内存模式下运行（--low-mem）。

如果没有给定条形码白名单文件，则 Chromap 将跳过条形码校正。当输入了条形码和白名单时，默认情况下 Chromap 将估计条形码的丰度，并使用该信息进行与白名单条形码的 Hamming 距离最多为 1 的条形码校正。通过将--bc-error-threshold 设置为 2，Chromap 能够对与白名单条形码的 Hamming 距离最多为 2 的条形码进行校正。用户还可以增加概率阈值以进行更可靠的校正，通过将--bc-probability-threshold（默认设置为 0.9）设为较大的值（例如 0.975）只进行可靠的校正。对于具有多个读取和条形码文件的 scATAC-seq 数据，您可以使用“,”来连接多个输入文件，如上面的示例所示。

Chromap 还支持用户定义的条形码格式，包括混合条形码和基因组数据情况。用户可以通过选项--read-format 指定序列结构。该值是一个逗号分隔的字符串，字符串中的每个字段也是一个由分号分割的字符串

起始位置和结束位置都是包含的，-1 表示读取的末尾。用户可以使用多个字段来指定非连续的片段，例如 bc:0:15,bc:32:-1。链'+'和'-'符号表示，如果是'-'，则在提取后的条形码上进行反向互补。如果是'+'，则可以省略链符号，并且在 r1 和 r2 上被忽略。例如，当条形码位于 read1 的前 16bp 时，可以使用选项-1 read1.fq.gz -2 read2.fq.gz --barcode read1.fq.gz --read-format bc:0:15,r1:16:-1。

Bulk 数据和单细胞数据的输出文件格式除前三列外是不同的。对于 Bulk 数据，列如下：

chrom chrom_start chrom_end N mapq strand

对于单细胞数据，列如下：

chrom chrom_start chrom_end barcode duplicate_count

与 CellRanger 中片段文件的定义相同。注意 chrom_end 是开区间。该输出片段文件可用作下游分析工具（如 MAESTRO、ArchR、signac 等）的输入。

此外，Chromap 可以将输入的细胞条形码转换为另一组条形码。用户可以通过选项--barcode-translate 指定翻译文件。翻译文件是一个两列的 tsv/csv 文件，第一列是翻译后的条形码，第二列是原始条形码。这在 10x Multiome 数据中非常有用，其中 scATAC-seq 和 scRNA-seq 数据使用不同的条形码集。该选项还支持组合式条形码，例如 SHARE-seq。Chromap 可以将第二列提供的每个条形码片段翻译为第一列中的 ID，并在输出中使用“-”连接 ID。

3）映射 Hi-C 短读取

要映射 Hi-C 数据，请使用以下命令：

chromap --preset hic -x index -r ref.fa -1 read1.fq.gz -2 read2.fq.gz -o aln.pairs

该命令会将 Hi-C read 映射到参考基因组，并生成包含映射配对的输出文件。

每个配对由两行表示，每行对应一对相互作用的限制性酶切位点。配对的格式为：

染色体 A 位点 A1 距离 A2 染色体 B 位点 B1 距离 B2 mapq

您还可以选择以 SAM 格式输出 Hi-C 的对齐结果：

chromap --preset hic -x index -r ref.fa -1 read1.fq.gz -2 read2.fq.gz --SAM -o aln.sam

这将生成一个包含对齐结果的 SAM 文件。

3.3 输出结果介绍

Number of reads: 46546534.

Number of mapped reads: 37253606.

Number of uniquely mapped reads: 36634844.

Number of reads have multi-mappings: 618762.

Number of candidates: 55612892.

Number of mappings: 37253606.

Number of uni-mappings: 36634844.

Number of multi-mappings: 618762.

uni-mappings: 9987234, # multi-mappings: 818524, total: 10805758.

Number of output mappings (passed filters): 9903273

四、数据过滤

本部分主要包括唯一比对筛选，去重质体比对结果和去除 PCR duplication 结果三部分的过滤。

4.1 软件安装

1、Samtools 主要用来对 SAM/BAM 格式的转换、排序、提取部分比对结果、过滤、建立索引或者合并，是一个处理比对文件非常好用的软件。

下载地址：

https://github.com/samtools/samtools/releases/download/1.9/samtools-1.9.tar.bz2

4.2 操作步骤

主要将 SAM 文件转换为 BAM 文件，去掉质体并做统计。将文件以 BAM 格式储存可以节省空间，而且许多下游的分析使用的是 BAM 格式的文件而不是 SAM 格式的文件。此处，我们指定输入文件为 SAM 格式（参数 -S），输出文件是 BAM 格式（参数 -b），同时筛选出比对质量大于 30 的比对结果，这也就是 unique map 的结果（参数 -o）。

```
grep -v -E 'Mt|Pt' ../02.Mapping/CK_rep1.aln.SAM | samtools view -bS -q 30 -o CK_rep1.aln.bam
samtools flagstat CK_rep1.aln.bam
```

9573383 + 0 in total (QC-passed reads + QC-failed reads)

9573383 + 0 primary

0 + 0 secondary

0 + 0 supplementary

0 + 0 duplicates

0 + 0 primary duplicates

9573383 + 0 mapped (100.00% : N/A)

9573383 + 0 primary mapped (100.00% : N/A)

9573383 + 0 paired in sequencing

4787296 + 0 read1

4786087 + 0 read2

9573383 + 0 properly paired (100.00% : N/A)

9573383 + 0 with itself and mate mapped

0 + 0 singletons (0.00% : N/A)

0 + 0 with mate mapped to a different chr

0 + 0 with mate mapped to a different chr (mapQ>=5)

```
grep -v -E 'Mt|Pt' ../02.Mapping/CK_rep1.aln.bed >CK_rep1.aln.bed
```

五、TSS 富集分析

主要展示 TSS 富集分析热图，评估数据的准确性。DeepTools 是一个用于高通量测序（HTS）数据分析和可视化的 Python 软件包，它提供了一系列强大的工具和函数。

5.1 DeepTools 安装

以下是在 Linux 环境下安装和使用 DeepTools 的基本介绍：

创建环境：为 DeepTools 创建一个独立的环境，以避免与其他软件包的冲突。打开终端，并运行以下命令来创建一个名为 deeptools 的环境：

conda create -n deeptools python=3

激活 deeptools 环境：

conda activate deeptools

安装 DeepTools，在激活的 deeptools 环境中，运行以下命令来安装

conda install -c bioconda deeptools

5.2 DeepTools 使用

1、计算信号覆盖度。

bamCoverage 是 DeepTools 中的一个工具，用于计算和生成测序数据的信号覆盖度（signal coverage）文件。它可以将对基因组进行比对的测序 reads 转换为每个基因组位置的信号强度，并将其保存为 BigWig 格式的文件，以便后续的可视化和分析。

以下是 bamCoverage 的详细用法说明：

bamCoverage --bam <alignment_file.bam> --outFileName <output.bw> [options]

必需参数：

--bam BAM 文件， 要处理的 BAM 文件 (默认值: None)

输出：

--outFileName 文件名, -o 文件名 输出文件名 (默认值: None)

--outFileFormat {bigwig,bedgraph}, -of {bigwig,bedgraph}输出文件类型。可选 "bigwig" 或 "bedgraph" (默认值: bigwig)

常用选项：

--binSize INT bp, -bs INT bp：输出 bigwig/bedgraph 文件的 bin 大小，以碱基为单位。（默认值：50）

--region CHR:START:END, -r CHR:START:END：限制操作范围的基因组区域-这在测试参数以减少计算时间时非常有用。格式为 chr:start:end，例如--region chr10 或--region chr10:456700:891000。（默认值：None）

--blackListFileName BED file [BED file ...], -bl BED file [BED file ...]：包含应从所有分析中排除的区域的 BED 或 GTF 文件。当前的处理方式是拒绝与条目发生重叠的基因组区块。因此，对于 BAM 文件，如果一个 read 部分重叠了黑名单区域或一个片段跨越了它，那么读取/片段可能仍然被考虑。请注意，如果相关，请调整有效基因组大小。（默认值：None）

--numberOfProcessors INT, -p INT：要使用的处理器数量。输入"max/2"使用最大处理器数量的一半，输入"max"使用所有可用的处理器。(默认值：1)

读取覆盖度归一化选项：

--effectiveGenomeSize EFFECTIVEGENOMESIZE：有效基因组大小是可映射的基因组部分。大部分基因组是由 NNNN 组成的伸展区域，应该丢弃。此外，如果重复区域未包含在读取的映射中，则需要相应调整有效基因组大小。这里提供了一个值表：http://deeptools.readthedocs.io/en/latest/content/feature/effectiveGenomeSize.html。（默认值：None）

--normalizeUsing {RPKM,CPM,BPM,RPGC,None}：使用输入的方法之一来归一化每个 bin 的读取数。默认情况下不进行归一化。RPKM (per bin) = number of reads per bin / (number of mapped reads (in millions) * bin length (kb)). CPM (per bin) = number of reads per bin / number of mapped reads (in millions). BPM (per bin)= number of reads per bin / sum of all reads per bin (in millions). RPGC (per bin) = number of reads per bin / scaling factor for 1x average coverage。None = 默认选项，相当于不设置此选项。

--smoothLength INT bp 平滑长度定义了一个窗口，大于 binSize，用于平均 reads 数量。例如，如果 --binSize 设置为 20，并且 --smoothLength 设置为 60，那么对于每个 bin，将考虑 bin 及其左右相邻 bin 的平均值。任何小于 --binSize 的值都将被忽略，不应用平滑。（默认值：None）

read 处理选项：

--extendReads [INT bp] 此参数允许将 reads 延伸到的片段大小。如果设置，每个 read 都会被延伸，没有例外。*注意*：通常不推荐对 spliced-read 数据（如 RNA-seq）使用此功能，因为它会将 reads 延伸到跳过的区域上。*单端*：需要指定最终片段长度的值。已经超过这个片段

长度的 reads 将不会被延伸。*成对端*：拥有配对的 reads 会被始终延伸到与两个读取配对定义的片段大小相匹配。无配对的 reads、映射间距过大的配对（>4 倍片段长度）或者映射到不同染色体的配对将被视为单端 reads。片段长度的输入是可选的，如果未指定值，则从数据中估计（所有配对的片段大小的平均值）。 （默认值：False）

--ignoreDuplicates 如果设置，将仅考虑具有相同方向和起始位置的 reads。如果 reads 是成对的，则也必须与 mate 的位置一致才能忽略该 read。 （默认值：False）

--minMappingQuality INT 如果设置，将仅考虑具有至少这个映射质量分数的 reads。 （默认值：None）

--centerReads 通过添加此选项，将 reads 以片段长度为基准居中。对于成对数据，read 在两个 fragment 端点所定义的片段长度处居中。对于单端数据，则使用给定的片段长度。此选项对于从富集区域获取更清晰的信号非常有用。 （默认值：False）

--samFlagInclude INT 根据 SAM flag 包括 reads。例如，要仅获取第一个 mate 的 reads，请使用标志 64。这对于只计算正确配对的 reads 非常有用，否则第二个 mate 也将被考虑在覆盖范围内。（默认值：None）

--samFlagExclude INT 根据 SAM flag 排除 reads。例如，要仅获取映射到正向链的 reads，请使用 --samFlagExclude 16，其中 16 是映射到反向链的 reads 的 SAM flag。 （默认值：None）

--minFragmentLength INT 读/配对包含所需的最小片段长度。此选项主要适用于 ATAC-seq 实验，用于过滤单核苷酸或二核苷酸片段。 （默认值：0）

--maxFragmentLength INT 读/配对包含所需的最大片段长度。（默认值：0）

示例用法：

```
samtools index ../03.Filter/CK_rep1.aln.bam
bamCoverage --bam ../03.Filter/CK_rep1.aln.bam --outFileName CK_rep1.bw --binSize 10 --centerReads --normalizeUsing RPKM
```

上述示例命令将基于alignment.bam文件生成coverage.bw的信号覆盖度文件。每个 10 bp 形成一个区间，对信号覆盖度进行归一化。

2、生成矩阵文件。

computeMatrix 是 DeepTools 中的一个工具，用于计算和生成基于测序数据的矩阵文件。它可以根据给定的参考基因组上的位置，从测序数据（如 BAM 文件、BigWig 文件）中提取信号强度，并将其整理成一个矩阵文件，以便后续的可视化和分析。

以下是 computeMatrix 的详细用法说明：

computeMatrix <command> [options]

常用命令：

- `reference-point`：根据参考点（如基因的转录起始位点）提取信号值。
- `scale-regions`：根据指定的区域（如基因体）大小和方向进行信号缩放。

必需参数:

--regionsFileName 文件 [文件 ...], -R 文件 [文件 ...]以 BED 或 GTF 格式提供的文件名或文件，包含要绘制的区域。如果给出多个 bed 文件，则每个文件被视为一个可以单独绘制的组。在 bed 文件中添加“ # ”符号会导致该符号之前的所有区域被视为一组。(默认: None)

--scoreFileName 文件 [文件 ...], -S 文件 [文件 ...]包含要绘制的分数的 bigWig 文件。多个文件应以空格分隔。可以使用 bamCoverage 或 bamCompare 工具获得 BigWig 文件。

输出选项:

--outFileName 保存由"plotHeatmap"和"plotProfile"工具所需的经过 gzip 压缩的矩阵文件的文件名。(默认: None)

--outFileNameMatrix 文件，如果给出此选项，则将使用指定的名称保存热图下面的值矩阵，例如 IndividualValues.tab。此矩阵可以轻松加载到 R 或其他程序中。(默认: None)

--outFileSortedRegions BED 文件,经过跳过零值或最小/最大阈值后保存区域的文件名。文件中的区域顺序遵循所选择的排序顺序。这在生成保持第一个热图排序的其他热图时非常有用。示例：Heatmap1sortedRegions.bed (默认: None)

可选参数:

--referencePoint {TSS,TES,center} 绘图的参考点可以是区域的起始点（TSS）、区域的结束点（TES）或区域的中心。请注意，无论您指定什么，plotHeatmap/plotProfile 都会默认使用“TSS”作为标签。(默认: TSS)

--beforeRegionStartLength 所选参考点上游的距离。(默认: 500)

--afterRegionStartLength 所选参考点下游的距离。(默认: 1500)

--nanAfterEnd 如果设置，将丢弃区域结束后的任何值。这对于在不使用比例区域模式且将参考点设置为 TSS 时可视化区域结束非常有用。(默认: False)

--binSize 非重叠 bin 的长度，用于对区域长度进行平均得分。(默认: 10)

--sortRegions {descend,ascend,no,keep}输出文件是否应按区域排序。默认情况下，不对区域排序。请注意，如果您计划自己绘制结果而不是使用 plotHeatmap 等，这只有用处。还要注意，未排序的输出将按照处理的区域的任意顺序进行，并且与输入文件中的顺序不匹配。如果您需要输出顺序与输入区域顺序匹配，则指定“keep”或使用 computeMatrixOperations 重新排序结果文件。(默认: keep)

--sortUsing {mean,median,max,min,sum,region_length} 指示应使用哪种方法进行排序。该值计算每行的数值。请注意，region_length 选项将导致在热图中出现一个点线，表示区域的结束。(默认: mean)

--sortUsingSamples 用于按--sortUsing 排序的样本编号列表（与矩阵中的顺序相同），不指定值将使用所有样本，例如：--sortUsingSamples 1 3 (默认: None)

--averageTypeBins {mean,median,min,max,std,sum} 定义应在 bin 大小范围内使用的统计类型。选项有：“mean”，“median”，“min”，“max”，“sum”和“std”。默认为"mean"。(默认: mean)

--missingDataAsZero　如果设置，缺失数据（NAs）将被视为零。默认情况下，将忽略这些情况，它们将以黑色区域显示在热图中。（有关其他选项，请参见 plotHeatmap 命令的--missingDataColor 参数）。(默认: False)

--skipZeros 是否应包含只有零分数的区域。默认是包括它们。(默认: False)

--minThreshold　数值。将跳过任何包含小于或等于此值的值的区域。这对于跳过任何 bin 中读取计数为零的基因非常有用。这可能是不可映射区域的结果，并可能使整体结果有偏差。(默认: None)

--maxThreshold 数值。将跳过任何包含大于或等于此值的值的区域。maxThreshold 适用于跳过少数具有非常高读取计数（例如，微卫星）的区域，这些区域可能导致平均值有偏差。(默认: None)

--blackListFileName BED 文件, -bl BED 文件,包含应从所有分析中排除的区域的 BED 文件。当前，这通过拒绝恰好与条目重叠的基因组块来工作。因此，对于 BAM 文件，如果一个读取部分重叠了黑名单区域，或者一个片段跨越该区域，则仍然可能考虑该读取/片段。(默认: None)

--samplesLabel 样本的标签。这将传递给 plotHeatmap 和 plotProfile。默认情况下，使用样本的文件名作为标签。样本标签应以空格分隔，并在标签本身包含空格时用引号括起来。例如：--samplesLabel label-1 "label 2" (默认: None)

--smartLabels 与手动指定 bigWig 和 BED/GTF 文件的标签不同，这会使 deepTools 使用去除路径和扩展名后的文件名。(默认: False)

--scale SCALE 如果设置，所有值都将乘以该数。 (默认: 1)

--numberOfProcessors 要使用的处理器数量。输入“max/2”使用最大处理器数量的一半，输入“max”使用所有可用处理器。(默认: 1)

GTF/BED12 选项：

--metagene 当使用 BED12 或 GTF 文件提供区域时，对合并的外显子执行计算，而不是使用由 BED 文件的第 2 列和第 3 列定义的基因组区间。如果使用 BED3 或 BED6 文件作为输入，则使用第 2 列和第 3 列作为一个外显子。(默认: False)

--transcriptID 当使用 GTF 文件提供区域时，只处理具有此值作为其特征（第 3 列）的条目作为转录本。(默认: transcript)

--exonID 当使用 GTF 文件提供区域时，只处理具有此值作为其特征（第 3 列）的条目作为外显子。CDS 可能是另一个常见的值。(默认: exon)

--transcript_id_designator 每个区域都有一个分配给它的 ID（例如 ACTB），对于 BED 文件来说，它要么是第 4 列（如果存在）要么是区间边界。对于 GTF 文件，它被存储在最后一列中，作为键值对（例如，'transcript_id "ACTB"'，其中键是 transcript_id，值是 ACTB）。在某些情况下，使用不同的标识符可能很方便。为了这样做，请将其设置为所需的键。(默认: transcript_id)

示例用法：

```
awk 'BEGIN{OFS="\t"}{if($3=="gene"||$3=="ncRNA_gene")print $1,$4,$5,$9,$9,$7}' ../02.Mapping/Arabidopsis_thaliana.TAIR10.gff3 >genes.bed

computeMatrix reference-point --referencePoint TSS -b 3000 -a 3000 -p 15 -R genes.bed
-S CK_rep1.bw CK_rep2.bw Treat_rep1.bw Treat_rep2.bw -o AT.TSS_distribute.gz
```

```
computeMatrix scale-regions -b 3000 -a 3000 --regionBodyLength 6000 -p 15 -R genes.bed -S CK_rep1.bw CK_rep2.bw Treat_rep1.bw Treat_rep2.bw -o AT.TSS-body-TES_distribute.gz
```

上述示例命令将基于 alignment.bam 文件根据转录起始位点（TSS）提取信号值，并以每 10 bp 为一个区间，从-3000 到+3000 bp 的范围内生成一个矩阵文件 AT.TSS_distribute.gz 和 AT.TSS-body-TES_distribute.gz 。

3、生成热图。

plotHeatmap 是 DeepTools 中的一个工具，用于生成热图（heatmap）来可视化基因组上的测序数据。它可以根据给定的矩阵文件或 BigWig 文件，将信号值映射为颜色，并以矩阵的形式展示在基因组的特定区域上。

以下是 plotHeatmap 的详细用法说明：

plotHeatmap -m <matrix.gz> [options]

常用选项：

翻译：

--matrixFile　computeMatrix 工具的矩阵文件。（默认值：无）

--outFileName 要保存图像的文件名。文件扩展名将用于确定图像格式。可用选项有："png"、"eps"、"pdf"和"svg"，例如 MyHeatmap.png。（默认值：无）

输出选项：

--outFileSortedRegions　跳过零值或最小/最大阈值后，将区域保存到的文件名。文件中的区域顺序遵循所选的排序顺序。这在生成其他热图时很有用，同时保持第一个热图的排序。示例：Heatmap1sortedRegions.bed（默认值：无）

--outFileNameMatrix 如果给定此选项，则使用此名称保存热图下层的数值矩阵，例如 MyMatrix.gz。（默认值：无）

--dpi DPI 设置保存图像的 DPI。（默认值：200）

聚类参数：

--kmeans 要计算的簇数。设置此选项时，使用 k-means 算法将矩阵分成簇。仅适用于未分组的数据，否则只会对第一组进行聚类。如果需要更具体的聚类方法，则保存底层矩阵并使用其他软件运行聚类。如果某个簇的成员数量与总区域数相比非常少，那么绘制聚类可能会失败并出现错误。（默认值：无）

--hclust 要计算的簇数。设置此选项后，使用层次聚类算法将矩阵分成簇，使用“ward linkage”。仅适用于未分组的数据，否则只会对第一组进行聚类。如果你有>1000 个区域，--hclust 可能非常慢。在这些情况下，您可能更喜欢--kmeans 或者如果需要更多聚类方法，您可以保存底层矩阵并使用其他软件运行聚类。如果某个簇的成员数量与总区域数相比非常少，那么绘制聚类可能会失败并出现错误。（默认值：无）

可选参数：

--plotType {lines,fill,se,std}“lines”基于选择的平均类型绘制轮廓线。“fill”填充零和轮廓曲线之间的区域。填充颜色是半透明的，以区分不同的轮廓。“se”和“std”用数据的标准误差或标准差来着色轮廓和区域之间的区域。（默认值：lines）

--sortRegions {descend,ascend,no,keep} 热图是否应按区域排序。默认按区域的平均值降序排序。注意，"keep"和"no"是同一回事。（默认值：descend）

--sortUsing {mean,median,max,min,sum,region_length}指示用于排序的方法。对于每行，都会计算方法。对于 region_length，根据适当的参考点（TSS 和中心）或区域起点（TES）绘制虚线。（默认值：mean）

--sortUsingSamples 用于排序的样本编号列表（与矩阵中的顺序相同），由--sortUsing 使用。如果未设置任何值，它将使用所有样本。示例：--sortUsingSamples 1 3（默认值：无）

--linesAtTickMarks 从所有刻度标记绘制虚线穿过热图。这类似于使用参考点和--sortUsing region_length 时在区域边界处绘制的虚线。（默认值：False）

--clusterUsingSamples 用于通过--kmeans 或--hclust 进行聚类的样本编号列表（与矩阵中的顺序相同）。如果不提供，将考虑所有样本进行聚类。示例：--ClusterUsingSamples 1 3（默认值：无）

--averageTypeSummaryPlot {mean,median,min,max,std,sum} 定义应在热图上方的摘要图中绘制的统计类型。选项有："mean"、"median"、"min"、"max"、"sum"和"std"。（默认值：mean）

--missingDataColor MISSINGDATACOLOR 如果未设置--missingDataAsZero，则此类情况将默认以黑色着色。使用此参数，可以设置不同的颜色。灰度色阶（黑色为 0）将使用介于 0 和 1 之间的值。有关可能的颜色名称列表，请参见：

http://packages.python.org/ete2/reference/reference_svgcolors.html。也可以使用#rrggbb 表示法指定其他颜色。（默认值：black）

--colorMap 用于热图的颜色映射。如果绘制多个热图，则可以单独输入每个热图的颜色（例如`--colorMap Reds Blues`）。如果颜色映射数目较少，将循环使用颜色映射。可以在此处查看可用值：http://matplotlib.org/users/colormaps.html

--alpha ALPHA：设置热力图的透明度通道（alpha 通道），默认值为 1.0，取值范围在 0 和 1 之间。

--colorList 设置创建颜色映射的颜色列表。可以使用颜色名称或十六进制颜色代码。如果需要为不同的热力图指定不同的颜色，需要将它们用空格分开，并使用双引号括起来。颜色列表的数量小于热力图数量时，会循环使用颜色列表。颜色过渡的数量由--colorNumber 选项定义。

--colorNumber COLORNUMBER：控制颜色过渡的数量。当--colorNumber 等于--colorList 中颜色的数量时，颜色之间没有过渡效果。

--zMin 设置热力图强度的最小值。可以为每个热力图设置多个值，当给定的值数量少于热力图数量时，将循环使用这些值。

--zMax 设置热力图强度的最大值。可以为每个热力图设置多个值，当给定的值数量少于热力图数量时，将循环使用这些值。

--heatmapHeight 设置热力图的高度，单位为厘米。默认值为 28，最小值为 3，最大值为 100。

--heatmapWidth 设置热力图的宽度，单位为厘米。默认值为 4，最小值为 1，最大值为 100。

--whatToShow {plot, heatmap and colorbar,plot and heatmap,heatmap only,heatmap and colorbar}：设置要显示的内容。默认情况下，会在热力图上方显示一个概要或剖析图以及热力图颜色条。其他选项包括："plot and heatmap"，"heatmap only"，"heatmap and colorbar"，和默认选项"plot, heatmap and colorbar"。

--boxAroundHeatmaps 默认情况下，在热力图周围绘制黑色边框。可以通过设置--boxAroundHeatmaps 为 no 来关闭。

--xAxisLabel 设置 x 轴标签的描述。

--startLabel [仅适用于 scale-regions 模式] 绘图中显示的起始区域标签。默认为 TSS（转录起始位点），但可以更改为任何标签，如"peak start"。

--endLabel 选项类似。

--refPointLabel [仅适用于 reference-point 模式] 绘图中显示的参考点标签。默认与选择的参考点相同（例如 TSS），但可以是任何标签，如"peak start"。

--labelRotation ： x 轴标签的旋转角度，单位为度。默认值为 0，正值表示逆时针旋转。

--regionsLabel ： 设置热力图中绘制的区域标签。如果绘制多个区域，则需要以空格分隔的标签列表。如果标签本身包含空格，则需要使用引号括起来。

--samplesLabel ：设置绘制的样本标签。默认为使用样本的文件名作为标签。样本标签应该以空格分隔，并在标签本身包含空格时加上引号。

--plotTitle ：设置绘图的标题，将显示在生成的图像顶部。留空表示没有标题。

--yAxisLabel ：设置顶部面板的 y 轴标签。

--yMin：Y 轴的最小值。可以为每个剖析图设置多个值，当给定的值数量少于剖析图数量时，将循环使用这些值。

--yMax ：Y 轴的最大值。可以为每个剖析图设置多个值，当给定的值数量少于剖析图数量时，将循环使用这些值。

--legendLocation{best,upper-right,upper-left,upper-center,lower-left,lower-right,lower-center,center,center-left,center-right,none}：概要图中图例的位置。注意，"none"选项在概要图中不起作用。

--perGroup：默认情况下，按样本将所有组别的区域绘制在一起。使用此选项可以按区域组绘制所有样本。只有在存在多个区域组时才有用。

--plotFileFormat：图像格式类型。如果给出此选项，则覆盖基于 plotFile 后缀的图像格式。可用选项包括："png"，"eps"，"pdf"，"plotly"和"svg"。

示例用法：

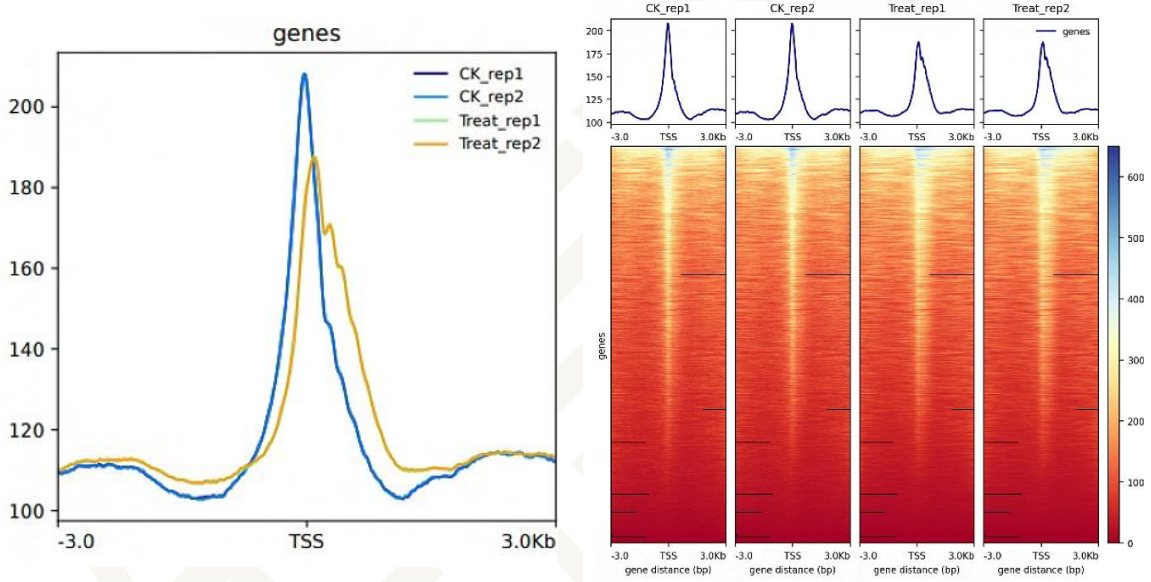
```
plotHeatmap --matrixFile AT.TSS_distribute.gz --dpi 200 --heatmapHeight 15 --outFileName AT.TSS_Heatmap.png
```

```
plotHeatmap --matrixFile AT.TSS-body-TES_distribute.gz --dpi 200 --heatmapHeight 15 --outFileName AT.TSS-body-TES_Heatmap.pdf
```

```
plotProfile --matrixFile AT.TSS_distribute.gz --dpi 200 --plotHeight 11 --perGroup
--outFileName AT.TSS_line.pdf
```

```
plotProfile --matrixFile AT.TSS-body-TES_distribute.gz --dpi 200 --plotHeight 11
--perGroup --outFileName AT.TSS-body-TES_line.pdf
```

上述示例命令将基于 matrix 文件生成一个使用 coolwarm 颜色映射方案的热图。



六、Peak 鉴定

6.1 MACS2 安装

安装 MACS2（Model-based Analysis of ChIP-Seq 2）需要以下步骤：

pip install MACS2

如果您使用的是 Python 3，并且已经存在 Python 2.x，则可能需要使用 pip3 进行安装。安装完成后，您就可以使用 MACS2 来进行 ChIP-Seq 数据的分析了。

6.2 MACS2 用法

macs2 callpeak 是 MACS2 提供的用于对 ChIP-Seq 数据进行峰识别和定位的命令。以下是 macs2 callpeak 的详细用法：

macs2 callpeak

以下是一些常用选项的说明：

`-t TREATMENT [TREATMENT ...]`：输入的 ChIP-Seq 数据文件。可以指定多个文件，以空格分隔。

`-c CONTROL [CONTROL ...]`：输入的对照组数据文件。可以指定多个文件，以空格分隔。

`-f {AUTO,BAM,BED,SAM}`：输入数据的格式。可选值为 AUTO（自动识别）、BAM、BED 和 SAM，默认为 AUTO。

`-g GSIZE`：基因组的大小。可以是常见物种的简写（如 mm、hs、ce），也可以是具体的基因组大小。

`--outdir OUTDIR`：输出结果的目录。

`--name NAME`：结果文件的名称前缀。

`--qvalue QVALUE`：用于 FDR 控制的 q-value 阈值，默认为 0.05。

`--broad`：执行广泛峰的分析。

`--nomodel`：不使用模型来计算片段长度，而是直接使用用户指定的长度（通过 --extsize 参数）。

`--pvalue PVALUE`：用于峰的显著性检验的 p-value 阈值，默认为 0.05。

`--keep-dup {auto,all,1st}`：决定是否保留 PCR 重复的方式，可选值为 auto、all 和 1st，默认为 auto。

`--call-summits`：在峰中心处调用 Summit。

这仅是 macs2 callpeak 命令的一些常用选项，还有其他选项可用于进一步控制分析过程和结果。您可以根据具体需求选择适当的选项，并根据输出结果进行进一步的分析和解释。

例子：

1)ATAC 数据

```
macs2 callpeak -t CK_rep1.aln.bed -g 119481543 -q 0.05 -f BED --nomodel --shift -100 --extsize 200 --keep-dup all --call-summits --nolambda --outdir CK_rep1 -n CK_rep1
```

注意：

--shift -100 --extsize 200 选项将一个 200 bp 的窗口居中于 Tn5 结合位点上，这样处理对于 ATAC-seq 数据更准确（相同的选项也适用于 DNase-seq 数据）。Read 的 5'端代表了 Tn5（和 DNase I）切割位点；因此，read 的 5'端代表了最容易访问的区域。如果您有成对的 ATAC-seq 数据，其中很多 read 对至少跨越一个核小体。如果您查看片段长度分布图，您应该会看到很多不跨越核小体（小于

150-200 bp）的片段，但您也会看到许多跨越核小体（大于 200-300 bp）的片段。因此，您不希望使用适用于 ChIP-seq 的默认模型构建参数来运行 MACS2。这些模型构建参数假设结合位点位于片段中间，这对于 ChIP-seq 是准确的，但对于 ATAC-seq 则不准确。如果您在 ATAC-seq 中使用 ChIP-seq 选项，那么您的很多峰值将以核小体为中心,而不是以可访问区域为中心。之所以经常使用 200 bp 的窗口（--shift 100 --extsize 200），是因为有理由认为 ATAC-seq 的峰值至少有 200 bp 长，因为这大约是一个去除了单个核小体的无核小体区域的大小。有些人可能使用--shift -75 --extsize 150，认为一个去除了单个核小体的可访问区域的长度约为 150 bp，这也是合理的。

2) Cut&Tag/ChIP-seq 数据

```
macs2 callpeak -t CK_rep1.aln.bed -g 119481543 -p 1e-5 -f BED --keep-dup all --outdir CK_rep1  -n CK_rep1
```

6.3 输出结果介绍

Peak Calling 结果文件:主要包含关于峰的信息,如位置、峰高度、p-value、q-value 等。

6.4 FRiP 计算

在峰区域内的 reads 占总 reads 的比例（FRiP 得分）应该大于 0.3，但是得分大于 0.2 的数据也可能可接受。

```
bedtools intersect -a CK_rep1.aln.bed -b CK_rep1/CK_rep1_peaks.narrowPeak| wc -l > FRiP_stat1
 wc -l CK_rep1.aln.bed >CK_rep1.aln_FRiP_stat2
FRip=1966382/4765637=0.41
```

七、Peak 注释

ChIPpeakAnno 是一个用于对 ChIP-Seq 数据进行峰注释和功能分析的 R 包。

7.1 软件安装

1、安装 ChIPseeker

打开 R 或 RStudio。在 R 终端中运行以下命令安装 ChIPseeker 包及其依赖：

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("ChIPseeker")
```

安装完成后，可以加载 ChIPseeker 包：

```
library(ChIPseeker)
```

7.2 Peak 注释

```
perl ../Scripts/make_sql.pl Arabidopsis_thaliana.TAIR10.gff3 AT
perl ../Scripts/get_seq_len.pl ../02.Mapping/Arabidopsis_thaliana.TAIR10.genome.fa AT.len
cut -f 1-4,8 CK_rep1_peaks.narrowPeak >CK_rep1_peaks.bed
/mnt/RAID-5/MD0/bioinformatics/soft/mamba/bin/Rscript ../Scripts/ChIPpeakAnnov2.r -p CK_rep1_peaks.bed --txdb TransDb/AT.sqlite --name CK_rep1 --outdir CK_rep1 --chrlen AT.len
```

八、差异 peak 分析

DiffBind 是一个用于分析 ChIP-Seq 数据中差异结合位点（DBP）的 R 包。

8.1 DiffBind 安装

1 安装 DiffBind

打开 R 或 RStudio。在 R 终端中运行以下命令安装 DiffBind 包及其依赖：

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("DiffBind")
```

安装完成后，可以加载 DiffBind 包：

```
library(DiffBind)
```

2、ClusterProfiler 是一个用于基因集富集分析和可视化的 R 包。

安装 ClusterProfiler，打开 R 或 RStudio。

在 R 终端中运行以下命令安装 ClusterProfiler 包及其依赖：

```
install.packages("BiocManager")
BiocManager::install("clusterProfiler")
```

安装完成后，可以加载 ClusterProfiler 包：

```
library(clusterProfiler)
```

8.2 DiffBind 使用

制作文件命名注意：CK1_CK2_vs_Treat1_Treat2.csv

```
CK1,CK1_CK2_vs_Treat1_Treat2,Factor,NonTreatment,Condition,1,CK_rep1.aln.bam,CK_rep1_peaks.bed,bed
CK2,CK1_CK2_vs_Treat1_Treat2,Factor,NonTreatment,Condition,2,CK_rep2.aln.bam,CK_rep2_peaks.bed,bed
```

```
Treat1,CK1_CK2_vs_Treat1_Treat2,Factor,Treatment,Condition,1,Treat_rep1.aln.bam,Treat_rep1_peaks.bed,bed
Treat2,CK1_CK2_vs_Treat1_Treat2,Factor,Treatment,Condition,2,Treat_rep2.aln.bam,Treat_rep2_peaks.bed,bed
/mnt/RAID-5/MD0/bioinformatics/soft/mambaforge2/bin/Rscript ../Scripts/DiffBind.r CK1_CK2_vs_Treat1_Treat2.csv FDR0.05 1.5 CK1_CK2_vs_Treat1_Treat2
```

差异热图展示：

```
cut -f 2-5,10 CK1_CK2_vs_Treat1_Treat2/CK1_CK2_vs_Treat1_Treat2.diff.xls |grep -v "Fold" |sort -k1,1 -k2,2n>CK1_CK2_vs_Treat1_Treat2/CK1_CK2_vs_Treat1_Treat2.diff.bed
computeMatrix reference-point --referencePoint center -b 3000 -a 3000 -R CK1_CK2_vs_Treat1_Treat2/CK1_CK2_vs_Treat1_Treat2.diff.bed -S CK1.bw CK2.bw Treat1.bw Treat2.bw -o CK1_CK2_vs_Treat1_Treat2/CK1_CK2_vs_Treat1_Treat2.diffpeak_distribute.gz
plotHeatmap --matrixFile CK1_CK2_vs_Treat1_Treat2/CK1_CK2_vs_Treat1_Treat2.diffpeak_distribute.gz --dpi 200 --heatmapHeight 15 --kmeans 6 --refPointLabel peak --whatToShow 'heatmap and colorbar' --outFileName CK1_CK2_vs_Treat1_Treat2/CK1_CK2_vs_Treat1_Treat2.diffpeak_Heatmap.png
plotHeatmap --matrixFile CK1_CK2_vs_Treat1_Treat2/CK1_CK2_vs_Treat1_Treat2.diffpeak_distribute.gz --dpi 200 --heatmapHeight 15 --kmeans 6 --refPointLabel peak --whatToShow 'heatmap and colorbar' --outFileName CK1_CK2_vs_Treat1_Treat2/CK1_CK2_vs_Treat1_Treat2.diffpeak_Heatmap.pdf
```

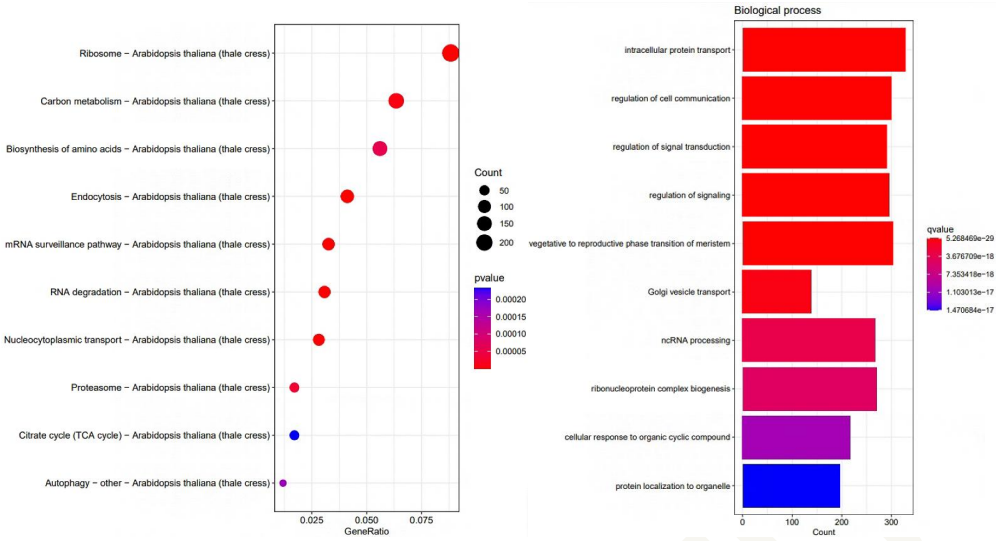
8.3 peak 注释

```
/mnt/RAID-5/MD0/bioinformatics/soft/mamba/bin/Rscript ../../Scripts/ChIPpeakAnnov2.r -p CK1_CK2_vs_Treat1_Treat2.diff.bed --txdb ../../06.Anno/TransDb/AT.sqlite --name CK1_CK2_vs_Treat1_Treat2 --outdir Anno --chrlen ../../06.Anno/AT.len
```

8.4 Peak 关联基因富集分析

GO 和 KEGG 富集分析，获得 id

```
cat Anno/CK1_CK2_vs_Treat1_Treat2.gene.ann|cut -f 14|sort|uniq >id
/mnt/RAID-5/MD0/bioinformatics/soft/mamba/bin/Rscript ../../Scripts/enrichment.r id AT 10
```



九、Motif 分析

HOMER（Hypergeometric Optimization of Motif EnRichment）是一个常用的用于预测 DNA 结合蛋白（TF）结合位点和寻找基因调控模体的软件。下面是使用 HOMER 预测 motif 的基本步骤：

9.1HOMER 安装

下载文件：

使用说明：下载 configureHomer.pl 文件（右键点击并选择"链接另存为"），并将其放置在专用于 HOMER 的目录中（例如 homer/）。通过键入 "perl configureHomer.pl" 运行该脚本

http://homer.ucsd.edu/homer/configureHomer.pl

9.2 预测 motif

使用以下命令运行 HOMER 的 `findMotifsGenome.pl` 脚本进行 motif 预测：

```
perl /mnt/RAID-5/MD0/bioinformatics/soft/bin/findMotifsGenome.pl CK1_CK2_vs_Treat1_Treat2.diff.bed ../02.Mapping/Arabidopsis_thaliana.TAIR10.genome.fa CK1_CK2_vs_Treat1_Treat2 -size given -len 6,8,10,12 -mset plants
```

9.3 解析结果

HOMER 将生成多个输出文件，包括一个 HTML 报告、富集分析结果和预测的 motif 序列。

- HTML 报告：在浏览器中打开 `output_directory_name/index.html` 文件查看报告，其中包含预测的 motif Logo 图和富集分析结果等信息。

- 富集分析结果：可以在 HTML 报告中查看或使用 `output_directory_name/homerMotifs.all.motifs` 文件获取。

- 预测的 motif 序列：可以在 `output_directory_name/homerResults` 目录中找到，以 `knownResults.txt` 和 `denovoResults.txt` 命名。

以上是使用 HOMER 进行 motif 预测的基本步骤。您可以根据实际需求和数据进行进一步的参数调整和结果解析。更详细的用法和选项可以参考 HOMER 的官方文档。

十、生物信息学软件安装方法

对于初学者软件使用非常重要。大家都是学生物的，软件一些参数用法结合一下生物学意义相对来说容易理解，但是可能对大家比较困难的是软件使用之前的工作--软件安装。由于不同的软件需要的依赖（包括种类和版本）不同或者使用的是公用计算机集群你根本无权限安装，导致使用某些软件，但软件安装困难。所以大家需要掌握一些软件安装的技巧与方法。另外在安装时候注意软件的一些版本问题，尽量选择最新版本。

本处主要讲你没有管理员权限安装软件的方法，即安装到自己目录下面方法。

~/soft_install/

10.1 perl 模块安装

直接运行程序：

perl get_quality_reportV2.pl

报错：

所以我们知道本处缺少 Counting.pm 模块，这样我们浏览器检索或者在 CPAN((http://search.cpan.org/))上直接下载 Counting.pm 模块即可

直接在程序中加入模块的路径即可。

```
BEGIN{
  unshift @INC, path/Math-Counting-0.1307/lib/";
}
```

10.2 R 包安装

我们运行一些 R 语言程序时经常出现找不到某个 package。对于这种报错，缺哪一个就下哪一个进行安装。

运行命令：

Rscript veen.R

报错信息：

Error in library(VennDiagram) : 不存在叫‘VennDiagram’这个名字的程辑包

首先下载所需要的 package，像本处为 VennDiagram，这样我们浏览器检索或者 bioconductor(http://www.bioconductor.org/)或者 CRAN(https://cran.r-project.org/)上下载 ggplot2 即可。

解决方案：

下载到 VennDiagram_1.6.20.tar.gz,然后用下面命令（针对无管理员权限，安装自己目录下）安装即可（默认版本的 R）。

```
R CMD INSTALL VennDiagram_1.6.20.tar.gz
```

非默认版本 R 的无权限安装：

打开 R:R

install.packages("VennDiagram_1.6.20.tar.gz")

10.3 Python 包安装

注意使用的 python 版本，Python2 与 Python3 差别较大，因此安装时要关注软件需要的 python 版本。

安装步骤：

解压：unzip pgltools-master.zip

cd pgltools-master/Module/

python setup.py install --user

也可以采用 pip 安装：

pip install -i http://mirrors.aliyun.com/pypi/simple/ --trusted-host mirrors.aliyun.com pip install --user mappy --user

10.4 C 包

一般安装方式，要看软件的 readme。通用的模式是

1. ./configure --prefix=$PWD

   建立 Makefile 文件

2. make。这一步就是编译，大多数的源代码包都经过这一步进行编译

   根据 Makefile 进行编译，生成可执行文件，可执行文件放在当前目录， 尚未被安装到预定安装目录中。

3. make install

   会根据 Makefile 中的 install 选项， 将上一步编译完的数据安装到默认目录中,一般可以指定安装目录

下载：

https://github.com/mummer4/mummer/releases/download/v4.0.0beta2/mummer-4.0.0beta2.tar.gz

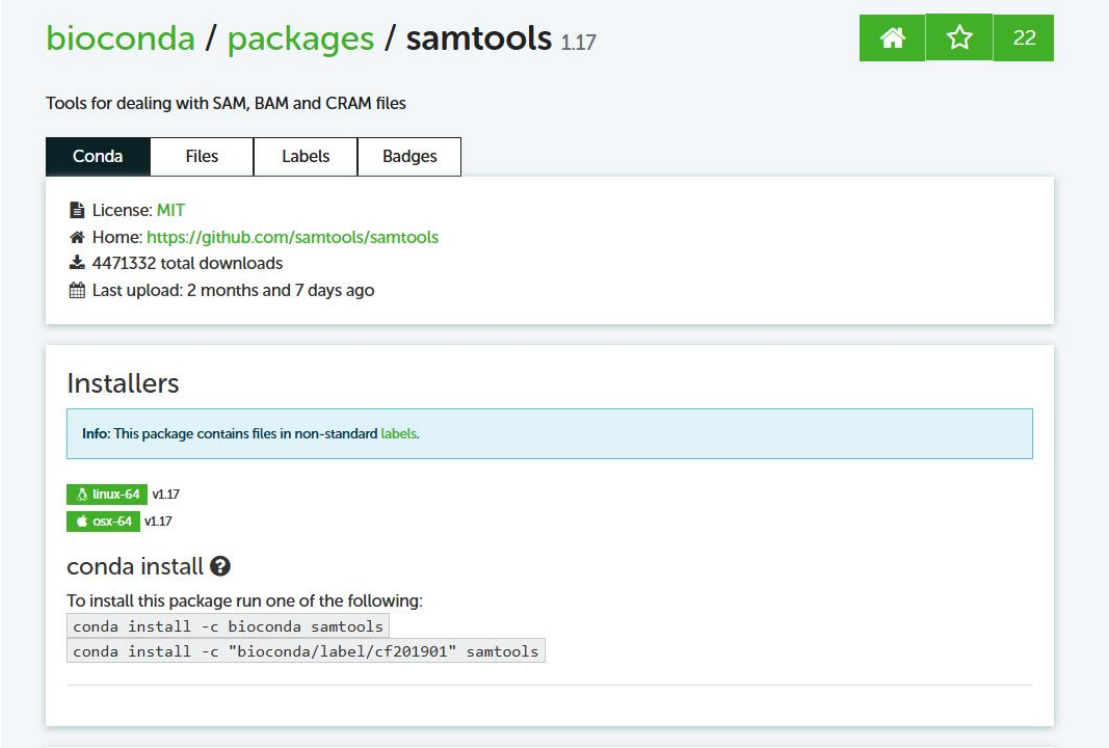
./configure --prefix=$PWD

make

make install

10.5 conda 安装

必应检索：conda samtools



10.6 环境变量

环境变量是软件运行所需要的依赖环境，在自定义安装软件的时候，经常需要配置环境变量，下面列举出各种对环境变量的配置方法。常用的添加环境变量方法：

当前界面生效。直接运行下面命令：

export PATH=path/mcl/current/bin/:$PATH

本账户永久有效。上述命令添加到~/.bashrc 这个文件中，并在添加完成后使用 source 使其生效。

source ~/.bashrc

1、PATH 类型

如果报下面类似的错误（ command not found）可能就是缺少环境变量（如果软件未安装，则可能缺少软件）：

```
bash: mcl: command not found
```

可以通过下面方式添加环境变量：

export PATH=path/mcl/current/bin/:$PATH

上述命令可以直接当前命令运行窗口执行(只在当前窗口有效），也可以写到 home 路径下的.bashrc 中（此账户永久有效）。当写入.bashrc 中，必须使用下面命令使环境变量生效：

source ~/.bashrc

LD_LIBRARY_PATH 类型

软件运行时候遇到的 lib××× not found 的问题，一般是缺少此种类型的环境变量或者版本过低。如下面错误：

trf/trf: /lib64/libc.so.6: version `GLIBC_2.14' not found (required by trf/trf)

其他，如报缺少 zlib 等报错，一般添加下面环境变量即可：

```
zlib_lib="/usr/lib32"##改为实际安装路径
zlib_inc="/usr/include"##改为实际安装路径
export CPPFLAGS="-I${zlib_inc} ${CPPFLAGS}"
export LD_LIBRARY_PATH="${zlib_lib}:${LD_LIBRARY_PATH}"
export LDFLAGS="-L${zlib_lib} -Wl,-rpath=${zlib_lib} ${LDFLAGS}"
```

Python 相关的环境变量：PYTHONPATH（如报缺少某一个 module)