ChIP-seq概述及分析实战
NGS data with Linux-R: ChIP-seq introduction and analysis

徐春晖
用户信息素养讲堂

中国 - 上海 - 生科院 - 2018年11月29号
CONTENT

01 / ChIP-seq overview
NGS sequencing methods review

02 / Chip-seq workflow
Tips for design an experiment

03 / Resources for ChIP-seq
How to get the extra data and information

04 / Analysis example
Analysis following a published paper
ChIP-seq overview

What is ChIP (Chromatin Immunoprecipitation)
NGS application brief overview
What is ChIP-seq technology

**Introduction**
A method used to analyze protein interactions with DNA.

**Tasks**
Mapping the loci of TFs, nucleosomes, histone modifications, chromatin remodeling enzymes, chaperones, and polymerases genome-wide.

**Extension methods**
- ChIP-exo sequencing
- DamID sequencing

**Subsequent analysis**
- Differential binding analysis
- Motif /GO /KEGG analysis
- Conjoint Analysis with RNA-seq/BS-seq/ATAC-seq
Chromatin immunoprecipitation (ChIP)

1. Activator proteins bind to pieces of DNA called enhancers. Their binding causes the DNA to bend, bringing them near a gene promoter, even though they may be thousands of base pairs away.

2. Other transcription factor proteins join the activator proteins, forming a protein complex which binds to the gene promoter.

3. This protein complex makes it easier for RNA polymerase to attach to the promoter and start transcribing a gene.

4. An insulator can stop the enhancers from binding to the promoter, if a protein called CTCF forms for the sequence CCCCTC, which occurs in all insulators, binds to it.

5. Methylation, the addition of a methyl group to the C nucleotides, prevents CTCF from attaching to the insulator, turning it off, allowing the enhancers to bind to the promoter.

Note: This diagram simplifies the DNA greatly—promoters, enhancers, and insulators can be thousands or even hundreds of base pairs long.
NGS application brief overview
An overview of ChIP-Seq technology

Chromatin ImmunoPrecipitation followed by Sequencing

1. Pretreatment of sample
   I. Cross-linking.
   II. Hydrolysis.
   III. Lyses of DNA fragments (shearing).
   IV. Add specific antibody (with beads) against the protein of interest.
   V. Separate DNA from the protein (unlinking), and purify the DNA
   VI. DNA fragments are amplified and fluorescently tagged, library construction

2. Data processing
   I. Reads preprocessing
   II. Quality filtering (trimming)
   III. FASTQ quality report
   IV. Alignments
   V. Peak calling
   VI. Peak annotation with genomic context
   VII. Differential binding analysis
   VIII. GO term or KEGG enrichment analysis
   IX. Motif analysis

3. Visualization
Q&A in ChIP-seq design

What is Control? Is it necessary?
These are essential to eliminate false positives as there is always some signal on open chromatin. Two types of controls are often used:

An “input” DNA sample (fragmented genomic DNA), the one that has been cross-linked and sonicated but not immuno-precipitated
An IgG “mock”-ChIP, using an antibody that will not bind to nuclear proteins to generate immuno-precipitated DNA that should be random. IP with non-specific antibody such as immunoglobulin G, IgG.

Motivation:
Why look for differential binding (DB) between biological conditions?
• The null hypothesis is easy to define, i.e., is binding equal between libraries?
• Results may be more relevant, as DB sites might contribute to biological differences.
• DB could answer biological questions in clinical investigation.
Chip-seq analysis workflow

A typical ChIP-Seq data analysis workflow
How to evaluate a ChIP-seq results
A typical ChIP-Seq data analysis workflow

1. Reads preprocessing
   1. Quality filtering (trimming)
   2. FASTQ quality report
2. Alignments: *Bowtie2* or *rsubread*
3. Alignment stats
4. Peak calling: *MACS2*, *BayesPeak*
5. Peak annotation with genomic context
6. Differential binding analysis
7. GO term or KEGG term enrichment analysis
8. Motif analysis
A typical ChIP-Seq data analysis workflow

**Sequencing Depth**
Mammalian TFs and chromatin modifications: 20M
Proteins with more binding sites (e.g., RNA Pol II): 60M
Sequencing depth rules of thumb:
>10M reads for narrow peaks
>20M for broad peaks

**Quality Metrics**
*FastQC* can be used for an overview of the data quality.
*Trimmomatic* can deal with fastq data.
Reads length is not important in ChIP-seq.

**Read Mapping**
It is important to consider the percentage of uniquely mapped reads reported by the mapper.
Most peak calling algorithms would ignore the multi-mapping reads (filtered out).

**Design**
Long & paired-end reads useful but not essential.
2 replicates couldn’t increase the result.
Control samples should be sequenced significantly deeper.

**Important considerations**
Resources for ChIP-seq
Tools used in ChIP-seq analysis
Forum about analysis discussion
Tools used in ChIP-seq analysis

Mapping to genome
- Bwa
- Bowtie1
- Bowtie2
- ...

Peaks calling
- CisGenome
- ERANGE
- FindPeaks
- F-Seq
- GLITR
- MACS14
- MACS2
- PeakSeq
- QuEST
- ...

Peaks calling
- BayesPeak
- PICS
- MOSAiCS
- iSeq
- ChIPseqR
- CSAR
- ChIP-Seq
- SPP
- NarrowPeaks
- ...

Annotation
- ChIPseeker
- ChIPpeakAnno
- clusterProfiler
- ...
Tools used in ChIP-seq analysis

Visualization in IGV/UCSC

- Costumer track
- Reference track
- Reference version
- Interest loci/gene
Tools used in ChIP-seq analysis

Motif discovery
The MEME Suite
Motif-based sequence analysis tools
Courses about ChIP-seq analysis

Basics of ChIP-seq data analysis

ChIP-seq Analysis with Bioconductor
https://www.r-bloggers.com/chip-seq-analysis-with-bioconductor/

NGS data analysis with R / Bioconductor: ChIP-Seq workflow
http://biocluster.ucr.edu/~rkaundal/workshops/R_feb2016/ChIPseq/ChIPseq.html
Resource for ChIP-seq analysis

Associated paper

Genome-Wide Mapping of in Vivo Protein-DNA Interactions.
(David et al – Science 2007)

Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.
(Tarjei et al – Nature 2007)

ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.
(Landt et al – Genome Research 2012)

Forum

Biostars: BIOINFORMATICS EXPLAINED
R-bloggers
GitHub
Csdn.net
Analysis example

Practice using a published paper
Plot paper-ready figures
Data acquisition

Data resources: GSE93164
Description: Examination of the changes of H3K27Ac in SKNO1 cells after knockdown of ASXL2 using ChIP-seq

Samples (3)
GSM2445788 SKNO1 shASXL2-3 H3K27ac
GSM2445789 SKNO1 shASXL2-2 H3K27ac
GSM2445790 SKNO1 control H3K27ac

SRA download software:
(1) Aspera
(2) sratoolkit (get SRR number)
(3) FTP

# data downloading
$prefetch SRR5145757
2018-05-31T09:56:56 prefetch.2.8.1: 1) Downloading 'SRR5145757'...
2018-05-31T09:56:56 prefetch.2.8.1:  Downloading via https...
2018-05-31T10:06:14 prefetch.2.8.1: 1) 'SRR5145757' was downloaded successfully

# data deposition
$ls ~/ncbi/public/sra/
SRR5145757.sra  SRR5145759.sra.lock  SRR5145759.sra.tmp.123.tmp
Decompression and QC

# decompression
fastq-dump --split-3 SRR5145757.sra
Fastq-dump --split-3 SRR5145759.sra

mkdir fastQC && 
fastqc -o ./fastQC -t 6 SRR5145757.fastq && 
fastqc -o ./fastQC -t 6 SRR5145759.fastq
Decompression and QC

**Summary**
- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

**Kmer Content**

---

Log2 Obs/Exp

---

GTCACGGA
GATCGTG
ACAGCTGC
CAGCTCAG
ATGACGG
CGTATAGC
Mapping to hg19 with bowtie2

# install
bioconda is recommended

# index
bowtie2-build [options]* <reference_in> <bt2_base>

# mapping
bowtie2 -x ~/reference/index/bowtie2/hg19 \
-N 1 -p 5 -U SRR5145757.fastq -S 57.sam 2>SRR5145757_align.log
# N seed mismatch
# pair-end:-1/-2 single-end:U number

# check alignment result
$cat 57_align.log
56665942 reads; of these:
  56665942 (100.00%) were unpaired; of these:
    3982220 (7.03%) aligned 0 times
    39884964 (70.39%) aligned exactly 1 time
    12798758 (22.59%) aligned >1 times
92.97% overall alignment rate
Call peaks with macs2

Control sample: “Input” or “IgG”
- Input: sonicated chromatin without immunoprecipitation
- IgG: “unspecific” IP

MACS2 – program(s)

INPUT DATA: aligned sequence reads
- ChIPed sample “treat”
- Input/gG “control”

OUTPUT FILES:
- pileup.bed
- peaks.xls
- peaks.narrowPeak
- summits.bed
- model.r
- model.pdf
- treat_pileup.bed
- control_lambda.bed

OUTPUT:
- filterdup
- randsample
- predictd
- pileup
- refinepeaks
- refinepeak.bed
- bdgpeakcall
- bdgbroadcall
- bdgcmp
- bdgdiff
- diffpeak
Scan the genome to find regions with tags more than \texttt{mfold} enriched relative to random tag distribution.

**Step 1**
Randomly sample 1000 of these (high quality peaks) and calculate the distance between the modes of their +/- peaks.

**Step 2**
Shift all the tags by $d/2$ toward the 3' end.

**Step 3**
Call peaks with \texttt{macs2}

\textbf{Command line}

\texttt{macs2 calipeak} [-h] -t TFILE [TFILE ...] [-c [CFILE [CFILE ...]]] [-f \{AUTO, BAM, SAM, BED, ELAND, ELANDMULTI, ELANDEXPORT, BOWTIE, BAMPE\}] [-g GSIZE] [--keep-dup KEEPDUPLICATES]

[-s TSIZE] [--bw BW] [-m MFOLD MFOLD] [--fix-bimodal]

[-n NAME] [--nomodel] [-shift SHIFT] [-extsize EXTSIZE]

[-q QVALUE] [-p PVALUE] [--to-large] [--ratio RATIO]

[-f down-sample] [-seed SEED] [--nolambda]

[-slocal SMALLLOCAL] [--local LARGELOCAL] [--broad]

[-broad-cutoff BROADCUTOFF] [--call-summits]
Call peaks with macs2

```bash
# sam to bam and index
samtools sort -O bam @ 5 -o 57.bam 57.sam && samtools indexe57.bam & rm 57.sam

# call peaks using macs2
# installed by bioconda
# macs2 use sample
# macs2 callpeak -t treatment.bam -c control.bam -g hs -B -f BAM -n prefix
# macs2 callpeak -t treatment.bam -c control.bam -g hs -B -f BAMPE -n prefix -q 0.00001
macs2 callpeak -t 57.bam -c 59.bam -g hs -n H3K27ac -B --SPMR 2> call_peak.log
```

参数解读：
- `t` 处理组
- `c` 对照组
- `g` 基因组的选择
- `B` 输出bgd文件，下游bigwig文件生成所需
- `f` 双端测序使用BAMPE，单端的话不需要加参数，默认是auto识别，但是识别不了BAMPE
- `q` 指定阈值
- `--SPMR` MACS will save fragment pileup signal per million reads
Call peaks with macs2

# output file includes:
prefix_model.r
prefix_peaks.narrowPeak
prefix_peaks.xls
prefix_summits.bed
prefix_treat_pileup.bdg

# prefix_model.r 为模型建立示意图，使用Rscript prefix_model.r 命令可以在Linux环境下作图
# prefix_peaks.xls 为包含了peak信息的表格，包括位置信息，堆积信号及pvalue等
# prefix_peaks.narrowPeak 为包含了peak位置及summit位置，p值，q值的文件，可以直接上传到UCSC browser，其中特定列的信息为:
   5th: 整数信号值
   7th: fold-change
   8th: -log10 pvalue
   9th: -log10 qvalue
   10th: summit距离peak起始位点的距离
# prefix_summits.bed 包含每个peak峰顶summit的位置，等同于从narrowPeak文件里面剥离出来的，方便用来寻找motif的binding，同样可以载入UCSC

# prefix_peaks.broadPeak & prefix_peaks.gappedPeak 文件都是为选择 -broad 参数后得到的文件，本质上和
narrowPeak文件一致，不过增加了broadregion，当然summit的定义也有区别，需要自己去指定

# prefix_treat_pileup.bdg 文件为输文件treat组的bedGraph file，prefixi__treat_pvalue.bdg为对照组的
bedGraph file
**Call peaks with macs2**

```bash
# This file is generated by MACS version 2.1.1.20160309
# Command line: callpeak -t 57.bam -c 59.bam -g hs -n H3K27ac -B -SPMR
# ARGUMENTS LIST:
#   name = H3K27ac
#   format = AUTO
#   ChiP-seq file = ['57.bam']
#   control file = ['59.bam']
#   effective genome size = 2.78e+09
#   band width = 500
#   maxdval fold = [5, 50]
#   qvalue cutoff = 5.0e-02
#   Larger dataset will be scaled towards smaller dataset.
#   Range for calculating regional lambda is: 10000 bps and 100000 bps
#   Broad region calling is off
#   Pair-end mode is off
#   MACS will save fragment pileup signal per million reads
#   tag size is determined at 58 bps
#   total tags in treatment: 52963722
#   tags after filtering in treatment: 47157631
#   maximum duplicate tags at the same position in treatment = 1
#   Redundant rate in treatment: 0.10
#   total tags in control: 35451636
#   tags after filtering in control: 28634337
#   maximum duplicate tags at the same position in control = 1
#   Redundant rate in control: 0.19
#   # d = 246
#   # alternative fragment length(s) may be 246 bps
```

<table>
<thead>
<tr>
<th>chr</th>
<th>start</th>
<th>length</th>
<th>ams_summit</th>
<th>pileup</th>
<th>-log10(pvalue)</th>
<th>fold_enrichment</th>
<th>-log10(qvalue)</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>1441375</td>
<td>1441631</td>
<td>257</td>
<td>10.39386</td>
<td>3.10460</td>
<td>1.64280</td>
<td>H3K27ac_peak_1</td>
<td></td>
</tr>
<tr>
<td>chr1</td>
<td>8021844</td>
<td>8022325</td>
<td>282</td>
<td>10.22131</td>
<td>4.49399</td>
<td>11.68888</td>
<td>H3K27ac_peak_2</td>
<td></td>
</tr>
<tr>
<td>chr1</td>
<td>99097155</td>
<td>99100228</td>
<td>233</td>
<td>9909935</td>
<td>27.32</td>
<td>10.20695</td>
<td>4.46632</td>
<td>5.57885</td>
</tr>
<tr>
<td>chr1</td>
<td>9708096</td>
<td>9710134</td>
<td>318</td>
<td>9709466</td>
<td>57.65</td>
<td>8.76356</td>
<td>2.07362</td>
<td>4.43868</td>
</tr>
<tr>
<td>chr1</td>
<td>11885341</td>
<td>11885625</td>
<td>285</td>
<td>11885577</td>
<td>68.01</td>
<td>5.84227</td>
<td>1.74083</td>
<td>1.61455</td>
</tr>
<tr>
<td>chr1</td>
<td>11886453</td>
<td>11886717</td>
<td>633</td>
<td>11886761</td>
<td>64.36</td>
<td>19.47732</td>
<td>3.84892</td>
<td>11.63859</td>
</tr>
<tr>
<td>chr1</td>
<td>11953070</td>
<td>11953478</td>
<td>257</td>
<td>11953478</td>
<td>18.82</td>
<td>5.85335</td>
<td>3.30839</td>
<td>2.28481</td>
</tr>
<tr>
<td>chr1</td>
<td>16513640</td>
<td>16513918</td>
<td>279</td>
<td>16513846</td>
<td>32.79</td>
<td>7.18288</td>
<td>2.81757</td>
<td>3.21371</td>
</tr>
<tr>
<td>chr1</td>
<td>16978764</td>
<td>16978614</td>
<td>251</td>
<td>16978864</td>
<td>181.40</td>
<td>8.29996</td>
<td>1.84951</td>
<td>4.95043</td>
</tr>
<tr>
<td>chr1</td>
<td>21582214</td>
<td>21582713</td>
<td>520</td>
<td>21582426</td>
<td>52.83</td>
<td>5.28589</td>
<td>1.92239</td>
<td>1.72756</td>
</tr>
<tr>
<td>chr1</td>
<td>25559391</td>
<td>25559401</td>
<td>301</td>
<td>25559561</td>
<td>73.47</td>
<td>11.82336</td>
<td>2.48232</td>
<td>6.23793</td>
</tr>
<tr>
<td>chr1</td>
<td>26594668</td>
<td>26594680</td>
<td>233</td>
<td>26594663</td>
<td>17.61</td>
<td>5.98381</td>
<td>3.42483</td>
<td>2.38641</td>
</tr>
<tr>
<td>chr1</td>
<td>27247564</td>
<td>27247817</td>
<td>316</td>
<td>27247783</td>
<td>64.97</td>
<td>19.70312</td>
<td>2.53794</td>
<td>5.98248</td>
</tr>
<tr>
<td>chr1</td>
<td>28564209</td>
<td>28564238</td>
<td>300</td>
<td>28564257</td>
<td>36.43</td>
<td>8.25000</td>
<td>2.87941</td>
<td>4.84808</td>
</tr>
<tr>
<td>chr1</td>
<td>28764823</td>
<td>28765531</td>
<td>708</td>
<td>28765080</td>
<td>35.83</td>
<td>8.39563</td>
<td>2.99002</td>
<td>4.15423</td>
</tr>
<tr>
<td>chr1</td>
<td>31537116</td>
<td>31537381</td>
<td>266</td>
<td>31537331</td>
<td>47.36</td>
<td>4.93333</td>
<td>1.92516</td>
<td>1.53674</td>
</tr>
<tr>
<td>chr1</td>
<td>32968371</td>
<td>32986642</td>
<td>272</td>
<td>32986851</td>
<td>25.50</td>
<td>8.87152</td>
<td>3.78009</td>
<td>5.52391</td>
</tr>
<tr>
<td>chr1</td>
<td>35656889</td>
<td>35657139</td>
<td>251</td>
<td>35657087</td>
<td>66.79</td>
<td>6.86996</td>
<td>1.99300</td>
<td>2.97660</td>
</tr>
<tr>
<td>chr1</td>
<td>35811254</td>
<td>35811499</td>
<td>246</td>
<td>35811389</td>
<td>17.00</td>
<td>8.02822</td>
<td>4.28817</td>
<td>3.86936</td>
</tr>
<tr>
<td>chr1</td>
<td>36234570</td>
<td>36234969</td>
<td>400</td>
<td>36234643</td>
<td>48.58</td>
<td>9.12372</td>
<td>2.66344</td>
<td>4.72957</td>
</tr>
</tbody>
</table>

**Input files and parameters setting**

- **Command line:** callpeak -t 57.bam -c 59.bam -g hs -n H3K27ac -B -SPMR
- **Arguments list:**
  - name = H3K27ac
  - format = AUTO
  - ChiP-seq file = ['57.bam']
  - control file = ['59.bam']
  - effective genome size = 2.78e+09
  - band width = 500
  - maxdval fold = [5, 50]
  - qvalue cutoff = 5.0e-02
- Larger dataset will be scaled towards smaller dataset.
- Range for calculating regional lambda is: 10000 bps and 100000 bps
- Broad region calling is off
- Pair-end mode is off
- MACS will save fragment pileup signal per million reads
- tag size is determined at 58 bps
- total tags in treatment: 52963722
- tags after filtering in treatment: 47157631
- maximum duplicate tags at the same position in treatment = 1
- Redundant rate in treatment: 0.10
- total tags in control: 35451636
- tags after filtering in control: 28634337
- maximum duplicate tags at the same position in control = 1
- Redundant rate in control: 0.19
- # d = 246
- alternative fragment length(s) may be 246 bps

**Peaks information**

- H3k27AC_peaks.xls

---

SICLS生命科学图书馆
Call peaks with macs2

<table>
<thead>
<tr>
<th>chr</th>
<th>H3K27ac_peaks.narrowPeak</th>
<th>Signal value</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>1441374</td>
<td>4141631</td>
</tr>
<tr>
<td>chr2</td>
<td>80016643</td>
<td>8601325</td>
</tr>
<tr>
<td>chr3</td>
<td>9898734</td>
<td>9980802</td>
</tr>
<tr>
<td>chr4</td>
<td>9970895</td>
<td>9971313</td>
</tr>
<tr>
<td>chr5</td>
<td>1186540</td>
<td>1186525</td>
</tr>
<tr>
<td>chr6</td>
<td>1186562</td>
<td>1186725</td>
</tr>
<tr>
<td>chr7</td>
<td>1186563</td>
<td>1186570</td>
</tr>
<tr>
<td>chr8</td>
<td>1563367</td>
<td>1563398</td>
</tr>
<tr>
<td>chr9</td>
<td>16970875</td>
<td>16971041</td>
</tr>
<tr>
<td>chr10</td>
<td>21592293</td>
<td>21592713</td>
</tr>
<tr>
<td>chr11</td>
<td>2055940</td>
<td>20559491</td>
</tr>
<tr>
<td>chr12</td>
<td>26498857</td>
<td>26469870</td>
</tr>
<tr>
<td>chr13</td>
<td>27274563</td>
<td>27274879</td>
</tr>
<tr>
<td>chr14</td>
<td>28556208</td>
<td>28556238</td>
</tr>
<tr>
<td>chr15</td>
<td>28764822</td>
<td>28765531</td>
</tr>
<tr>
<td>chr16</td>
<td>31573115</td>
<td>31573781</td>
</tr>
<tr>
<td>chr17</td>
<td>32986370</td>
<td>32986642</td>
</tr>
<tr>
<td>chr18</td>
<td>35656888</td>
<td>35657139</td>
</tr>
<tr>
<td>chr19</td>
<td>35811215</td>
<td>35811499</td>
</tr>
<tr>
<td>chr20</td>
<td>36234569</td>
<td>36234969</td>
</tr>
<tr>
<td>chr21</td>
<td>37948529</td>
<td>37948673</td>
</tr>
<tr>
<td>chr22</td>
<td>41158614</td>
<td>41159066</td>
</tr>
<tr>
<td>chr23</td>
<td>41175809</td>
<td>41175962</td>
</tr>
<tr>
<td>chr24</td>
<td>41832499</td>
<td>41832751</td>
</tr>
<tr>
<td>chr25</td>
<td>44841737</td>
<td>44844716</td>
</tr>
<tr>
<td>chr26</td>
<td>44847692</td>
<td>44857680</td>
</tr>
<tr>
<td>chr27</td>
<td>46271718</td>
<td>46272614</td>
</tr>
<tr>
<td>chr28</td>
<td>47798888</td>
<td>47799068</td>
</tr>
<tr>
<td>chr29</td>
<td>51425197</td>
<td>51425461</td>
</tr>
<tr>
<td>chr30</td>
<td>51426345</td>
<td>51426735</td>
</tr>
<tr>
<td>chr31</td>
<td>51437864</td>
<td>51438035</td>
</tr>
<tr>
<td>chr32</td>
<td>51438186</td>
<td>51438378</td>
</tr>
<tr>
<td>chr33</td>
<td>51434635</td>
<td>51434908</td>
</tr>
<tr>
<td>chr34</td>
<td>51983288</td>
<td>51983581</td>
</tr>
<tr>
<td>chr35</td>
<td>52195679</td>
<td>52196150</td>
</tr>
<tr>
<td>chr36</td>
<td>55351202</td>
<td>55352218</td>
</tr>
<tr>
<td>chr37</td>
<td>55356092</td>
<td>55356297</td>
</tr>
<tr>
<td>chr38</td>
<td>63832766</td>
<td>63839201</td>
</tr>
<tr>
<td>chr39</td>
<td>63899457</td>
<td>63899722</td>
</tr>
<tr>
<td>chr40</td>
<td>67316739</td>
<td>67316925</td>
</tr>
<tr>
<td>chr41</td>
<td>67398052</td>
<td>67398999</td>
</tr>
<tr>
<td>chr42</td>
<td>74663838</td>
<td>74663667</td>
</tr>
<tr>
<td>chr43</td>
<td>78147613</td>
<td>78148240</td>
</tr>
</tbody>
</table>
bigwig file generation

Deeptools

For visualizing continuous data, e.g. in the UCSC Genome Browser or IGV, bigWig files come in really handy.

Remember that there are 2 deepTools for bam → bigWig conversion:

- `bamCoverage`: for individual files (like those shown here)
- `bamCompare`: to normalize two files to each other
bigwig file generation

# generate individually
bamCoverage --bam 57.bam -o 57.bw --binSize 10
bamCoverage --bam 59.bam -o 59.bw --binSize 10

# --binSize set the resolution

# generate using two bam files
bamCompare -b1 57.bam -b2 59.bam -o IGV.bw --scaleFactorsMethod readCount --ratio ratio
bigwig file generation

Deeptools-plot

```bash
computeMatrix scale-regions -p 10 \
  -R hg19.bed \
  -S ATAC_1.bw \
  -b 3000 -a 3000 \
  --regionBodyLength 5000 \
  --skipZeros -o heatmap.gz \
  --outFileNameMatrix Values.tab 2> computeMatrix.log

plotHeatmap -m heatmap.gz -out heatmap.png --heatmapWidth 10 --sortUsing mean
```
Binding factor motifs discovery

HOMER is designed to find differential motifs, which means the users need to provide two sets of data: one for experimental data, and one for control data.

It uses ZOOPS (zero or once occurrence per sequence) and hypergeometric enrichment algorithms to perform the analysis, initially designed to connect ChIP-seq analysis and promoter data analysis, but now widely used.

It has several modules:
- findMotifs.pl
- findMotifsGenome.pl
- homer2
- scanMotifsGenomeWide.pl

# installed by bioconda
# install interest genome
perl configureHomer.pl -install hg19

#
findMotifsGenome.pl H3K27ac_peaks.narrowPeak hg19 ./motifs/ -size given

# R logo plot
ggseqlogo package
Annotate peaks with genomic context

Input file: H3K27ac_peaks.narrowPeak

Packages: ChIPseeker org.Hs.eg.db
          TxDb.Hsapiens.UCSC.hg19.knownGene

```
> setwd("D:/Xu chunhui/ChIP-seq introduction")
> library(ChIPseeker)
> library(VennDiagram)
> library(org.Hs.eg.db)
> require(TxDB.Hsapiens.UCSC.hg19.knownGene)
> txdb = TxDb.Hsapiens.UCSC.hg19.knownGene
> peaks <- readPeakFile("H3K27ac_peaks.narrowPeak")
> peaks_anno = annotatePeak(peaks, tssRegion=c(-1000, 1000), TxDb=txdb, annoDb = "org.Hs.eg.db")

>> preparing features information... 2018-06-01 13:40:39
>> identifying nearest features...    2018-06-01 13:40:39
>> calculating distance from peak to TSS...  2018-06-01 13:40:40
>> assigning genomic annotation...    2018-06-01 13:40:40
>> adding gene annotation...         2018-06-01 13:40:54

'select()' returned 1:many mapping between keys and columns
>> assigning chromosome lengths      2018-06-01 13:40:54
>> done...                           2018-06-01 13:40:54
```
Annotate peaks with genomic context

- `plotAnnoPie(peaks_anno)`
- `vennpie(peaks_anno)`
- `upsetplot(peaks_anno, text.scale = c(2, 2, 2, 2, 2, 2))`
- `plotDistToTSS(peaks_anno, title="Distribution of transcription factor-binding loci relative to TSS")`
- `plotAnnoBar(peaks_anno)`

`clusterProfiler + gglot2 + org.Hs.eg.db`
Reference:


公众号：生信媛

互课：第二期

徐春晖

https://www.jianshu.com/u/fe854ffa1f9e
谢谢

THANK YOU