系统生物学课程数据分析实践 1

小组共同完成以下实践作业。该作业为基于下述文献的数据分析复现，分为 3 个部分：RNA-seq 数据分析，ChIP-seq 数据分析，及 ATAC-seq 数据分析。请首先阅读该文献了解其生物学背景和研究内容，然后按要求合作完成作业，并于 4 月 9 日之前以小组为单位提交作业。

（注： 该作业工作量偏大，建议尽快开始。）

项目背景：

Comprehensive characterization of three classes of Arabidopsis SWI/SNF chromatin remodelling complexes.

https://doi.org/10.1038/s41477-022-01282-z

数据来源：

NCBI ： GEO Datasets ： GSE193397

注：鉴于原始分析涉及数据量较大，实际分析不要求使用全部数据，仅需使用部分数据完成必要的分析和作图即可。

1）对于 RNA-seq 的上游分析（服务器端）仅处理以下样本即可：

brd1/2/13：SRR17446281，SRR17446282

syd：SRR17446275，SRR17446276

minu1/2：SRR17446266，SRR17446267

swp73b：SRR17446260，SRR17446261

an3：SRR17446254，SRR17446255

对于 RNA-seq 的下游分析（R 中），可使用之前自己处理的上游分析结果，或直接从 NCBI 下载原始分析处理好的各组突变体和野生型的基因表达矩阵：

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE193095

2）对于 ChIP-seq，全部分析仅处理以下样本即可：

BRD1：SRR17487760，SRR17487761，SRR17487762

SYD：SRR17487754，SRR17487755，SRR17487756

MINU2：SRR17487745，SRR17487746，SRR17487747

3）对于 ATAC-seq，全部分析仅处理以下样本即可：

brd1/2/13：SRR18125496，SRR18125497

syd：SRR17535790，SRR17535791

minu1/2：SRR18125498，SRR18125499

wild type：SRR17535796，SRR17535797

工具参考：

1） 服务器端：

sra-tools (3.0.0)

samtools (1.16.1)

idr (2.0.4.2)

HISAT2（2.1.0）
subread（2.0.0）
bowtie2（2.3.5.1）
MACS2（2.2.7.1）
DeepTools（3.4.3）
Trim Galore（0.6.6）
Homer（4.11.1）

2） R 包（4.1.0 for R）：
edgeR（3.28.1）
Pheatmap（1.0.12）
gplots（3.0.3）
ChIPseeker (1.22.1)
DiffBind（2.16.0）
clusterProfiler（3.16.1）

3） 其他软件：
JBrowse (https://github.com/GMOD) / IGV

前期准备：
1 搭建环境
1.1） 在服务器自行安装 anaconda3 或其他 conda：
https://repo.anaconda.com/archive/Anaconda3-2022.10-Linux-x86_64.sh
构建 conda 环境，安装包括如下包
conda create --name sys_bio python=3.7 matplotlib
conda activate sys_bio

conda install -c bioconda hisat2=2.1.0
conda install -c bioconda subread=2.0.0
conda install -c bioconda bowtie2=2.3.5.1
conda install -c bioconda macs2=2.2.7.1
conda install -c bioconda deeptools=3.4.3
conda install -c bioconda trim-galore=0.6.6
conda install -c bioconda homer=4.11
conda install -c bioconda sra-tools=3.0.0
conda install -c bioconda samtools=1.16.1
conda install -c bioconda idr=2.0.4.2

conda deactivate

1.2）在本地或服务器或安装 R 及上述的 R 包

1.3）在本地或服务器安装基因组可视化工具 JBrowse，或 IGV

2.数据下载

2.1 拟南芥（TAIR10）参考基因组及注释（来源：Ensembl plant）：

参考基因组：

http://ftp.ebi.ac.uk/ensemblgenomes/pub/release-52/plants/fasta/arabidopsis_thaliana/dna/Arabidopsis_thaliana.TAIR10.dna.toplevel.fa.gz

参考基因组注释：

gff3 格式：

http://ftp.ebi.ac.uk/ensemblgenomes/pub/release-52/plants/gff3/arabidopsis_thaliana/Arabidopsis_thaliana.TAIR10.52.gff3.gz

gtf 格式：

http://ftp.ebi.ac.uk/ensemblgenomes/pub/release-52/plants/gtf/arabidopsis_thaliana/Arabidopsis_thaliana.TAIR10.52.gtf.gz

根据参考基因注释构建 BED 文件：

从 gff 注释文件中提取（也可在 R 中完成）

```
awk -F ';' '{print $1}' Arabidopsis_thaliana.TAIR10.52.gff3 | awk -F "\t" '{if($3~/gene/) print $1"\t"$4"\t"$5"\t"$9}' | sed 's/ID=gene://g' > Arabidopsis_thaliana.TAIR10.52.bed
```

参考基因组 Symbol 对应基因 ID 转换：

从 gtf 注释文件中提取（也可在 R 中完成）

```
awk -F "\t" '{if ($3~/gene/) print $9}' Arabidopsis_thaliana.TAIR10.52.gtf | awk -F "" '{print $2"\t"$4}' > Arabidopsis_thaliana.TAIR10.52.Symbol.txt
grep "araport11" Arabidopsis_thaliana.TAIR10.52.Symbol.txt |cut -f 1 > nosymbol.txt
paste nosymbol.txt nosymbol.txt > AGI2symbol.txt
sed '/araport11/d' Arabidopsis_thaliana.TAIR10.52.Symbol.txt > havesymbol.txt
cat havesymbol.txt AGI2symbol.txt | sort > Arabidopsis_thaliana.TAIR10.52.Symbol.txt
rm AGI2symbol.txt havesymbol.txt nosymbol.txt
```

2.2 原始数据下载：

NCBI ： GEO Datasets ： GSE193397 ：

注意，本项目中的 RNA-Seq 为单端测序

主要流程参考：

1、RNA-Seq:

1.1）构建参考基因组的 HISAT2 索引（HISAT2）

1.2）测序数据解压、质控和过滤（sra-tools，fastqc，trim-galore）

1.3）利用 HISAT2 进行序列比对（HISAT2）

1.4）比对结果文件排序和格式转换（samtools）

1.5）利用 featureCounts 对比对结果进行 Reads 计数 （subread）

1.6）在 R 中利用 edgeR 等包对各组样品进行 MDS 分析、差异表达分析等并做图 （R）

1.7）根据差异基因相对表达水平，分析突变体间的皮尔森相关性，并利用 pheatmap，gplots 等包绘制相关热图 （R）

2、ChIP-Seq:

2.1) 构建参考基因组的 bowtie2 索引 (bowtie2)

2.2) 测序数据解压、质控和过滤（sra-tools，fastqc，trim-galore）

2.3）利用 bowtie2 进行序列比对 (bowtie2)

2.4）比对结果文件排序和格式转换 (samtools)

2.5） 利用 bamCoverage 函数将 bam 格式转 bigWig 格式，利用 multiBamSummary 和 plotCorrelation 函数分析样本间的 Spearman 相关性 （deeptools）

2.6）利用 MACS2 对排序后的 bam 文件进行 Peak Calling （注：本项目含有对照组，在 NCBI 可查看测序数据为对照组还是实验组。）(MACS2)

2.7）评估每个实验组的两个生物学重复得到的 peaks 的一致性，保留在两个生物学重复中均检测到(irreproducible discovery rate ≤0.05)的 peaks （idr）

2.8) 利用 ChIPseeker 等包对 Peaks 注释, 包括 genomic location annotation , nearest gene annotation 等，根据注释到基因组位置信息，统计不同实验组间的 specified chromatin features 差异并做堆叠条形图。 根据注释到的基因信息，统计不同实验组间的 bound genes 差异并做韦恩图（R ）

2.9）利用 deeptools 的 plotHeatmap 函数, 绘制不同实验组的基因富集热图（deeptools）

2.10）在基因组可视化工具内，对 bigWig 格式文件进行可视化，观察不同实验组的 Reads 分布区域差异。(JBrowse or IGV)

3、ATAC-Seq:

3.1) 构建参考基因组的 bowtie2 索引 (bowtie2)

3.2) 测序数据解压、质控和过滤（sra-tools，fastqc，trim-galore）

3.3）利用 bowtie2 进行序列比对 (bowtie2)

3.4）比对结果文件排序和格式转换 (samtools)

3.5） 利用 bamCoverage 函数将 bam 格式转 bigWig 格式。(deeptools)

3.6）利用 MACS2 对排序后的 bam 文件进行 Peak Calling (MACS2)

3.7）评估每个实验组的两个生物学重复得到的 peaks 的一致性，保留在两个生物学重复中均检测到(irreproducible discovery rate ≤0.05)的 peaks （idr）

3.8) 利用 ChIPseeker 等包对 Peaks 注释。(R)

3.9）利用 DiffBind 包对 Peaks 进行主成分分析并做图 (R)

3.10）利用 DiffBind 包对 Peaks 进行差异 peaks 分析，鉴定不同突变体相对于野生型的 differentially accessible regions（DARs），并分析这些 DAR 相关的基因 (R)

3.11) 统计不同突变体找到的所有 DARs 相关的基因的并集, 定量这些突变体在这些基因上的 peak reads，并进行皮尔森相关分析，利用 pheatmap，gplots 等包绘制相关热图（R）

3.12）利用 ChIPseeker 包，对 DARs 进行 Peaks 注释，根据注释到基因组位置信息，统计不同突变体的下调 DARs 间的 specified chromatin features 差异并做堆叠条形图，并比较各个突变体内下调 DARs 中各种 specified chromatin features 占该突变体全部 ARs 的各种 specified chromatin features 的比例。(R)

3.13）根据各组突变体中找到的下调 DAR 相关基因，利用 deeptools 的 plotProfile 功能绘制各组 ATAC-seq 数据在这些基因的 profile plot （deeptools）

3.14）根据各组突变体中找到的下调 DAR 相关基因，利用 deeptools 的 plotProfile 功能绘制之前各组 ChIP-seq 数据在这些基因的 profile plot （deeptools）

作业要求：

1， 通过分别进行 RNA-seq，ChiP-seq，ATAC-seq 数据分析，完成对原文 Fig. 3 – Fig. 5 的复现（其中部分 panels 不做要求）。结合原文，解释得到的结果可以支持文章的对应结论。

2， 对以下各主要步骤，提供必要的代码、脚本和注释。对于在线工具或本地程序，提供必要的运行截图。

RNA-seq (Fig. 3):

1） 构建参考基因组索引
2） 数据的质控和过滤
3） 序列比对
4） Reads 计数
5） 样本间的 MDS 分析 **（Fig3. a）**
6） 不同突变体和 WT 间的差异表达分析，以及 DEGs 数目统计 **（Fig3. b）**
7） DEGs 在不同突变体中的 log_2FC 表达热图和聚类可视化 **（Fig3. c）**
8） 不同突变体间皮尔森相关性分析的热图可视化 **（Fig3. d）**
9） 部分代表性基因在在不同突变体中的 log_2FC 表达热图可视化 **（Fig3. e）**

ChIP-seq (Fig. 4，其中 Fig. 4e 不做要求):

1） 构建参考基因组索引
2） 数据的质控和过滤
3） 序列比对
4） 样本间相关性分析 **（Fig4. a）**
5） Peak Calling
6） 可视化各组 peaks 的 genomic location annotation **（Fig4. b）**
7） 各组 peaks 的 bound genes 重叠分析 **（Fig4. c）**
8） 不同实验组的富集热图 **（Fig4. d）**
9） 各个实验组的 bw 文件在基因组可视化工具内的展示 **（Fig4. f）**

ATAC-seq (Fig. 5，其中 Fig. 5c 不做要求):

1） 构建参考基因组索引
2） 数据的质控和过滤
3） 序列比对
4） Peak Calling
5） 使用 DiffBind 包对 Peaks 进行主成分分析 **（Fig5. a）**
6） 不同突变体间皮尔森相关性分析的热图可视化 **（Fig5. b）**
7） 可视化各组突变体下调 DARs 的 genomic location annotation 类型 **（Fig5. d）**
8） 可视化各组突变体下调 DARs 的 genomic location annotation 相对于整体 ARs 的百分比 **（Fig5. e）**
9） 各组 ATAC-seq 数据在各组下调 DAR 相关基因上的 profile plot **（Fig5. f）**
10） 各组 ChIP-seq 数据在各组下调 DAR 相关基因上的 profile plot **（Fig5. g）**

3， 对于全部分析，不限制程序、工具或编程语言。完成上述主要流程，得到最终需要提交的图和结论即可。

4， 作业以小组为单位打包提交，包含： 1）主要的图和其对应的分析流程简述，并在该文档中说明组内分工情况。 2）必要的代码和/或运行截图等其他内容。