第五课 表观基因组计算方法与数据分析

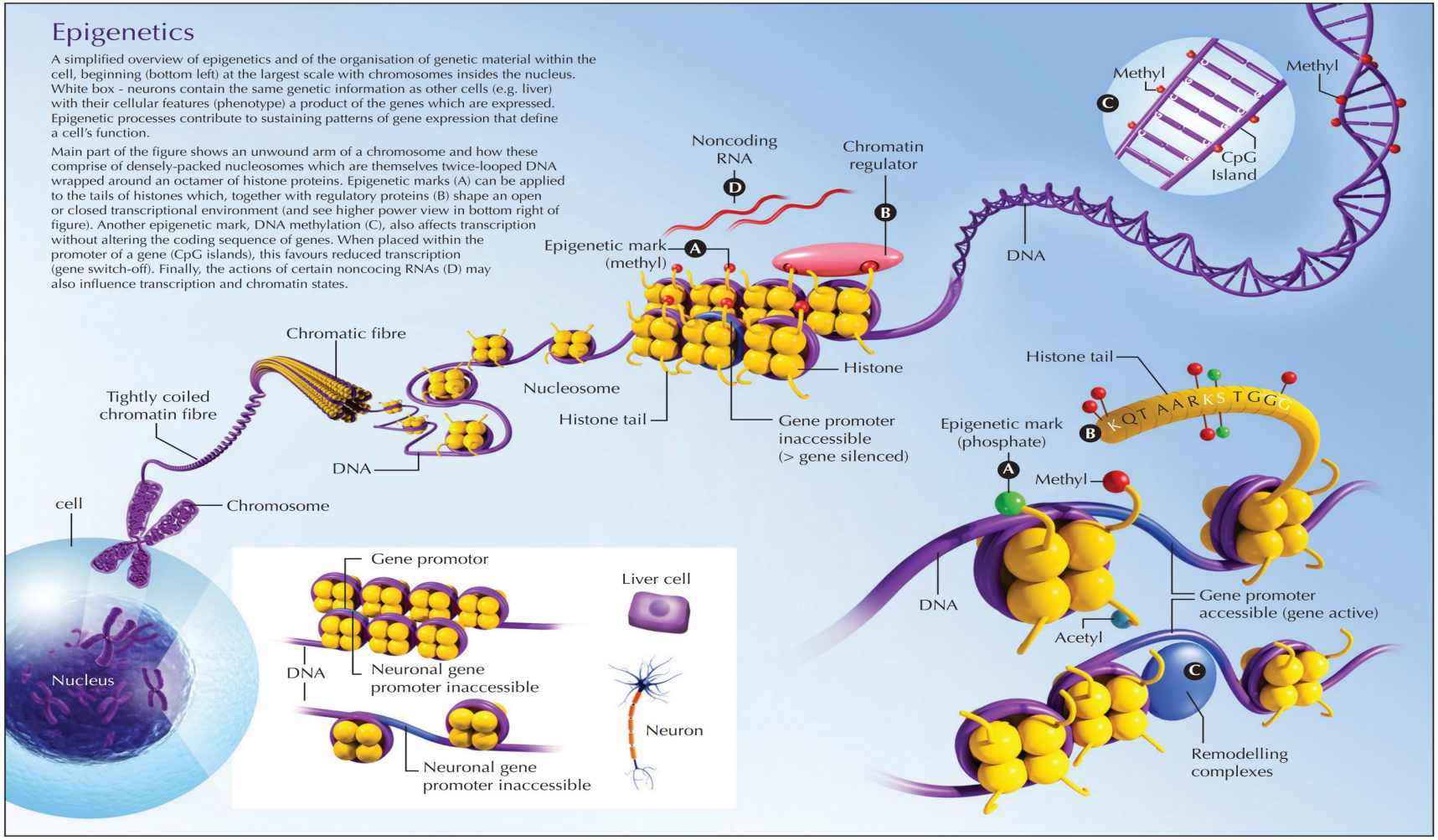
中国医学科学院基础医学研究所

陈阳

yc@ibms.pumc.edu.cn

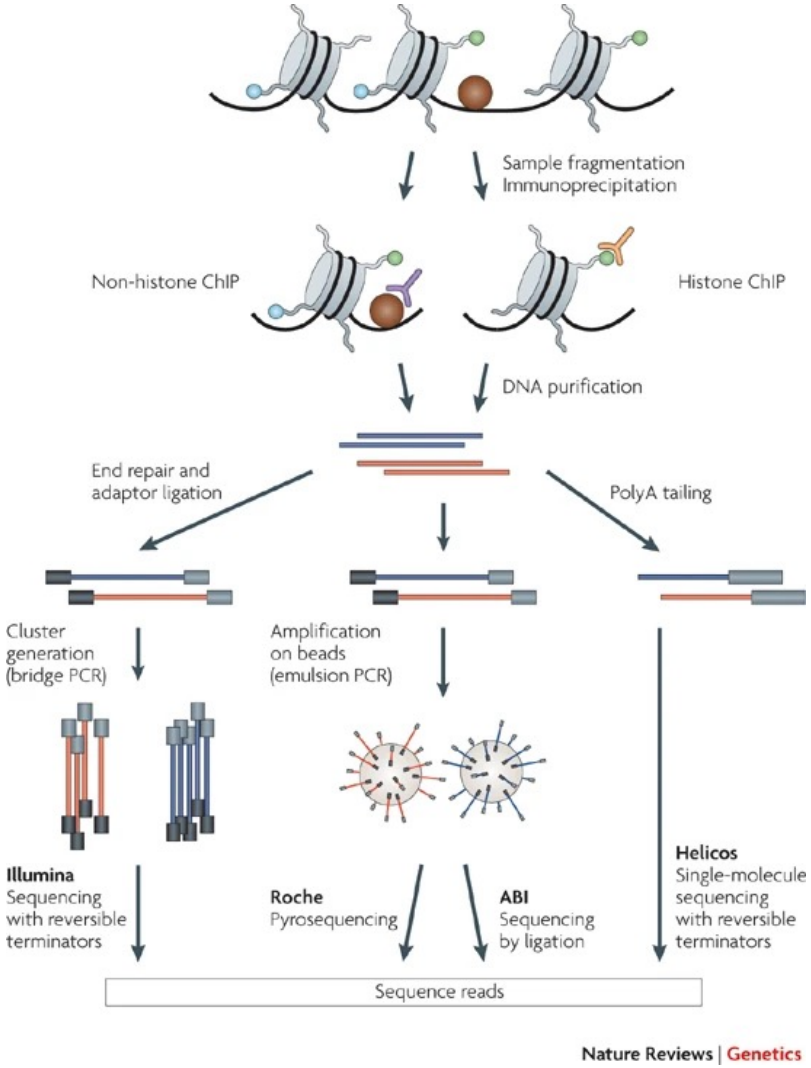
20250408

表观遗传调控



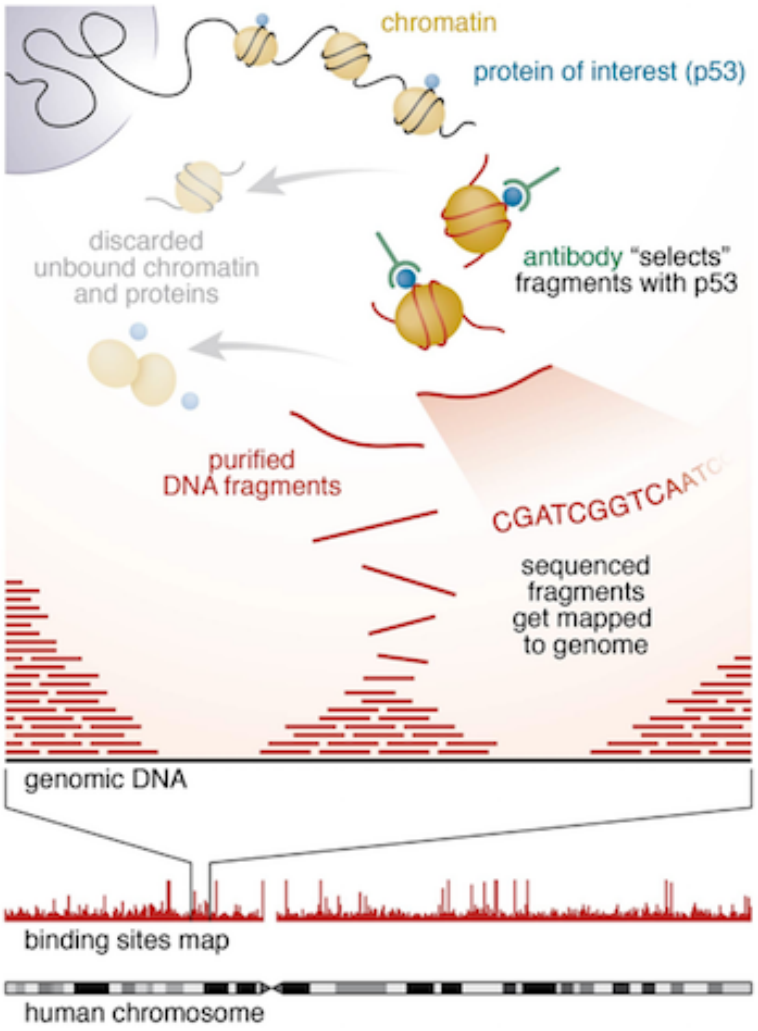
https://onlinelibrary.wiley.com/doi/full/10.1684/epd.2020.1143

ChIP-seq简介



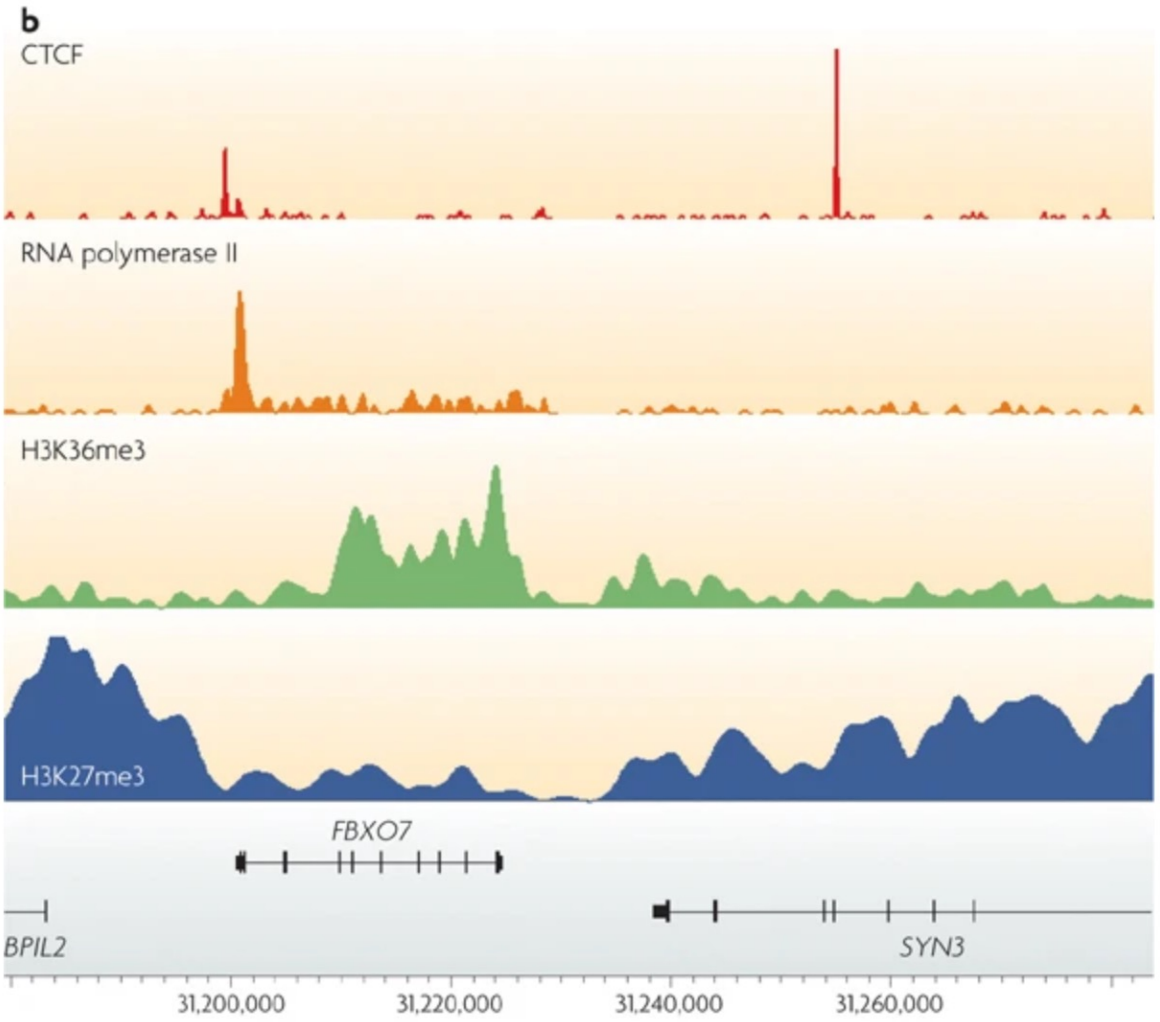
https://www.nature.com/articles/nrg2641

ChIP–seq: advantages and challenges of a maturing technology



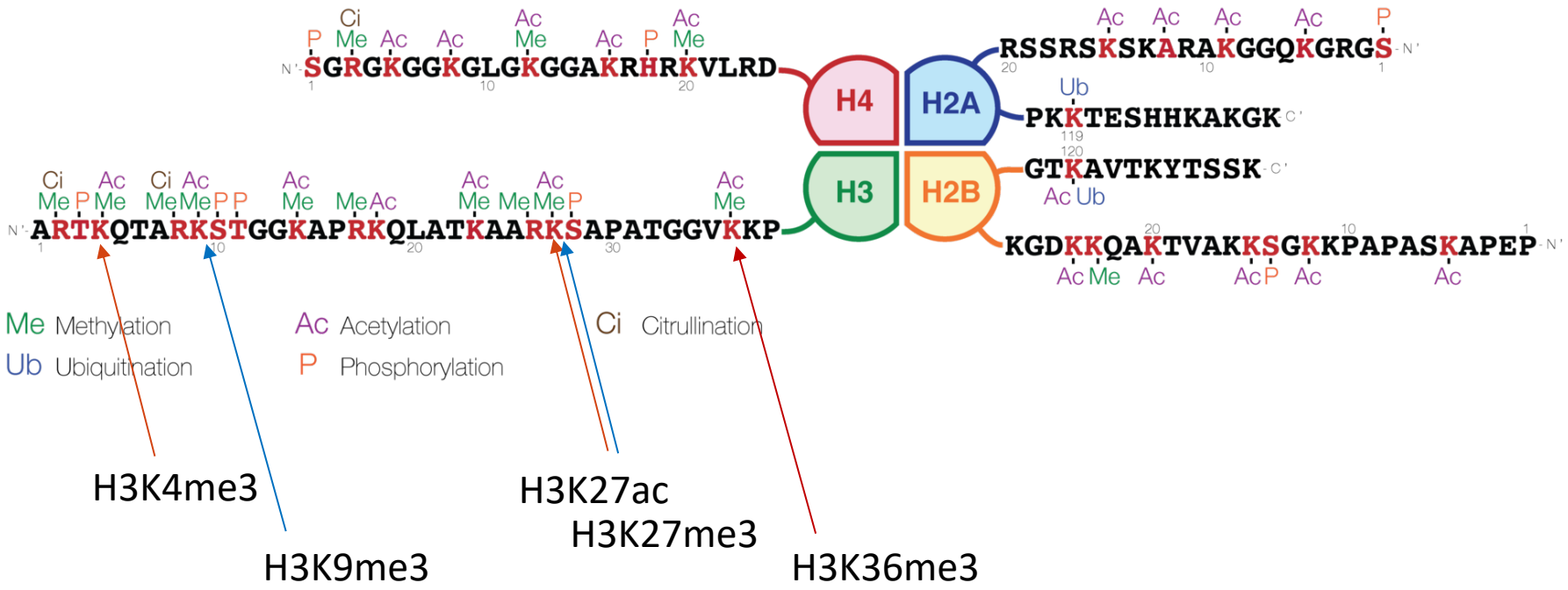
https://www.nature.com/articles/nrg2641

ChIP profiles



https://www.nature.com/articles/nrg2641

Histone modification

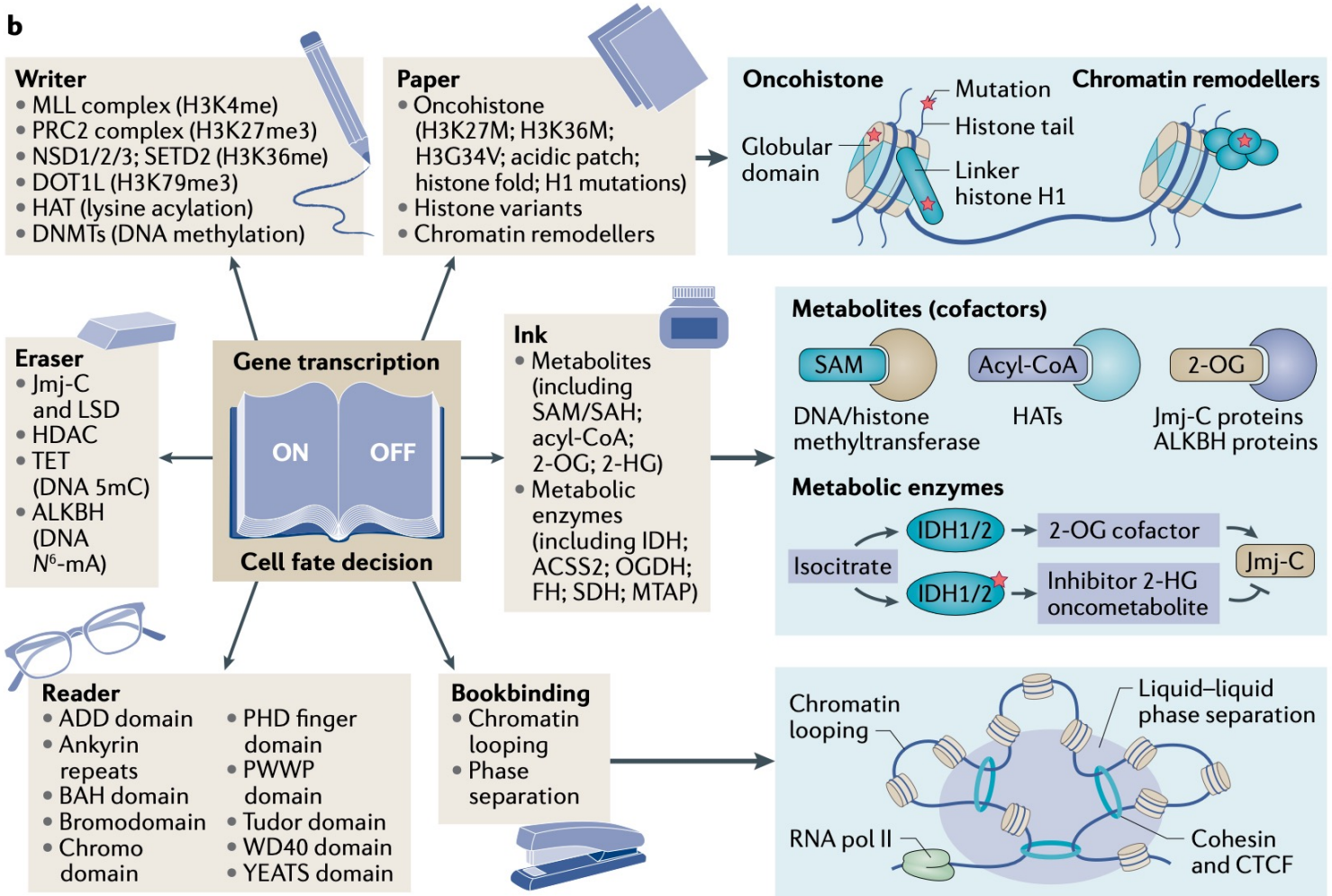


Me 甲基化
Ac 乙酰化
Ci 瓜氨酸化
Ub 泛素化
P 磷酸化

S Serine/丝氨酸
R Arginine/精氨酸
K Lysine/赖氨酸
H Histidine/组氨酸
T Threonine/苏氨酸

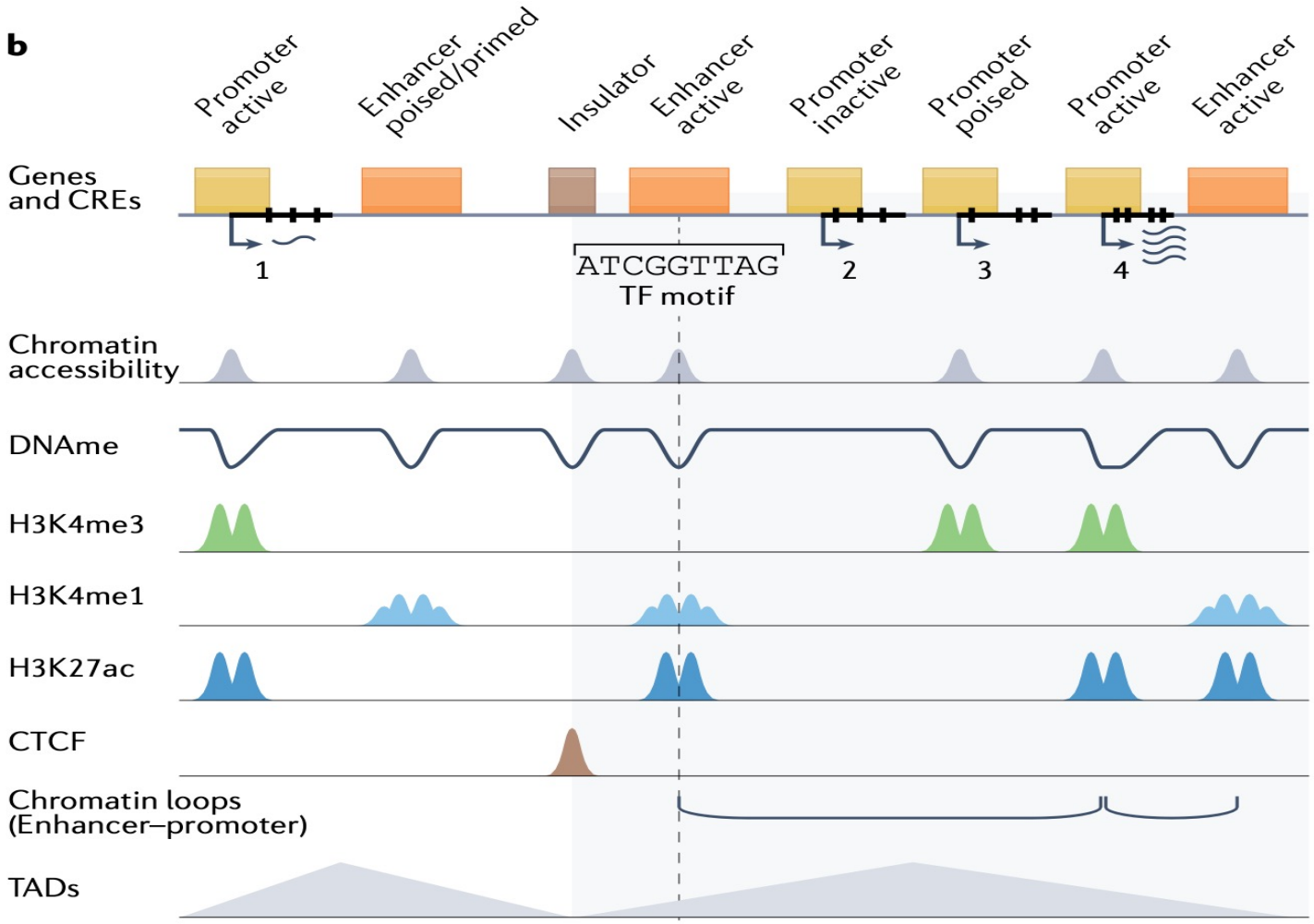
https://en.wikipedia.org/wiki/

‘language’ of chromatin modification

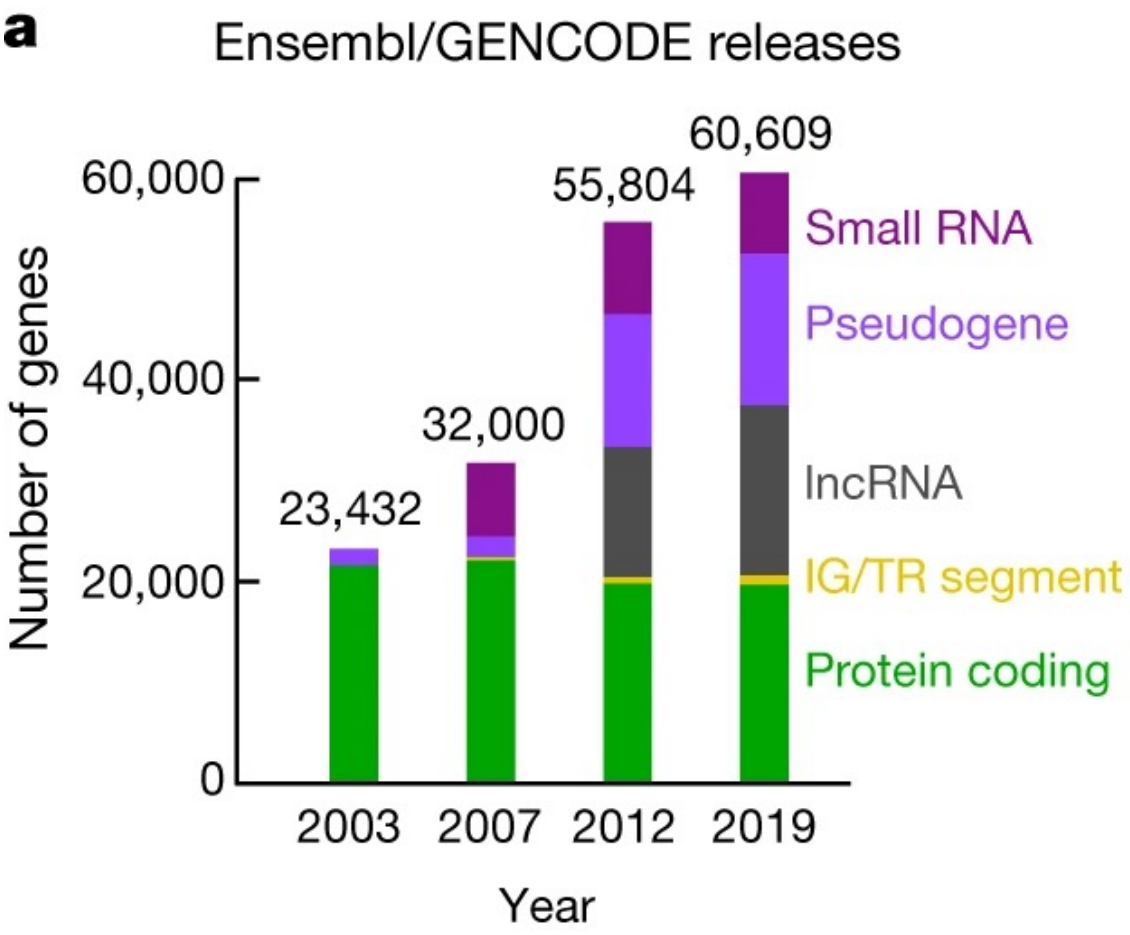


https://doi.org/10.1038/ s41568-021-00357-x

epigenomic marks at cRes and their association with gene expression

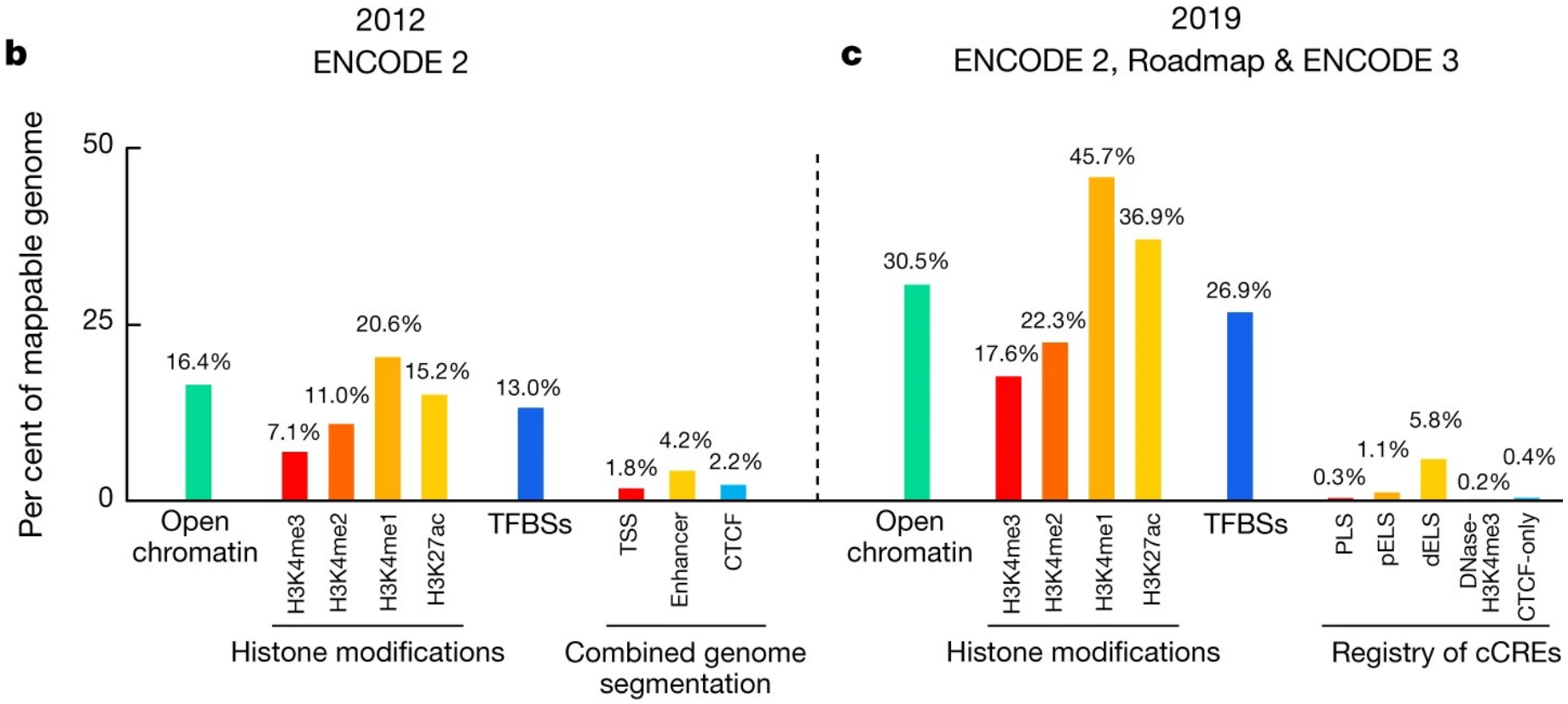


Improvement of gene annotations in the past 15 years



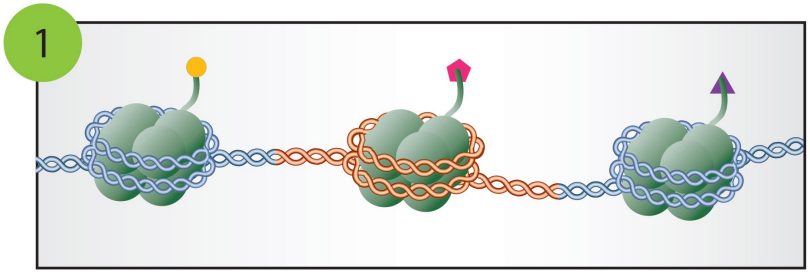
https://www.nature.com/articles/s41586-020-2449-8

ENCODE annotations in 2019 with ENCODE 2, Roadmap, and ENCODE 3 data

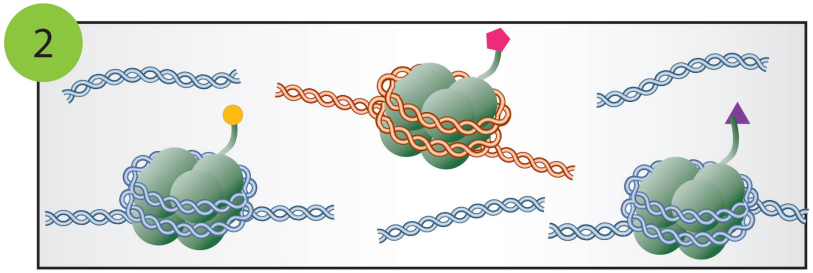


https://www.nature.com/articles/s41586-020-2449-8

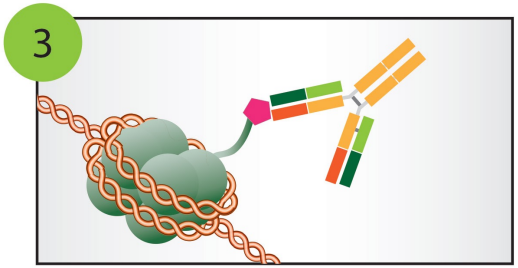
ChIP–seq: advantages and challenges of a maturing technology



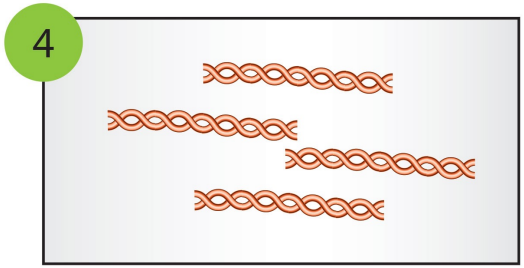
1. DNA-protein cross-linking



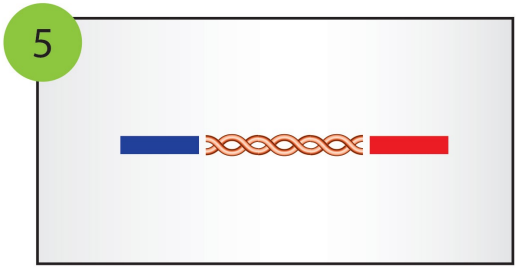
2. Chromatin extraction and DNA fragmentation



3. Immunoprecipitation of target protein-bound fragments



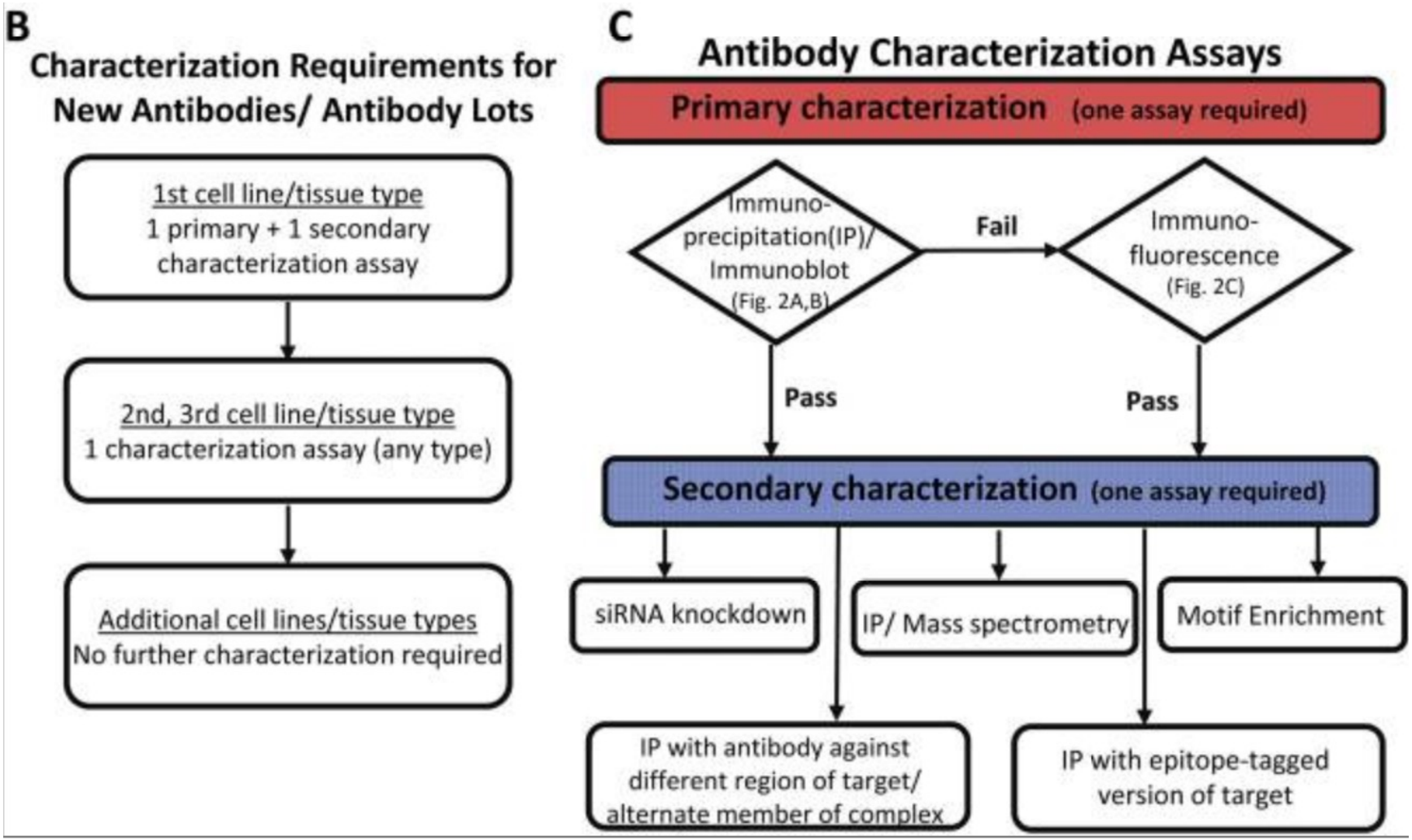
4. Cross-link reversal, protein digestion, and DNA isolation



5. Adaptor ligation for DNA library construction

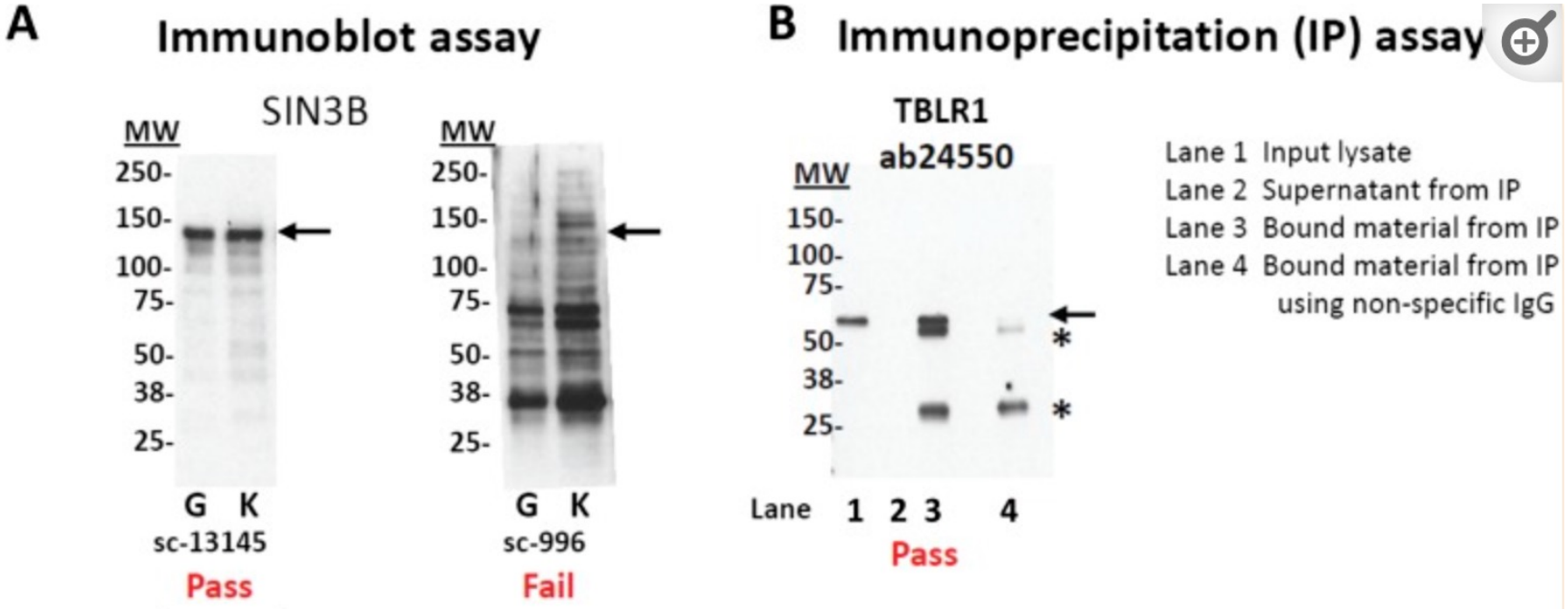
https://www.nature.com/articles/nrg2641

ChIP-seq技术基础：Antibody for ChIP-seq



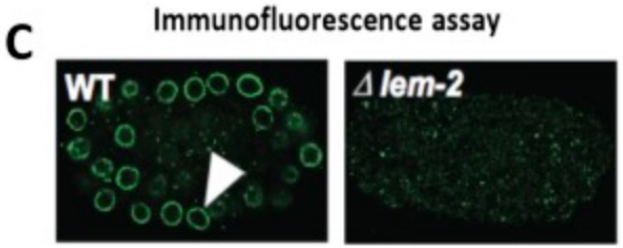
https://genome.cshlp.org/content/22/9/1813.abstract

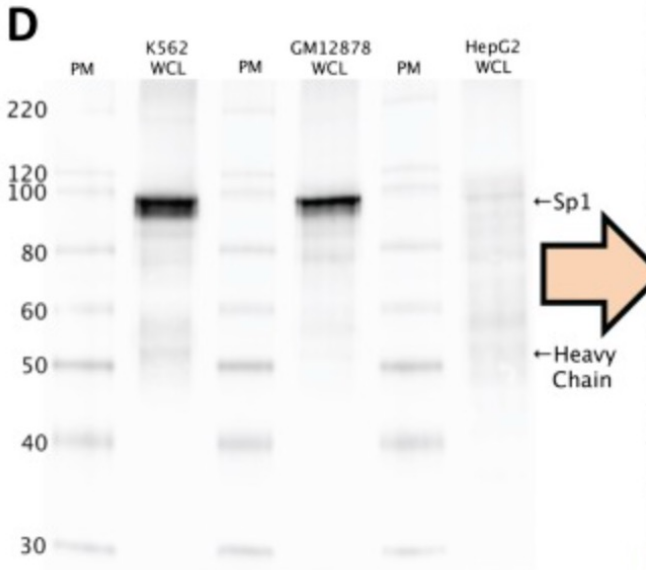
Representative results from antibody characterization assays



https://genome.cshlp.org/content/22/9/1813.abstract

Representative results from antibody characterization assays

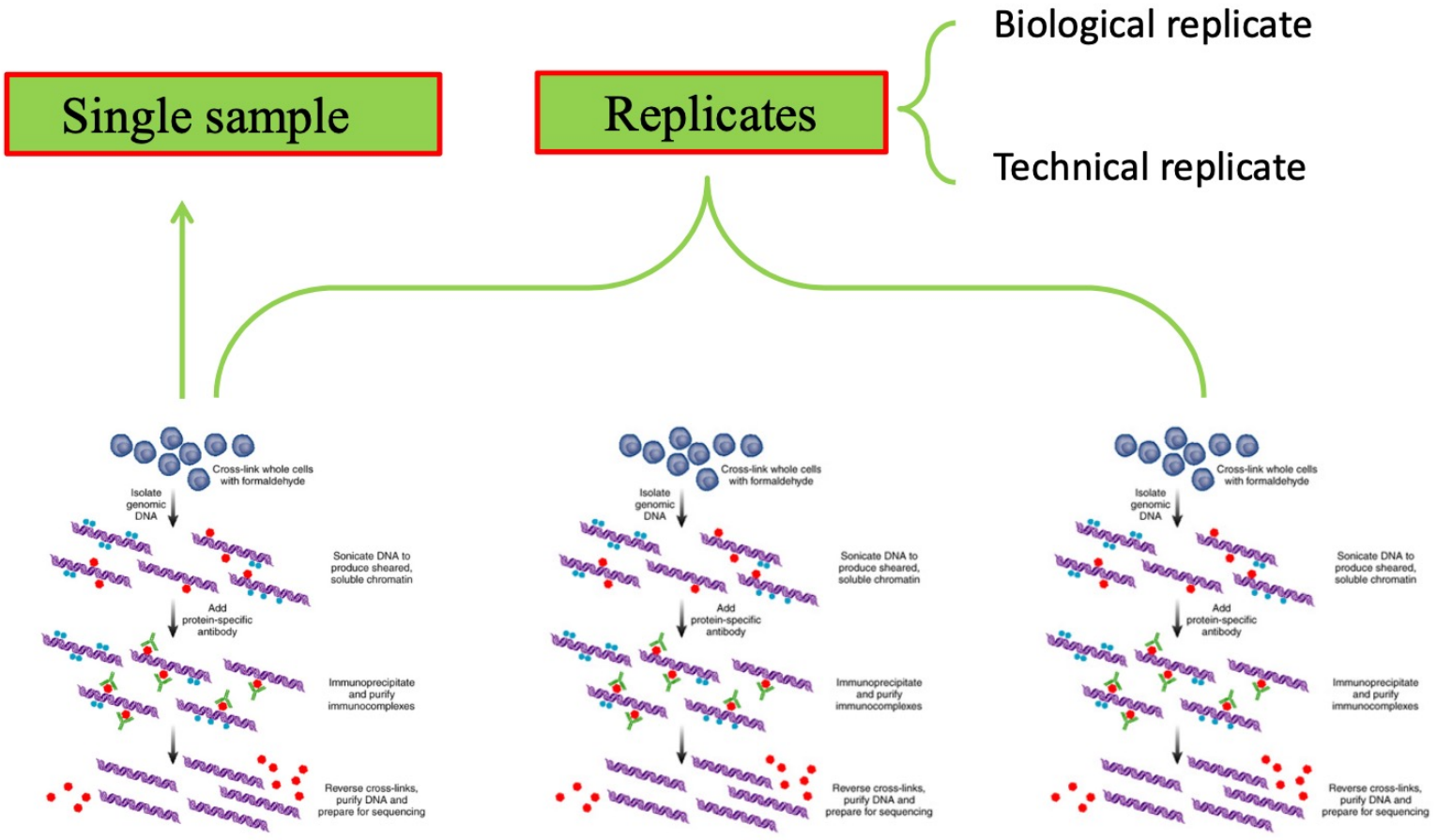




Description	Unique Peptides	Protein Identification Probability (%)
Heat shock protein HSP 90-beta	39	100
Nucleolar RNA helicase 2	24	100
Elongation factor 2	18	100
Heat shock protein HSP 90-alpha	18	100
TBC1 domain family member 5	16	100
Endoplasmin	14	100
DNA replication licensing factor MCM7	13	100
DNA replication licensing factor MCM5	10	100
Transitional endoplasmic reticulum ATPase	10	100
26S proteasome non-ATPase regulatory subunit 2	10	100
X-ray repair crs-complementing protein 5	9	100
Methionyl-tRNA synthetase, cytoplasmic	9	100
Transferrin receptor protein 1	9	100
Programmed cell death 6-interacting protein	9	100
Neutral alpha-glucidase AB	9	100
Transportin-1	7	100
Probable ATP-dependent RNA helicase DDX27	7	100
Transportin-3	6	100
N-alpha-acetyltransferase 15, NatA auxiliary subunit	6	100
Coatomer subunit gamma	5	100
6-phphofructokinase type C	5	100
Transcription factor Sp1	**5**	**100**
Keratin, type II cytkeletal 1	4	100
ATP-dependent RNA helicase DDX1	4	100
Glutaminyl-tRNA synthetase	4	100

https://genome.cshlp.org/content/22/9/1813.abstract

ChIP-Seq experimental design



https://biohpc.cornell.edu/lab/doc/CHIPseq_workshop_20150504_lecture1.pdf (2)

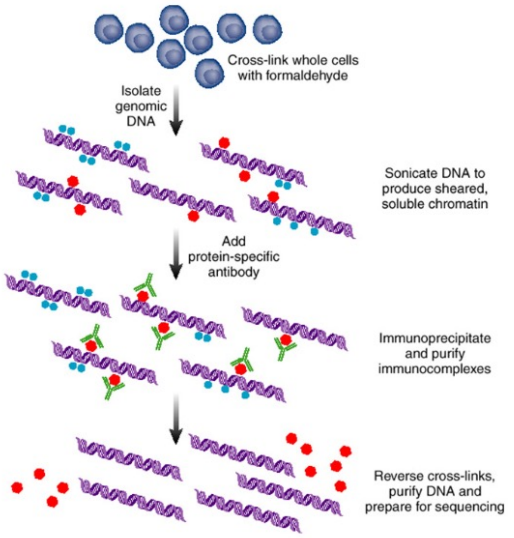
ChIP-Seq experimental design

Controls for ChIP-seq

Most experimental protocols involve a control sample that is processed the same way as the test sample except that no immunoprecipitaion step or no specific antibody

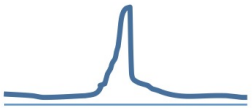
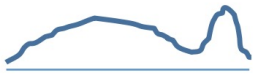
Input DNA & IgG

- Input DNA does not demonstrate "flat" or random (Poisson) distribution.
- Open chromatin regions tend to be fragmented more easily during shearing.
- Amplification bias.
- Mapping artifacts-increased coverage of more "mappable" regions (which also tend to be promotor regions) and repetitive regions due inaccuracies in number of copies in assembled genome.



https://biohpc.cornell.edu/lab/doc/CHIPseq_workshop_20150504_lecture1.pdf

Sequencing depth depends on data type

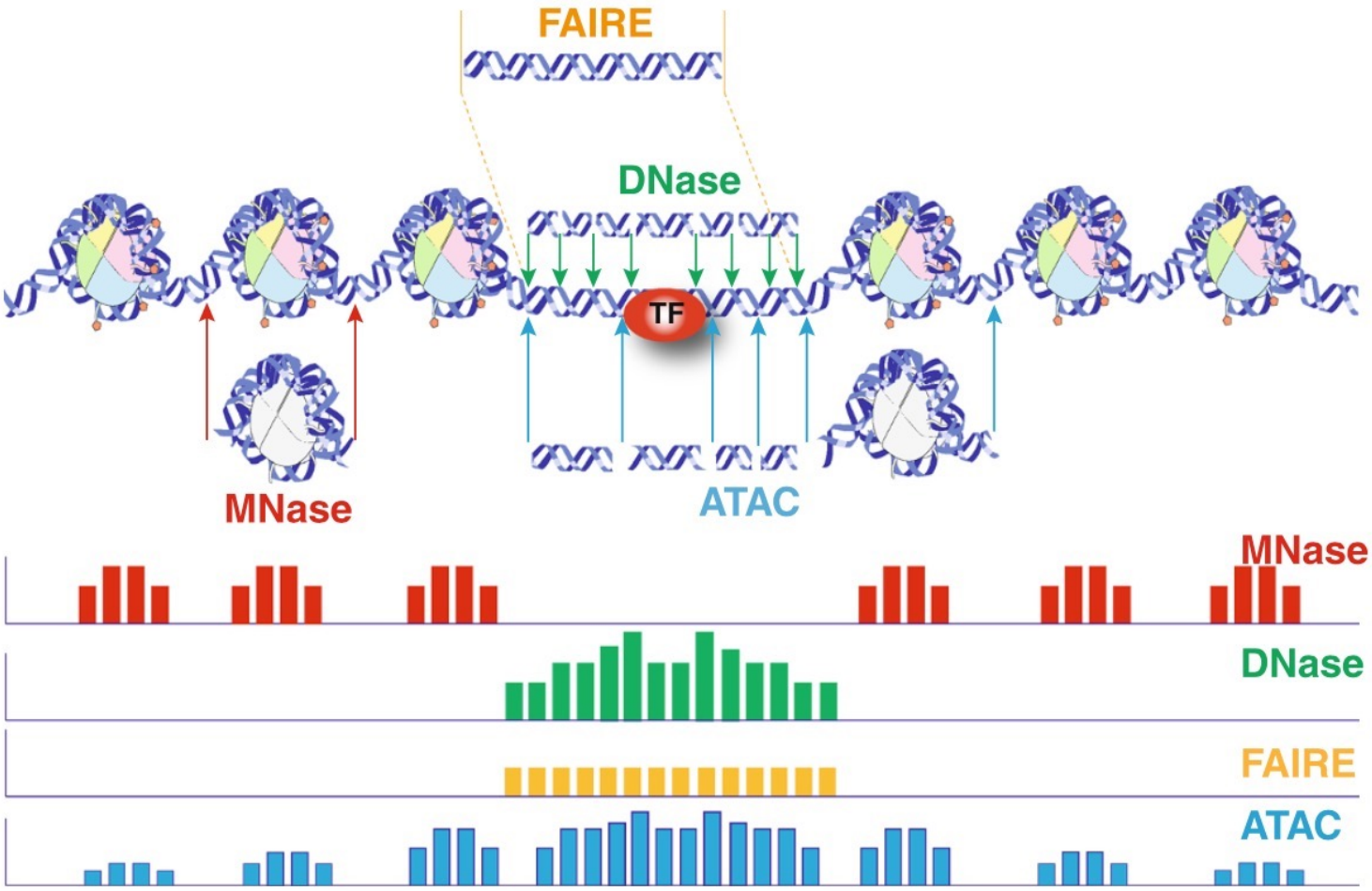
	Transcription Factors	Chromatin Remodellers Histone marks	Chromatin Remodellers Histone marks RNA polymerase II
			
	point-source	mixed signal	broad signal
Human:	TF: 20 M H3K4me3: 25 M	? H3K36me3: 35 M	? H3K27me3: 40 M H3K9me3: >55 M

No clear guidelines for mixed and broad type of peaks

Source: The ENCODE consortium; Jung et al, NAR 2014

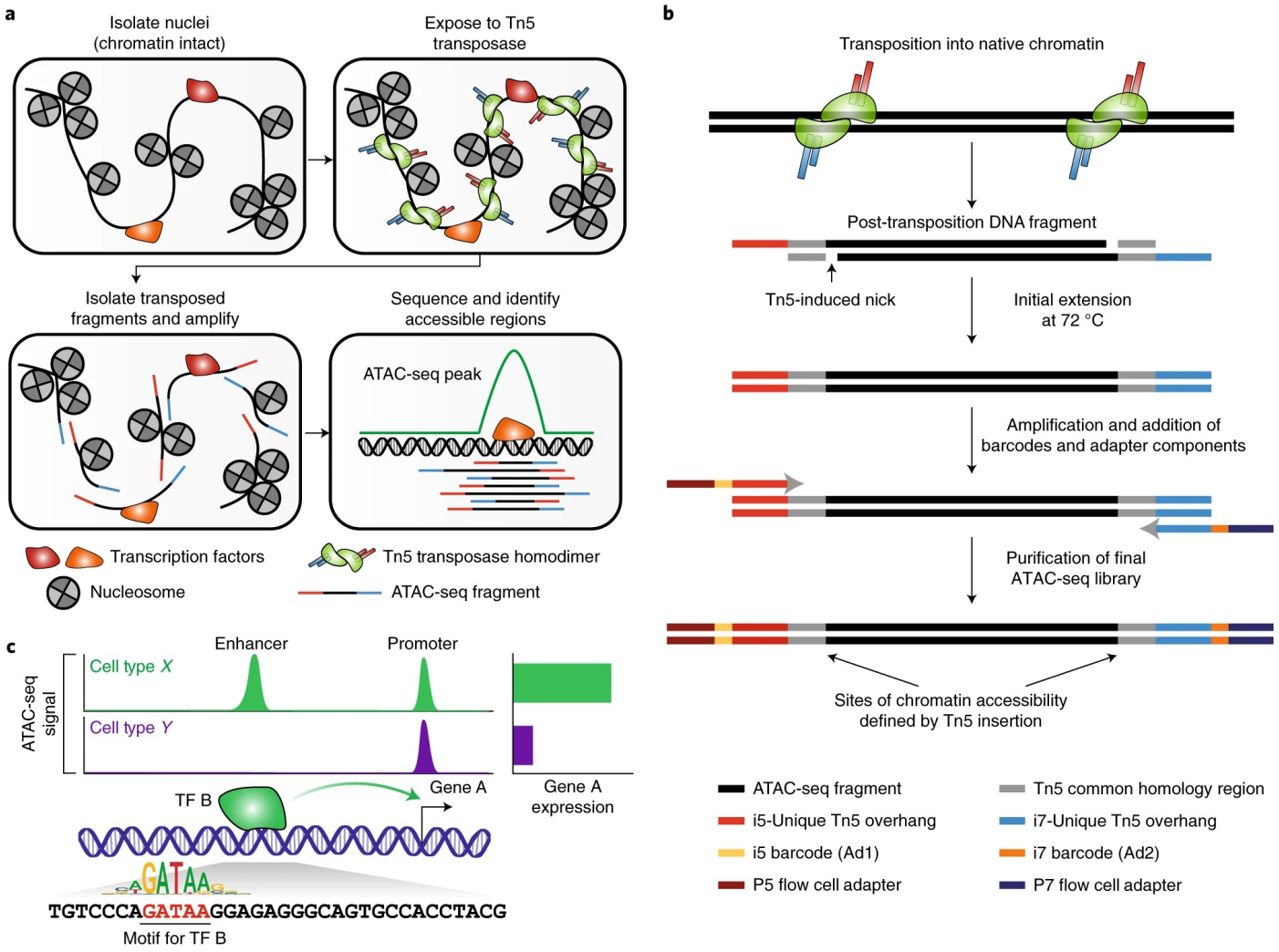
https://nbisweden.github.io/workshop-archive/workshop-ChIP-seq/2018-11-07/lectures/slides-main_as-2018.pdf (2)

表观基因组技术进阶：不同酶的功能



David C. Klein & Sarah J. Hainer, Chromosome Res, 2020

ATAC-seq



https://www.nature.com/articles/s41596-022-00692-9

Cut&Run

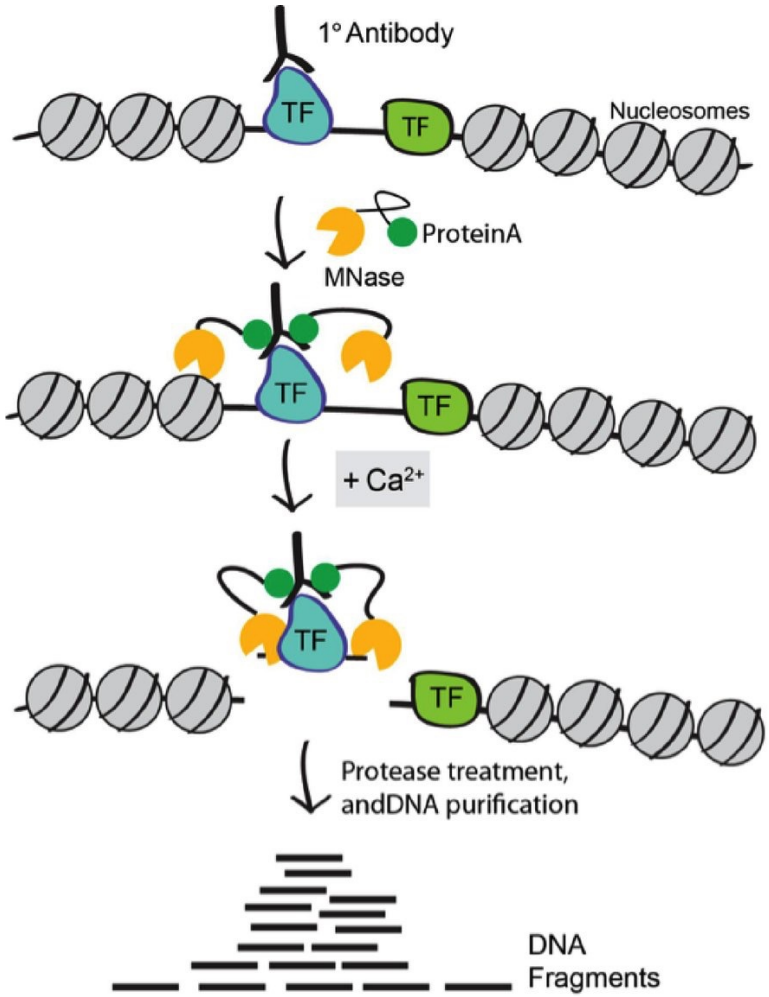
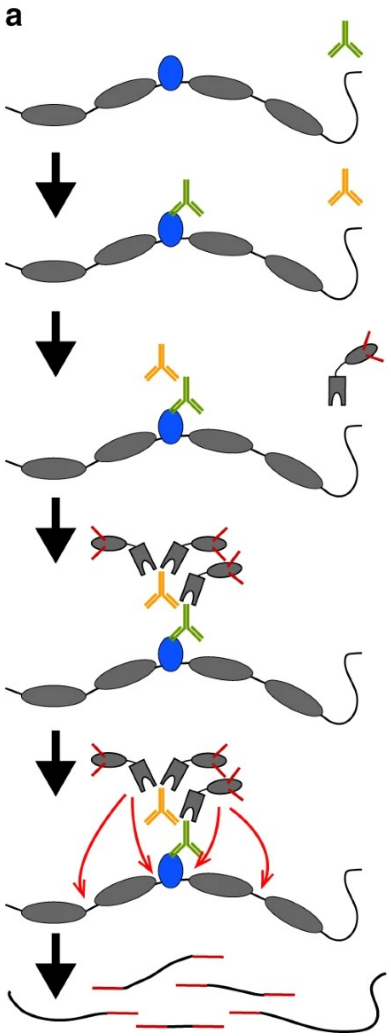


Figure 1. CUT&RUN schematic (see text for details).

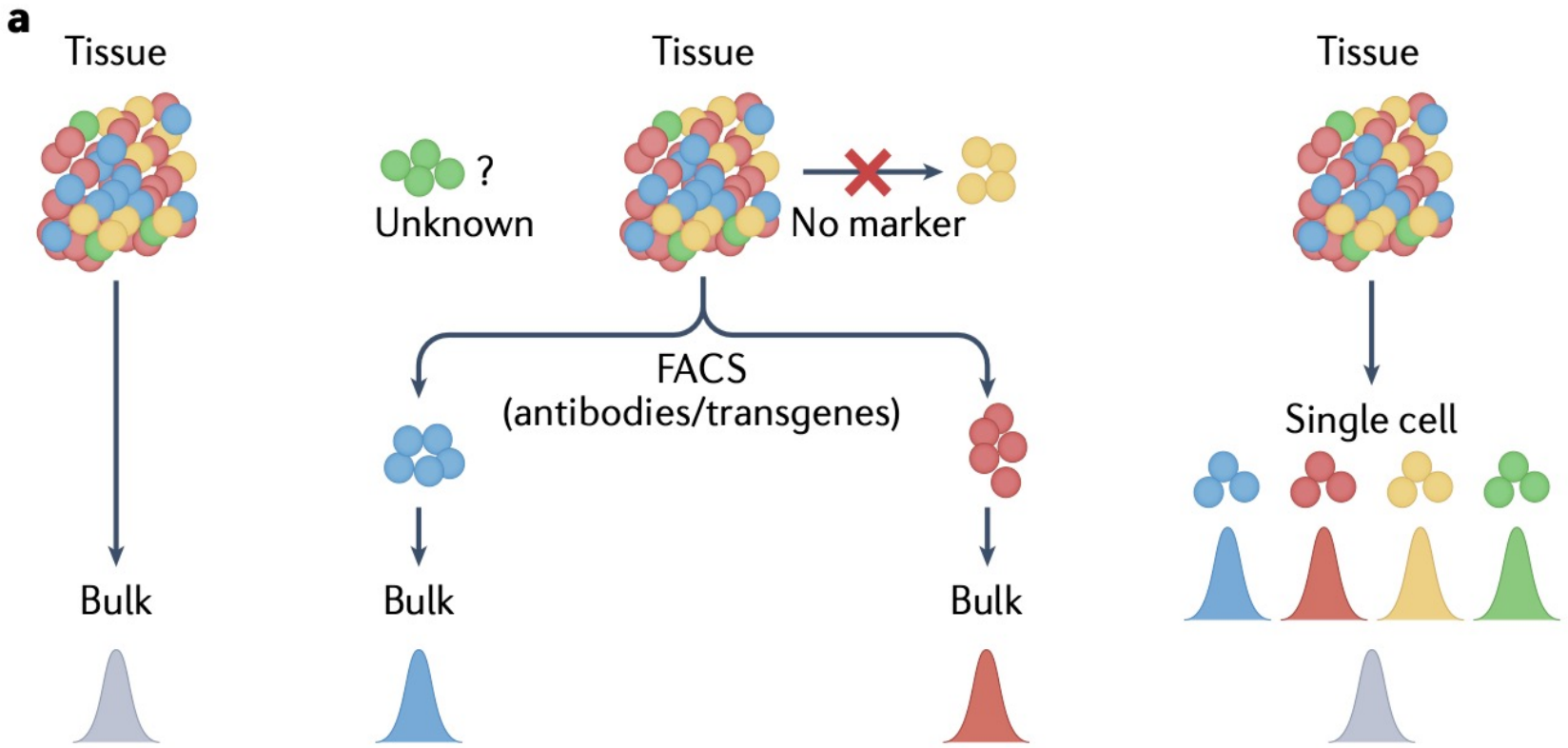
https://blog.addgene.org/cutrun-a-improved-method-for-studying-protein-dna-interactions

CUT&Tag



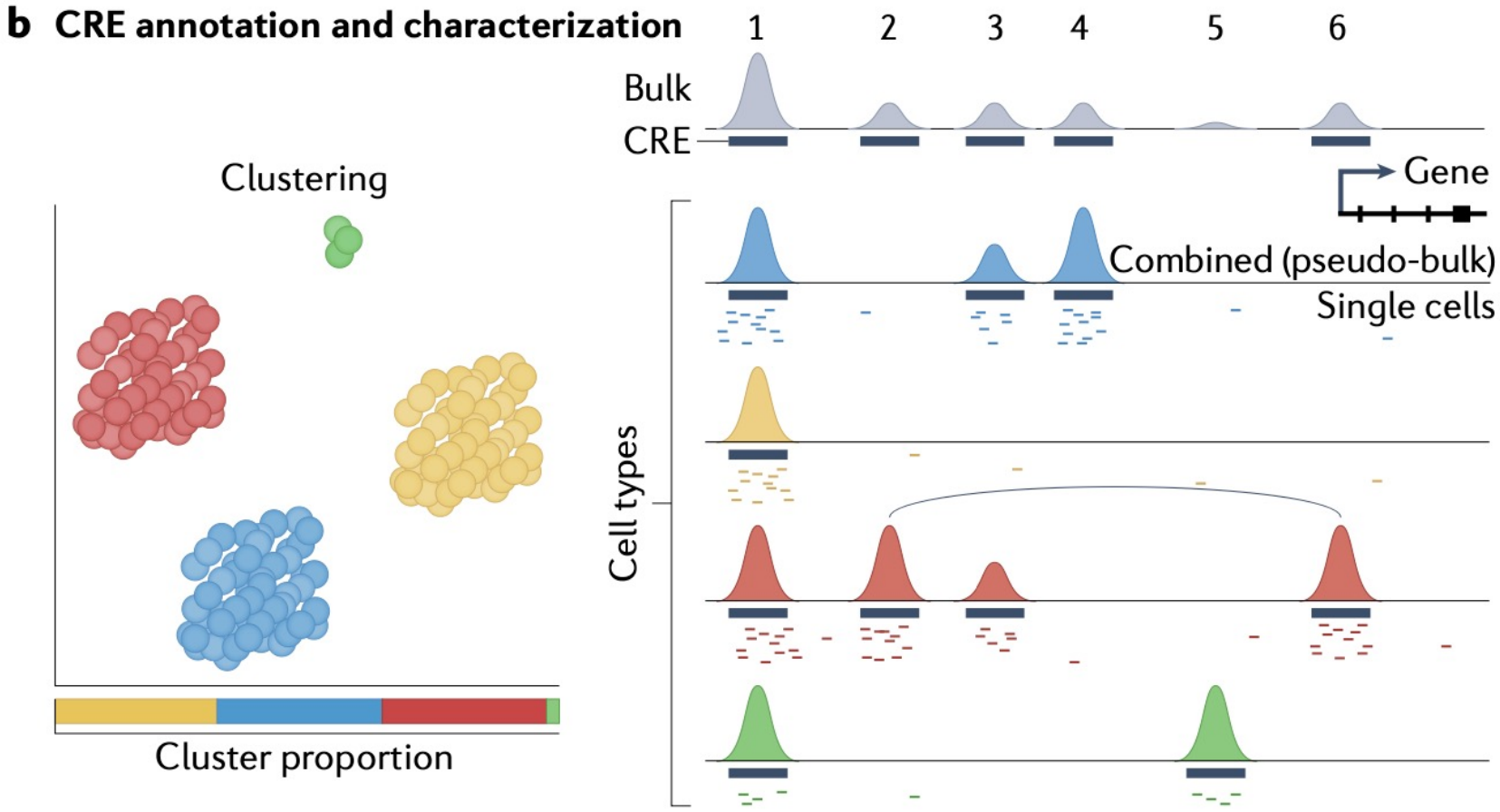
https://www.nature.com/articles/s41467-019-09982-5

Single-cell epigenomic profiling



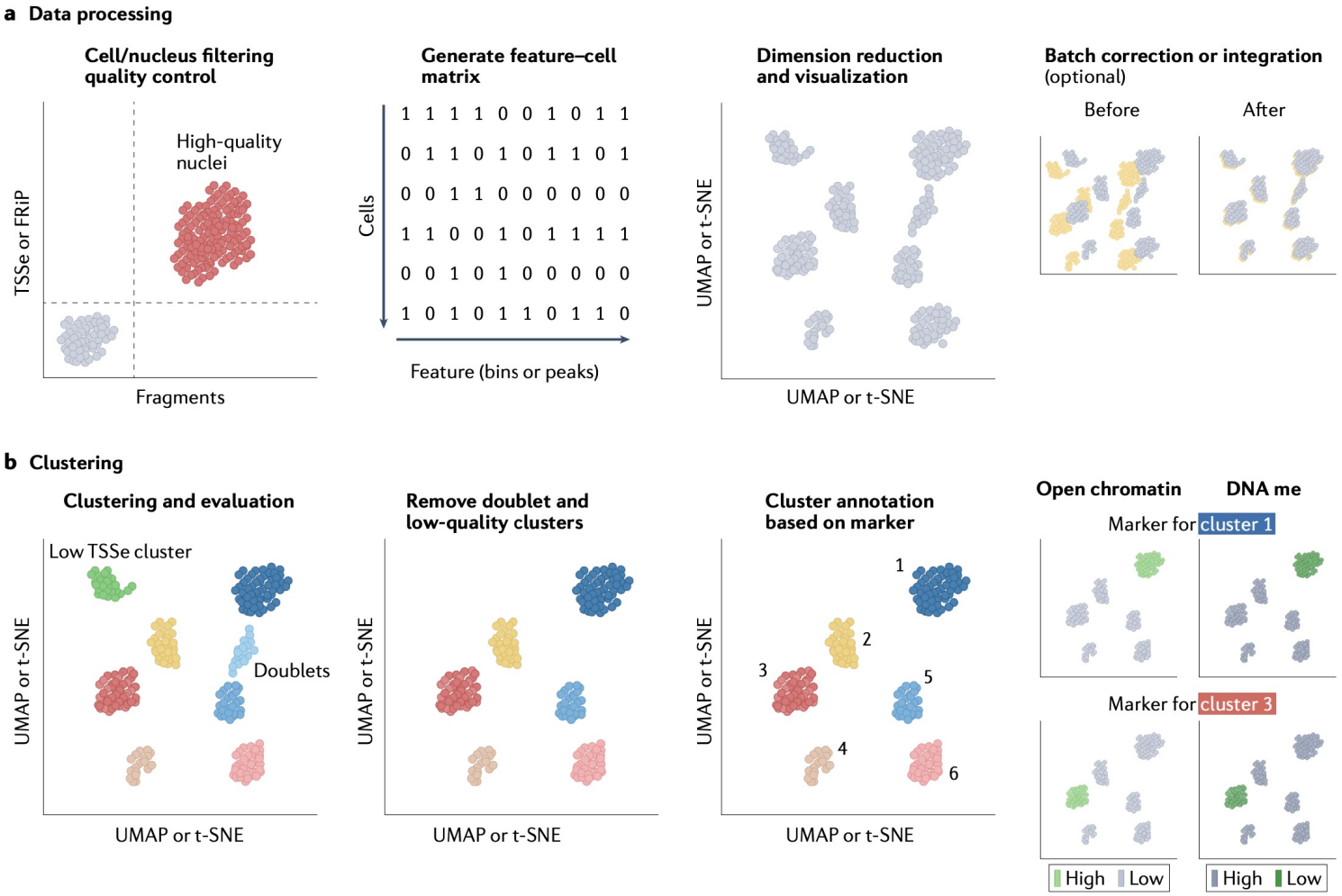
https://doi.org/10.1038/s41576-022-00509-1

Single-cell epigenomic profiling



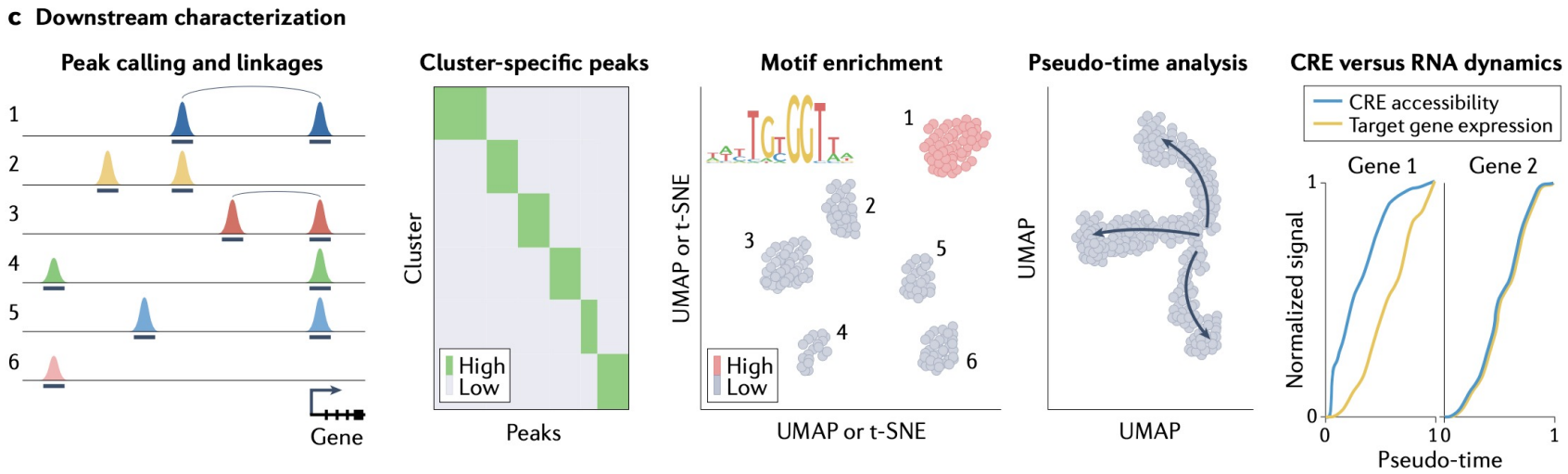
https://doi.org/10.1038/s41576-022-00509-1

General workflow for the analysis of single-cell epigenomic dataset



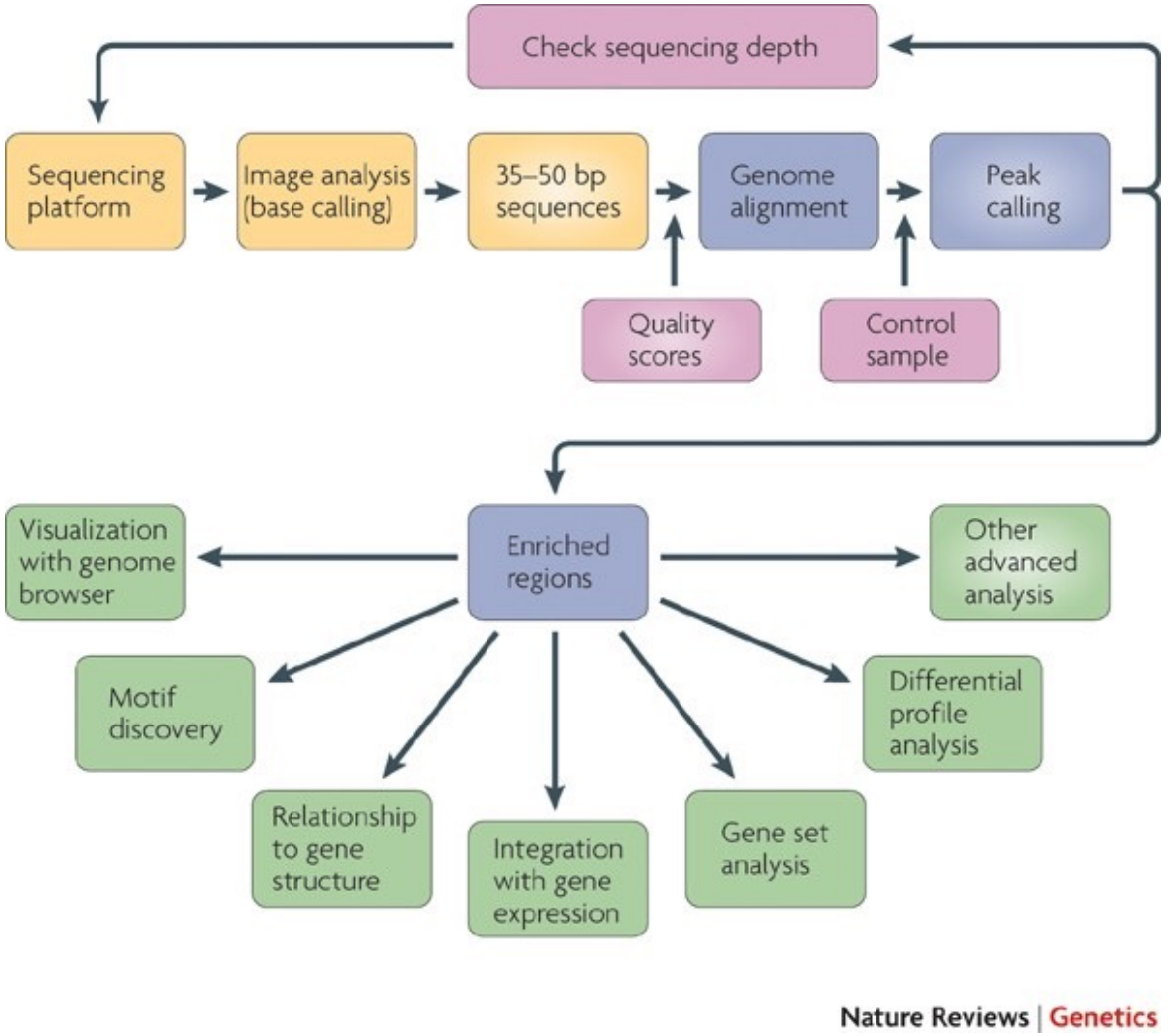
https://doi.org/10.1038/s41576-022-00509-1

General workflow for the analysis of single-cell epigenomic dataset



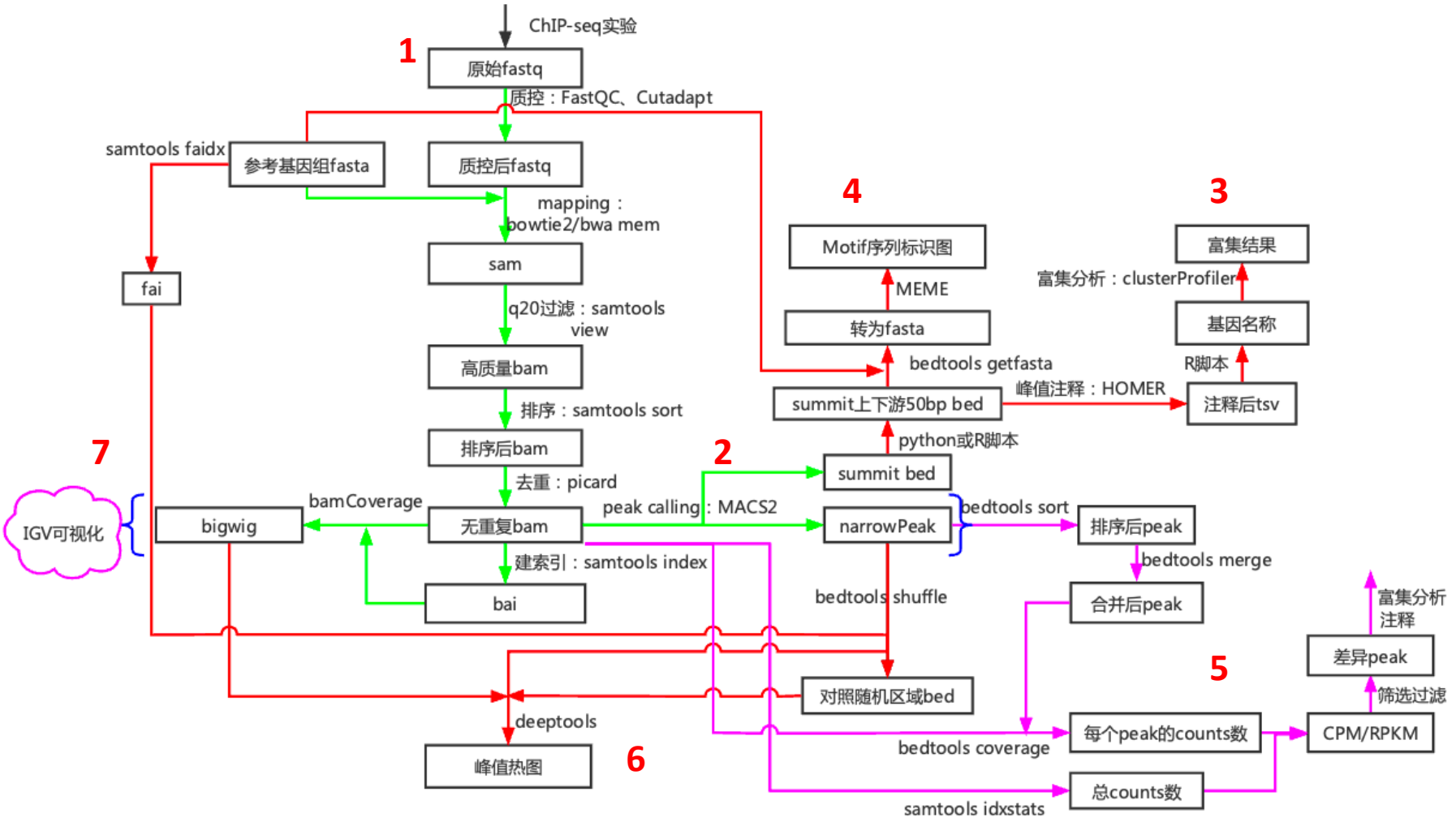
https://doi.org/10.1038/s41576-022-00509-1

Overview of ChIP–seq data analysis

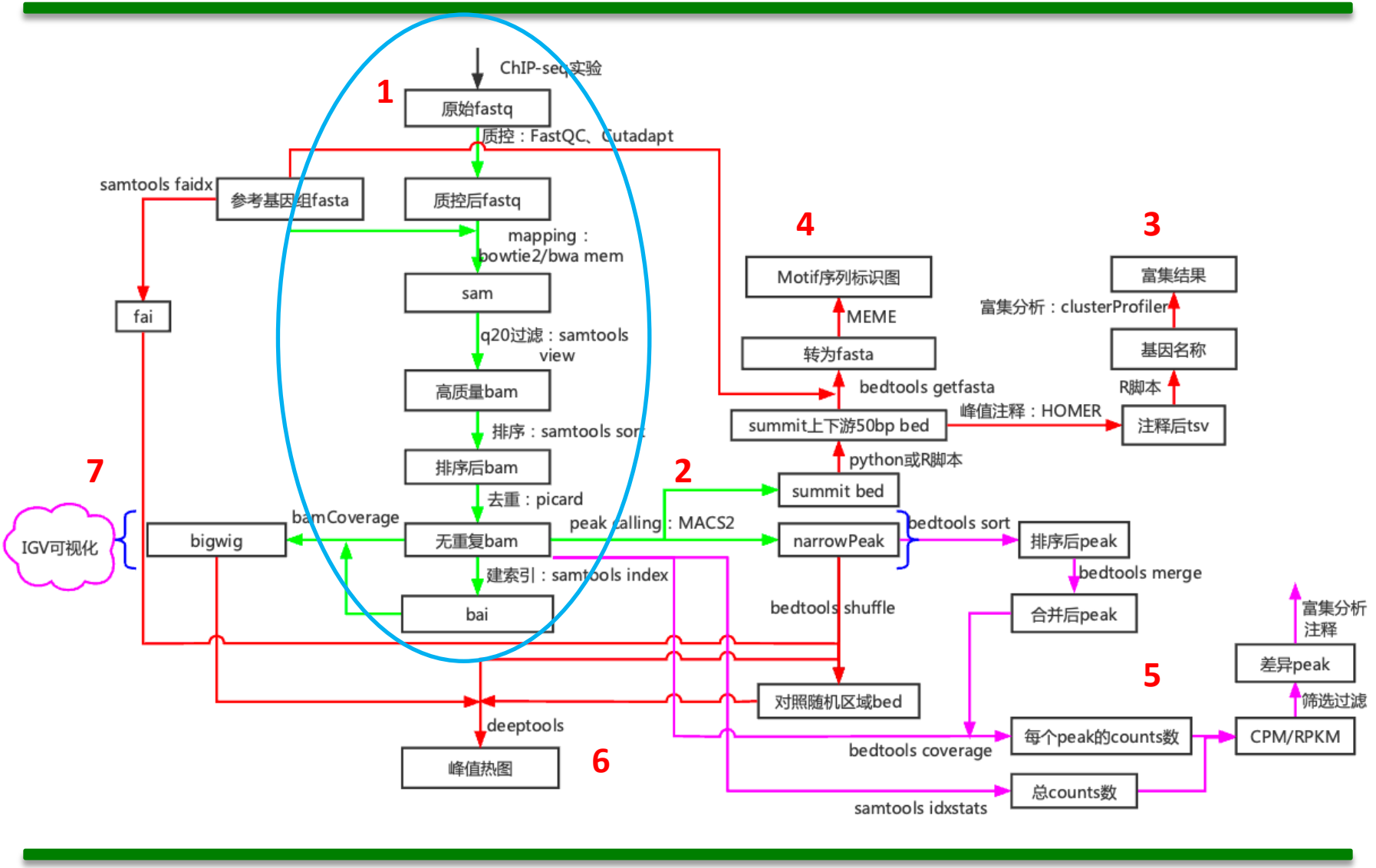


https://www.nature.com/articles/nrg2641

ChIP-seq数据详细分析流程

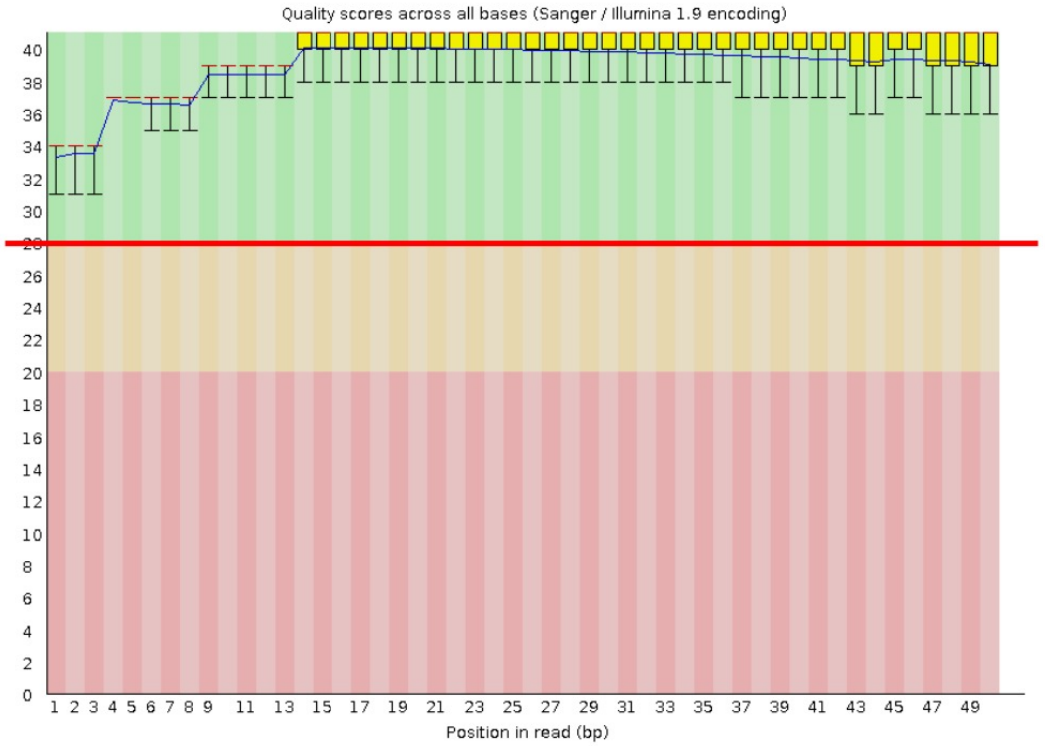


ChIP-seq数据详细分析流程



Quality metrics of sequencing reads

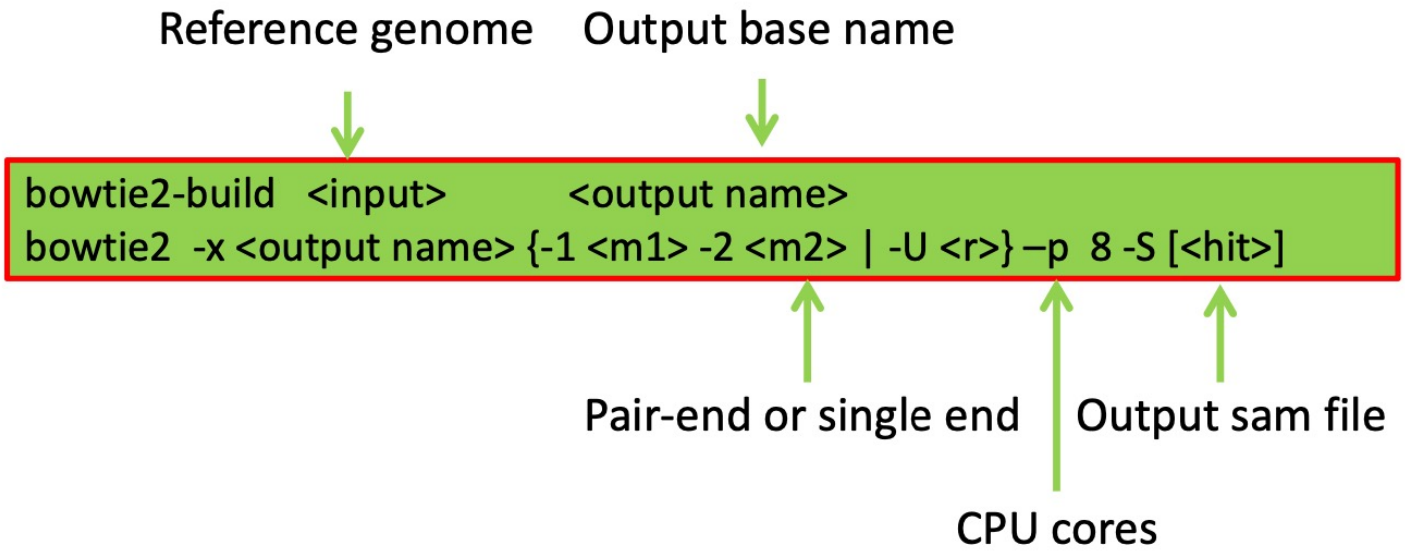
- FastQC can be used for an overview of the data quality
- Phred quality scores used for trimming low quality bases
- *P* = 10^(-*Q*/10); *Q*=30 base is called incorrectly 1 in 1000



https://biohpc.cornell.edu/lab/doc/CHIPseq_workshop_20150504_lecture1.pdf (2)

Reads mapping

Most popular software: Bowtie, BWA, MAQ etc



- Multiple mapping hits were discarded

https://biohpc.cornell.edu/lab/doc/CHIPseq_workshop_20150504_lecture1.pdf

Reference genome; FASTA format: 2 lines for each read (“>name”, sequence)

```
>I
GCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAA
GCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAA
GCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAA
GCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAA
GCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAA
```

Illumina raw data; FASTQ format: 4 lines per read (“@name”, sequence, “+”, quality string)

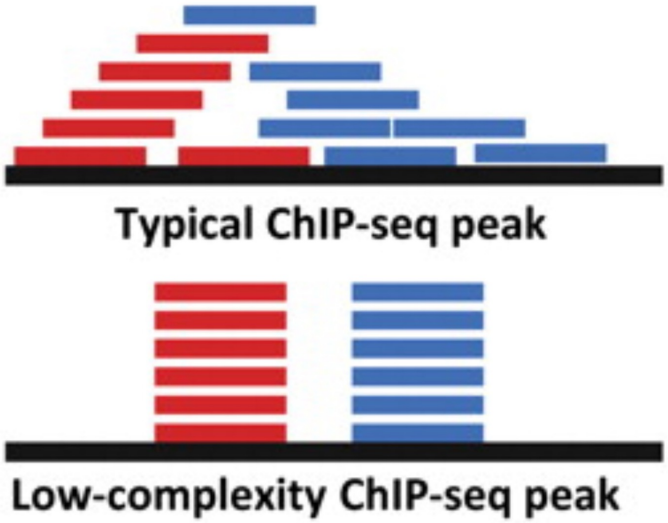
```
@ILLUMINA:405:C269YACXX:1:1101:3833:1996 1:N:0:NAAA
CAATGGAAGAACAGACACTACATATATTGAGCACATTATCATGTTA
+
FFFFHHHFHJJJIHGIGGIGGGIJGIJJJIGEHIGIIJIIHIIHII
```

SAM output Sequence Alignment Map

```
ILLUMINA:405:C269YACXX:1:2115:4141:76972        0       I       350017  42      46M     *       0
0       AAATAGATAATTCTTTGAAAATGAATAATTTTTTGGAAAAATCTTT  FFFFHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJIJJJ
JJJJJJ  AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:46 YT:Z:UU
ILLUMINA:405:C269YACXX:1:2203:10724:54489       0       I       350046  42      46M     *       0
0       TTTTTGGAAAAATCTTTTTTTCTTATTTTTCTAGACGTACACCTTT  FFFFHHHHJJJJJJJJJJJJJJJJIJJJJJJJJJJJJHIJ
JJJJJJ  AS:i:0  XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:46 YT:Z:UU
```

https://biohpc.cornell.edu/lab/doc/CHIPseq_workshop_20150504_lecture1.pdf

Quality Control： NRF



- Nonredundant fraction (NRF)

$$\mathrm{NRF} = \frac{\#\text{unique start positions of uniquely mappable reads}}{\#\text{uniquely mappable reads}}$$

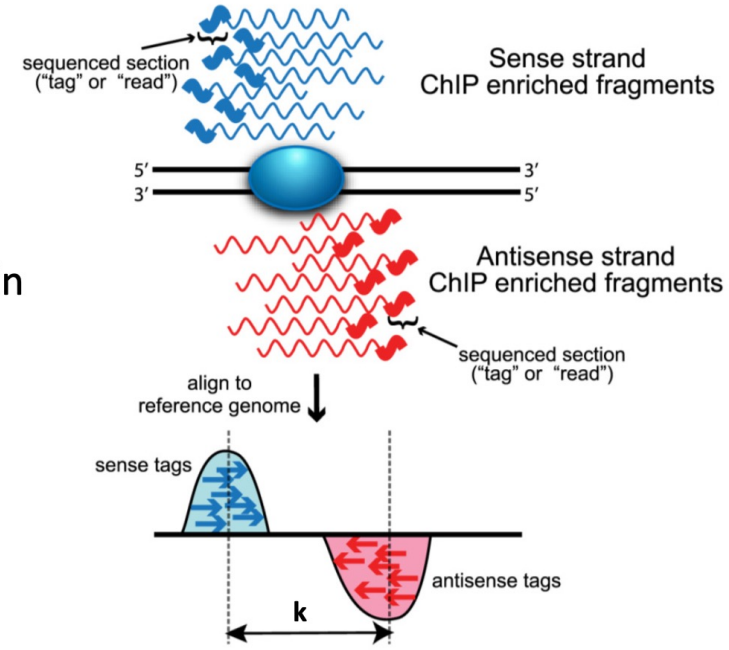
ENCODE recommends target of NRF 0:8 for 10 million uniquely mapped reads

https://biohpc.cornell.edu/lab/doc/CHIPseq_workshop_20150504_lecture1.pdf

Cross-correlation

DNA fragments from a chromatin immunoprecipitation experiment are sequenced from the 5' end.

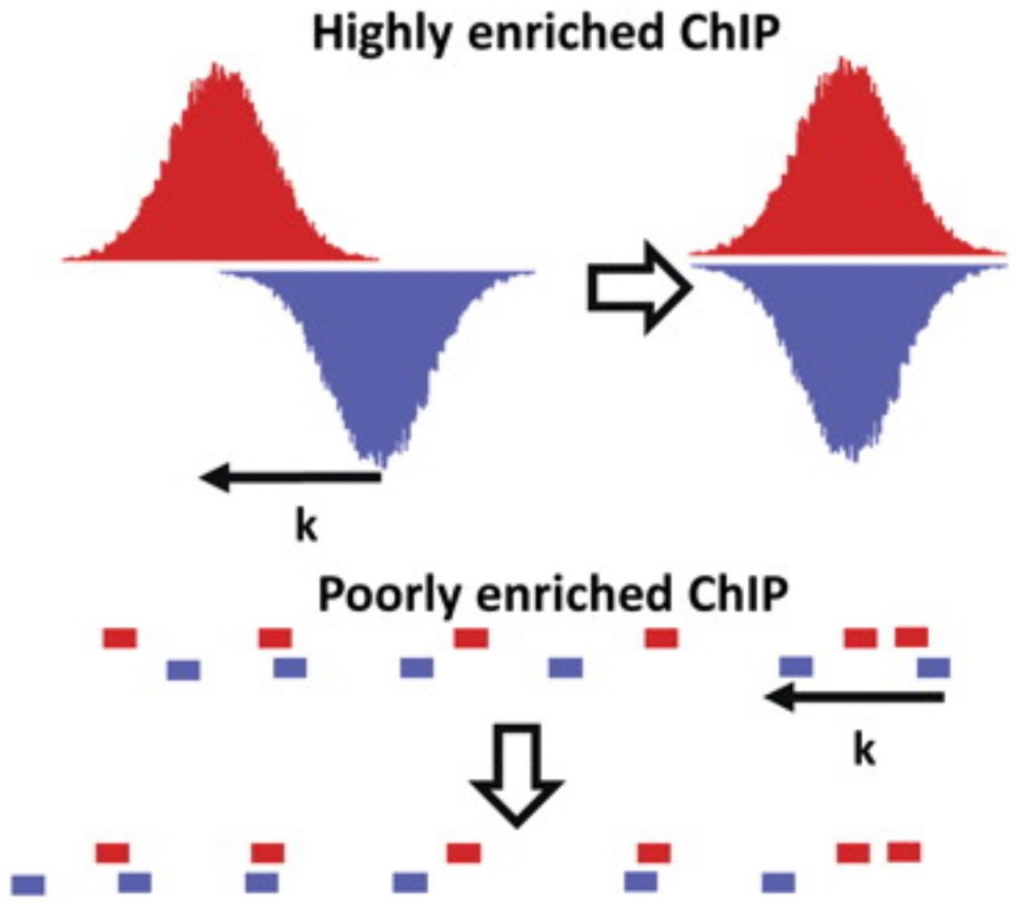
- With ChIP-seq, the alignment of the reads to the genome results in two peaks (one on each strand) that located on flanking sides of the protein or nucleosome of interest.
- The distance between strands specific peaks (k) represents the average sequenced fragment.



Wilbanks EG (2010) Evaluation of Algorithm Performance in ChIP-Seq Peak Detection PLoS ONE 5:e11471

https://biohpc.cornell.edu/lab/doc/CHIPseq_workshop_20150504_lecture1.pdf

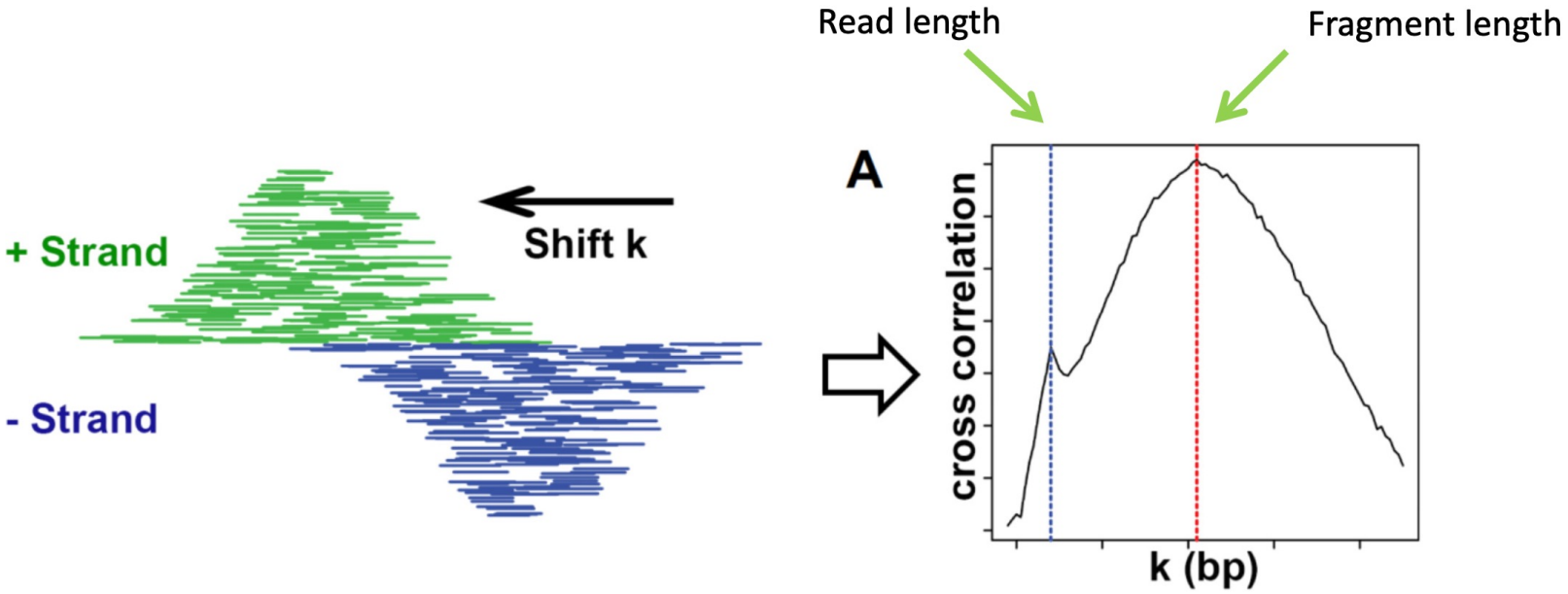
Cross-correlation



Stephen G. Landt (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia Genome Res 22: 1813-1831

https://biohpc.cornell.edu/lab/doc/CHIPseq_workshop_20150504_lecture1.pdf

Cross-correlation



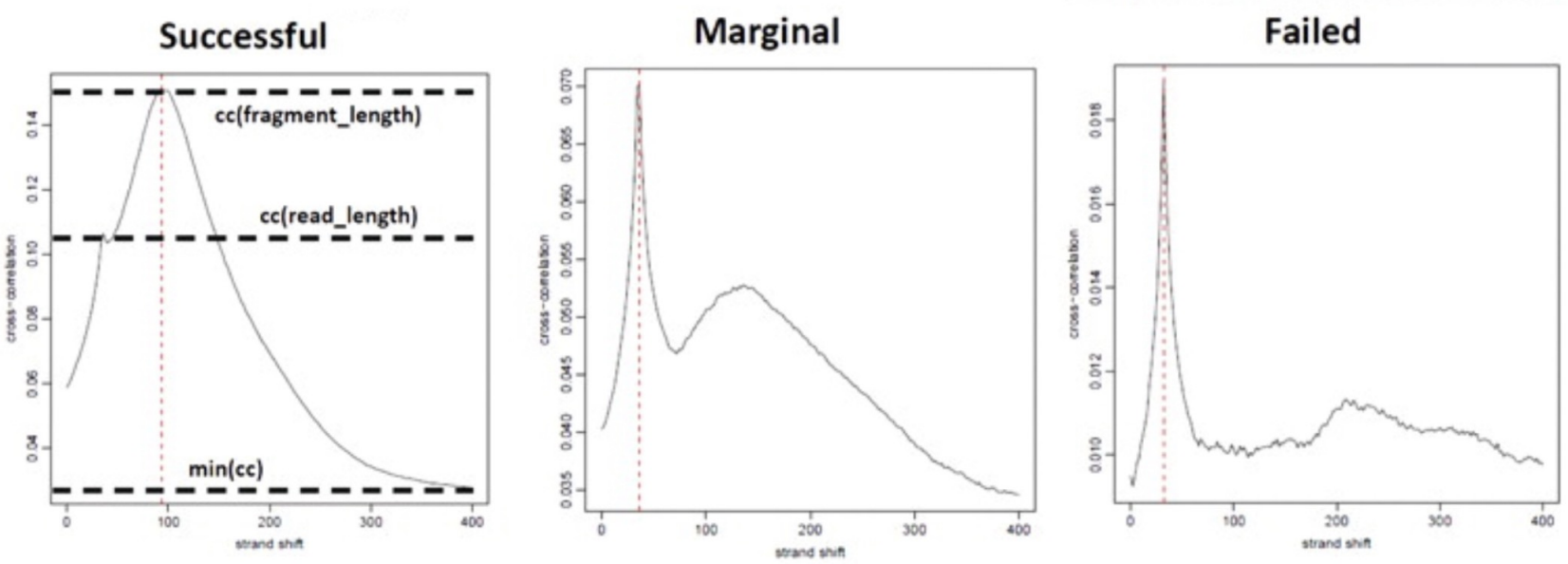
Strand cross-correlation is computed as the Pearson correlation between the positive and the negative strand profiles at different strand shift distances, k

https://sites.google.com/a/brown.edu/bioinformatics-in-biomed/spp-r-from-chip-seq

Bailey, et al (2013). Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data, PLOS Computational Biology

https://biohpc.cornell.edu/lab/doc/CHIPseq_workshop_20150504_lecture1.pdf

Cross-correlation



$$NSC = \frac{cc(fragment\ length)}{min(cc)} \qquad RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$

NSC values < 1.05 and RSC values < 0.8

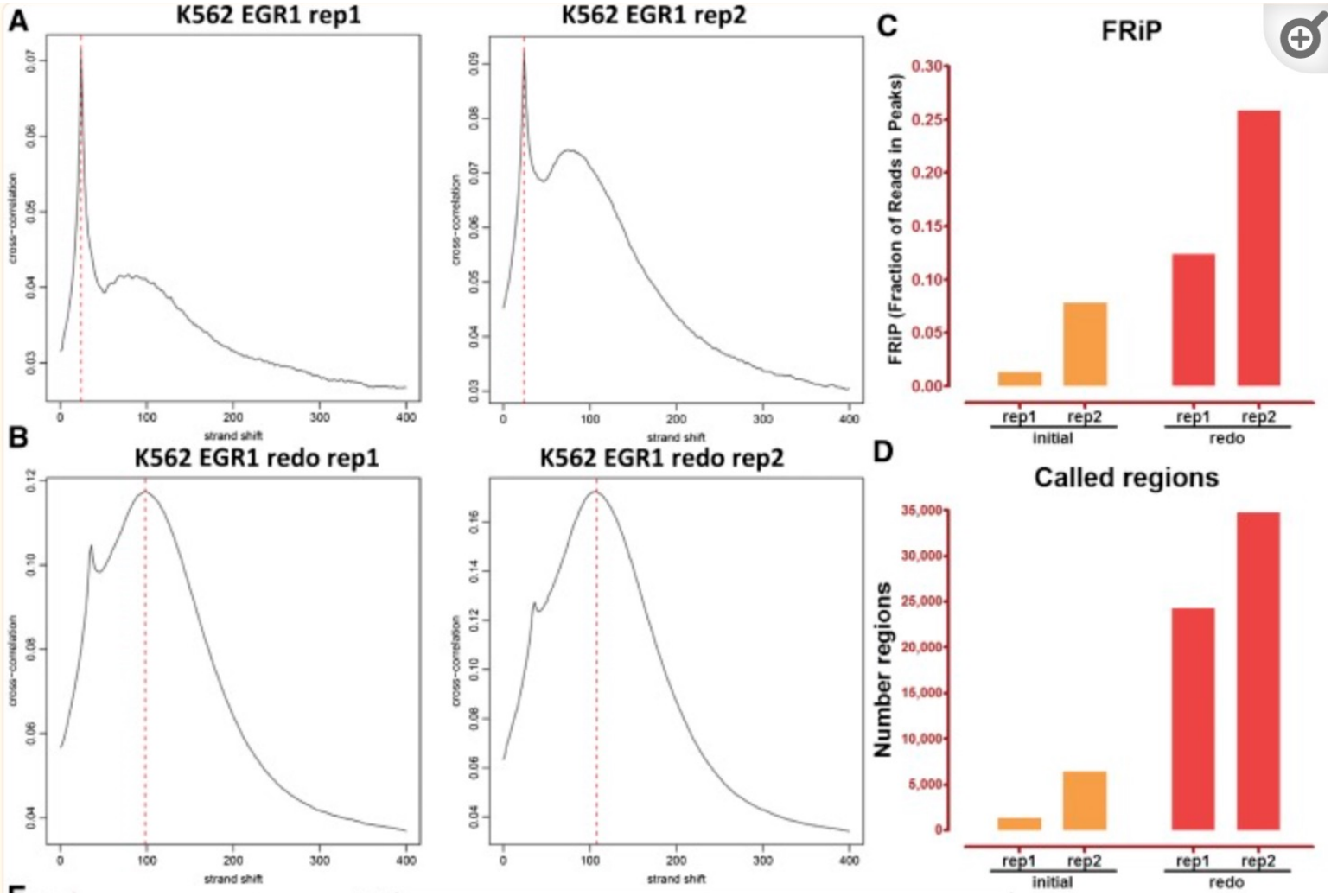
http://code.google.com/p/phantompeakqualtools/

NSC（Normalized Strand Cross-Correlation）

Stephen G. Landt (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia Genome Res 22: 1813-1831

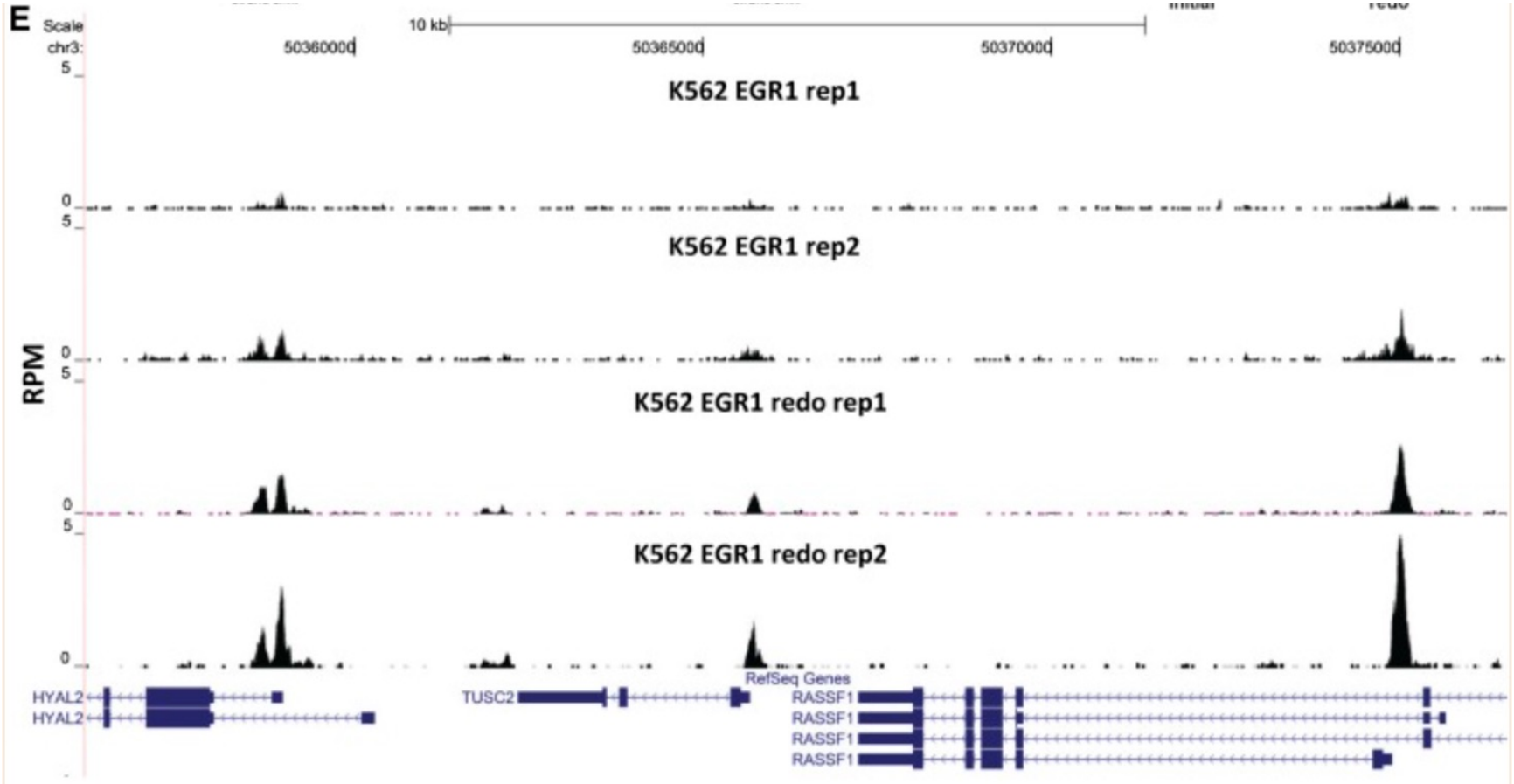
https://biohpc.cornell.edu/lab/doc/CHIPseq_workshop_20150504_lecture1.pdf

Quality control of ChIP-seq data sets in practice



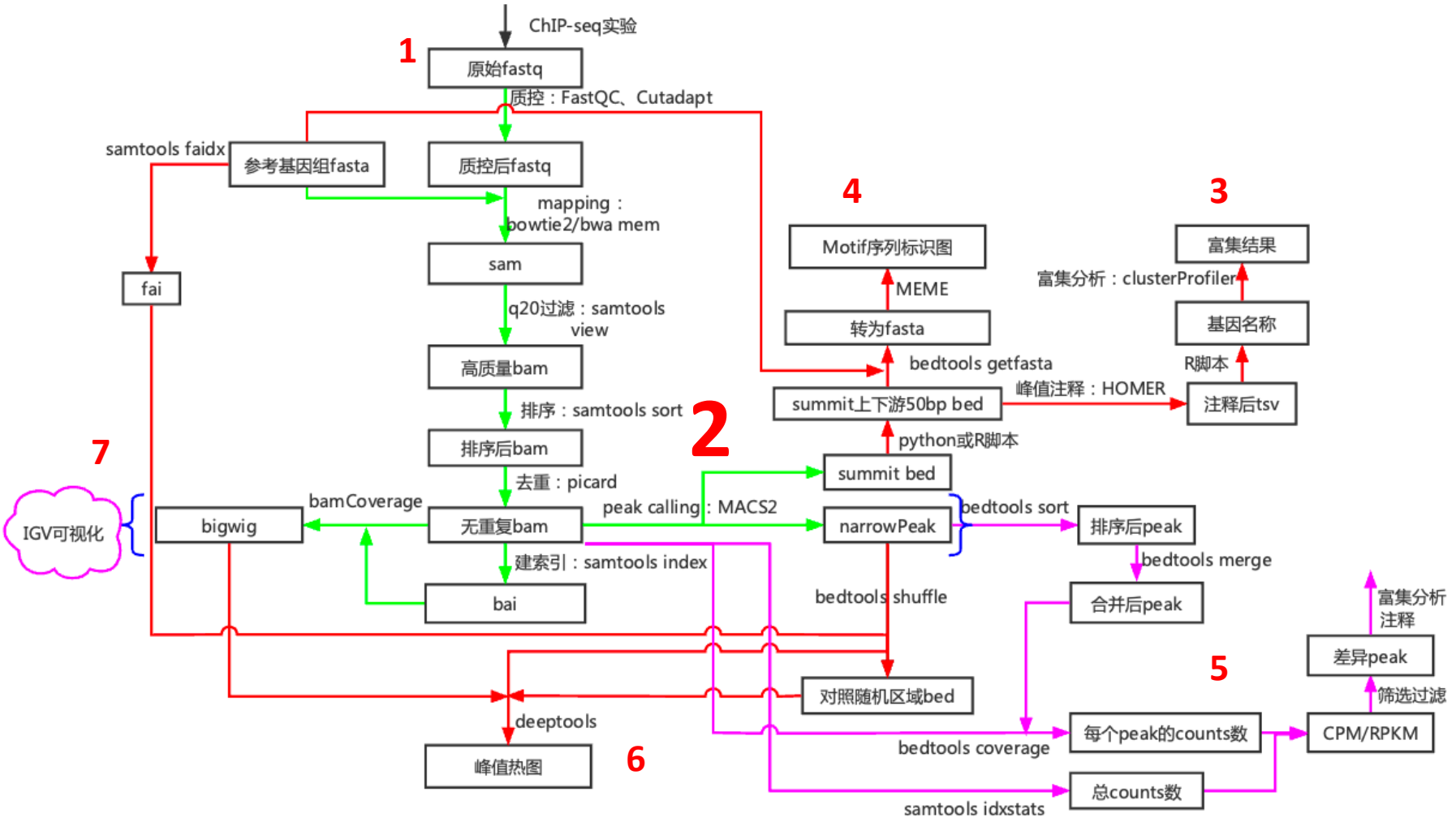
https://genome.cshlp.org/content/22/9/1813.abstract

Quality control of ChIP-seq data sets in practice

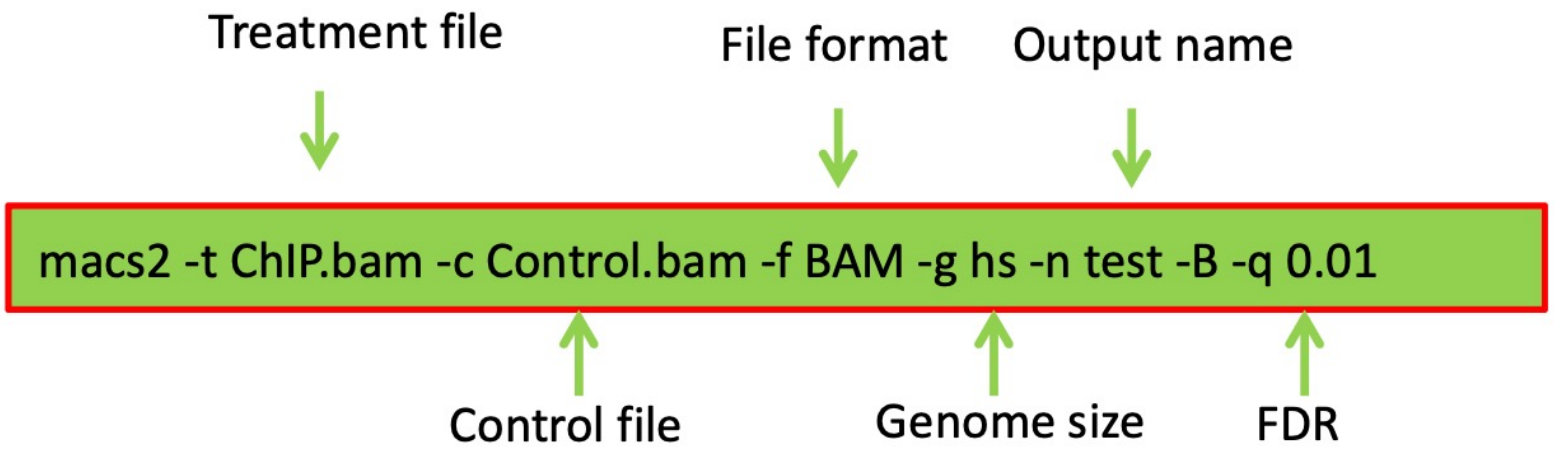


https://genome.cshlp.org/content/22/9/1813.abstract

ChIP-seq数据详细分析流程



Command of MACS2



Option:

-s TSIZE, tsize=TSIZE

-m MFOLD, --mfold=MFOLD

--bw=BW

--nomodel

--shiftsize=SHIFTSIZE

https://biohpc.cornell.edu/lab/doc/CHIPseq_workshop_20150504_lecture1.pdf

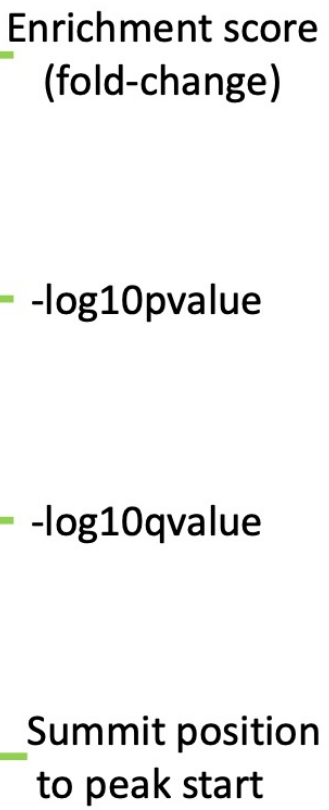
峰值定量-BED文件

- BED文件用于记录染色体的区域位置
- 常见的BED为3-6列
  - chrom
  - chromStart
  - chromEnd
  - name
  - score(A score between 0 and 1000)
  - strand (Defines the strand. Either “.” (=no strand) or “+” or “-”)

chr7	127471196	127472363	Pos1	0	+
chr7	127472363	127473530	Pos2	0	+
chr7	127473530	127474697	Pos3	0	+
chr7	127474697	127475864	Pos4	0	+
chr7	127475864	127477031	Neg1	0	-

Output of MACS2

```
[mingh@cbsumm11 H3K4]$ more rep3_D2K4_H3_peaks.encodePeak
track type=narrowPeak nextItemButton=on
I       4059    4291    MACS_peak_1     136     .       3.92    15.47   13.65   88
I       16620   16867   MACS_peak_2     102     .       3.45    11.98   10.24   107
I       24153   24398   MACS_peak_3     130     .       3.84    14.87   13.06   113
I       24562   24868   MACS_peak_4     325     .       5.86    34.67   32.50   140
I       26424   27627   MACS_peak_5     462     .       7.11    48.60   46.21   275
I       28283   28442   MACS_peak_6     61      .       2.55    7.72    6.13    71
I       30981   31131   MACS_peak_7     62      .       2.84    7.84    6.24    86
I       31801   32130   MACS_peak_8     84      .       2.98    10.08   8.41    97
I       33712   33899   MACS_peak_9     61      .       2.55    7.72    6.13    44
I       34605   35205   MACS_peak_10    74      .       2.59    9.08    7.44    451
I       35353   35741   MACS_peak_11    97      .       3.38    11.43   9.71    78
I       36168   36391   MACS_peak_12    68      .       2.78    8.48    6.86    143
I       39389   39878   MACS_peak_13    148     .       4.07    16.71   14.86   345
I       40039   40344   MACS_peak_14    71      .       2.99    8.81    7.18    214
I       40930   41090   MACS_peak_15    53      .       2.69    6.91    5.35    69
I       46949   47213   MACS_peak_16    180     .       4.45    19.92   18.00   135
I       47288   47607   MACS_peak_17    124     .       3.76    14.27   12.47   203
I       70140   70613   MACS_peak_18    354     .       6.30    37.65   35.43   135
I       93000   93232   MACS_peak_19    62      .       2.84    7.84    6.24    100
I       97597   98073   MACS_peak_20    305     .       5.15    32.72   30.59   292
I       98224   98465   MACS_peak_21    180     .       4.45    19.92   18.00   123
I       107919  108073  MACS_peak_22    60      .       2.78    7.62    6.04    65
I       108184  109005  MACS_peak_23    202     .       4.34    22.16   20.20   604
I       109091  111927  MACS_peak_24    820     .       9.26    85.23   82.07   2142
I       171398  171571  MACS_peak_25    45      .       2.53    6.03    4.51    79
I       182407  182922  MACS_peak_26    307     .       5.83    32.90   30.76   378
I       237822  237986  MACS_peak_27    81      .       3.15    9.83    8.16    65
I       288519  289415  MACS_peak_28    496     .       7.60    52.09   49.64   299
I       310449  310912  MACS_peak_29    148     .       4.07    16.71   14.86   149
I       310963  311271  MACS_peak_30    90      .       3.12    10.75   9.06    145
I       314136  315758  MACS_peak_31    659     .       8.98    68.73   65.96   782
I       315988  316213  MACS_peak_32    81      .       3.15    9.83    8.16    90
```



https://biohpc.cornell.edu/lab/doc/CHIPseq_workshop_20150504_lecture1.pdf

Consistency of replicates: IDR

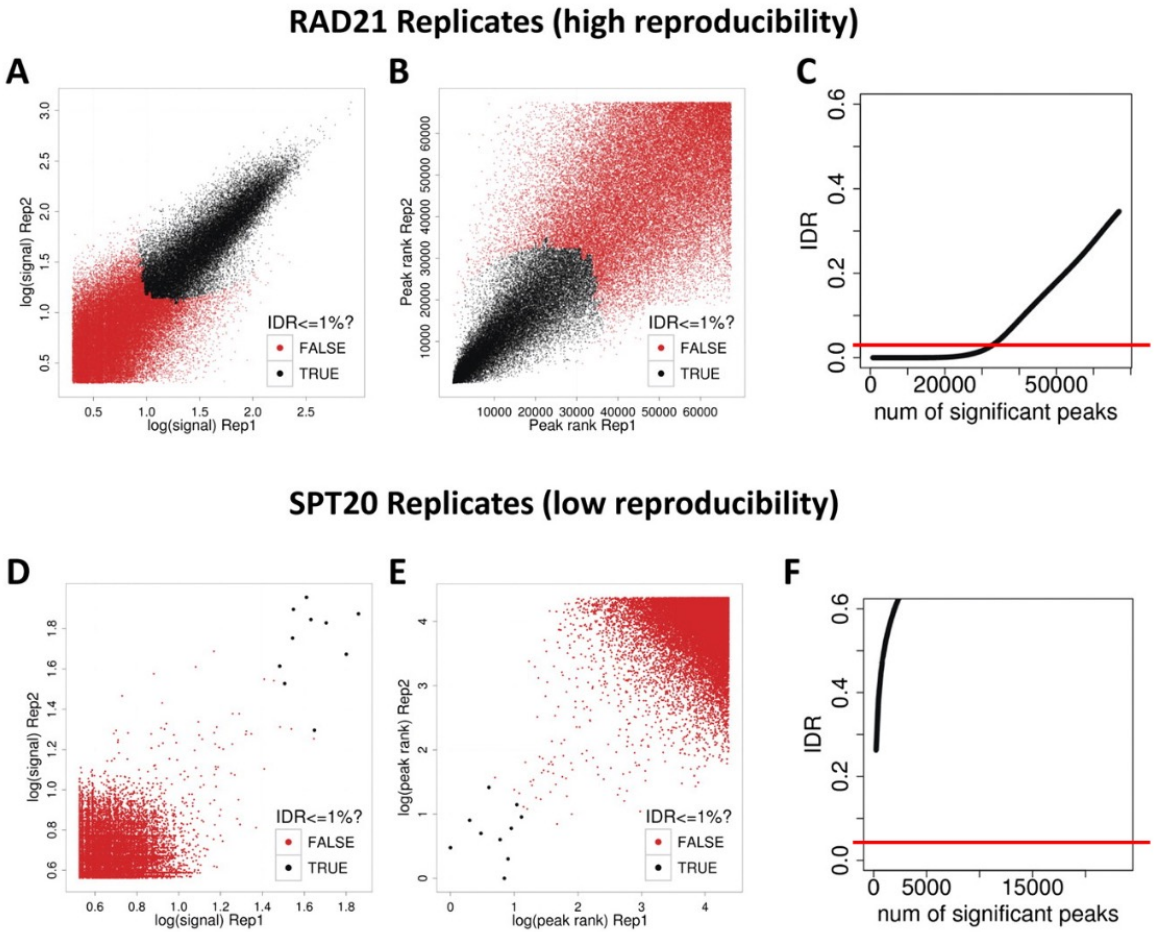
- IDR (Irreproducible Discovery Rate) 是一种用于比较不同 ChIP-seq 数据集之间一致性的方法。它可以帮助鉴定高质量的信号和峰值，避免因峰值重复而导致的假阳性结果。

- 在 ChIP-seq 分析中，研究人员通常会对同一个蛋白质结合在不同实验条件下得到的 ChIP-seq 数据进行比较。IDR 分析可以用于确定两个 ChIP-seq 数据集之间共同检测到的信号的数量，即 IDR 阈值。只有通过 IDR 阈值过滤的峰值才会被认为是高可重复性和高可靠性的峰值。

- IDR 分析是一种广泛使用的技术，特别是在研究转录因子和组蛋白修饰等染色质相关蛋白的结合时，可帮助鉴定高度一致和可重复的峰值，并减少假阳性的结果。

https://genome.cshlp.org/content/22/9/1813.abstract

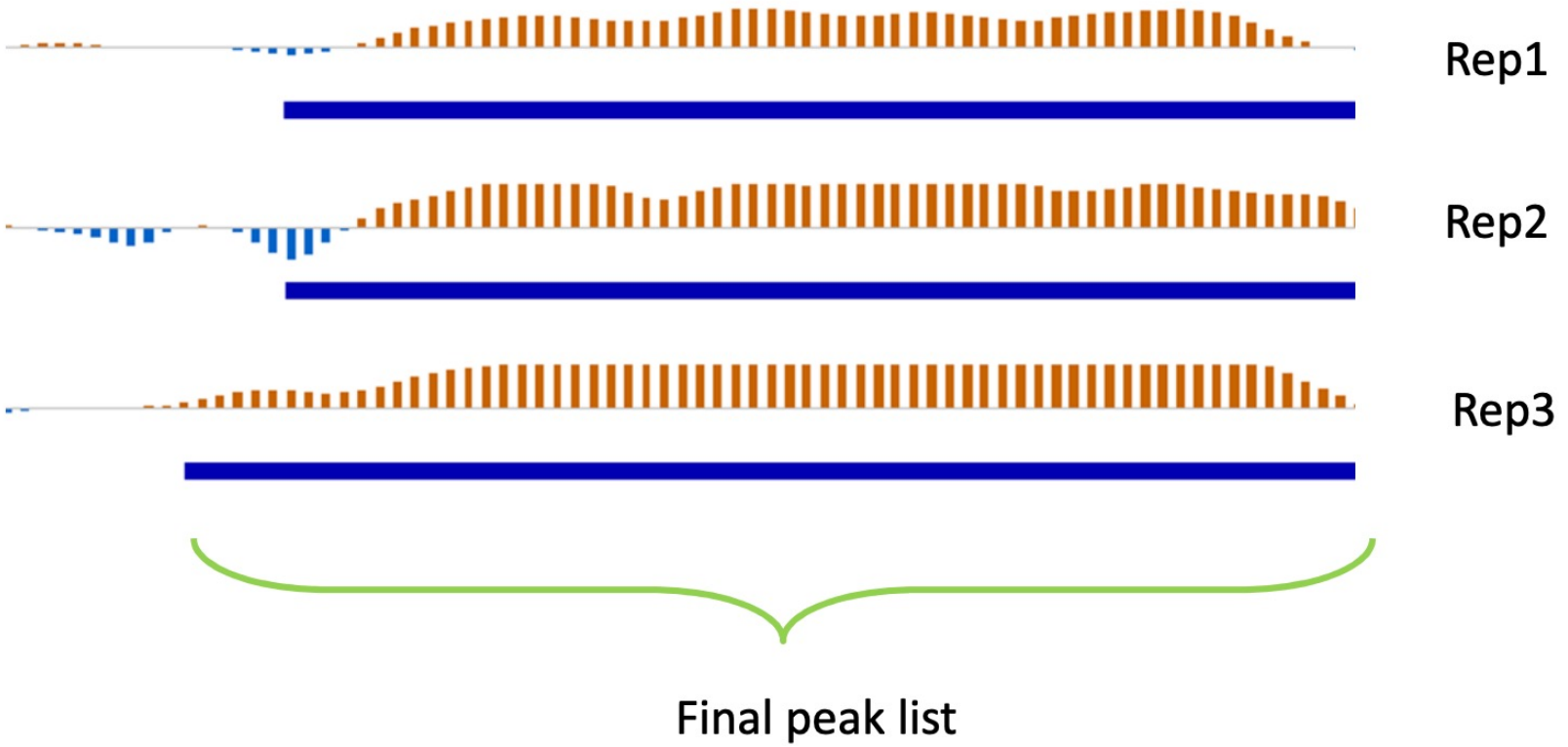
Consistency of replicates: IDR



Stephen G. Landt (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia Genome Res 22: 1813-1831

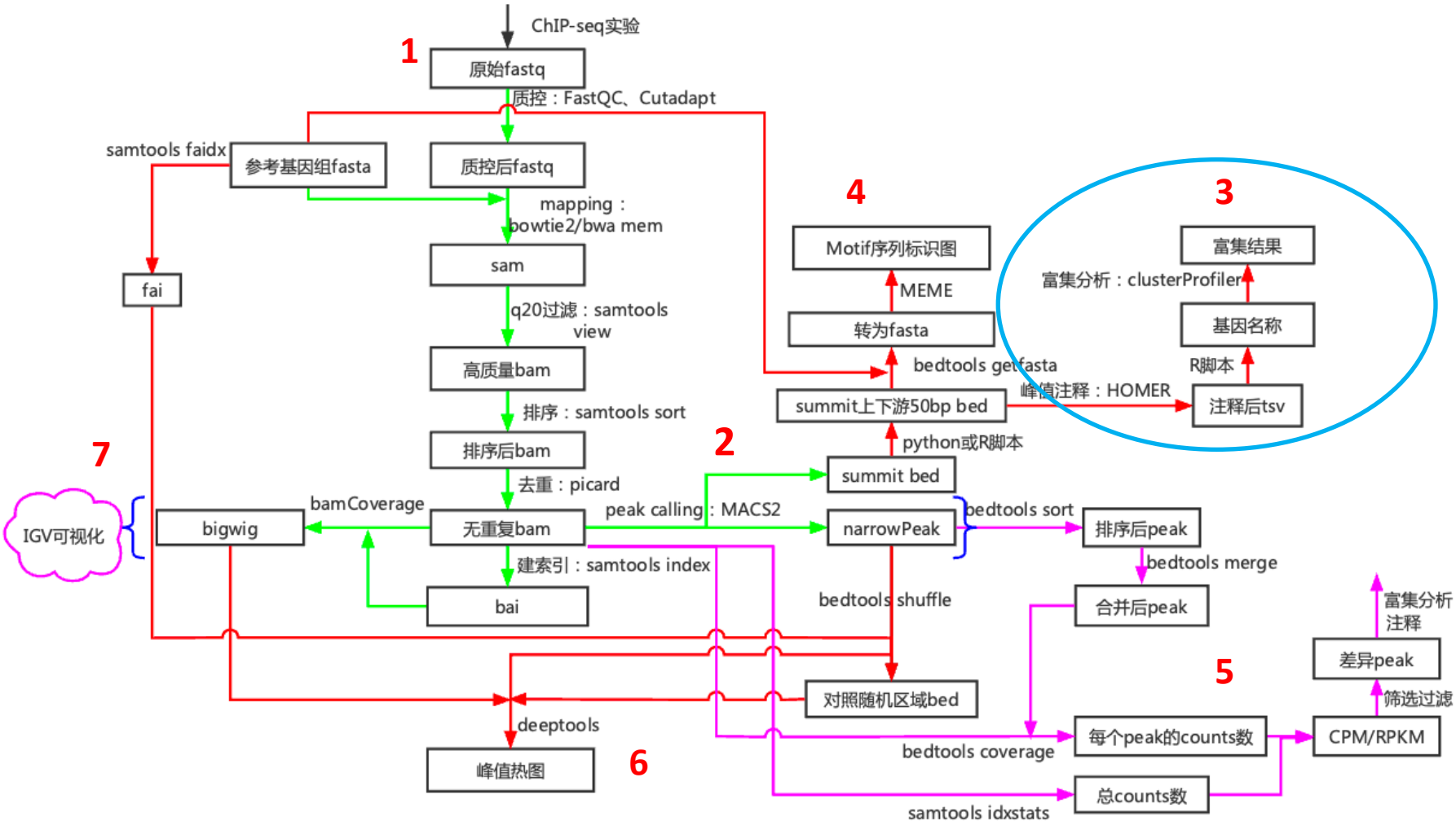
https://genome.cshlp.org/content/22/9/1813.abstract

Peak region merging



https://biohpc.cornell.edu/lab/doc/CHIPseq_workshop_20150504_lecture1.pdf

ChIP-seq数据详细分析流程



ChIP-seq peak的注释

■使用HOMER软件对富集峰进行注释

HOMER

Software for motif discovery and next-gen sequencing analysis

---

Installation Guide:

Basic requirements

Homer is computationally intensive collection of programs. The following are minimum hardware requirements for running promoter analysis (ChIP-Seq in parenthesis).

- Unix-style operating system (UNIX/LINUX/Mac/Cygwin)
- 1 Gb of RAM (4+ Gb)
- 1 Gb of Hard Drive Space (>10Gb)

While running Homer is designed to be as simple as possible, some basic knowledge of UNIX commands is required. If you are new to UNIX, try googling "UNIX tutorial" for a more formal introduction.

Required UNIX tools (fairly standard) and recommended NGS software

The following UNIX utilities are required to use HOMER. You might need to run your UNIX package manager to install them if missing (i.e. "sudo apt-get install wget"):

- gcc
- g++
- make
- perl
- zip/unzip
- gzip/gunzip
- wget

http://homer.ucsd.edu/homer/introduction/install.html

ChIP-seq peak的注释

- Peak文件从ENCODE数据库下载，如ENCFF151CRB
- 注释命令：
  - annotatePeaks.pl ENCFF151CRB.bed hg38 -annStats ann_stats.log > ENCFF151CRB.ann.tsv
- 主要结果
  - ENCFF151CRB.ann.tsv [注释结果文件]
  - ann_stats.log [peak分布统计情况]

ChIP-seq peak的注释

■结果解读

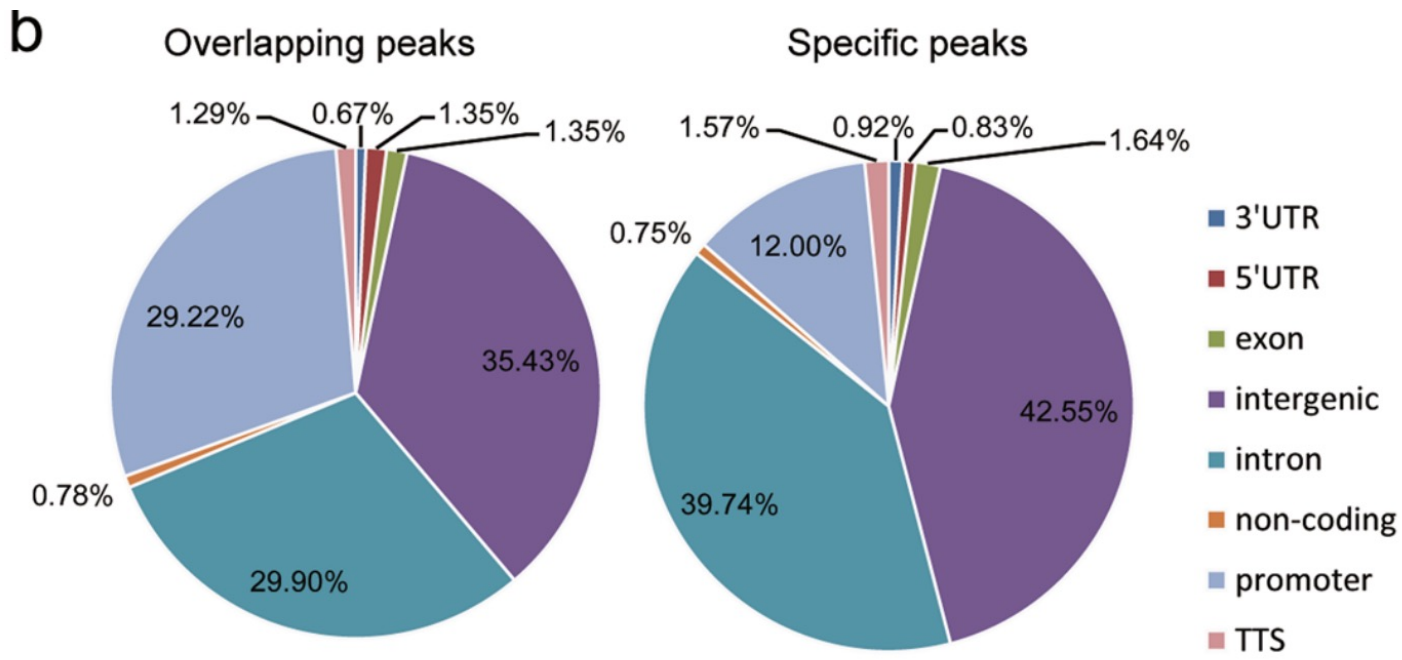
	A	B	C	D	E	F	G	H	I
1	PeakID	Chr	Start	End	Strand	Peak Score	Focus Ratio/Region Size	Annotation	Detailed Annotation
2	1-38726	chr19	44955374	44955777	+	0	NA	intron (NM_001294, intron 1 of 13)	CpG
3	1-114866	chr12	10128768	10129171	+	0	NA	intron (NR_125336, intron 1 of 6)	intron (NR_125336, intron 1 of 6)
4	1-64735	chr14	20495800	20496203	+	0	NA	Intergenic	Intergenic
5	1-278925	chrX	78629154	78629557	+	0	NA	Intergenic	L1PA7|LINE|L1
6	1-107601	chr15	92642372	92642775	+	0	NA	intron (NM_207446, intron 1 of 2)	THE1C|LTR|ERVL-MaLR
7	1-133298	chr14	104665908	104666311	+	0	NA	promoter-TSS (NR_135293)	promoter-TSS (NR_135293)
8	1-92085	chr12	111834479	111834882	+	0	NA	Intergenic	L1PA3|LINE|L1
9	1-190675	chr15	36228128	36228531	+	0	NA	Intergenic	L1MD|LINE|L1
10	1-169192	chr10	107218697	107219100	+	0	NA	Intergenic	L1ME3|LINE|L1
11	1-27042	chr2	18514948	18515351	+	0	NA	Intergenic	Intergenic
12	1-271331	chrX	74489733	74490136	+	0	NA	intron (NM_006517, intron 1 of 5)	MER4B-int|LTR|ERV1
13	1-262785	chr2	179210154	179210557	+	0	NA	intron (NM_178123, intron 1 of 17)	L1PB1|LINE|L1
14	1-272194	chr2	25925897	25926300	+	0	NA	TTS (NM_002254)	TTS (NM_002254)
15	1-60733	chr1	167192922	167193325	+	0	NA	intron (NR_110811, intron 2 of 3)	intron (NR_110811, intron 2 of 3)
16	1-213669	chr5	160480369	160480772	+	0	NA	intron (NR_132748, intron 1 of 1)	L2a|LINE|L2
17	1-25524	chr12	8373753	8374156	+	0	NA	intron (NR_024420, intron 1 of 1)	intron (NR_024420, intron 1 of 1)
18	1-168526	chr3	4271841	4272244	+	0	NA	Intergenic	Intergenic
19	1-38643	chrX	41085942	41086345	+	0	NA	intron (NM_001039591, intron 1 of 44)	CpG-29609
20	1-282156	chr8	92430509	92430912	+	0	NA	Intergenic	MSTB1|LTR|ERVL-MaLR
21	1-40713	chr11	1554309	1554712	+	0	NA	3' UTR (NM_004420, exon 7 of 7).2	3' UTR (NM_004420, exon 7 of 7).2
22	1-20464	chr1	58804868	58805271	+	0	NA	intron (NR_108106, intron 1 of 1)	L2a|LINE|L2
23	1-219131	chr17	43982414	43982817	+	0	NA	intron (NM_004160, intron 1 of 6)	intron (NM_004160, intron 1 of 6)
24	1-294648	chr2	179531269	179531672	+	0	NA	intron (NM_001352811, intron 4 of 10)	AluSx1|SINE|Alu
25	1-96454	chr19	43564786	43565189	+	0	NA	intron (NM_006297, intron 2 of 16)	MLT1D|LTR|ERVL-MaLR
26	1-220439	chr5	177635363	177635766	+	0	NA	intron (NR_026921, intron 1 of 5)	L1PA5|LINE|L1

ChIP-seq peak的注释

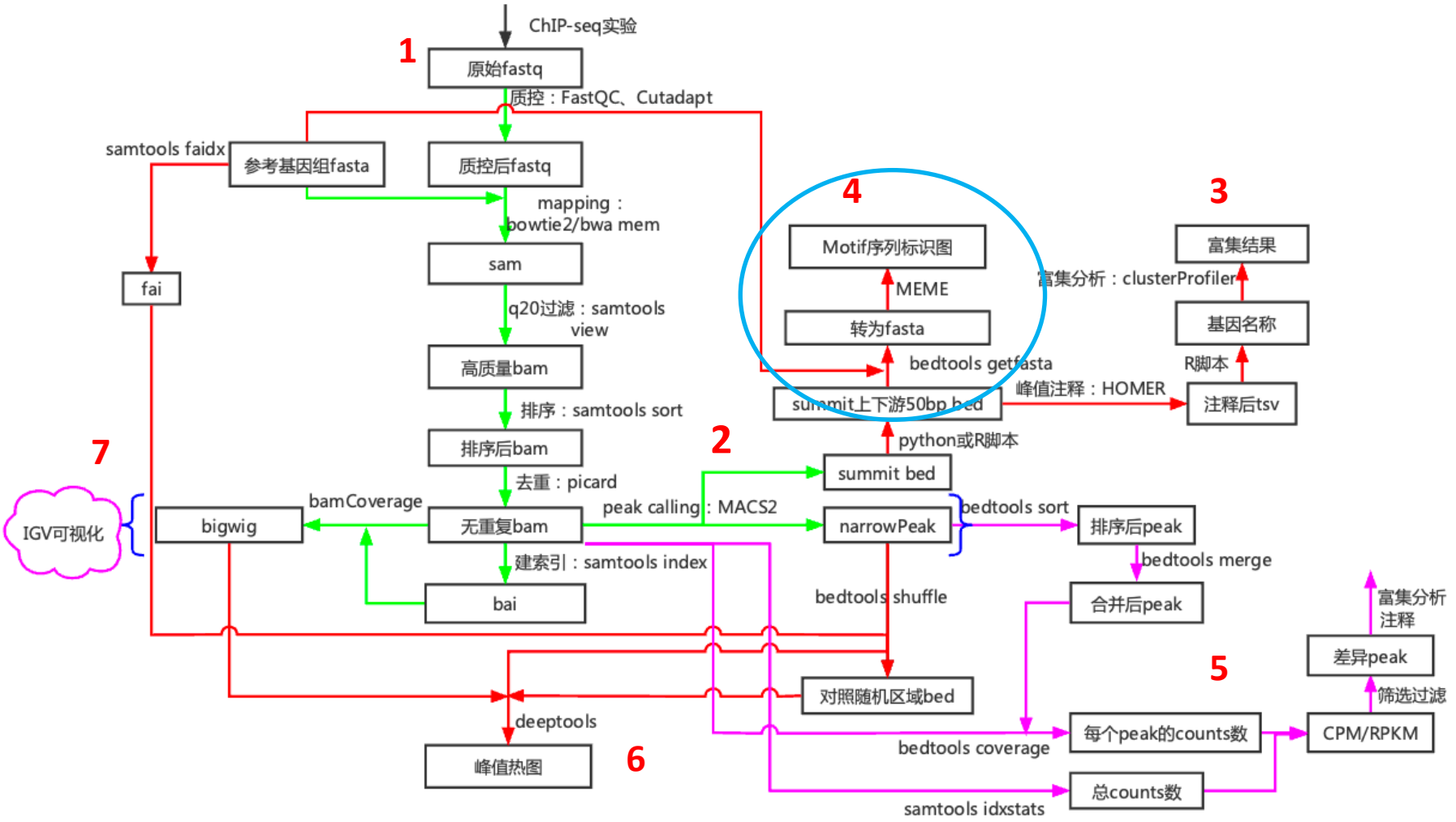
■结果解读

J	K	L	M	N	O	P
Distance to TSS	Nearest PromoterID	Entrez ID	Nearest Unigene	Nearest Refseq	Nearest Ensembl	Gene Name
195	NM_001294	1209	Hs.444441	NM_001294	ENSG00000104853	CLPTM1
1230	NR_125336	64581	Hs.143929	NM_022570	ENSG00000172243	CLEC7A
-14471	NM_001012975	338879	Hs.451057	NM_001012975	ENSG00000182545	RNASE10
29973	NM_152694	203430	Hs.134873	NM_152694	ENSG00000179300	RTL3
13202	NM_207446	400451	Hs.27373	NM_207446	ENSG00000185442	FAM174B
-551	NR_135293	101929634	Hs.532502	NR_135293	ENSG00000260792	LINC02280
7250	NR_152605	51275	Hs.333120	NR_015404	ENSG00000234608	MAPKAPK5-AS1
301473	NR_039735	100616293		NR_039735	ENSG00000265098	MIR4510
-54192	NM_052918	114815	Hs.591915	NM_052918	ENSG00000108018	SORCS1
45504	NM_020905	57665	Hs.120319	NM_020905	ENSG00000240857	RDH14
68441	NM_006517	6567	Hs.75317	NM_006517	ENSG00000147100	SLC16A2
54477	NM_178123	91404	Hs.30977	NM_178123	ENSG00000187231	SESTD1
-47611	NM_018263	55252	Hs.119815	NM_018263	ENSG00000143970	ASXL2
2682	NR_110811	101928484	Hs.568496	NR_110811		LINC01363
-4782	NR_029701	406938		NR_029701	ENSG00000283733	MIR146A
16798	NR_024420	389634	Hs.434403	NM_001012988	ENSG00000226091	LINC00937
-31262	NM_001243723	6419	Hs.475300	NM_006515	ENSG00000170364	SETMAR
698	NM_001039590	8239	Hs.77578	NM_004652	ENSG00000124486	USP9X
224784	NR_125827	102724710	Hs.125714	NR_125827	ENSG00000253634	LOC102724710
17761	NM_004420	1850	Hs.41688	NM_004420	ENSG00000184545	DUSP8
19918	NR_034015	100131060	Hs.680604	NR_034014	ENSG00000234807	LINC01135
21854	NM_004160	5697	Hs.169249	NM_004160	ENSG00000131096	PYY
31118	NR_148055	151126	Hs.655005	NM_152520	ENSG00000144331	ZNF385B
10591	NM_006297	7515	Hs.98493	NM_006297	ENSG00000073050	XRCC1
35432	NM_007255	11285	Hs.455109	NM_007255	ENSG00000027847	B4GALT7

ChIP-seq peak的注释



ChIP-seq数据详细分析流程



Motif分析

■定义

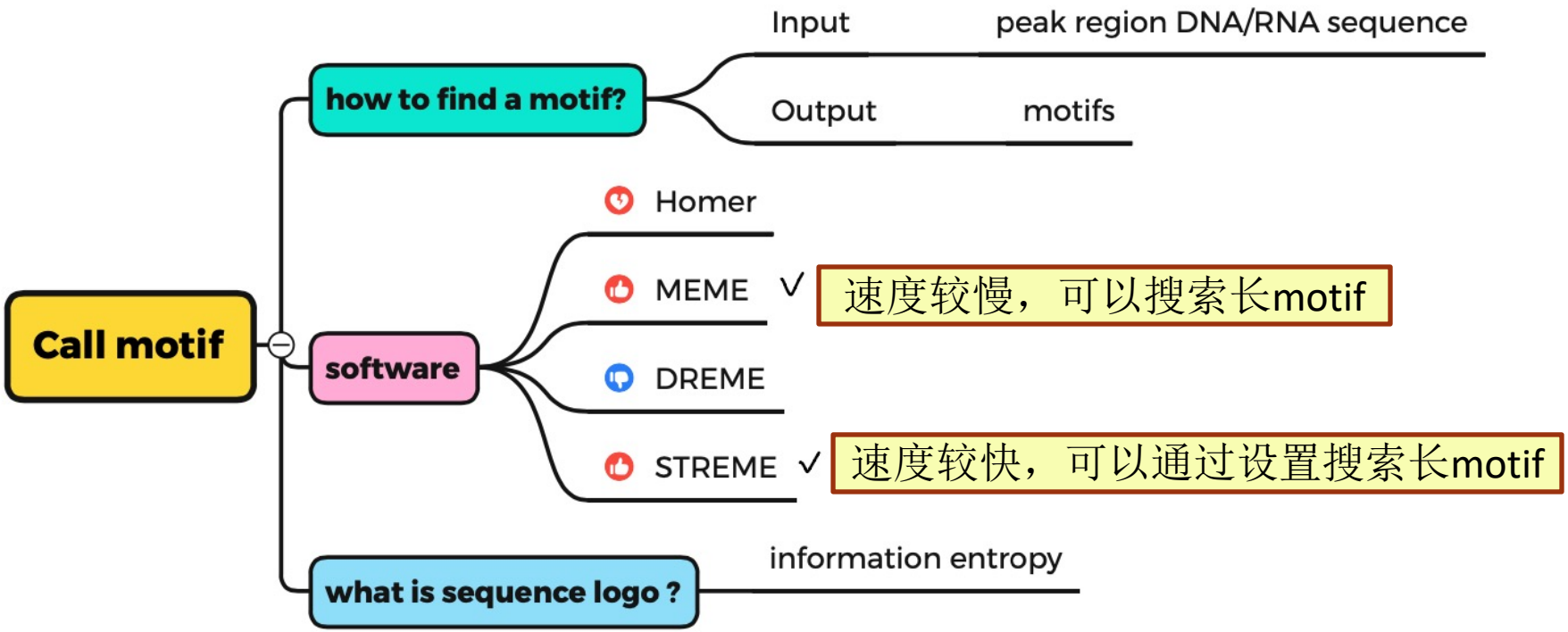
– 核酸或蛋白质序列中的一些特定模式的序列片段

■目的

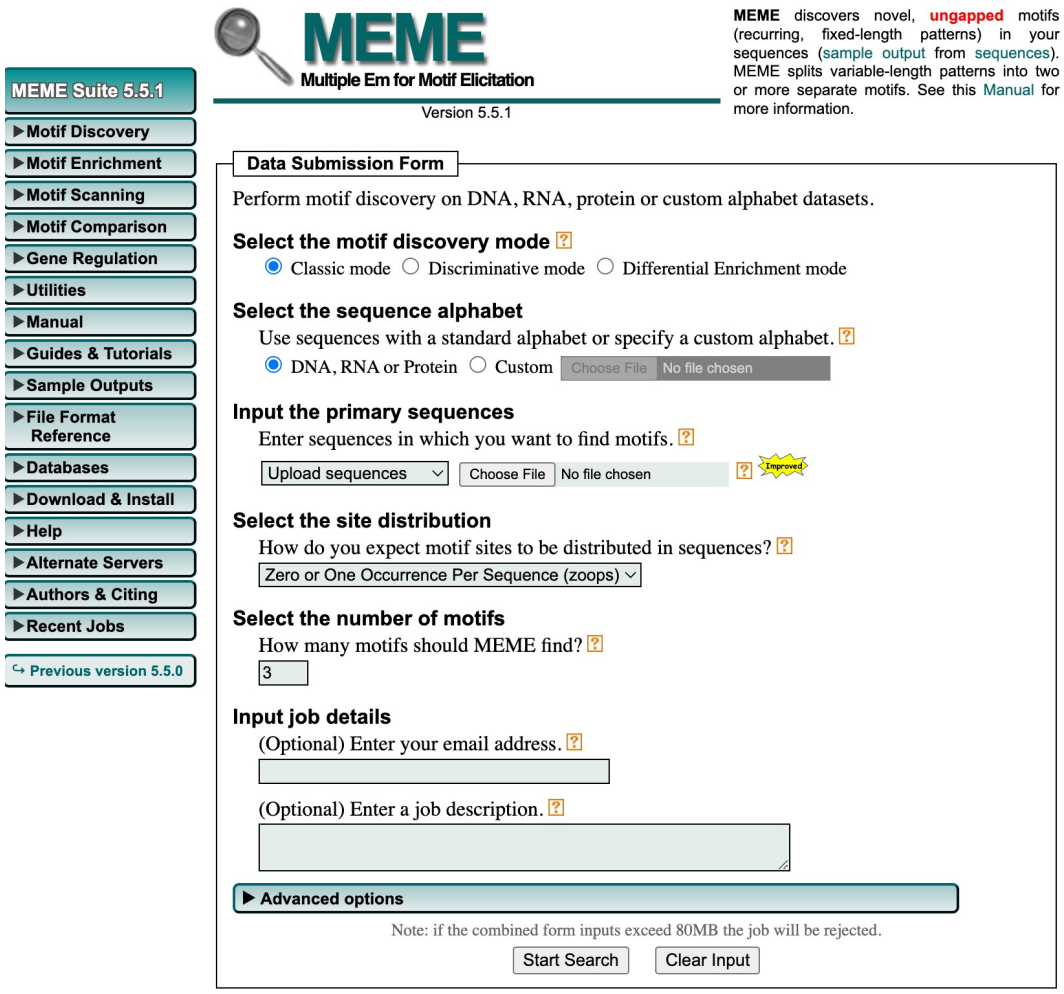
– 根据序列寻找显著富集的motif

– 根据已有motif进行搜库

Motif分析-常用软件

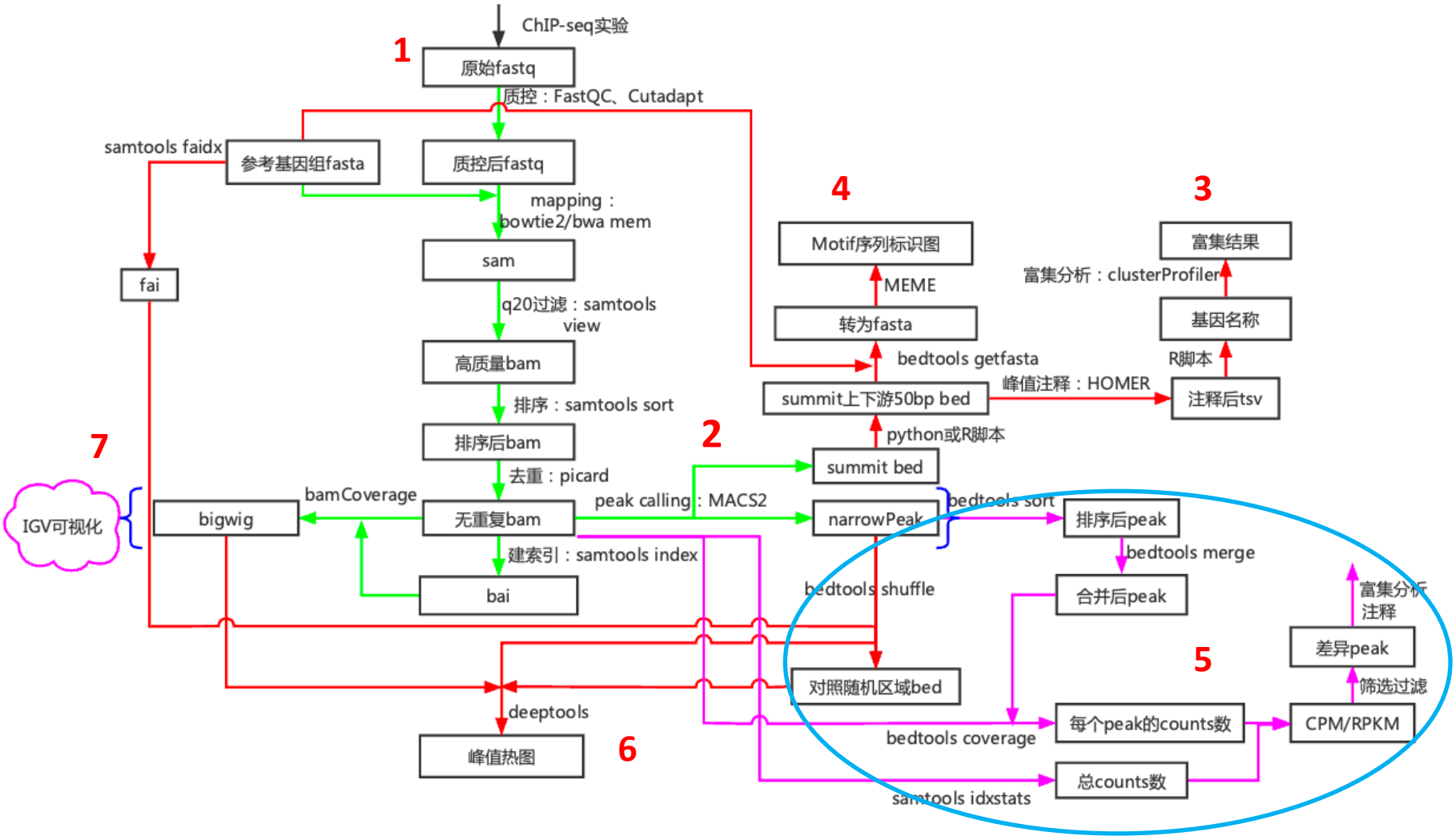


MEME discovers novel motifs



https://meme-suite.org/meme/tools/meme

ChIP-seq数据详细分析流程



差异峰值分析

■ 分析思路

– 直接计算RPKM，计算fold change直接进行筛选

– 使用类似DESeq2,edgeR的工具进行差异分析

- 可以提供一个所谓的统计检验值
- 潜在的基本假设可能并不相符

差异峰值分析

■分析流程

– 每个个样品单独call peak

– 合并样品间的peak

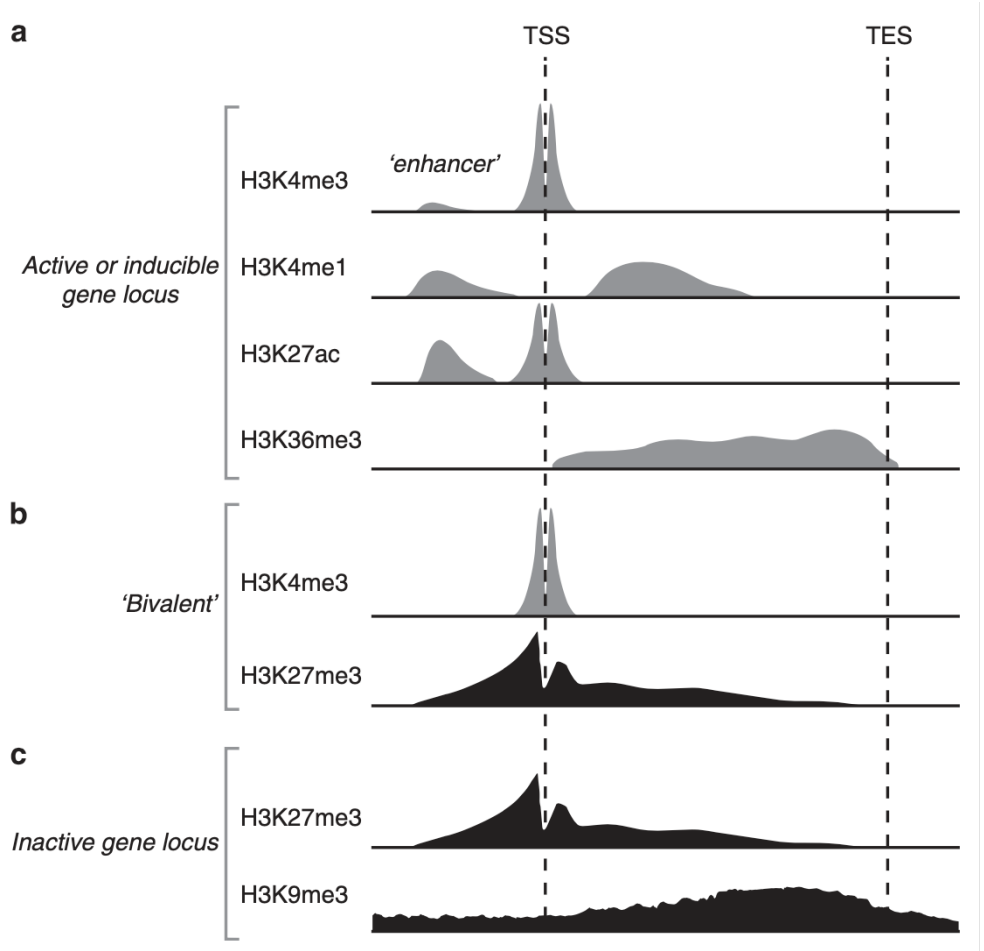
– 计算每个peak的reads count

– 计算RPKM和fold change

– 使用DESeq2或edgeR计算p value

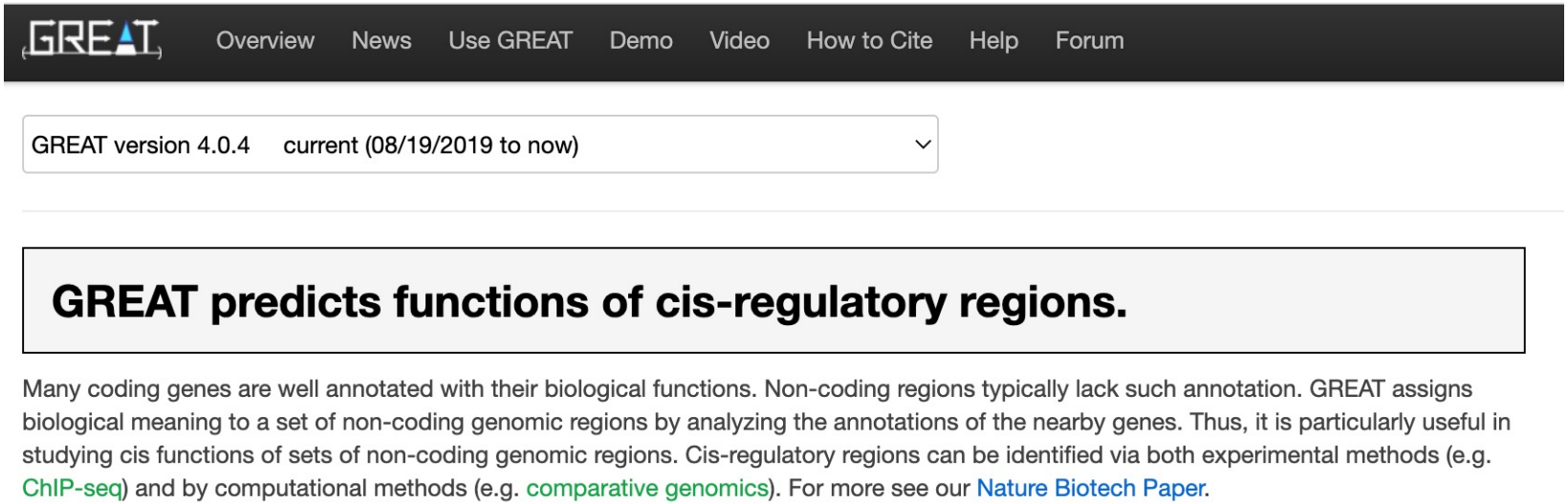
差异峰值分析

- 常用的broad peak
  - H3K27me3
  - H3K36me3
  - H3K9me3



功能富集分析

- 直接在网站里输入bed文件
- 转录因子结合位置的位置信息，只需要三列
  - **cut -f 1-3 D4_CpG_STAT3_ChIP_peaks.narrowPeak > D4_CpG_STAT3_ChIP_peaks.bed**



http://great.stanford.edu/public/html/

功能富集分析

■结果展示

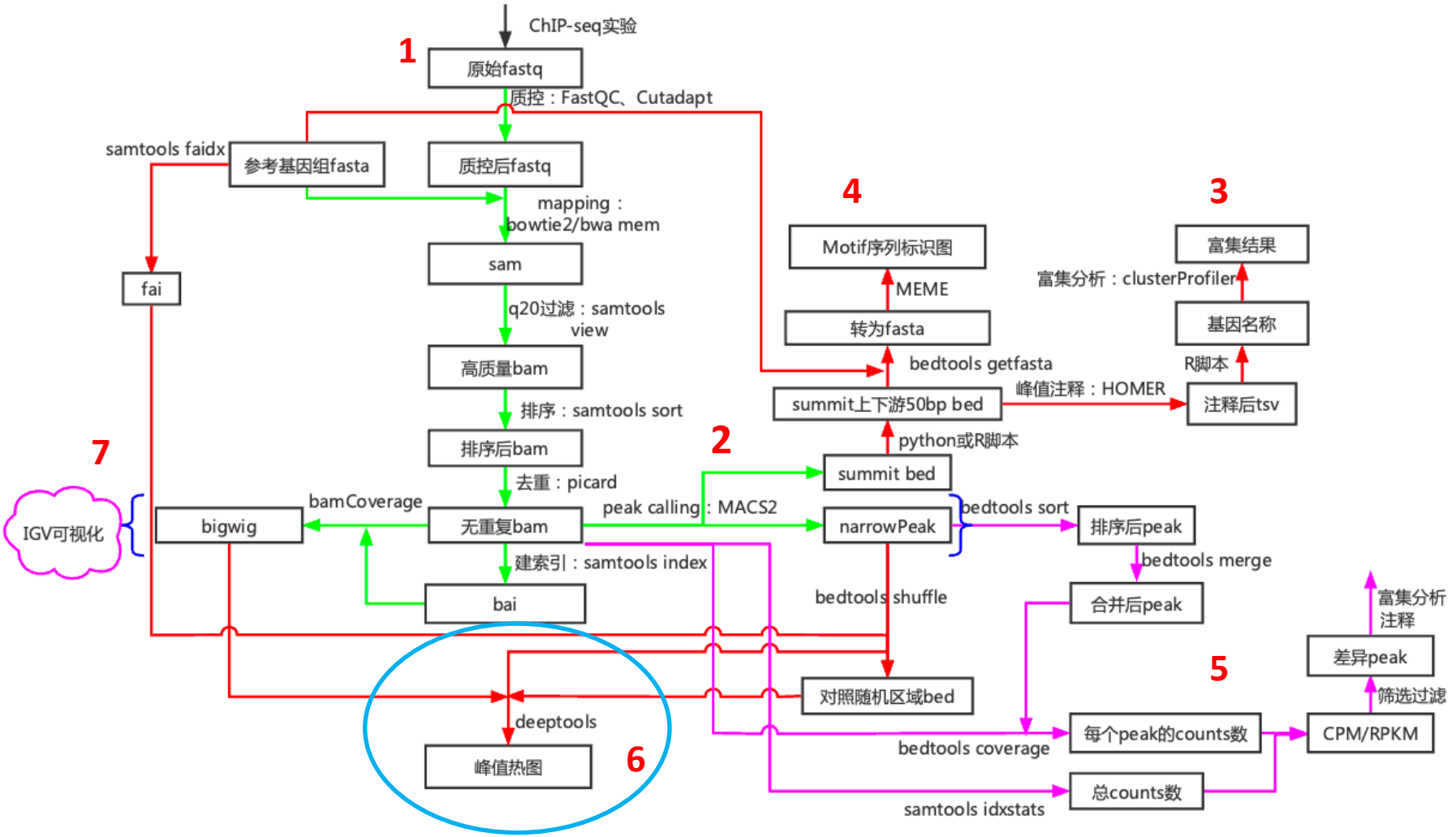
Term Name	Binom Rank	Binom Raw P-Value	*Binom FDR Q-Val*	Binom Fold Enrichment	Binom Observed Region Hits	Binom Region Set Coverage	Hyper Rank	*Hyper FDR Q-Val*	Hyper Fold Enrichment
immune response	2	**3.0822e-31**	**2.0258e-27**	2.0369	288	20.59%	37	**1.2661e-8**	1.5416
regulation of immune response	4	**2.8068e-27**	**9.2238e-24**	2.2078	215	15.37%	13	**3.3294e-10**	1.7980
leukocyte activation	5	**2.2363e-25**	**5.8793e-22**	2.1418	213	15.23%	42	**2.4771e-8**	1.7240
defense response	7	**2.0615e-23**	**3.8712e-20**	2.0206	223	15.94%	153	**3.3361e-5**	1.4659
cytokine-mediated signaling pathway	8	**5.6389e-23**	**9.2654e-20**	2.5610	138	9.86%	97	**2.8187e-6**	1.8000
lymphocyte activation	9	**6.0443e-23**	**8.8281e-20**	2.6156	133	9.51%	22	**1.6666e-9**	2.3215
immune effector process	10	**1.6670e-22**	**2.1913e-19**	2.1431	189	13.51%	128	**1.4070e-5**	1.5616
response to other organism	11	**7.9781e-21**	**9.5339e-18**	2.1662	171	12.22%	143	**2.0092e-5**	1.6020
response to external biotic stimulus	13	**1.7331e-20**	**1.7525e-17**	2.1491	171	12.22%	146	**2.4130e-5**	1.5958
regulation of innate immune response	16	**1.5462e-19**	**1.2703e-16**	2.6235	112	8.01%	77	**9.1992e-7**	1.9765

功能富集分析

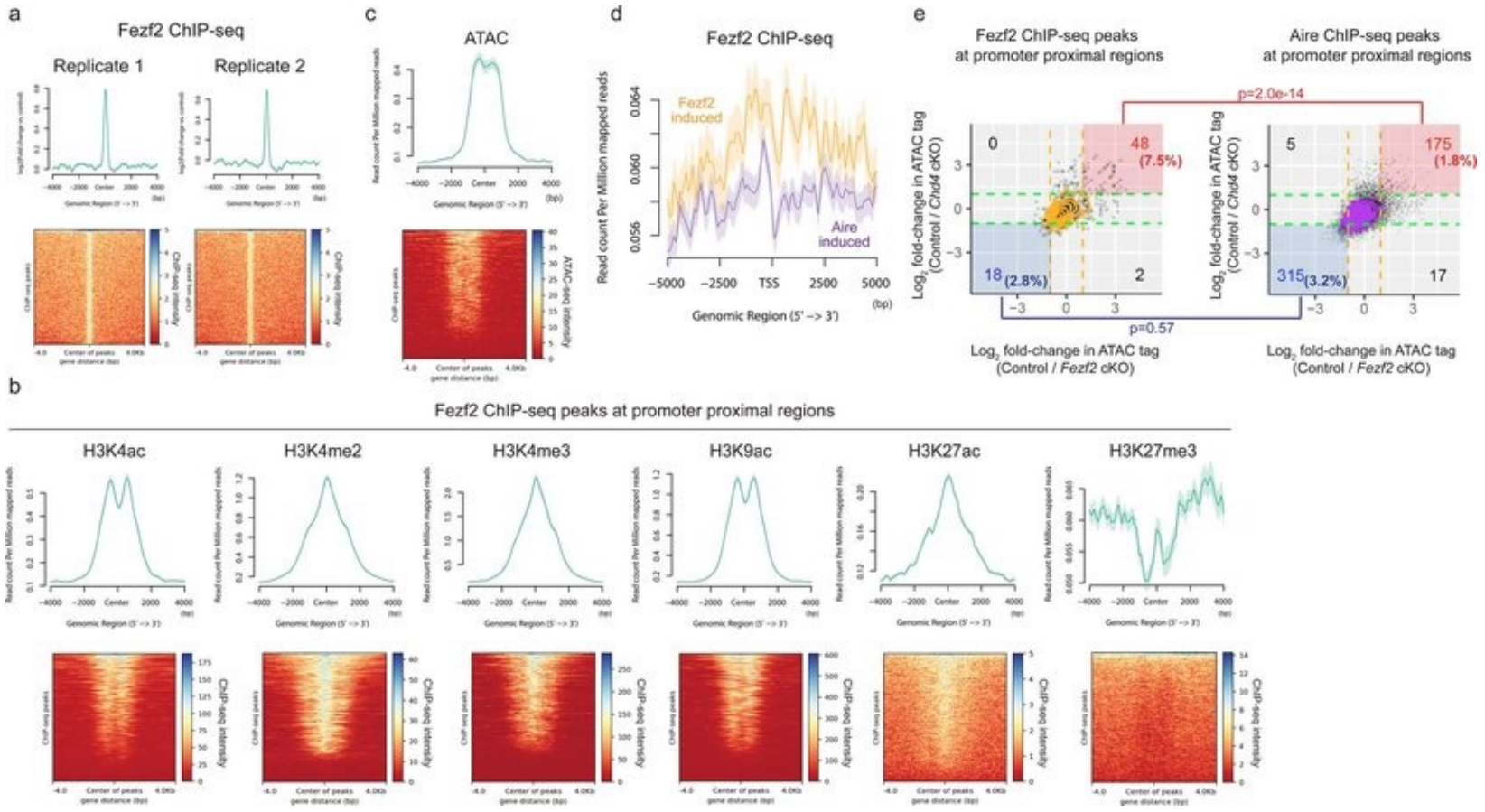
- 找到转录因子结合位点所对应的基因名字，使用DAVID网站
  - https://david.ncifcrf.gov/tools.jsp
- 将基因名字输入David网站结果

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☐	GOTERM_BP_DIRECT	positive regulation of transcription from RNA polymerase II promoter	RT		23	13.2	9.4E-6	1.1E-2
☐	GOTERM_BP_DIRECT	JAK-STAT cascade	RT		5	2.9	1.7E-4	7.5E-2
☐	GOTERM_BP_DIRECT	regulation of transcription from RNA polymerase II promoter	RT		25	14.4	2.5E-4	7.5E-2
☐	GOTERM_BP_DIRECT	negative regulation of ERK1 and ERK2 cascade	RT		6	3.4	2.7E-4	7.5E-2
☐	GOTERM_BP_DIRECT	regulation of cell proliferation	RT		7	4.0	6.9E-4	1.4E-1
☐	GOTERM_BP_DIRECT	positive regulation of transcription, DNA-templated	RT		14	8.0	8.3E-4	1.4E-1
☐	GOTERM_BP_DIRECT	T-helper 2 cell differentiation	RT		3	1.7	8.5E-4	1.4E-1
☐	GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter	RT		16	9.2	1.5E-3	1.8E-1
☐	GOTERM_BP_DIRECT	defense response	RT		5	2.9	1.7E-3	1.8E-1
☐	GOTERM_BP_DIRECT	germinal center formation	RT		3	1.7	1.8E-3	1.8E-1
☐	GOTERM_BP_DIRECT	JAK-STAT cascade involved in growth hormone signaling pathway	RT		3	1.7	1.8E-3	1.8E-1
☐	GOTERM_BP_DIRECT	cellular response to tumor necrosis factor	RT		6	3.4	2.6E-3	2.4E-1
☐	GOTERM_BP_DIRECT	response to drug	RT		8	4.6	3.0E-3	2.6E-1
☐	GOTERM_BP_DIRECT	cytokine-mediated signaling pathway	RT		6	3.4	3.5E-3	2.8E-1
☐	GOTERM_BP_DIRECT	transforming growth factor beta receptor signaling pathway	RT		5	2.9	4.2E-3	3.1E-1
☐	GOTERM_BP_DIRECT	cellular response to leukemia inhibitory factor	RT		5	2.9	4.5E-3	3.2E-1
☐	GOTERM_BP_DIRECT	response to peptide hormone	RT		4	2.3	5.7E-3	3.7E-1
☐	GOTERM_BP_DIRECT	positive regulation of inflammatory response	RT		5	2.9	5.9E-3	3.7E-1
☐	GOTERM_BP_DIRECT	cellular response to interleukin-6	RT		3	1.7	8.8E-3	5.2E-1
☐	GOTERM_BP_DIRECT	follicular dendritic cell differentiation	RT		2	1.1	1.3E-2	7.2E-1
☐	GOTERM_BP_DIRECT	positive regulation of interferon-gamma production	RT		4	2.3	1.5E-2	7.5E-1
☐	GOTERM_BP_DIRECT	negative regulation of transcription, DNA-templated	RT		10	5.7	1.5E-2	7.5E-1
☐	GOTERM_BP_DIRECT	regulation of cell shape	RT		5	2.9	1.8E-2	7.7E-1

ChIP-seq数据详细分析流程



峰值可视化-Deeptools



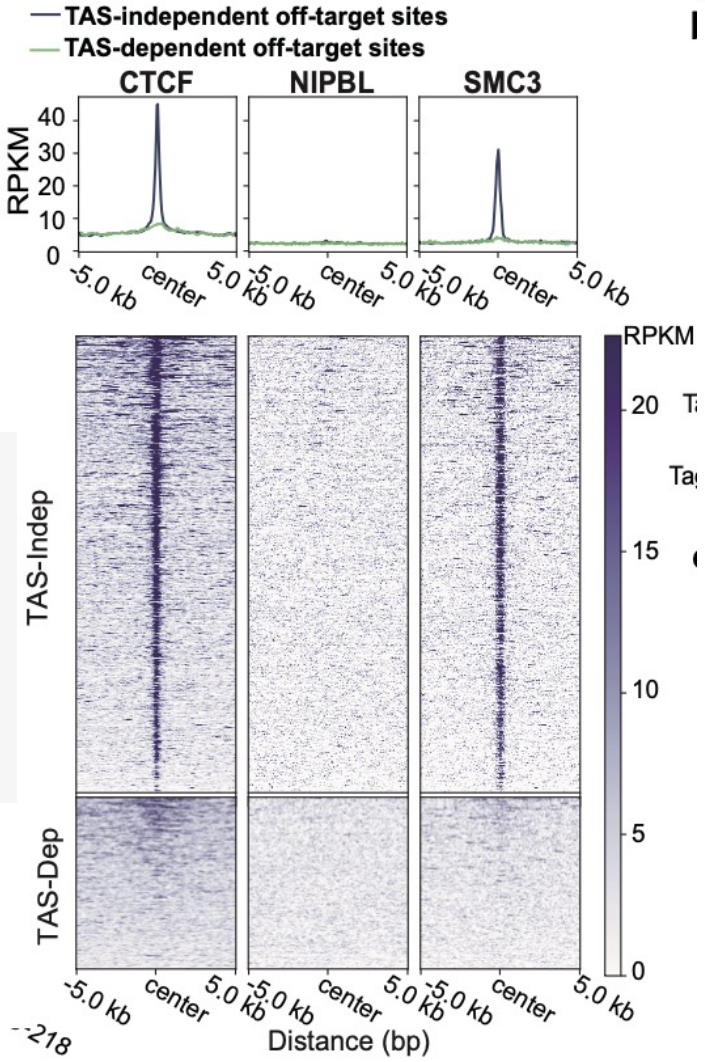
https://www.nature.com/articles/s41590-020-0717-2

峰值可视化-Deeptools

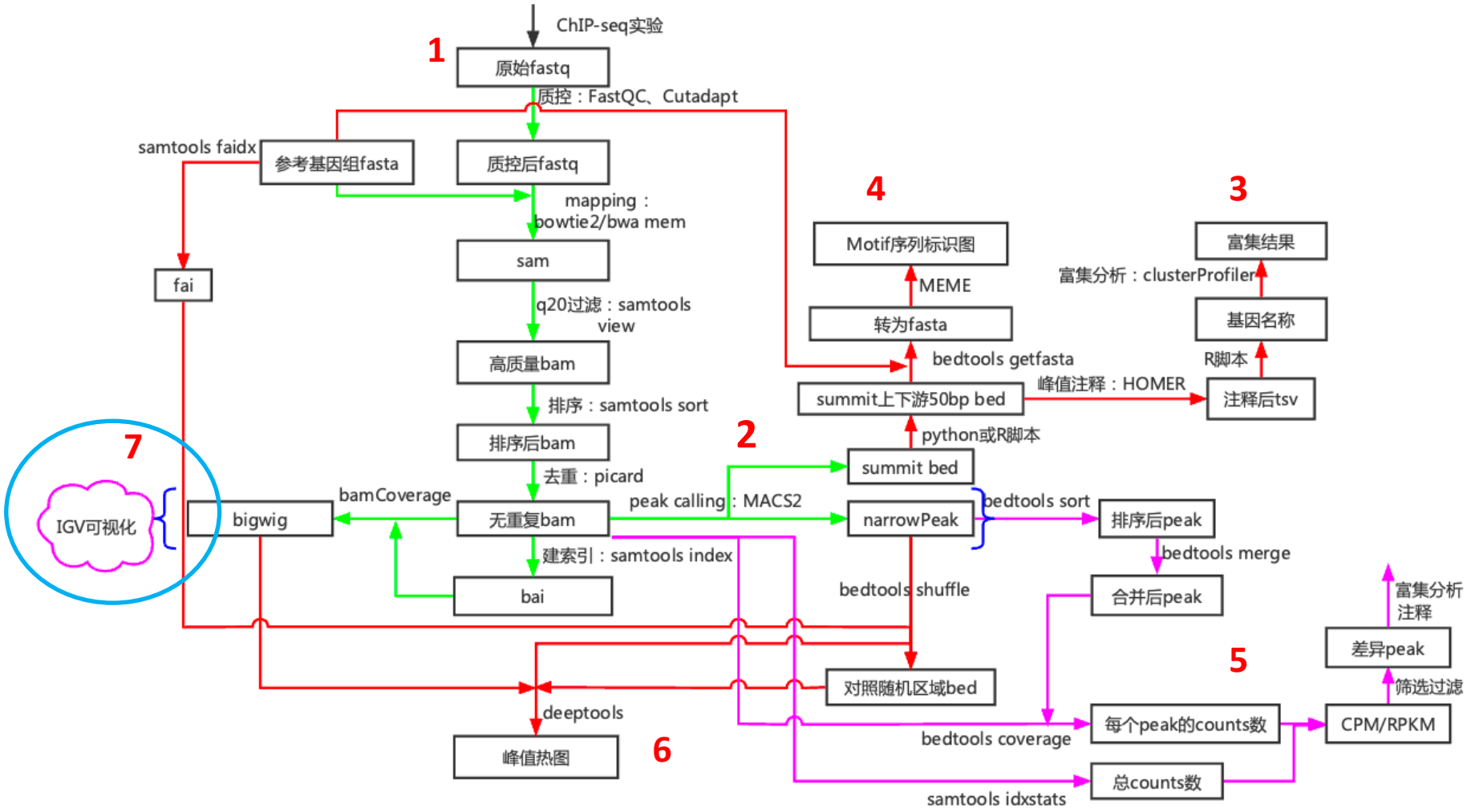
■热图

- --kmeans 聚类的数量
- --colorList 修改颜色
- --whatToShow 修改显示的图形

```
plotHeatmap -m matrix_two_groups.gz \
   -out ExampleHeatmap2.png \
   --colorMap RdBu \
   --whatToShow 'heatmap and colorbar' \
   --zMin -3 --zMax 3 \
   --kmeans 4
```



ChIP-seq数据详细分析流程

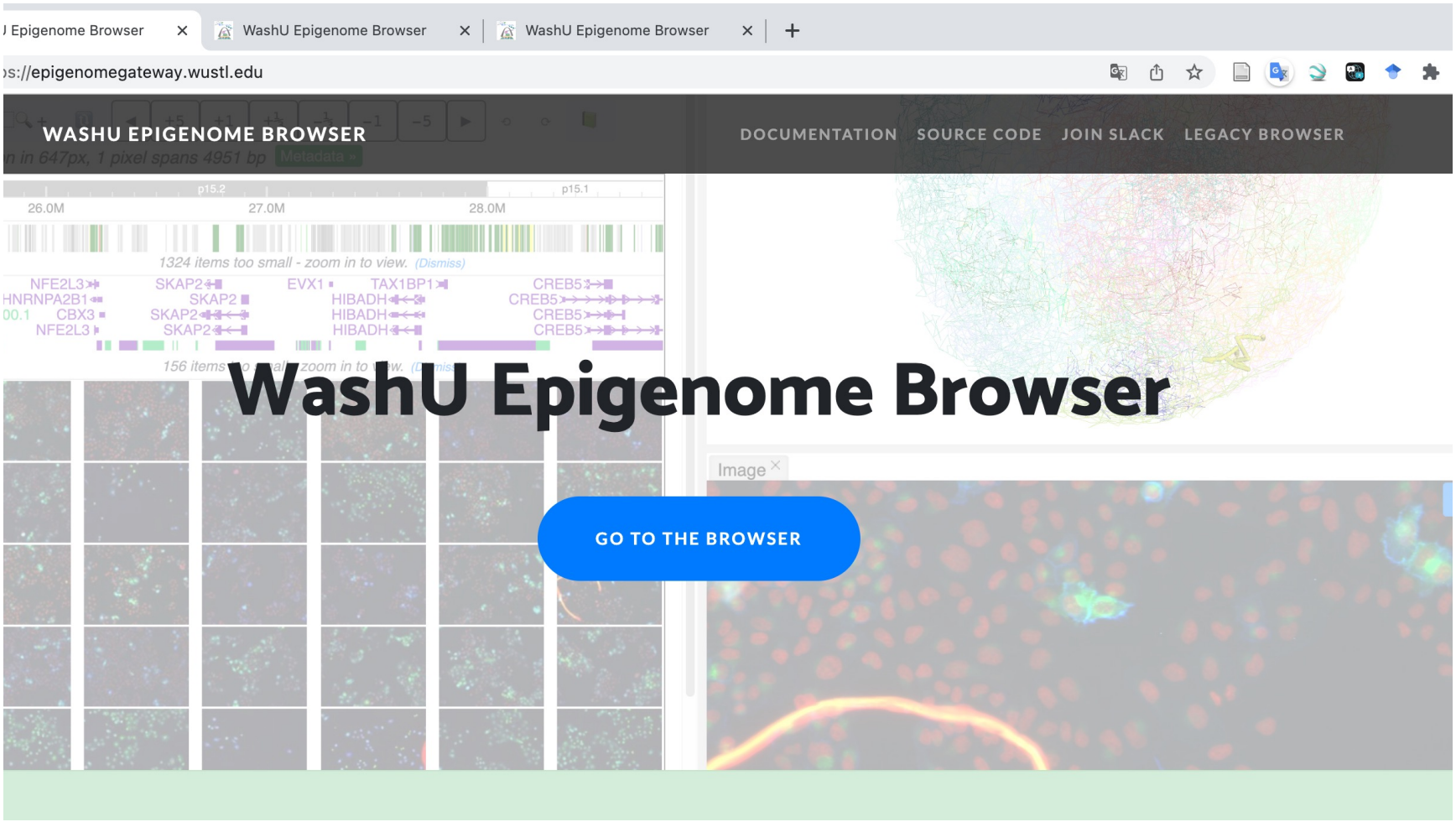


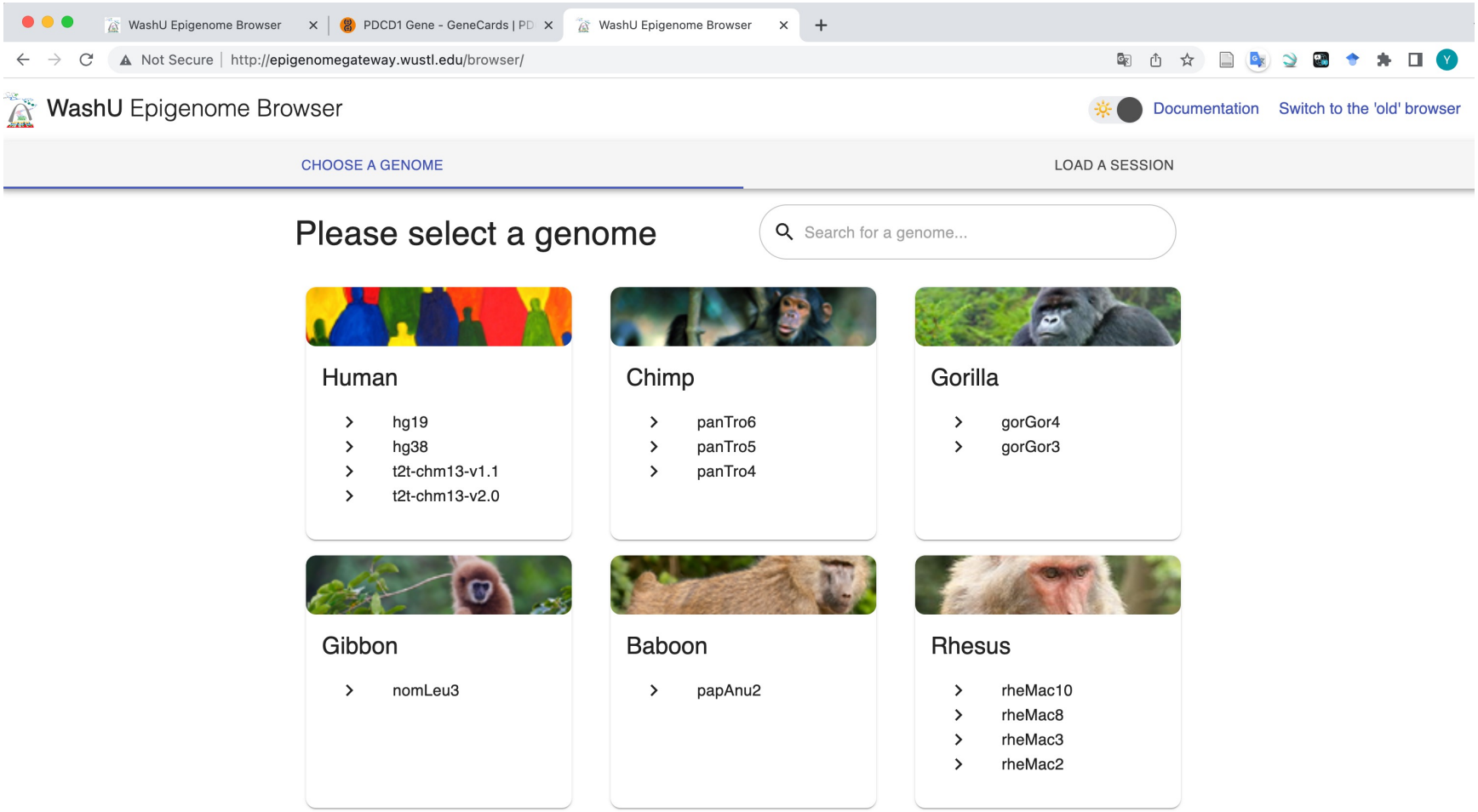
峰值可视化-IGV的使用

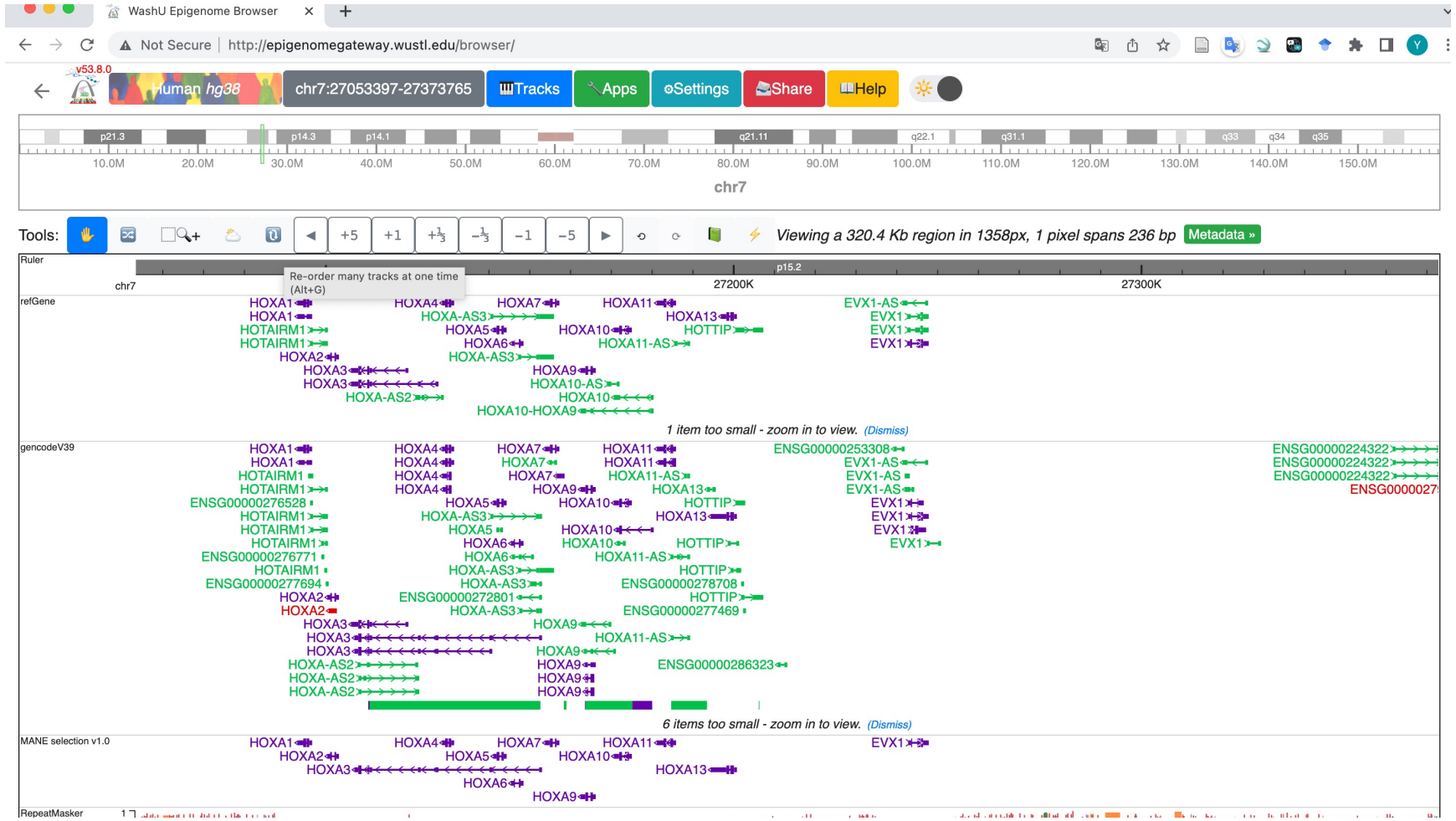
- 下载并安装IGV

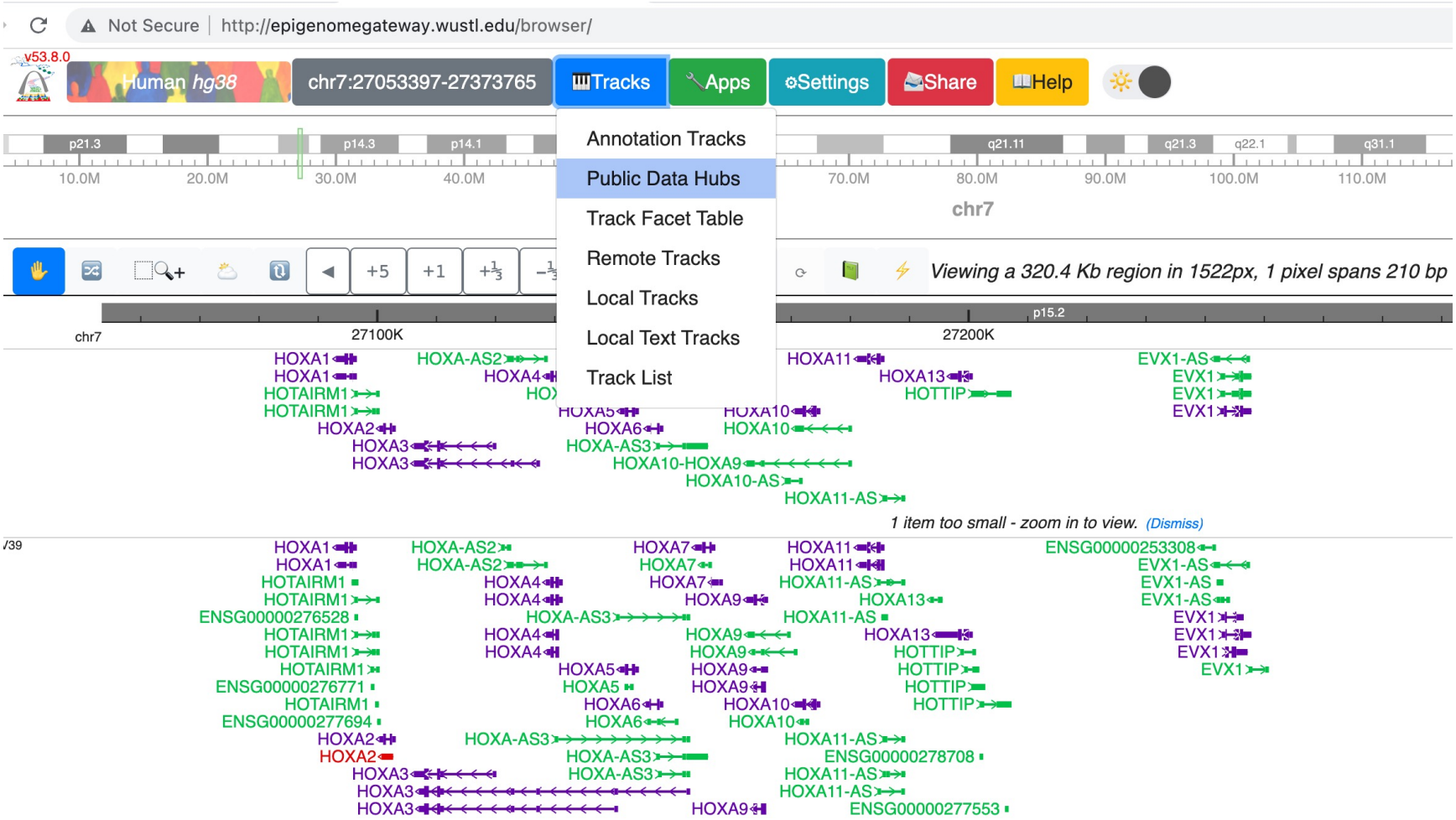
  https://software.broadinstitute.org/software/igv/download

- 加载BAM文件，并以不同的排列方式进行展示
  - 不同颜色的含义
  - Reads pair的表示
  - Reads alignment细节的探索
  - 不同排序方式
- 载入BED文件，学习如何使用注释文件
- IGV的截图
- IGV的命令行截图


Epigenome Browser
WashU Epigenome Browser
WashU Epigenome Browser
://epigenomegateway.wustl.edu
WASHU EPIGENOME BROWSER
DOCUMENTATION
SOURCE CODE
JOIN SLACK
LEGACY BROWSER
26.0M
27.0M
28.0M
1324 items too small - zoom in to view.
NFE2L3
SKAP2
EVX1
TAX1BP1
CREB5
HNRNPA2B1
CBX3
HIBADH
WashU Epigenome Browser
GO TO THE BROWSER
Image


WashU Epigenome Browser
PDCD1 Gene - GeneCards | PD
WashU Epigenome Browser
Not Secure | http://epigenomegateway.wustl.edu/browser/
WashU Epigenome Browser
Documentation
Switch to the 'old' browser
CHOOSE A GENOME
LOAD A SESSION
Please select a genome
Search for a genome...
Human
hg19
hg38
t2t-chm13-v1.1
t2t-chm13-v2.0
Chimp
panTro6
panTro5
panTro4
Gorilla
gorGor4
gorGor3
Gibbon
nomLeu3
Baboon
papAnu2
Rhesus
rheMac10
rheMac8
rheMac3
rheMac2


WashU Epigenome Browser
Not Secure | http://epigenomegateway.wustl.edu/browser/
v53.8.0
Human hg38
chr7:27053397-27373765
Tracks
Apps
Settings
Share
Help
chr7
Tools:
Viewing a 320.4 Kb region in 1358px, 1 pixel spans 236 bp
Metadata »
Ruler
Re-order many tracks at one time (Alt+G)
27200K
27300K
refGene
1 item too small - zoom in to view. (Dismiss)
gencodeV39
6 items too small - zoom in to view. (Dismiss)
MANE selection v1.0
RepeatMasker


Not Secure | http://epigenomegateway.wustl.edu/browser/
v53.8.0
Human hg38
chr7:27053397-27373765
Tracks
Apps
Settings
Share
Help
Annotation Tracks
Public Data Hubs
Track Facet Table
Remote Tracks
Local Tracks
Local Text Tracks
Track List
p21.3
p14.3
p14.1
q21.11
q21.3
q22.1
q31.1
10.0M
20.0M
30.0M
40.0M
70.0M
80.0M
90.0M
100.0M
110.0M
chr7
Viewing a 320.4 Kb region in 1522px, 1 pixel spans 210 bp
p15.2
chr7
27100K
27200K
HOXA1
HOXA-AS2
HOXA11
EVX1-AS
HOTAIRM1
HOXA4
HOXA13
EVX1
HOTTIP
HOXA2
HOXA6
HOXA10
HOXA3
HOXA-AS3
HOXA10-HOXA9
HOXA10-AS
HOXA11-AS
1 item too small - zoom in to view. (Dismiss)
HOXA7
ENSG00000253308
ENSG00000276528
HOXA-AS3
HOXA11-AS
ENSG00000276771
HOXA5
HOXA9
ENSG00000277694
HOXA2
ENSG00000278708
ENSG00000277553

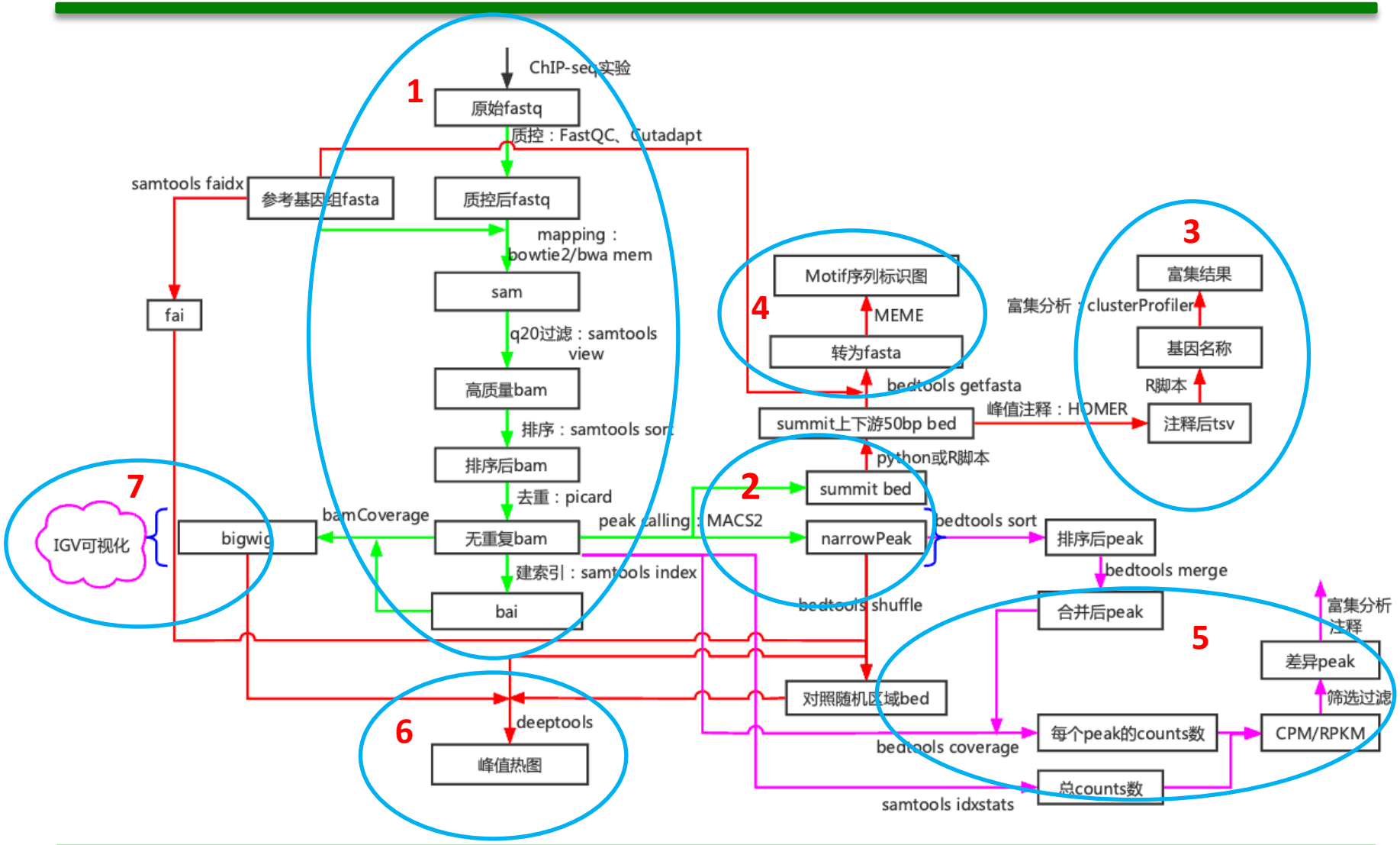
Public data hubs

	Genome	Collection	Hub name	Tracks	Add
▸	hg38	Reference human epigenomes from Roadmap Epigenomics Consortium	All Chromatin states tracks	352	+
▸	hg38	Reference human epigenomes from Roadmap Epigenomics Consortium	Roadmap ChIP-seq datasets	12494	+
▸	hg38	Reference human epigenomes from Roadmap Epigenomics Consortium	Roadmap RNA-seq, WGBS etc. datasets	5586	+
▸	hg38	Image collection	IDR image data	28	+
▸	hg38	Image collection	4dn image data	1	+
▸	hg38	Encyclopedia of DNA Elements (ENCODE)	ENCODE signal of unique reads	5230	+
▸	hg38	Encyclopedia of DNA Elements (ENCODE)	ENCODE signal of all reads	5230	+
▸	hg38	4D Nucleome Network	4DN datasets	2876	+
▸	hg38	Encyclopedia of DNA Elements (ENCODE)	Human ENCODE from ENCODE data portal	38092	+
▸	hg38	Encyclopedia of DNA Elements (ENCODE)	Human ENCODE HiC from ENCODE data portal	20	+

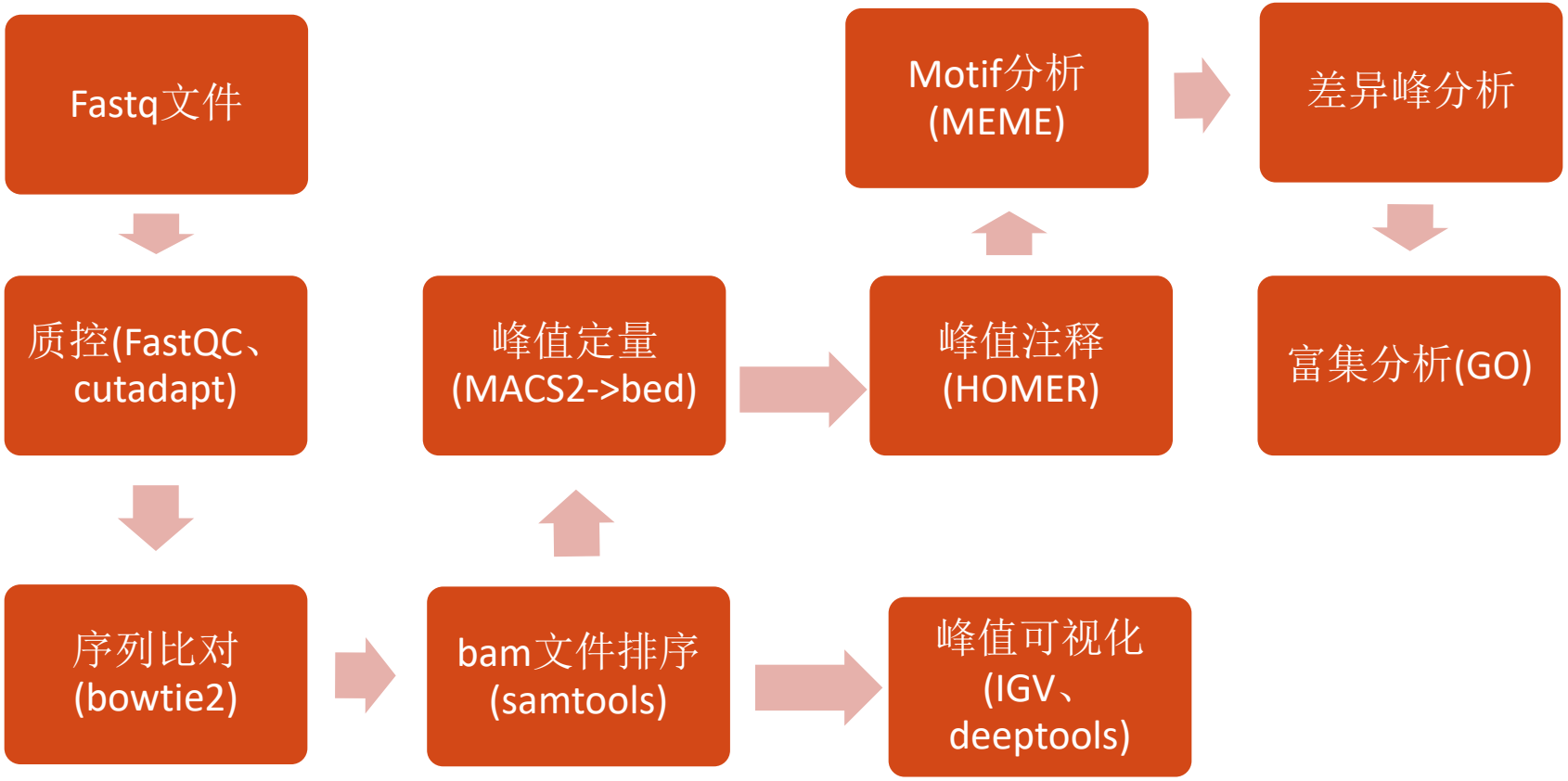
Previous Page 1 of 2 10 rows Next

No tracks from data hubs yet. Load a hub first.

ChIP-seq数据详细分析流程



Thanks for your attention !


Fastq文件
质控(FastQC、cutadapt)
序列比对(bowtie2)
bam文件排序(samtools)
峰值定量(MACS2->bed)
峰值注释(HOMER)
Motif分析(MEME)
差异峰分析
富集分析(GO)
峰值可视化(IGV、deeptools)