

chip-seq信息分析



chip-seq

染色质免疫共沉淀技术

（Chromatin Immunoprecipitation, **ChIP**）也称结合位点分析法，是研究体内蛋白质与**DNA**相互作用的有力工具，通常用于转录因子结合位点或组蛋白特异性修饰位点的研究。

Chip-seq将**ChIP**与第二代测序技术相结合的**ChIP-Seq**技术，能够高效地在全基因组范围内检测与组蛋白、转录因子等互作的**DNA**区段。



主要内容

- 🔊 数据库准备和下载原始数据
- 🔊 质控
- 🔊 比对
- 🔊 peak finding
- 🔊 peak可视化
- 🔊 peak注释
- 🔊 motif discovery



数据库准备(hg19)

UCSC(fasta、 bed)

<http://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz>

```
$ bowtie2-build hg19.fa hg19
```



下载fastq数据

<http://www.ebi.ac.uk/>

"chip-seq FoxA1" --> Nucleotide sequences (74)

<http://www.ncbi.nlm.nih.gov/>

SRA --> "chip-seq FoxA1"



质控

```
$ fastqc ERR499.read1.fq -t 2 -o qcOutdir
```

●质量值



过滤

1. 过滤接头。对含接头的reads去除接头序列。
2. 一条reads上N（未能确定出具体的碱基类型）的比例大于5%，则过滤掉该reads。
3. 过滤低质量reads，过滤掉 $Q20 < 70\%$ reads。

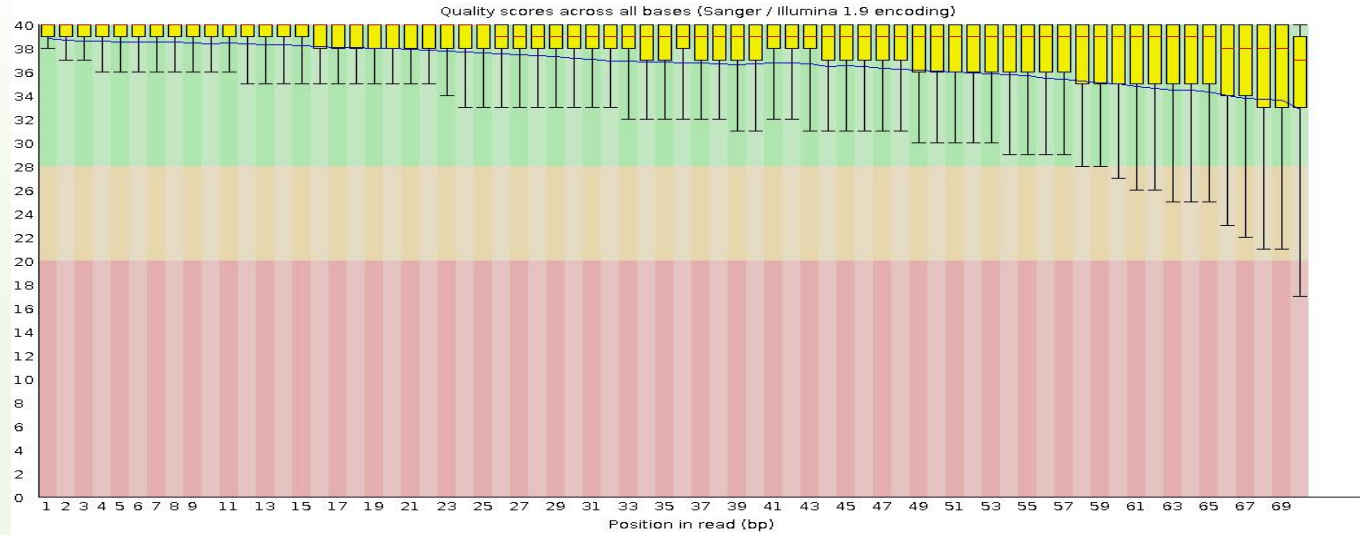


过滤统计

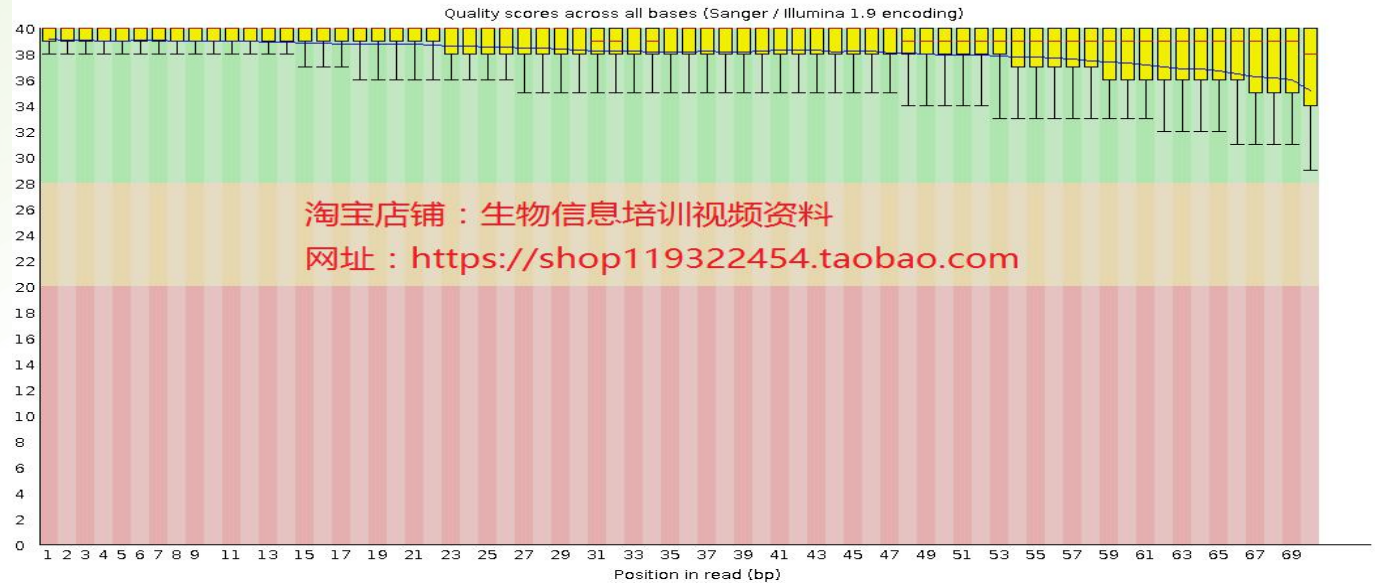
read	raw	adapter	N	Low qual	clean
read1	28701483 (100%)	53761(0.19%)	29(0.00%)	1849520 (6.44%)	25779623 (89.82%)

过滤前后质量值

过滤前：



过滤后：



比对

1. 比对

```
$ bowtie2 -p 15 -x /home/zhiming/software/rna/rnaSeqSoftware/hg19/hg19 ERR990.clean1.fq -S ERR990.sam
```

statistics	Input Reads	mapped	Multiple	Unique
Percentage	48948102(100%)	48212358(98.50%)	561991(1.15%)	47650367(97.35%)

Peak finding

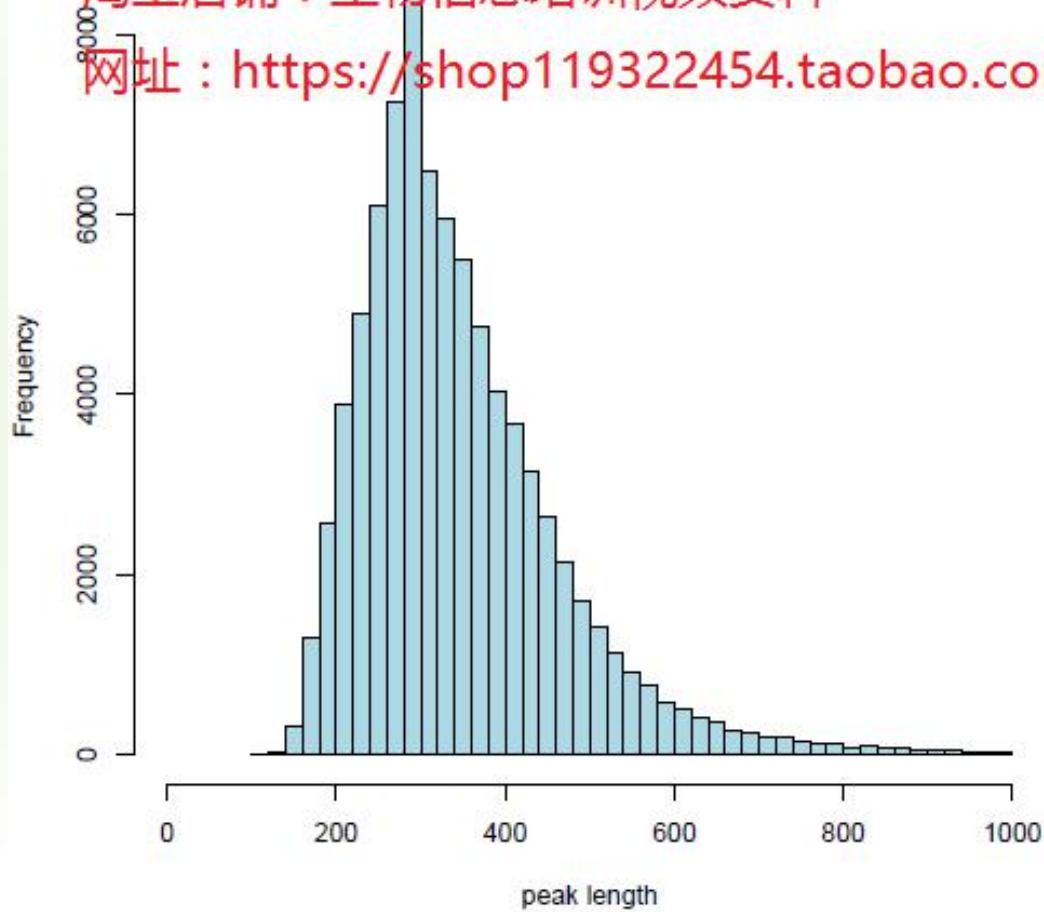
chr	start	end	length	summit	tags	$-10 \cdot \log_{10}(\text{pvalue})$	fold_enrichment
chr1	565230	565418	189	94	21	96.33	9.78
chr1	569311	569508	198	99	51	304.78	17.41
chr1	662532	662763	232	96	25	210.79	29.34
chr1	715028	715371	344	153	13	91.66	23.29
chr1	811850	812221	372	273	12	77.92	23.29

Peak finding

Histogram of peak length

淘宝店铺：生物信息培训视频资料

网址：<https://shop119322454.taobao.com>



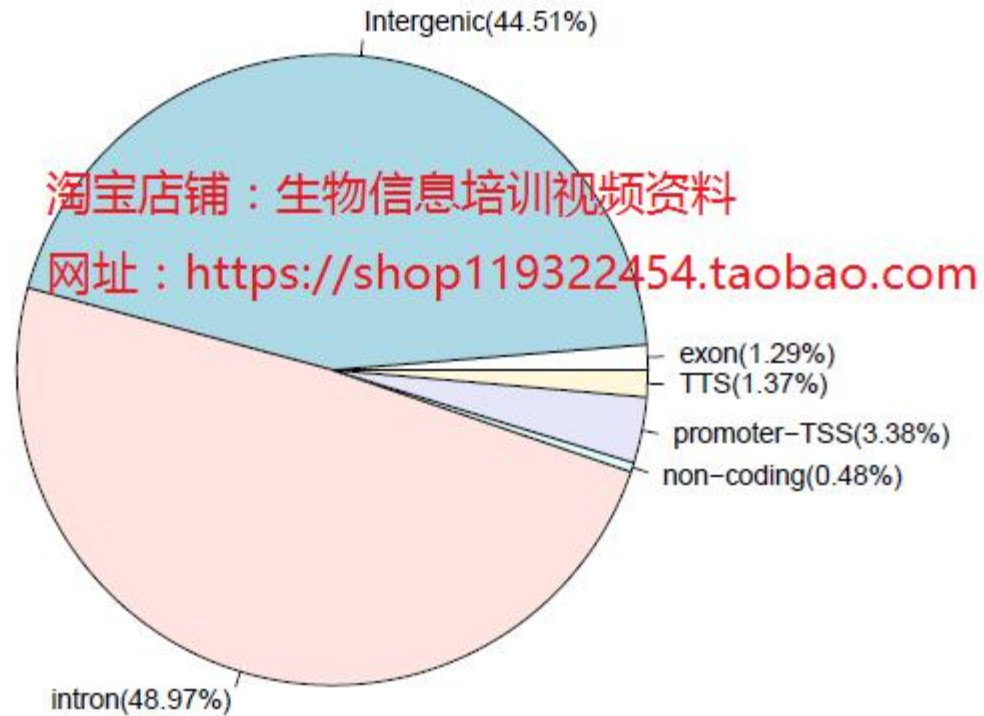
Peak view










Peak注释

PeakID	Chr	Start	End	Annotation	Distance to TSS	Nearest PromoterID	Entrez ID	Gene Name
MACS_peak_5762	chr1	210893963	210894257	intron (NM_172362, intron 10 of 10)	391470	NM_001170580	55733	HHAT
MACS_peak_48704	chr20	12998002	12998354	intron (NM_018327, intron 1 of 11)	8551	NM_018327	55304	SPTLC3
MACS_peak_18650	chr13	91925944	91926343	Intergenic	-62191	NR_047004	100874150	LINC00379
MACS_peak_38942	chr2	24160550	24160821	Intergenic	-2691	NM_181713	165324	UBXN2A

Peak注释

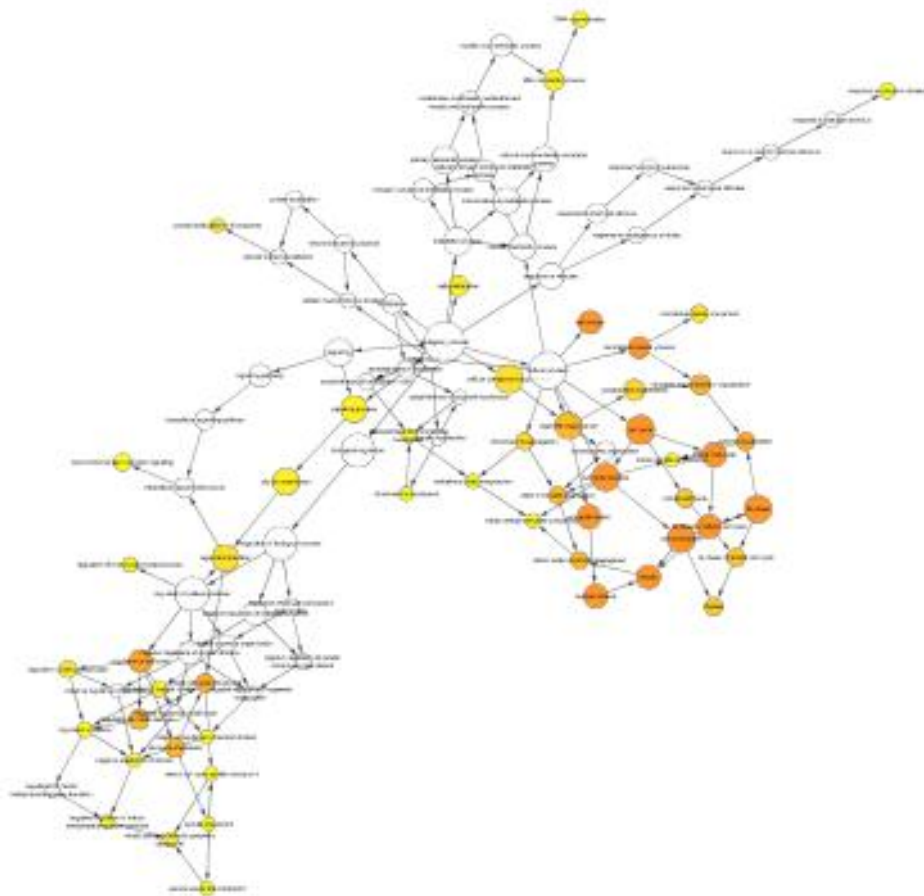


Motif

Rank	Motif	P-value	log P-pvalue	% of Targets	% of Background
1		1e-42572	-9.803e+04	65.00%	6.20%
2		1e-705	-1.625e+03	8.62%	4.14%
3		1e-537	-1.238e+03	23.21%	16.50%
4		1e-453	-1.044e+03	21.38%	15.40%
5		1e-444	-1.024e+03	19.78%	14.05%
6		1e-389	-8.977e+02	1.14%	0.19%
7		1e-349	-8.044e+02	28.51%	22.54%

GO分析

该有向无环图(DAG)为差异基因GO富集分析的结果图形化展示方式，分支代表包含关系，箭头方向从上之下所定义的功能范围越来越小，并通过包含关系，将相关的GO Term一起展示，颜色深浅代表富集程度，越深富集水平越高，反之，则越低。



KEGG分析

