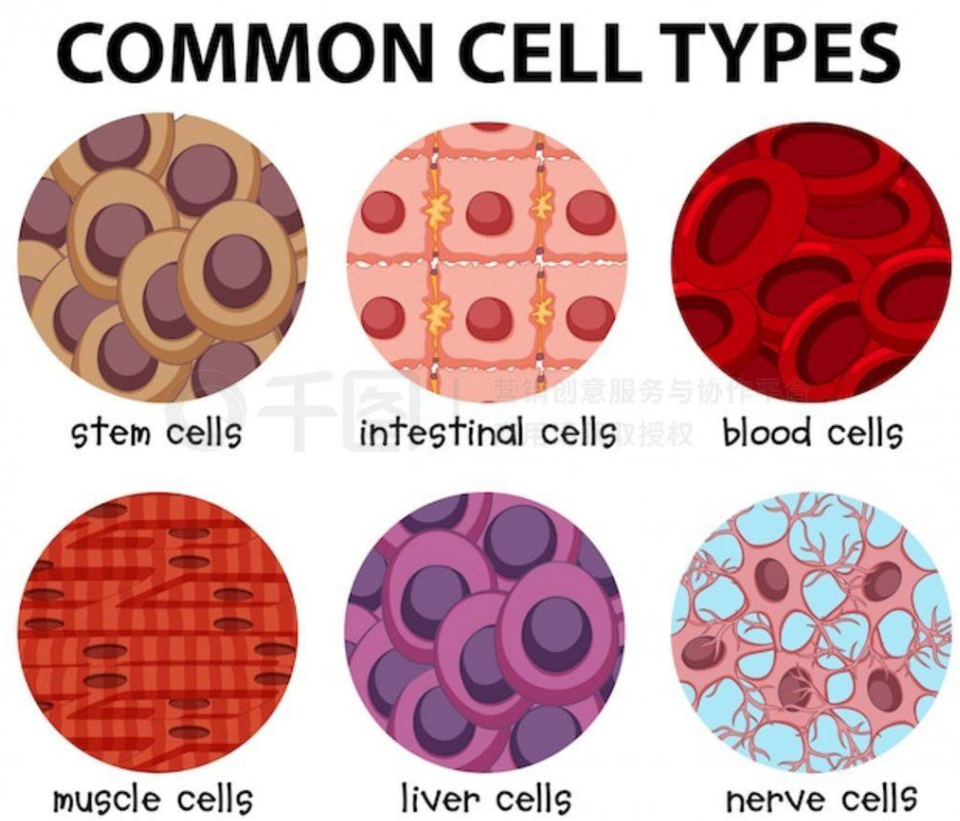
生物信息学

转录调控

Yu Hou (侯 宇) Zhejiang University 2025

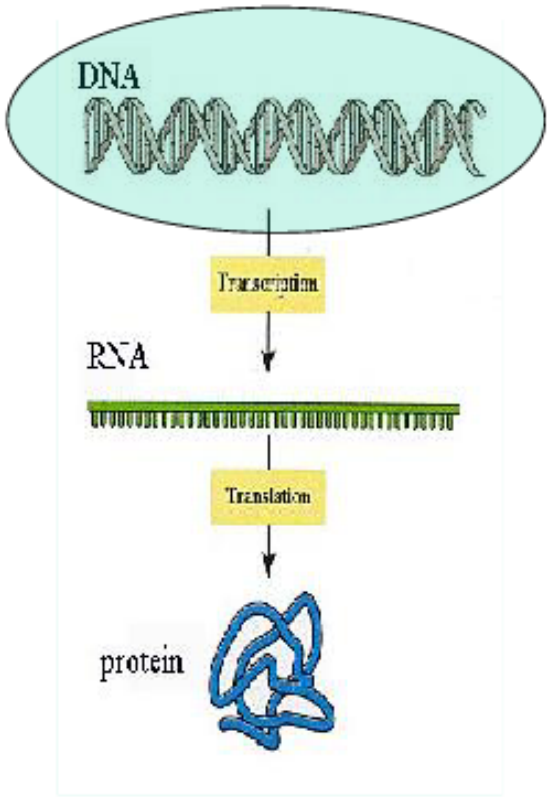
基因表达调控



细胞类型	特异表达基因（
① 干细胞（Stem cells）	OCT4 (POU5F1)、SOX2、NANOG、KLF4、LIN28A
② 小肠上皮细胞（Intestinal epithelial cells）	LGR5、VIL1 (Villin)、MUC2、CDX2、ALPI (碱性磷酸酶)
③ 血液细胞（以造血细胞/白细胞为例）	CD45 (PTPRC)、CD3E（T细胞）、CD19（B细胞）、MPO（髓系）、GATA1（红系）
④ 骨骼肌细胞（Skeletal muscle cells）	MYOD1、MYOG (Myogenin)、ACTA1 (肌动蛋白α1)、TNNT3 (肌钙蛋白T3)、CKM (肌酸激酶M型)
⑤ 肝脏细胞（Hepatocytes）	ALB (白蛋白)、CYP3A4、TAT (酪氨酸氨基转移酶)、AFP (甲胎蛋白)、FGA (纤维蛋白原α链)
⑥ 神经元（Neurons）	MAP2、NEUN (RBFOX3)、SYN1 (Synapsin I)、TUBB3 (βIII-tubulin)、GRIN1 (NMDA受体亚基)

为什么同一个基因在不同组织中表达差异巨大?

基因表达调控



转录前调控
- 染色质的活化
- 基因扩增
- 基因重排
- 基因丢失

转录调控
- 顺式作用元件
- 反式作用因子

转录后调控
- mRNA加工成熟
- mRNA的转运

翻译调控
- 蛋白质合成过程

翻译后调控
- 蛋白质的加工成熟

基因表达调控

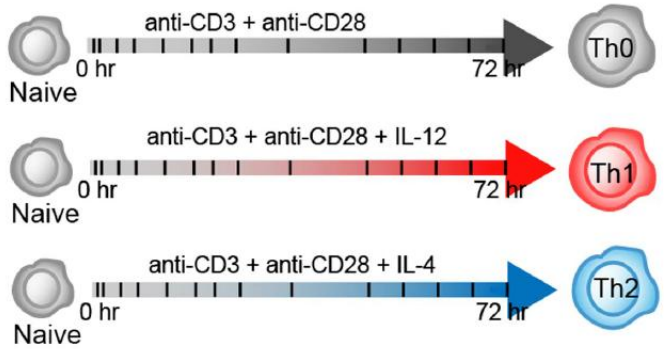
基因表达的特异性

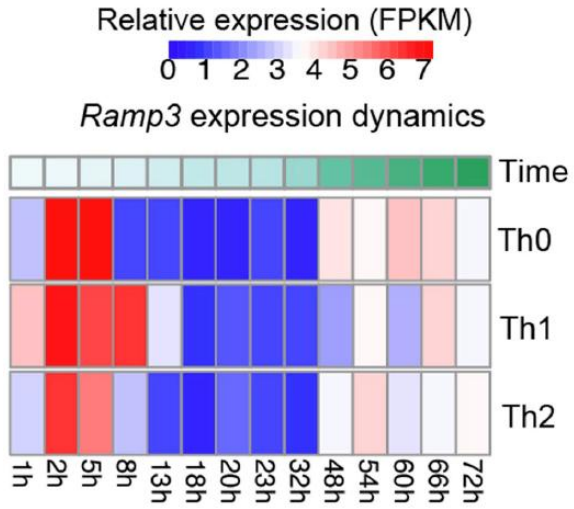
（一）时间特异性

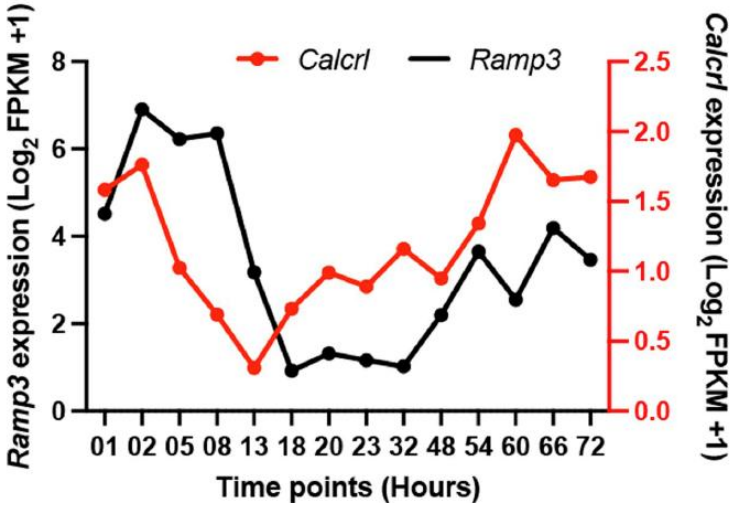
往往与细胞或个体的特定分化、发育阶段相适应，又称阶段特异性。

（二）空间特异性

由细胞在各组织器官的分布差异所决定的，故又称为细胞特异性或组织特异性。







基因表达调控

基因表达的方式

（一）组成性表达

管家基因

在生命全过程都是必需的、且在一个生物个体的几乎所有细胞中持续表达的基因。（GAPDH, ACTB)

组成性基因表达

管家基因较少受环境因素的影响，在个体发育的任一阶段都能在大多数细胞中持续表达。

基因表达调控

基因表达的方式

（二）诱导和阻遏表达

诱导表达

在特定环境信号刺激下，基因表现为开放或增强，表达产物增加。

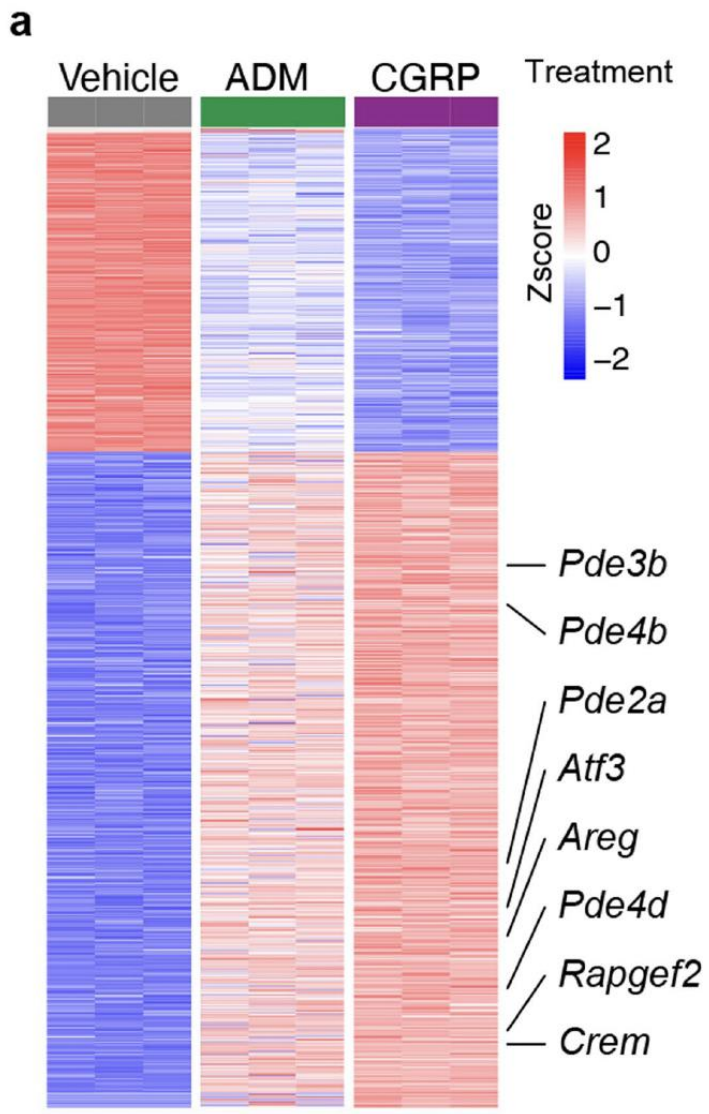
阻遏表达

在特定环境信号刺激下，基因被抑制，从而使表达产物减少。

（三）协调表达

协调表达

在一定机制控制下，功能相关的一组基因，协调一致，共同表达。



基因表达调控

基因表达调控的生物学意义

（一）适应环境变化、维持细胞增殖、分化

（二）维持个体生长、发育

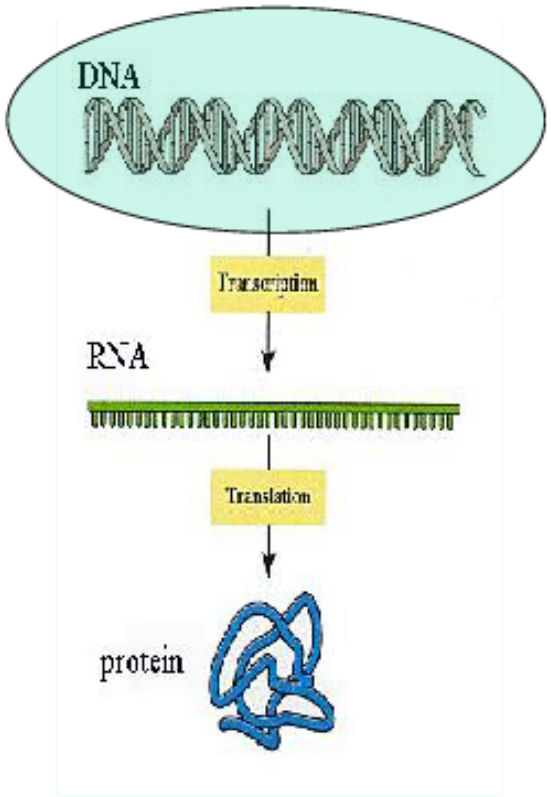
基因转录激活调节基本要素

（一） 特异DNA序列

（二） 调节蛋白

（三） RNA聚合酶

基因表达调控



转录前调控
- 染色质的活化
- 基因扩增
- 基因重排
- 基因丢失

转录调控
- 顺式作用元件
- 反式作用因子

转录后调控
- mRNA加工成熟
- mRNA的转运

翻译调控
- 蛋白质合成过程

翻译后调控
- 蛋白质的加工成熟

基因表达调控

调控层次	主要机制	关键分子或过程	特点与示例
表观遗传调控（Epigenetic Regulation）	调控DNA可及性和染色质结构	DNA甲基化、组蛋白修饰、染色质重塑复合物	决定基因是否“可转录”；如H3K27me3导致基因沉默；胚胎发育、细胞分化关键
转录水平调控（Transcriptional Regulation）	控制RNA合成的起始与速率	转录因子、增强子、抑制子、RNA聚合酶II	决定基因mRNA的合成量，是最核心、最广泛的调控层次
转录后调控（Post-transcriptional Regulation）	影响mRNA的加工与成熟	剪接（splicing）、5’端加帽、3’端加poly(A)尾、RNA编辑	产生不同转录本（可变剪接），影响翻译潜能，如Bcl-xL/Bcl-xS剪接变体
mRNA稳定性与运输调控（mRNA Stability and Localization）	控制mRNA降解速率与定位	RNA结合蛋白（RBP）、miRNA、AU-rich elements	决定mRNA在细胞中的半衰期与空间分布；如miRNA介导mRNA降解
翻译水平调控（Translational Regulation）	控制蛋白质合成速率	起始因子（eIFs）、tRNA、mTOR通路	响应营养和应激状态；如mTOR激活促进翻译、细胞生长
翻译后调控（Post-translational Regulation）	调控蛋白质的修饰、定位与降解	磷酸化、乙酰化、泛素化、SUMO化等	决定蛋白质功能活性与稳定性，如p53的磷酸化激活、蛋白质泛素化降解

基因转录调控

1. 顺式作用元件

原核生物的特异DNA序列（操纵子）

真核生物的特异DNA序列（启动子，终止子，增强子，衰减子，绝缘子，沉默子）

2. 反式作用因子

按功能分类（转录因子，转录调节因子，共调节因子）

按结构域不同分类 （螺旋-转角-螺旋；锌指蛋白；亮氨酸拉链；螺旋-环-螺旋）

01. 基因顺式作用元件

1. 顺式作用元件

原核生物的特异DNA序列（操纵子）

真核生物的特异DNA序列（启动子，终止子，增强子，衰减子，绝缘子，沉默子）

2. 反式作用因子

按功能分类（转录因子，转录调节因子，共调节因子）

按结构域不同分类 （螺旋-转角-螺旋；锌指蛋白；亮氨酸拉链；螺旋-环-螺旋）

01. 基因顺式作用元件

1.1. 原核生物的特异DNA序列

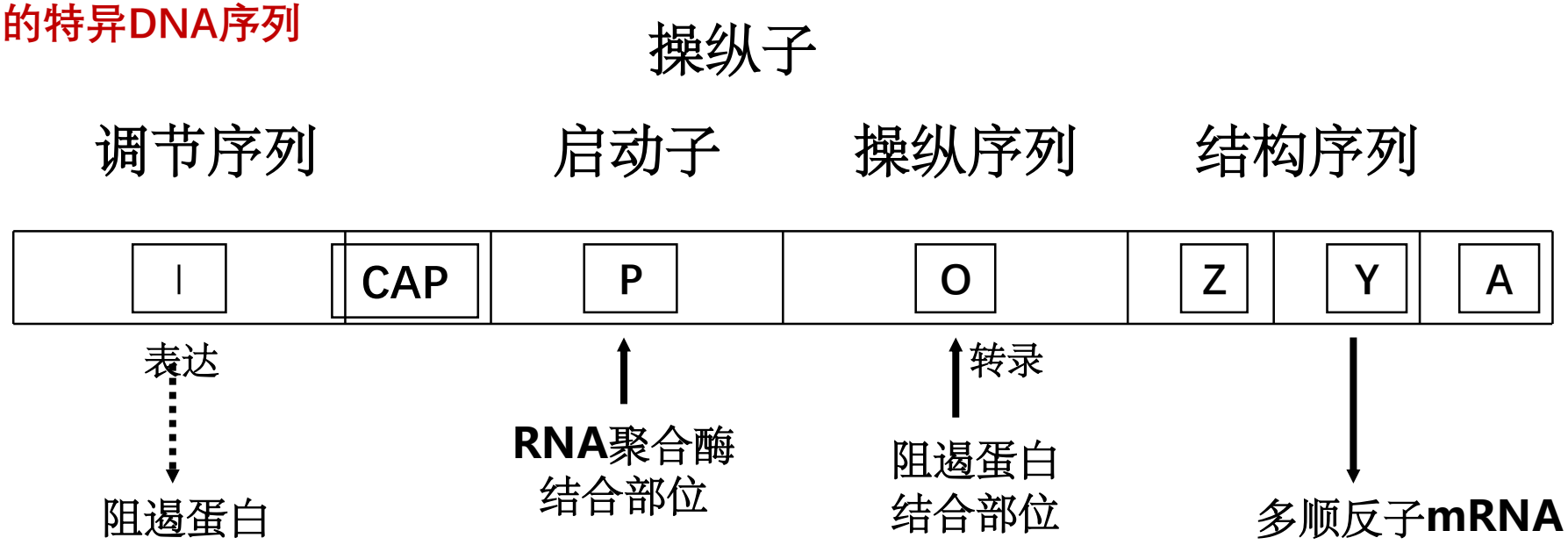
原核生物的基因表达与调控是通过操纵子机制实现的。

操纵子

是由功能上相关联的多个编码序列（2个以上）及其上游的调控序列（包括操纵序列、启动序列和调节序列）等成簇串联在一起，构成的一个转录协调单位。

01. 基因顺式作用元件

1.1. 原核生物的特异DNA序列



编码序列

规定蛋白质结构，又称结构基因；

多顺反子**mRNA**

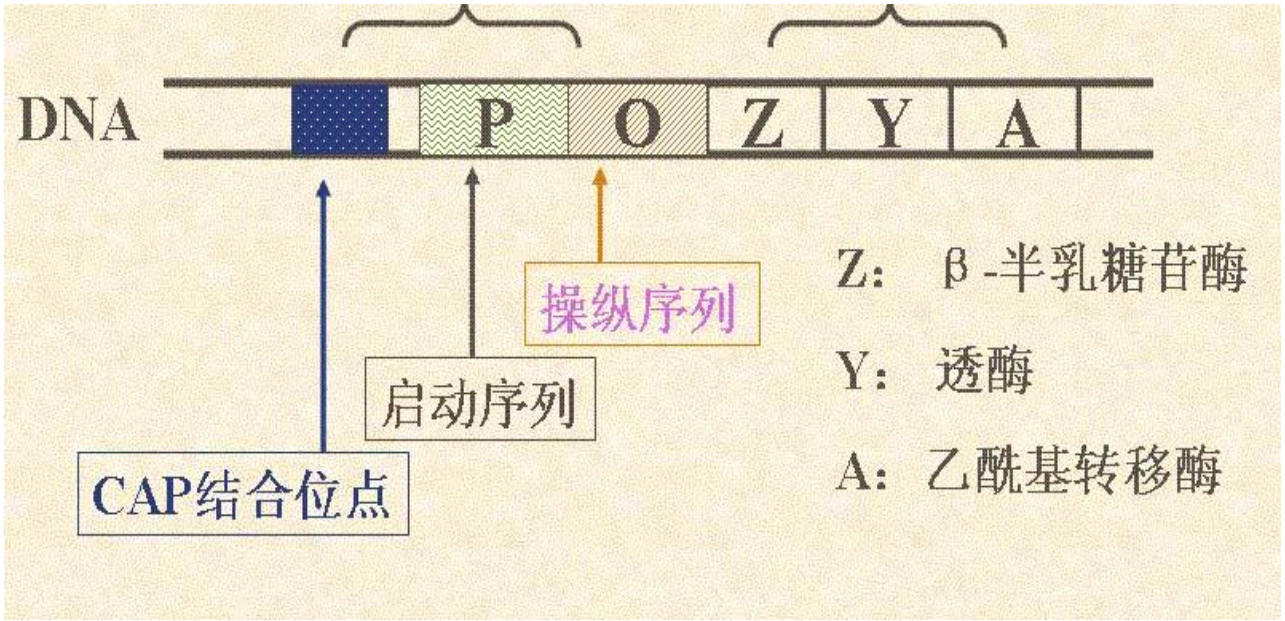
由多个结构基因串联在一起，受同一个启动序列调控，转录生成一个**mRNA,** 翻译生成多个蛋白质,称此为多顺反子**mRNA.**

01. 基因顺式作用元件

1.1. 原核生物的特异DNA序列

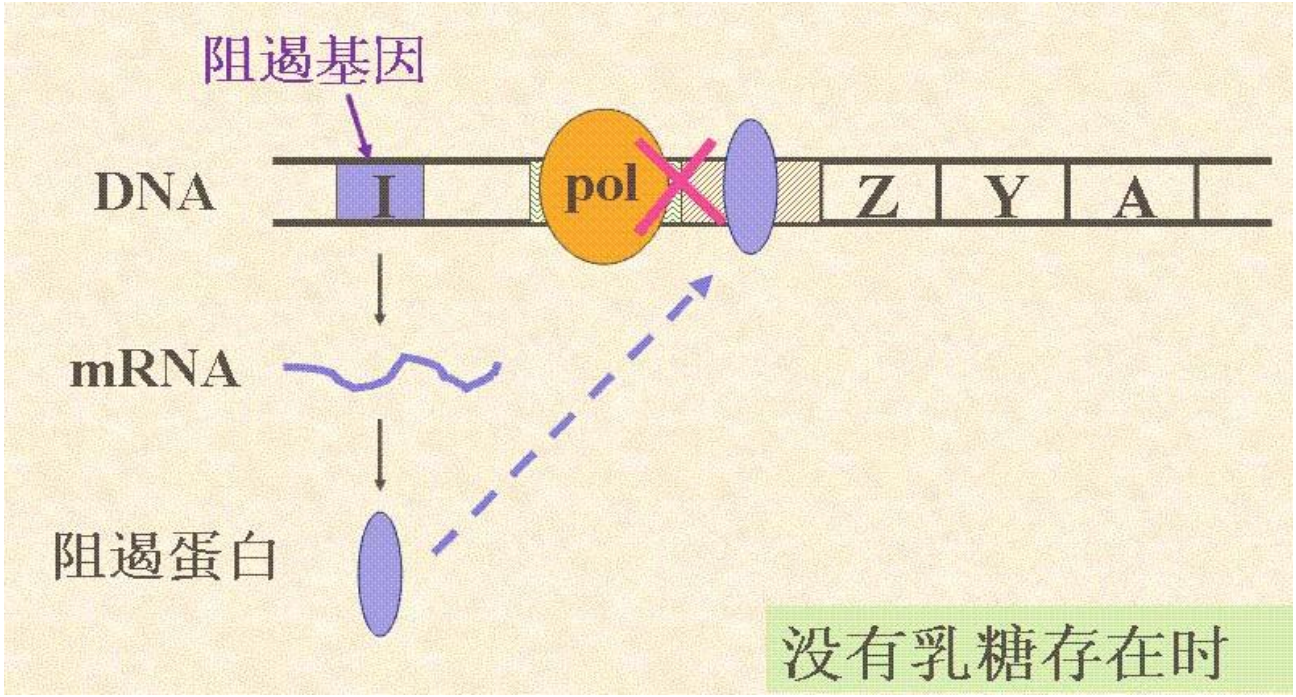
乳糖操纵子

P 启动序列：**RNA**聚合酶的结合位点
O 操纵序列：阻遏物的结合位点，当阻遏物附着在操纵基因上时，结构基因无法转录。
I: 调节序列（阻遏基因）：编码阻遏物



01. 基因顺式作用元件

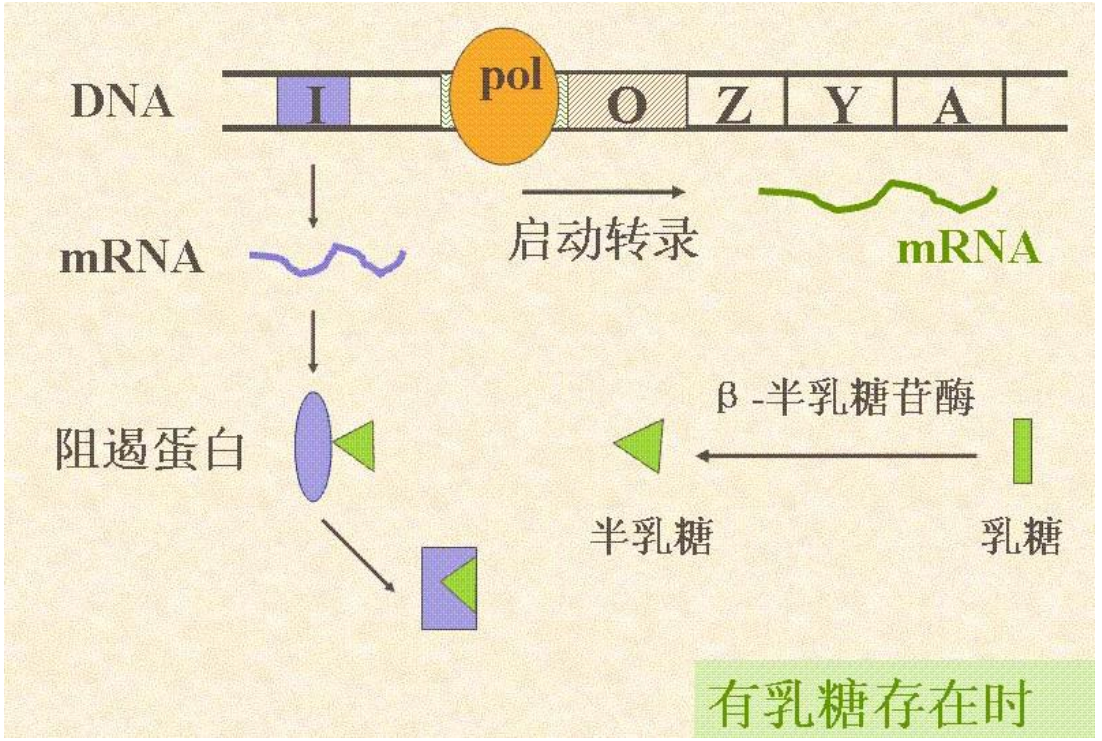
1.1. 原核生物的特异DNA序列



没有乳糖的时候，阻遏基因编码产生的阻遏蛋白结合在操纵基因序列上，使得RNA聚合酶无法对下游的结构基因进行转录

01. 基因顺式作用元件

1.1. 原核生物的特异DNA序列



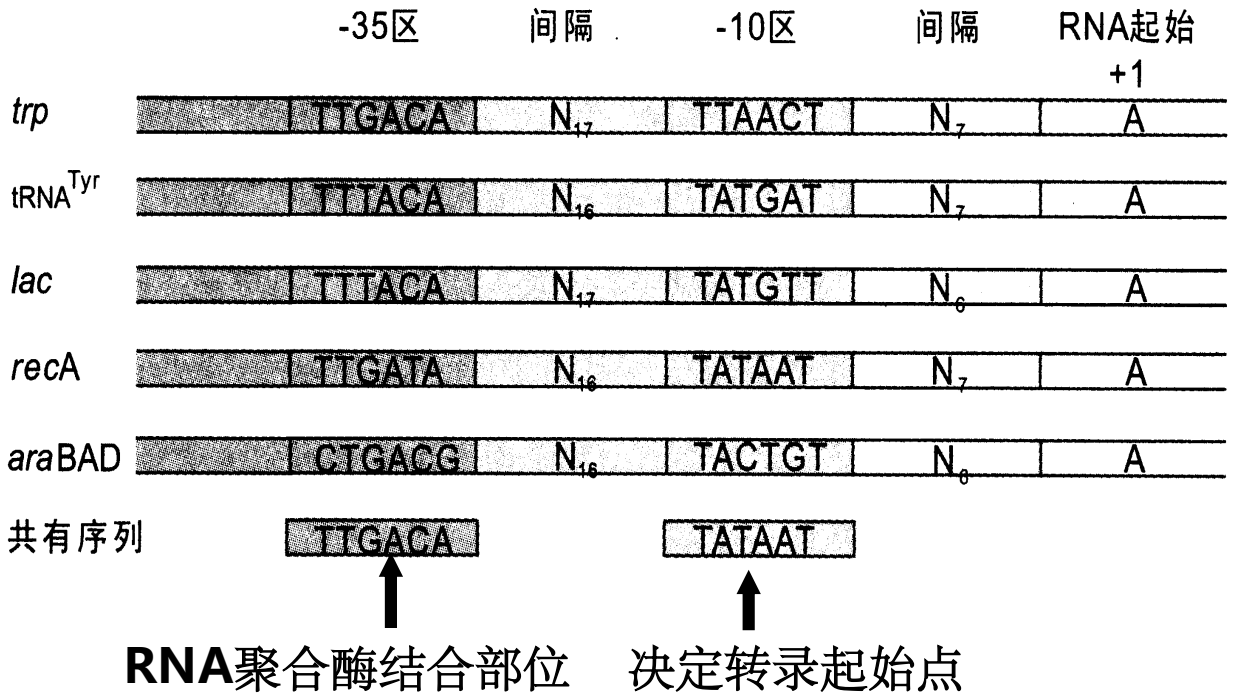
01. 基因顺式作用元件

1.1. 原核生物的特异DNA序列

原核生物的启动序列(RNA聚合酶结合位点)，在距离转录起始点－10区和－35区往往含有一些重要的保守序列（共有序列）。

－10区：含TATAAT序列，又称Pribnow盒。

－35区：含TTGACA序列。



01. 基因顺式作用元件

1.2 真核生物的特异DNA序列

真核生物基因组中含有可以调控自身基因表达活性的特异**DNA**序列，称为顺式作用元件。

顺式作用元件能够被转录调节蛋白特异识别和结合，从而影响基因表达活性。

启动子，增强子，绝缘子，沉默子

01. 基因顺式作用元件

1.2 真核生物的特异DNA序列

启动子 （Promoter)

是RNA聚合酶结合位点及其周围的一组转录调控元件，所对应的是RNA聚合酶，能指导RNA聚合酶结合在启动子上，活化RNA聚合酶使得转录进行。

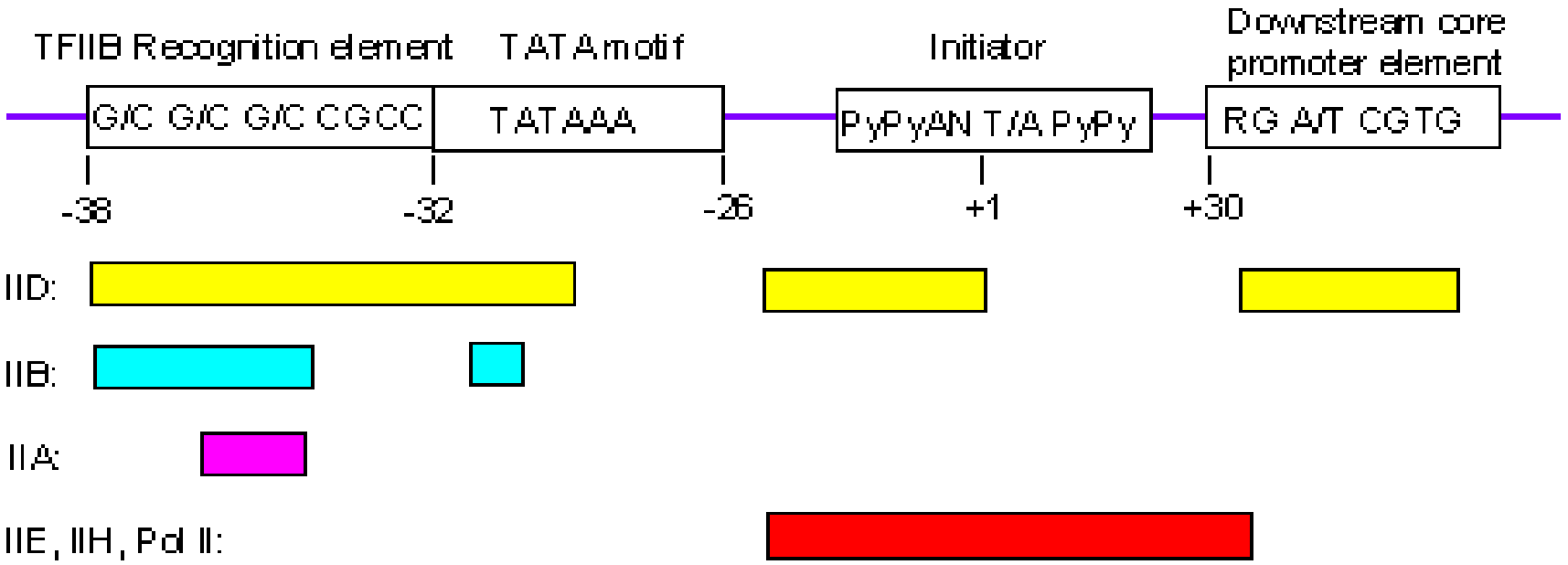
真核生物的启动子序列，在距离转录起始点-30bp区和-75bp区往往含有一些重要的保守序列：

-30区： TATA框（TATA frame): 其一致序列为TATAA(T)AA(T)。 相当于原核生物的-10序列（pribnow box）

-75区：CAAT框（ CAAT frame)：一致序列为GGC(T)CAATCT。CAAT框可能控制着转录起始的频率。

01. 基因顺式作用元件

1.2 真核生物的特异DNA序列



Sequence elements in a typical core promoter

01. 基因顺式作用元件

1.2 真核生物的特异DNA序列

增强子 （enhancer）

增强子又称为远上游序列（far upstream sequence)。它是远距离调节启动子以增加转录速率的DNA序列，其增强作用与序列的方向无关，与它在基因的上下游位置无关。

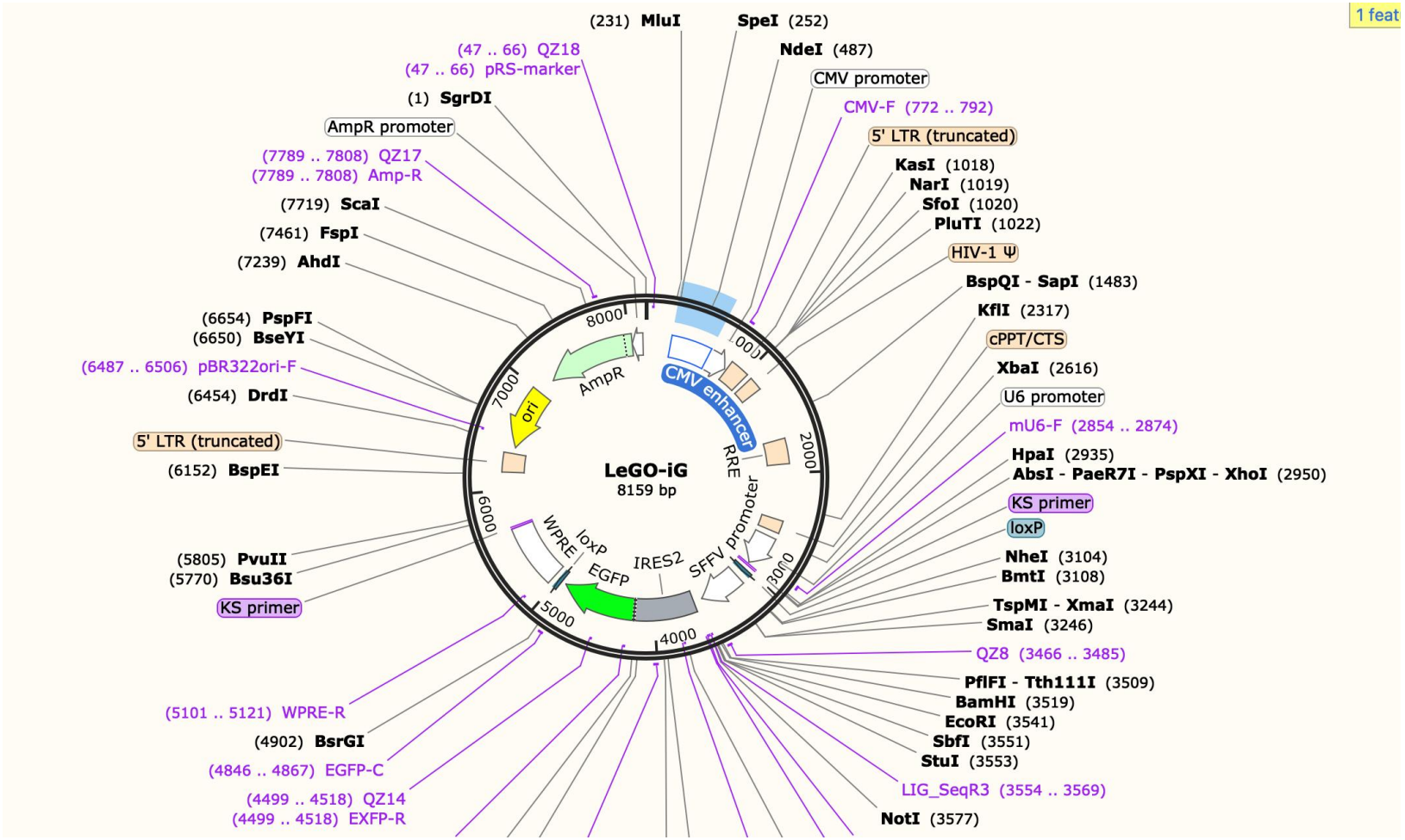
增强子有强烈的细胞类型选择，即不同细胞类型，增强作用不同。

（1）它能通过启动子大幅度地增加同一条DNA链上靶基因转录的频率，一般能增加 10～200倍，有的甚至可达千倍。
（2）增强子的作用对同源或异源的基因同样有效，如把SV40 的增强子连接到兔β-珠蛋白的基因上，可使转录强度增大100倍；
（3）增强子的位置可在基因5′上游、基因内或其3′下游的序列中，而其作用与所在基因旁侧部位的方向似无关系，因为无论正向还是反向，它都具有增强效应；

01. 基因顺式作用元件

1.2 真核生物的特异DNA序列

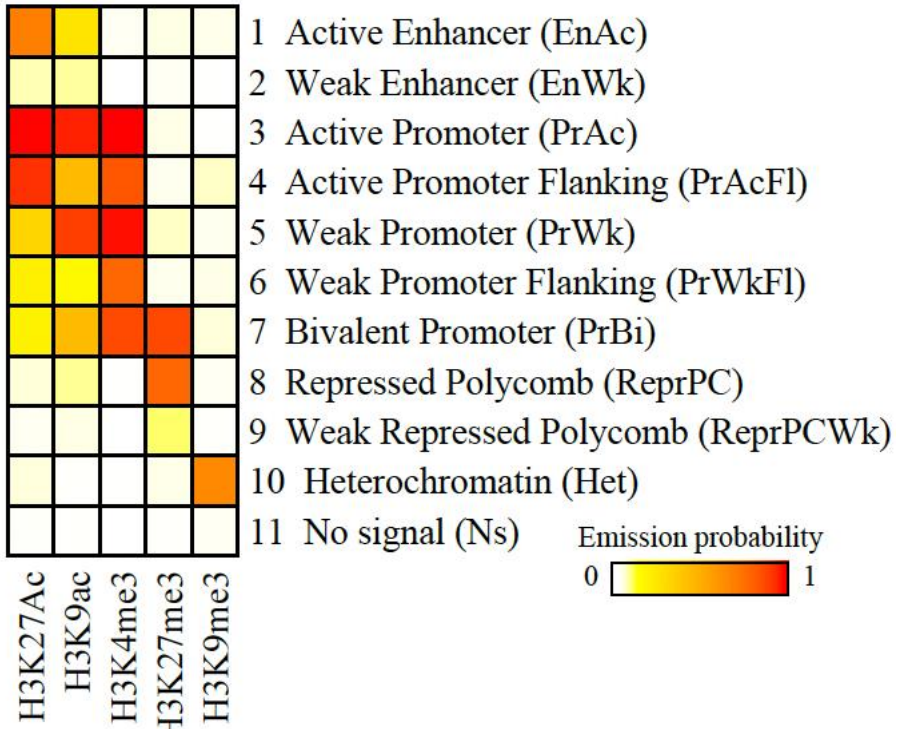
增强子 （enhancer）

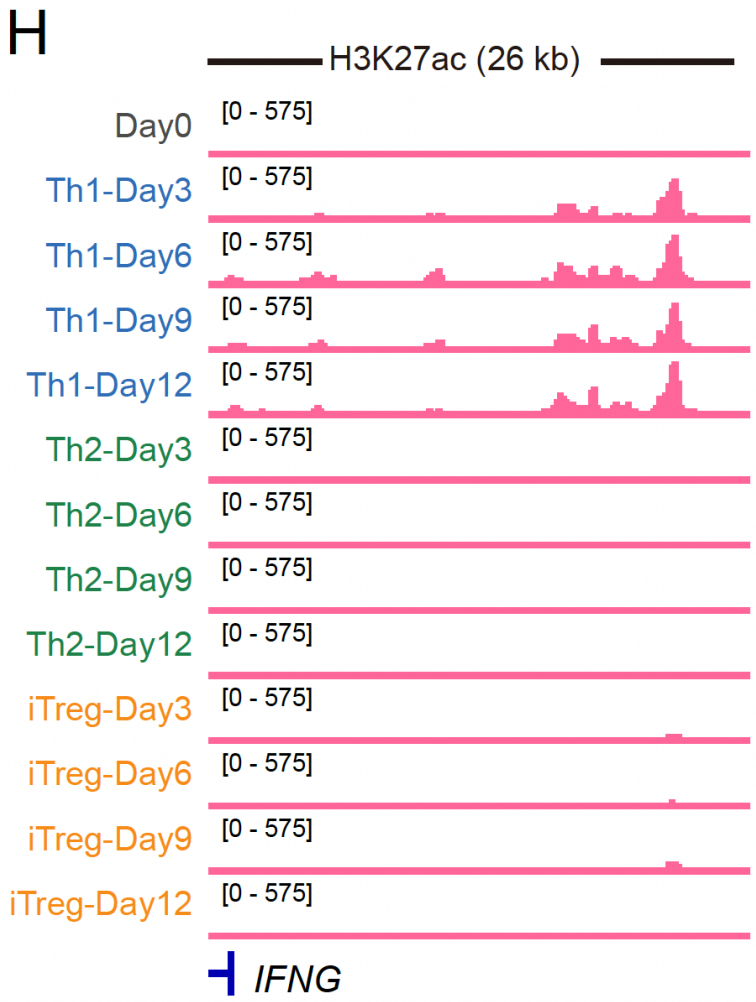


01. 基因顺式作用元件

1.2 真核生物的特异DNA序列

增强子 （enhancer）

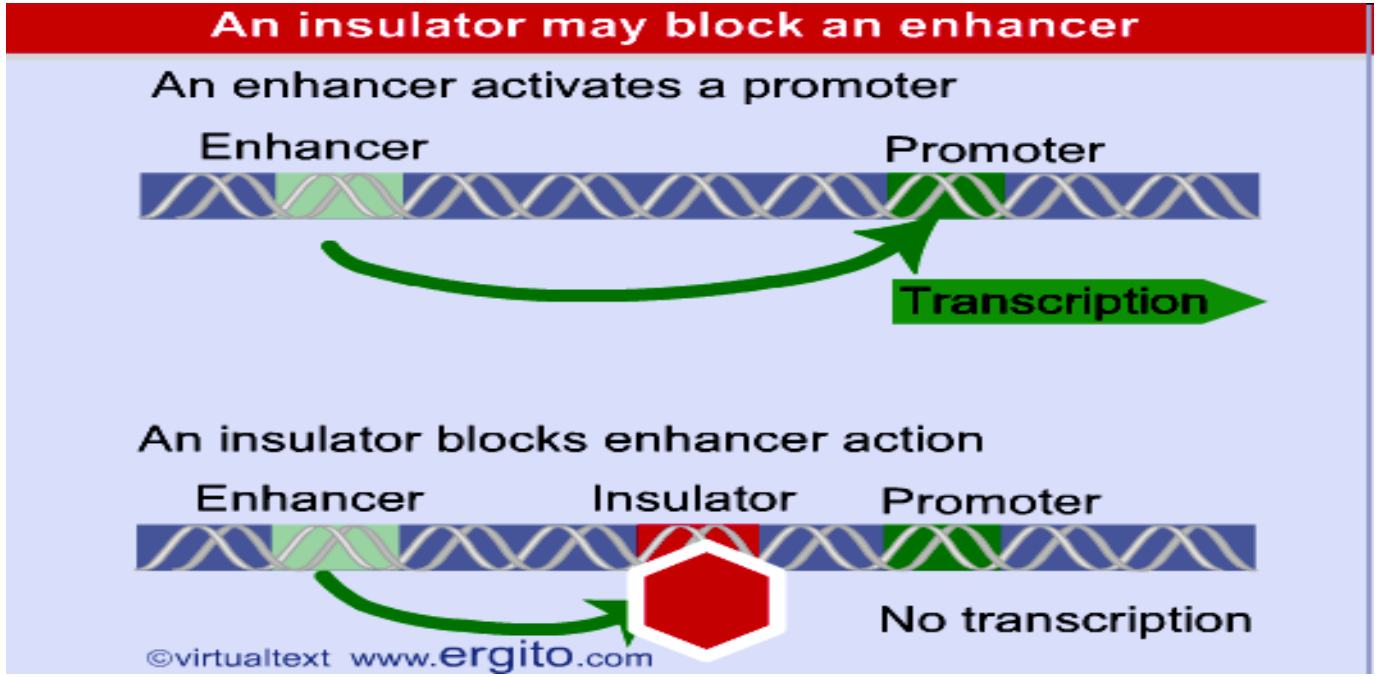




01. 基因顺式作用元件

1.2 真核生物的特异DNA序列

绝缘子：一种间接调控元件，本身对基因没有正向或是负向的直接作用，但可以使得其它调控元件对被调控基因的调控失去作用。



01. 基因顺式作用元件

1.2 真核生物的特异DNA序列

沉默子

是其中的一种负调控元件，它能够同反式因子协同作用，抑制靶基因的转录活性，在基因表达调控中发挥重要作用

01. 基因顺式作用元件

1.2 真核生物的特异DNA序列

应答元件 (response element) 是位于基因上游能被转录因子识别和结合，从而调控基因专一性表达的DNA序列

热激应答元件(heat shock response element，HSE)

金属应答元件(metal response element，MRE)

糖皮质激素应答元件(glucor-ticoid response element，GRE)

血清应答元件(serum response element，SRE)等。

02. 反式作用因子

1. 顺式作用元件

原核生物的特异DNA序列（操纵子）

真核生物的特异DNA序列（启动子，增强子，绝缘子，沉默子）

2. 反式作用因子

反式作用因子分类

转录因子结合位点 – Motif

3. 转录因子相关数据数据分析

转录因子全基因组结合位点测序 （ChIP-seq, CUT&Tag, CUT&RUN）

预测转录因子靶基因

预测转录因子活性

02. 反式作用因子

2.1 反式作用因子分类

反式作用因子

能直接或间接与顺式作用元件相互作用，进而调控基因转录的一类调节蛋白，统称为反式作用因子。

反式作用因子的结构特征

1、DNA识别或DNA结合结构域

2、激活基因转录的功能结构域

3、结合其他蛋白或调控蛋白的调节结构域

按其功能不同，常有以下三类：

转录因子：蛋白质-DNA

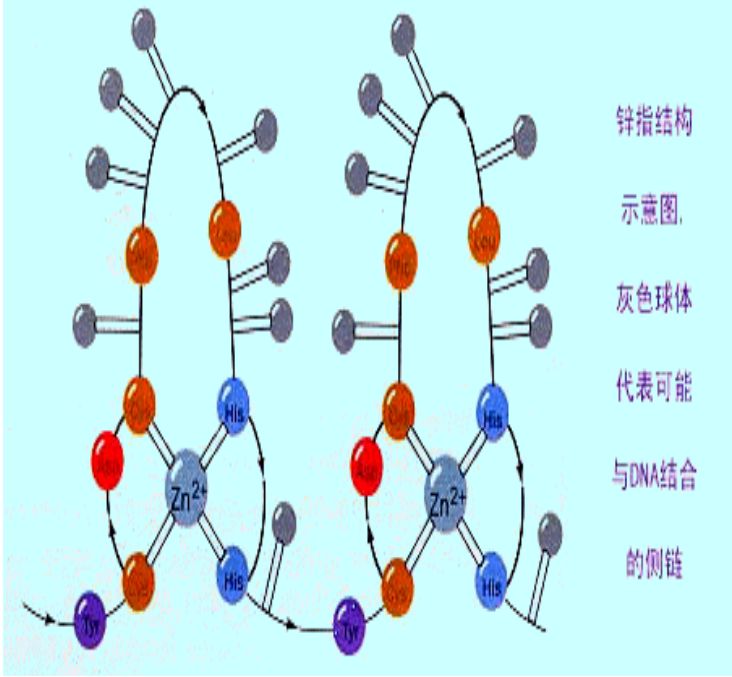
共调节因子：蛋白质-蛋白质

02. 反式作用因子

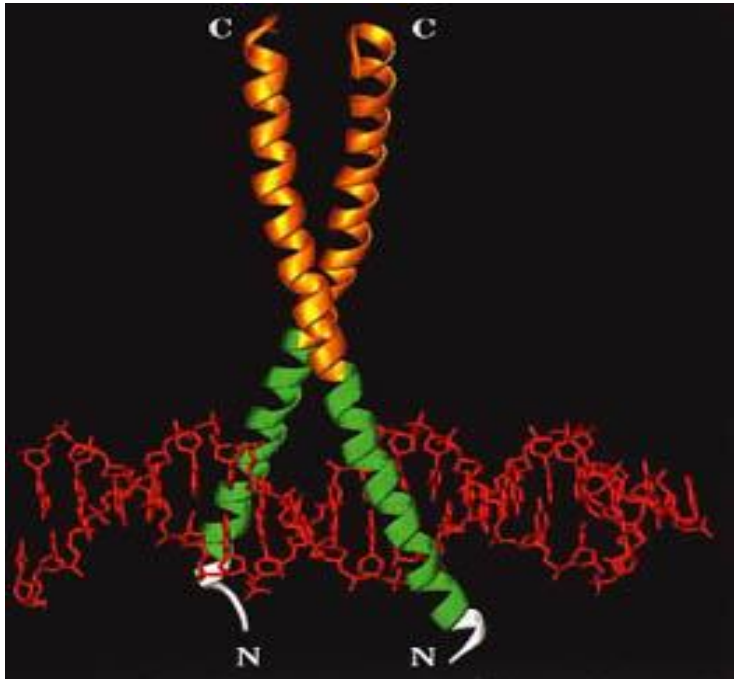
2.1 反式作用因子分类

反式作用因子的特殊功能域

DNA结合域；转录激活域；结合其他蛋白质的功能域。



锌指结构



亮氨酸拉链结构

02. 反式作用因子

2.1 反式作用因子分类

四种主要的**反式作用因子**结构域

1．螺旋-转角-螺旋结构 螺旋-转角-螺旋（helix-turn-helix）

2．锌指结构 锌指（zinc finger）是由一小群氨基酸与一个锌原子结合，在蛋白质中形成相对独立的一个结构域，故而得名

3 亮氨酸拉链结构 亮氨酸拉链（leucine zipper，ZIP）结构也是转录因子DNA结合区的一种结构模式

4．螺旋-环-螺旋结构 螺旋-环-螺旋（helix-loop-helix，HLH）是新近发现的一种DNA结合区的结构模式

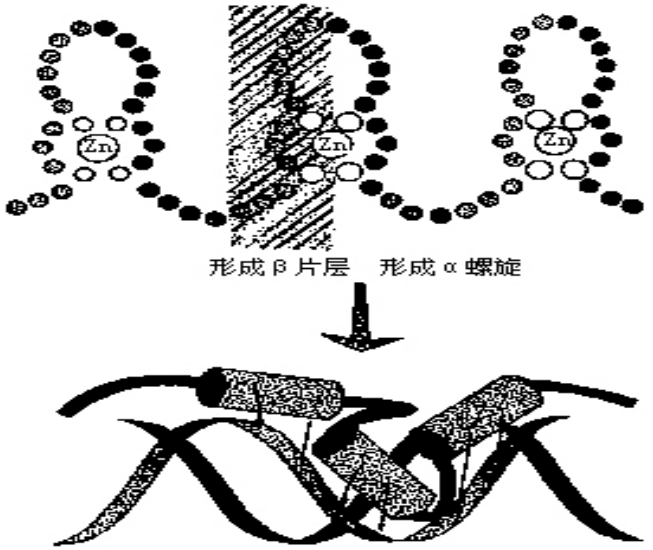


图10-7 锌指及其与DNA的结合

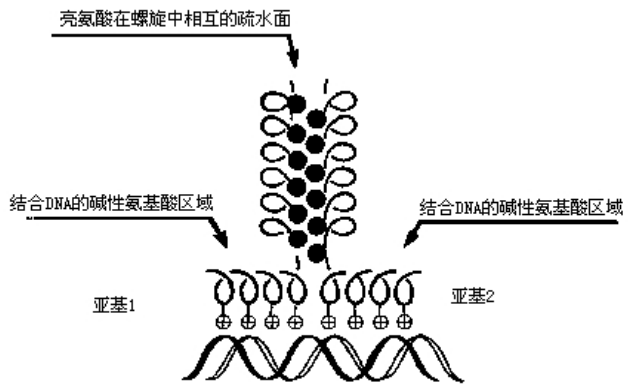
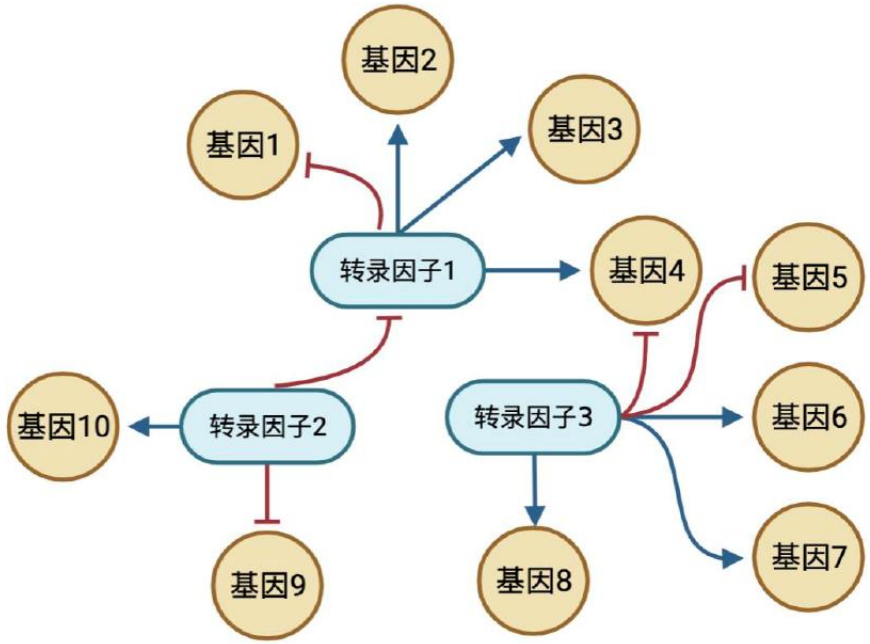


图10-8 亮氨酸拉链及其与DNA的结合模式图

02. 反式作用因子

2.2 转录因子结合位点 – Motif

- 识别转录因子结合位点是研究转录调控机制和建立转录调控网络的关键
  - 转录因子结合模体表示方法
  - 转录因子结合模体从头发现
  - 转录因子ChIP-seq数据分析



基因转录调控网络

02. 反式作用因子

2.2 转录因子结合位点 – Motif

模体（Motif）是指序列中局部的保守区域，或者是一组序列中共有的一小段序列模式；与转录因子、组蛋白等结合。在蛋白质、DNA、RNA序列中都存在，同源序列中不同位点的保守程度不同，一般对于功能影响较大的序列，通常比较保守。通俗来讲Motif就是反复出现的模式，并且假设其具有生物学功能。转录因子通常通过识别motif，与基因的调控区域结合，发挥转录调控作用

➢ 表示方法1：DNA共有序列（consensus sequence）

C C G G C A G C G G G T G G C G C T G
G A T C C T G A A G A T G G C G C T G
C T G C C A A C A G G A G G C G C T G
C T A C C T G C T G G T G G C G C T G
T G G G C A G C A G G A G G C A G T G
T G G C C T G T A G G A G G C A G C A
T C T C C A G C A G G G G G A G A G C
C T G A C A C T A G A T G G C G C T T
A C A C C A C T T G G T G G C G C T C
C C A C C A G C A G G A G G A G G G A G
C G C A C T G A A G G G G G C G C T C

DNA共有序列 N N N V C W V H D G R D G G M R V N N

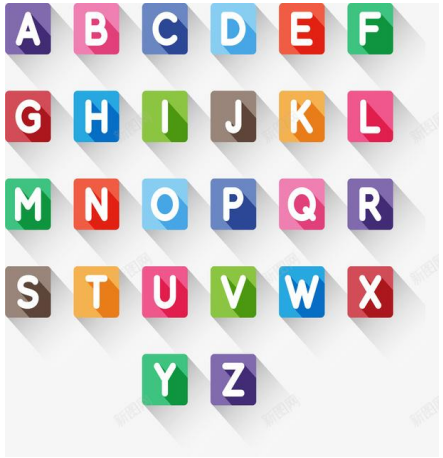
02. 反式作用因子

2.2 转录因子结合位点 – Motif

➢ IUPAC简并码

List of symbols

Symbol[2]	Meaning/derivation	Possible bases					Complement
A	Adenine	A				1	T (or U)
C	Cytosine		C				G
G	Guanine			G			C
T	Thymine				T		A
U	Uracil				U		A
W	Weak	A			T	2	S
S	Strong		C	G			W
M	aMino 胺	A	C				K
K	Keto 酮			G	T		M
R	puRine 嘌呤	A		G			Y
Y	pYrimidine 嘧啶		C		T		R
B	not A (B comes after A)		C	G	T	3	V
D	not C (D comes after C)	A		G	T		H
H	not G (H comes after G)	A	C		T		D
V	not T (V comes after T and U)	A	C	G			B
N	any Nucleotide (not a gap)	A	C	G	T	4	N
Z	Zero					0	Z



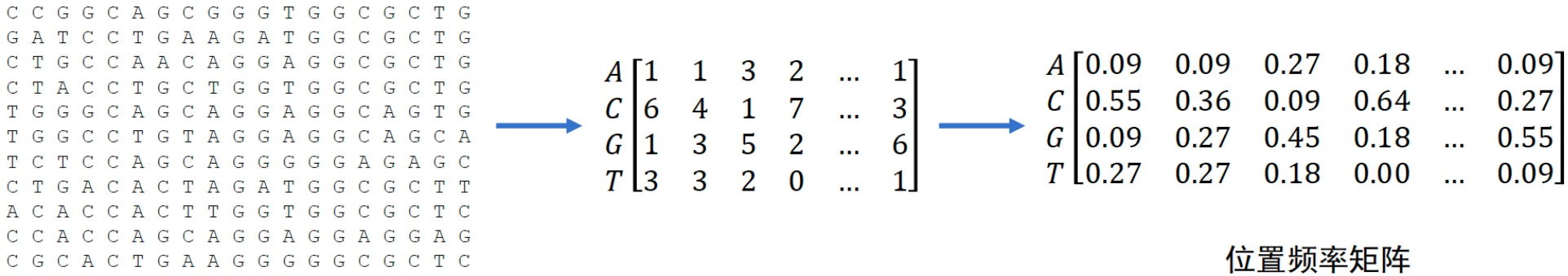
S = G or C
•S represents bases that form strong hydrogen bonds, specifically G (Guanine) and C (Cytosine). G and C pair together with three hydrogen bonds, which is stronger compared to the two hydrogen bonds between A and T. Hence, S is used to denote this stronger pairing.

W = A or T
•W represents bases that form weaker hydrogen bonds, specifically A (Adenine) and T (Thymine). Since A and T form only two hydrogen bonds, which are considered weaker compared to the three hydrogen bonds between G and C, W is used to represent these bases.

02. 反式作用因子

2.2 转录因子结合位点 – Motif

- 表示方法2：位置频率矩阵（position frequency matrix，PFM）



位置频率矩阵

Position Probability Matrix (PPM)

02. 反式作用因子

2.2 转录因子结合位点 – Motif

基于已知模体的转录因子结合位点预测

- 位置权重矩阵（position weight matrix，PWM）
  - ✓ 由于DNA序列碱基组成具有一定偏好性，进行转录因子结合位点预测时需要将位置频率矩阵转换为位置权重矩阵。

$$\begin{bmatrix} q_{A,1}, q_{A,2}, \cdots, q_{A,n} \\ q_{C,1}, q_{C,2}, \cdots, q_{C,n} \\ q_{G,1}, q_{G,2}, \cdots, q_{G,n} \\ q_{T,1}, q_{T,2}, \cdots, q_{T,n} \end{bmatrix}$$

位置频率矩阵

$$S_{i,j} = log_2(\frac{q_{i.j}}{b_i})$$

b_i是碱基i在DNA序列中出现频率

$$\begin{bmatrix} S_{A,1}, S_{A,2}, \cdots, S_{A,n} \\ S_{C,1}, S_{C,2}, \cdots, S_{C,n} \\ S_{G,1}, S_{G,2}, \cdots, S_{G,n} \\ S_{T,1}, S_{T,2}, \cdots, S_{T,n} \end{bmatrix}$$

位置权重矩阵

02. 反式作用因子

2.2 转录因子结合位点 – Motif

位置权重矩阵 position weight matrix（**PWM**）

A	C	G	T
0.142857	0.000000	0.000000	0.857143
0.857143	0.000000	0.071429	0.071429
0.000000	1.000000	0.000000	0.000000
0.000000	0.000000	0.000000	1.000000
0.000000	0.000000	1.000000	0.000000
0.000000	0.000000	0.071429	0.928571
1.000000	0.000000	0.000000	0.000000
0.000000	0.071429	0.000000	0.928571
0.928571	0.000000	0.071429	0.000000
0.214286	0.000000	0.000000	0.785714
0.642857	0.071429	0.214286	0.071429
0.357143	0.285714	0.000000	0.357143
1.000000	0.000000	0.000000	0.000000
0.357143	0.285714	0.000000	0.357143
0.500000	0.428571	0.000000	0.071429
0.000000	1.000000	0.000000	0.000000
1.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.785714	0.214286

公式中k表示A/C/G/T，j表示位点，b表示背景碱基频率，M表示PPM矩阵中的碱基频率。

$$M_{k,j} = log_2\left(\frac{M_{k,j}}{b_k}\right)$$

Name	Freq.	Bg.			Bg.	Freq.	Name
Adenine	0.29	0.29	A	~	T	0.29	0.29	Thymine
Cytosine	0.21	0.21	C	~	G	0.21	0.21	Guanine

第一行第一列为例：log2(0.142857/0.29)=-1.02148117

02. 反式作用因子

2.2 转录因子结合位点 – Motif

PWM矩阵

A	C	G	T
-1.02148117	-Inf	-Inf	1.563483014
1.563483014	-Inf	-1.555807499	-2.021471071
-Inf	2.251538767	-Inf	-Inf
-Inf	-Inf	-Inf	1.785875195
-Inf	-Inf	2.251538767	-Inf
-Inf	-Inf	-1.555807499	1.678959325
1.785875195	-Inf	-Inf	-Inf
-Inf	-1.555807499	-Inf	1.678959325
1.678959325	-Inf	-1.555807499	-Inf
-0.436515303	-Inf	-Inf	1.437951367
1.148444953	-1.555807499	0.029148269	-2.021471071
0.300448945	0.444182402	-Inf	0.300448945
1.785875195	-Inf	-Inf	-Inf
0.300448945	0.444182402	-Inf	0.300448945
0.785875195	1.029144903	-Inf	-2.021471071
-Inf	2.251538767	-Inf	-Inf
1.785875195	-Inf	-Inf	-Inf
-Inf	-Inf	1.903614939	-0.436515303

根据PWM矩阵可以计算序列的得分，即将碱基在各个位点的在PWM矩阵中的值进行加和，从而可以判断该序列是一段随机序列还是功能位点。

例如上述序列TACTGTATATAHAHMCAG的的得分为1.56+1.56+2.25+1.79+2.25+1.68+……，

该得分如果大于0，则认为该序列是一个潜在的功能位点，预测到该Motif；

如果得分小于0，则认为该序列是一段随机序列；如果等于0，则认为各有50%的概率。

02. 反式作用因子

2.2 转录因子结合位点 – Motif

基于已知模体的转录因子结合位点预测

- 预测一段DNA序列中某一转录因子的潜在结合位点
  - ✓ 滑动窗口（长度为n）；
  - ✓ 应用位置权重矩阵对每个窗口进行打分$S = \sum_{i=1}^{n} S_{t_i,j}$
  - ✓ 基于阈值筛选

GTTATTACGCTGGCCACTAGCGGGCGCGTTGTAACGCTG

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
A	0.2038	0.0483	0.000	0.6218	0.0567	0.0924	0.9034	0.0990	0.3866	0.0231	0.0063
C	0.0710	0.8650	0.9950	0.0350	-0.0550	0.5520	0.0180	0.2000	0.0000	0.0330	0.0020
G	0.5966	0.0525	0.0042	0.2647	0.3697	0.0777	0.0588	-0.0500	0.6134	0.6996	0.9916
T	0.1282	0.0336	0.000	0.0777	0.0168	0.2773	0.0189	0.4970	0.0000	0.2437	0.0000

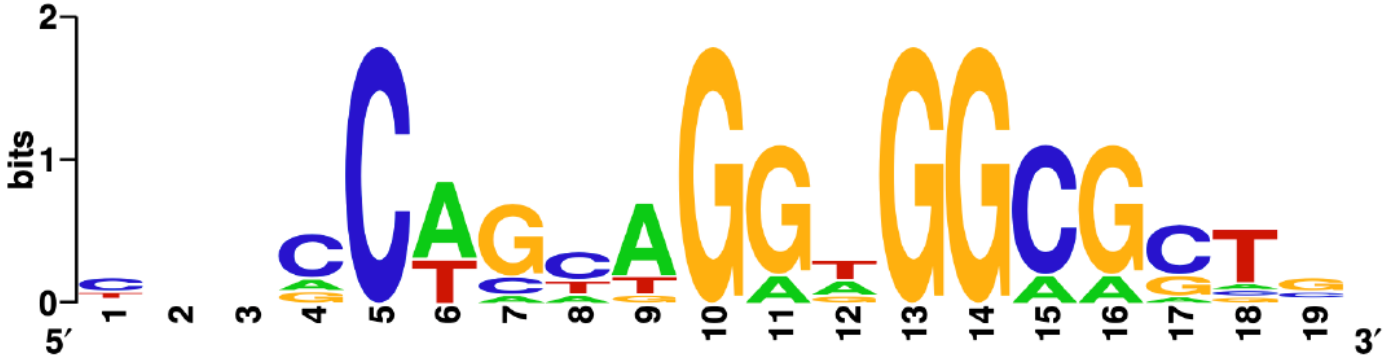
0.5966 + 0.8650 + 0.9950 + 0.6218 - 0.055 + 0.2773 + 0.9034 - 0.0500 + 0.0000 + 0.6996 + 0.9916 = 5.8453

应用位置权重矩阵预测转录因子潜在结合位点

02. 反式作用因子

2.2 转录因子结合位点 – Motif

➢ 表示方法3：序列标识图（sequence logo）



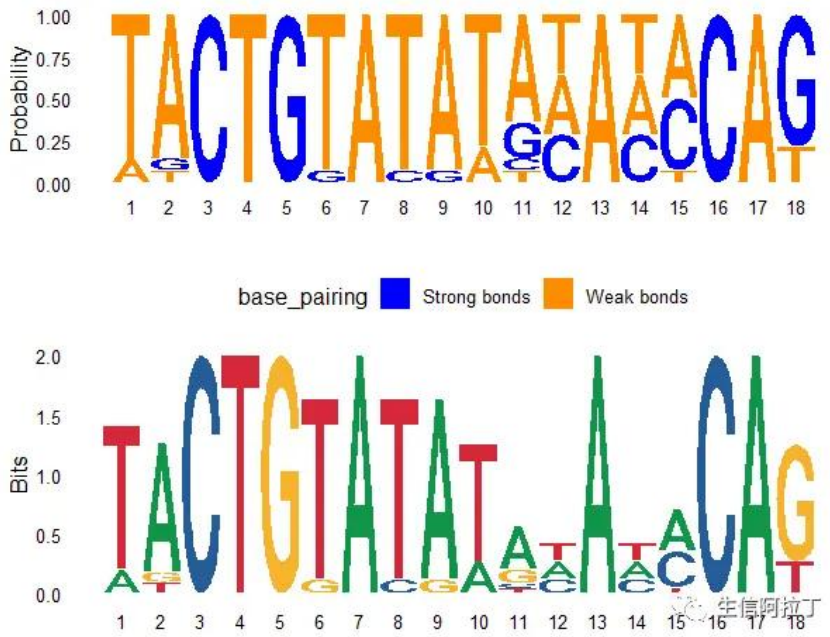
序列标识图

字母越大—序列越保守

02. 反式作用因子

2.2 转录因子结合位点 – Motif

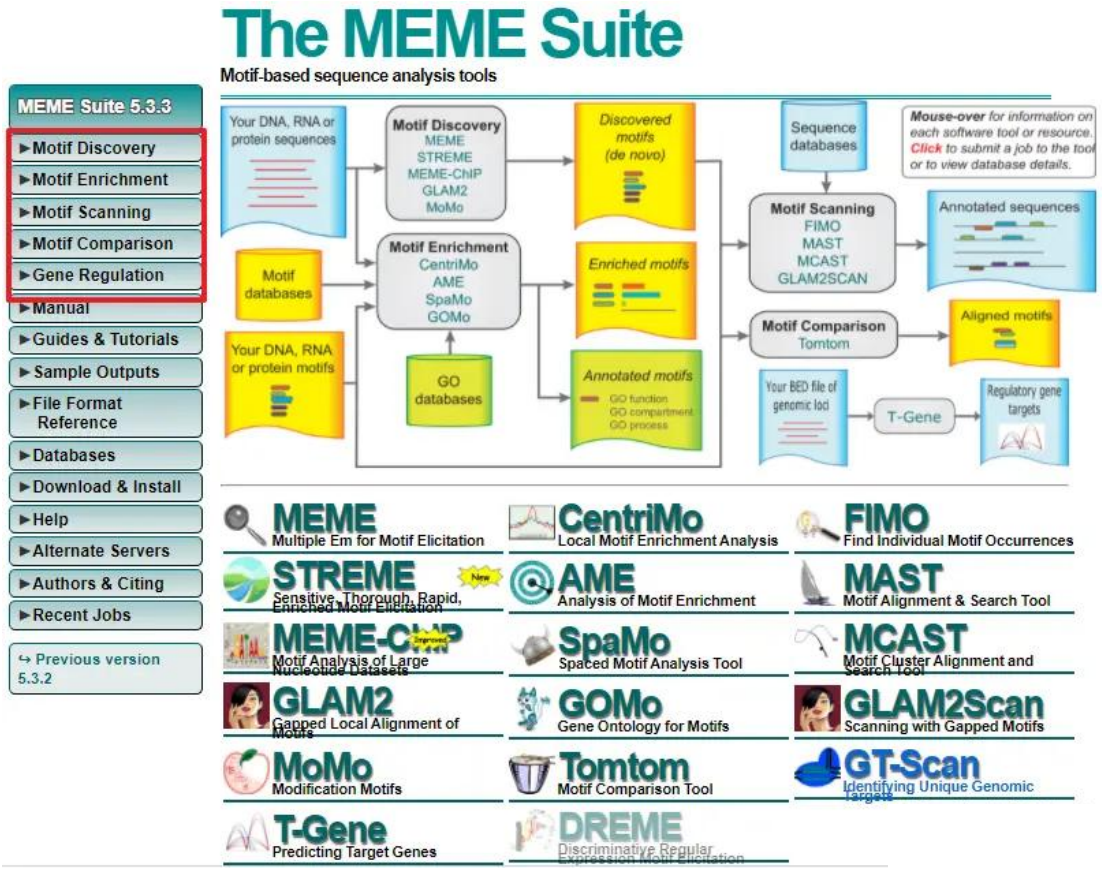
```
library(ggseqlogo)
library(ggplot2)
Motif <- t(read.table("ppm.txt"))
rownames(Motif) <- c("A","C","G","T")
 ##list_col_schemes(v = T)查看配色
p1 <- ggseqlogo(Motif,method="prob",col_scheme="base_pairing")
p2 <- ggseqlogo(Motif,method="bits",col_scheme="nucleotide")
gridExtra::grid.arrange(p1,p2)
```



02. 反式作用因子

2.2 转录因子结合位点 – Motif

- 转录因子结合模体从头发现
  - ✓ 通过收集多条相关的DNA序列，在其中寻找具有统计显著性的短片段模式，预测为该转录因子潜在的结合模体



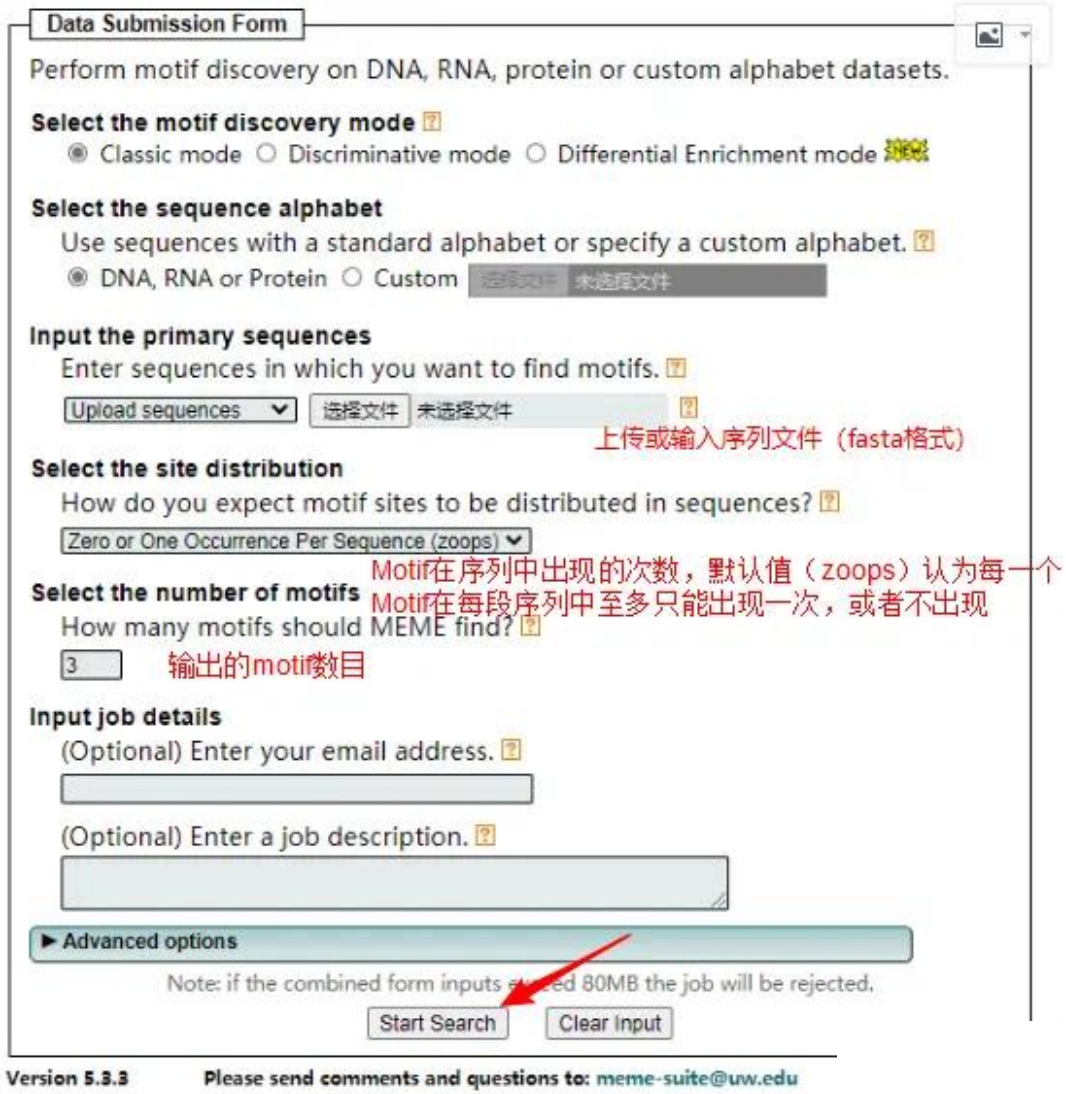
MEME Suite 大礼包

02. 反式作用因子

2.2 转录因子结合位点 – Motif

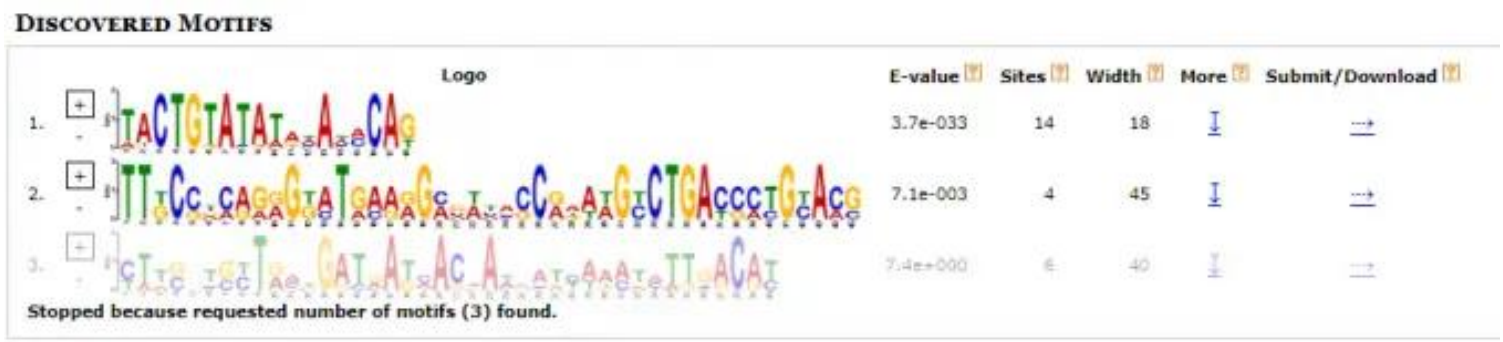
主要可以实现五个功能：
Motif Discovery
Motif Enrichment
Motif Scanning
Motif Comparison
Gene Regulation
用于预测输入序列上的motif信息，它们支持DNA、RNA或蛋白序列的分析

以MEME为例：



02. 反式作用因子

2.2 转录因子结合位点 – Motif



输出的Motif结果，包含Logo，E-value、Sites、Width、More和Submit/Download六列。
Sequence Logo展示的是Motif的一致性序列，字母的高度表示该碱基在Site Count各序列中出现的频率，结果中彩色的logo表示显著，不显著的为灰色。
第一个motif的Sites为14，即输入的序列中，该Motif出现了14次。
Width表示该Motif的长度。

02. 反式作用因子

2.2 转录因子结合位点 – Motif

Motif在序列上的位置信息。每一行为输入的序列序号以及名称，每个Block表示一个Motif，以颜色进行区分，Block的高度表示其重要性，高度越高P值越小，即越显著。上方Block表示该Motif位于正链，下方则位于负链。将光标置于Block上方时可展示该Motif的详细信息。

02. 反式作用因子

2.2 转录因子结合位点 – Motif

INPUTS & SETTINGS

Sequences

Role	Source	Alphabet	Sequence Count	Total Size
Primary Sequences	lex0.fna	DNA	16	3067

Background Model

Source: built from the (primary) sequences

Order: 0

Name	Freq.	Bg.				Bg.	Freq.	Name
Adenine	0.29	0.29	A	~	T	0.29	0.29	Thymine
Cytosine	0.21	0.21	C	~	G	0.21	0.21	Guanine

Other Settings

Motif Site Distribution	ZOOPS: Zero or one site per sequence
Objective Function	E-value of product of p-values
Starting Point Function	E-value of product of p-values
Site Strand Handling	Sites may be on either strand
Maximum Number of Motifs	3
Motif E-value Threshold	no limit
Minimum Motif Width	8
Maximum Motif Width	50
Minimum Sites per Motif	2
Maximum Sites per Motif	16

Show Advanced Settings

03. 转录因子相关测序数据分析

1. 顺式作用元件

原核生物的特异DNA序列（操纵子）

真核生物的特异DNA序列（启动子，增强子，绝缘子，沉默子）

2. 反式作用因子

反式作用因子分类（转录因子，转录调节因子，共调节因子）

转录因子结合位点 – Motif

3. 转录因子相关测序数据分析

转录因子全基因组结合位点测序 （ChIP-seq, CUT&Tag, CUT&RUN）

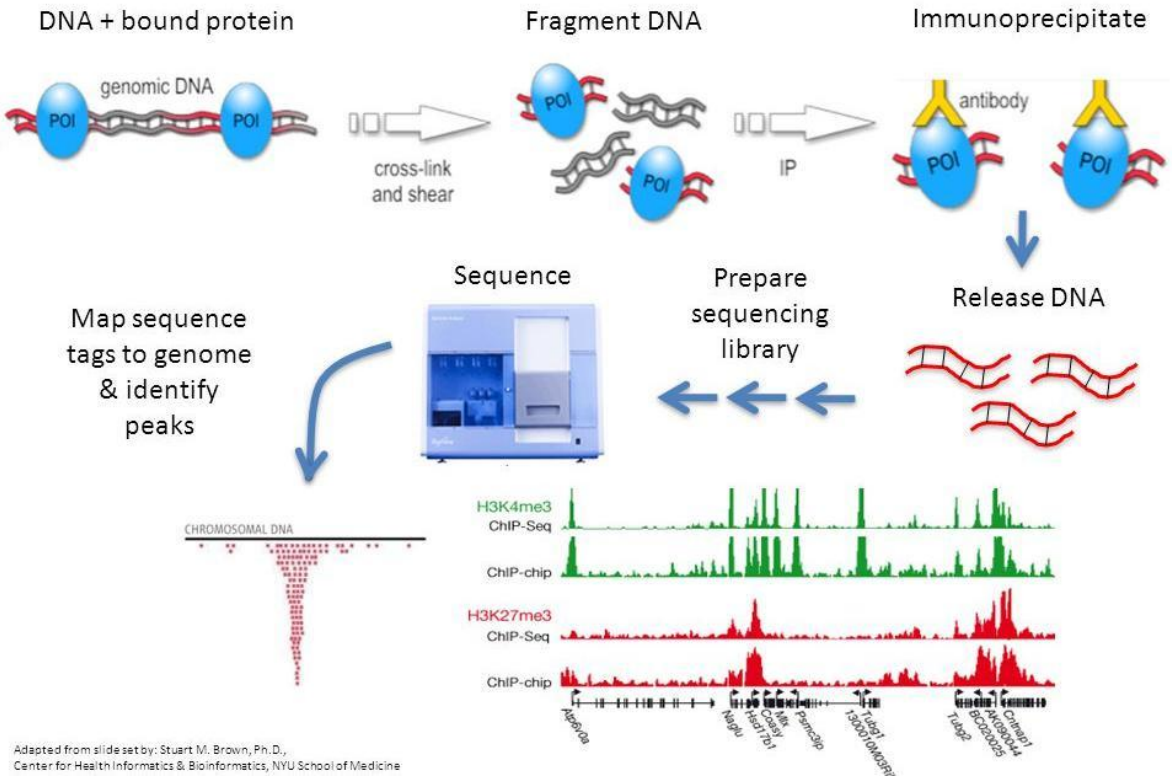
预测转录因子靶基因

03. 转录因子相关测序数据分析

3.1 转录因子结合位点测序

ChIP-seq

Chromatin Immunoprecipitation Sequencing (ChIP-seq) is a powerful technique used to study protein-DNA interactions on a genome-wide scale. It combines chromatin immunoprecipitation (ChIP) with next-generation sequencing (NGS) to identify binding sites of DNA-associated proteins, such as transcription factors, histones, or other chromatin-modifying enzymes.



Workflow:

1.Chromatin crosslinking & shearing (via sonication or MNase digestion).
2.Immunoprecipitation (IP) with modification-specific antibodies.
3.Library preparation & sequencing (typically Illumina short reads).

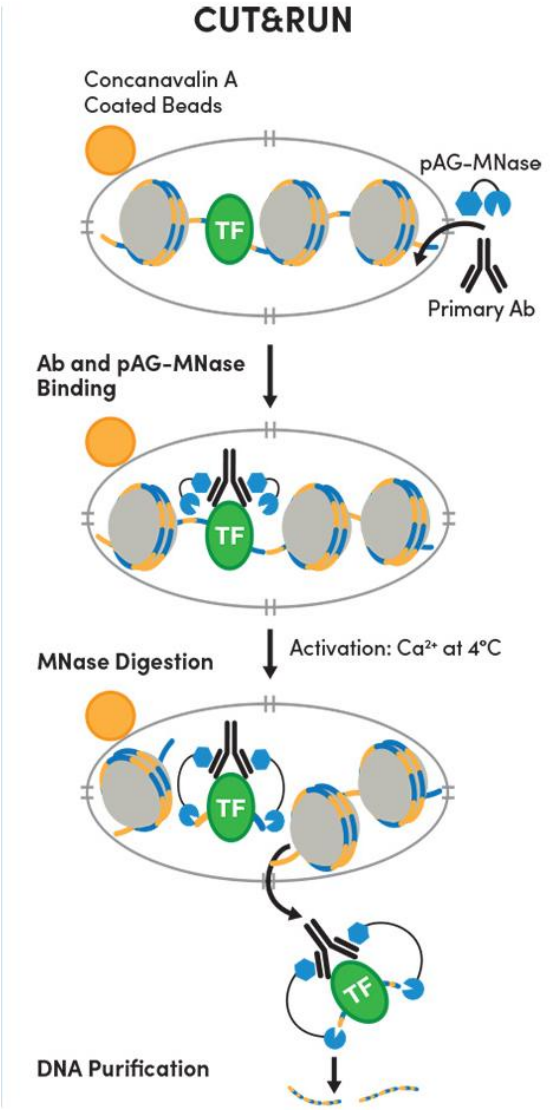
03. 转录因子相关测序数据分析

3.1 转录因子结合位点测序

CUT&RUN

CUT&RUN (Cleavage Under Targets & Release Using Nuclease) : is an innovative, high-resolution method for mapping protein-DNA interactions in the genome. It is more efficient and requires fewer cells than traditional **ChIP-seq**, making it ideal for studying rare cell populations or low-abundance proteins.

1.Permeabilization of Cells/Nuclei
2.Antibody Binding
3.Targeted DNA Cleavage
4.DNA Release & Sequencing



03. 转录因子相关测序数据分析

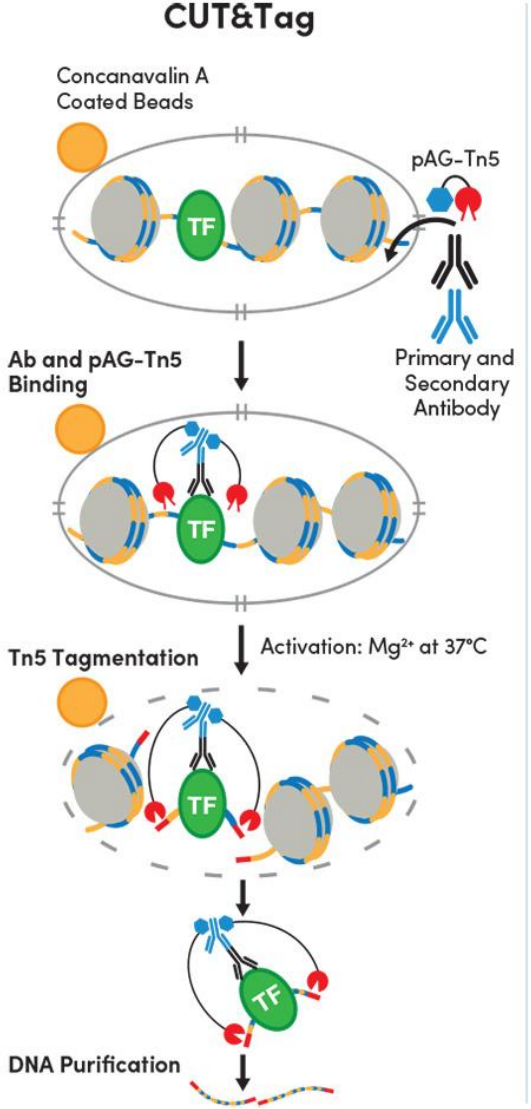
3.1 转录因子结合位点测序

CUT&Tag

CUT&Tag (Cleavage Under Targets & Tagmentation)

is an advanced epigenomic profiling technique that builds upon CUT&RUN, offering even higher sensitivity and scalability for mapping protein-DNA interactions. Instead of using Micrococcal Nuclease (MNase) like CUT&RUN, CUT&TAG employs a fusion protein of Protein A-Tn5 transposase (pA-Tn5) to simultaneously cleave and tag DNA at protein-binding sites.

1.Cell Permeabilization
2.Antibody Binding
3.Tn5 Transposase Tagmentation
4.DNA Release & Amplification



03. 转录因子相关测序数据分析

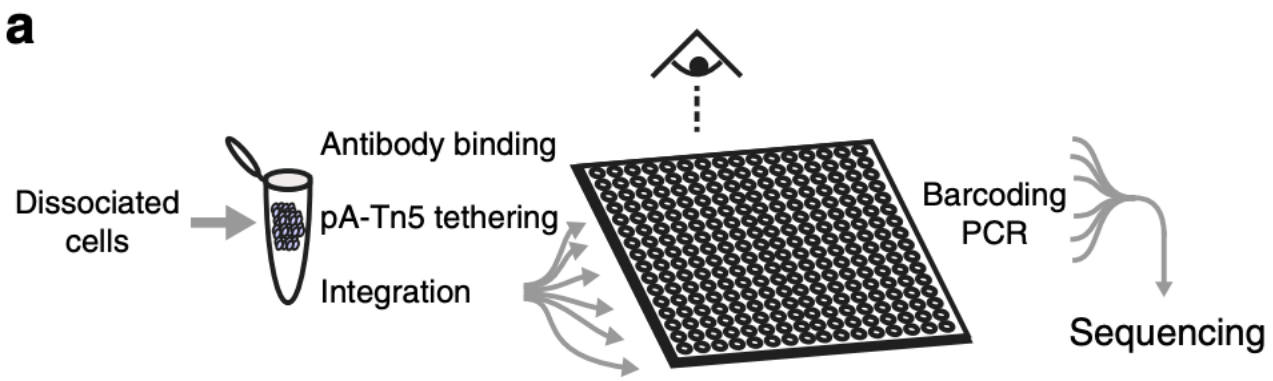
3.1 转录因子结合位点测序

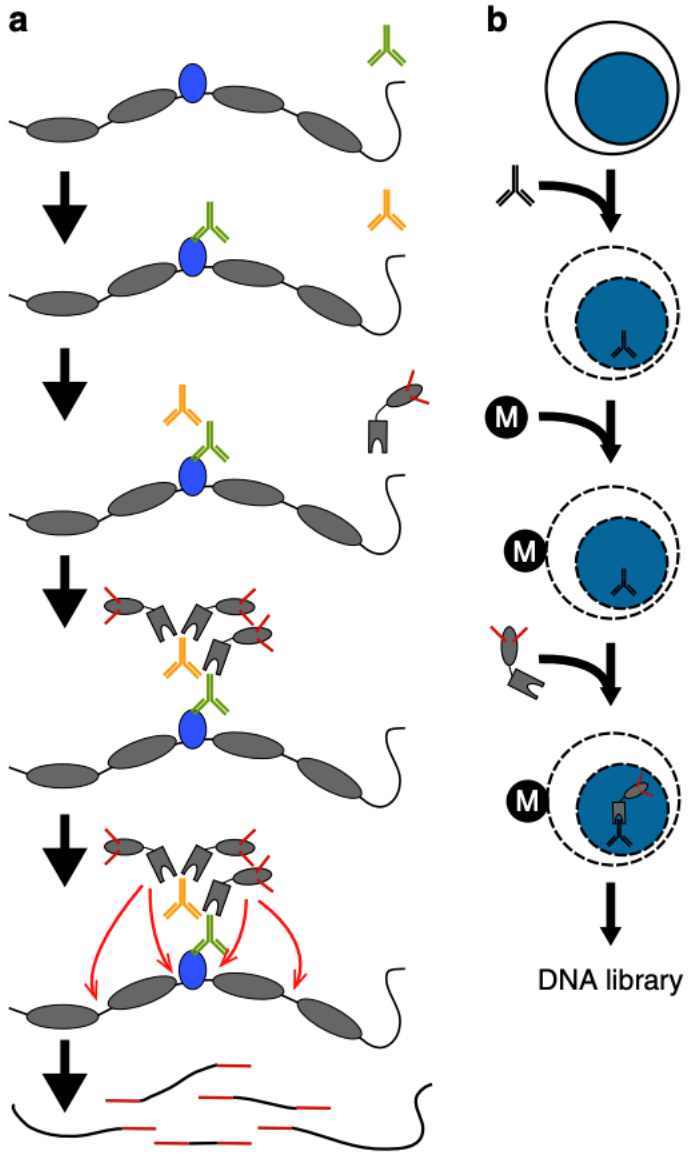
Method	Input (Cells/DNA)	Resolution	Advantages	Disadvantages
ChIP-seq	High (~1M cells)	~200 bp	Well-established, widely used	High background noise, requires crosslinking
CUT&RUN	Low (~100-1000 cells)	~300 bp	High signal-to-noise, no crosslinking	
CUT&TAG	Ultra-low (~10-100 cells)	~100 bp	Direct tagmentation, high resolution	Tn5 bias

03. 转录因子相关测序数据分析

3.1 转录因子结合位点测序

Single cell CUT&Tag





03. 转录因子相关测序数据分析

3.2 预测转录因子靶基因

转录因子**ChIP-seq**实验设计

1. 高质量的特异性抗体是ChIP-seq实验的前提，对于缺乏高质量商业化抗体的转录因子，研究人员需要制备抗体或表达转录因子与标签序列的融合蛋白
2. 实验设计时应确定测序深度对于人类转录因子，ChIP-seq实验通常需要约20M测序reads
3. 需要进行生物学重复和对照实验以增加可靠度

03. 转录因子相关测序数据分析

3.2 预测转录因子靶基因

转录因子**ChIP-seq**数据质量控制

- 测序读长层面质量控制
- 信号峰层面质量控制

  ChIP-seq信号可视化展示　　测序读长位于信号峰的比例

  互相关分析　　不可重复发现率

- 注释层面质量控制

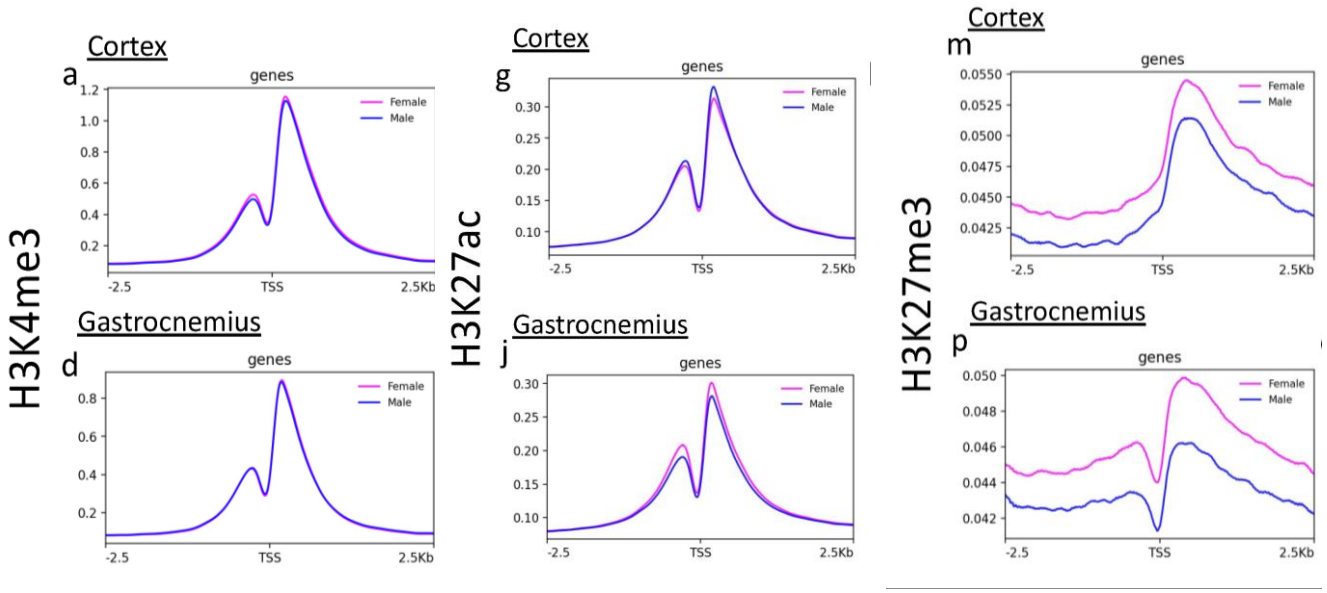
  基因组分布特征　　序列保守性分析

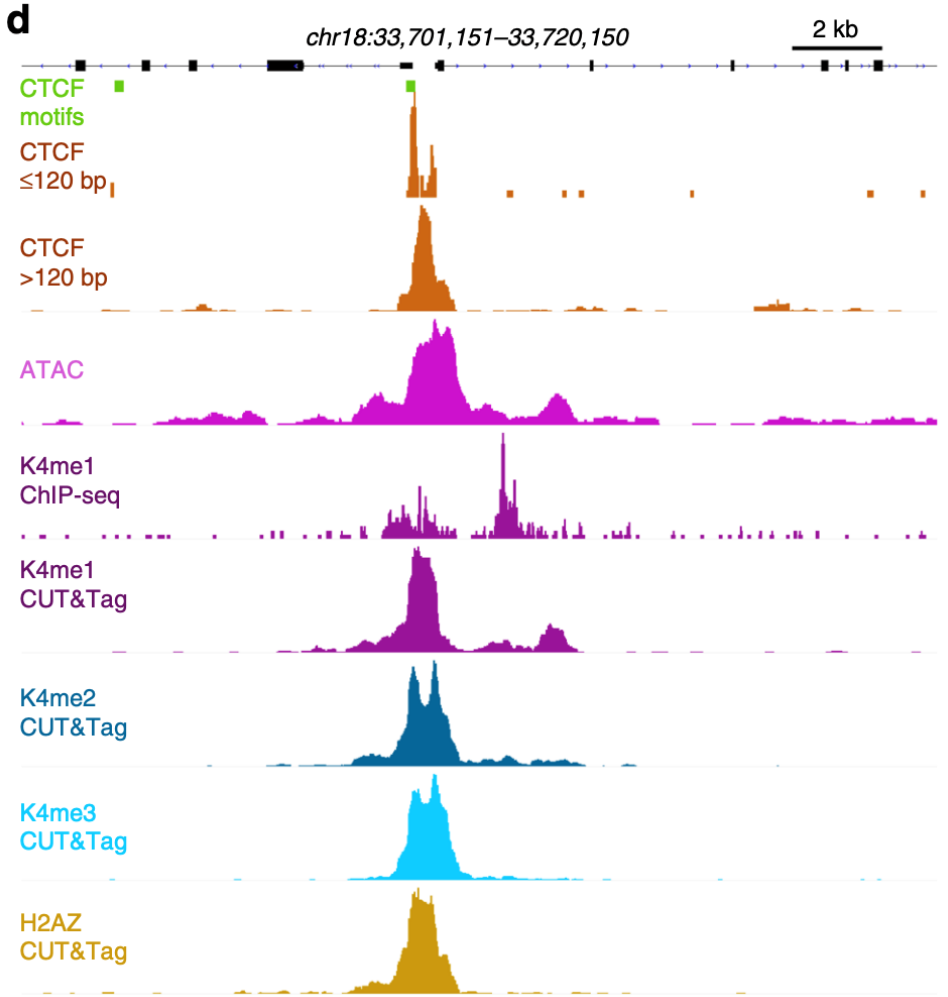
  DNA模体分析

03. 转录因子相关测序数据分析

3.2 预测转录因子靶基因

Data analyses





03. 转录因子相关测序数据分析

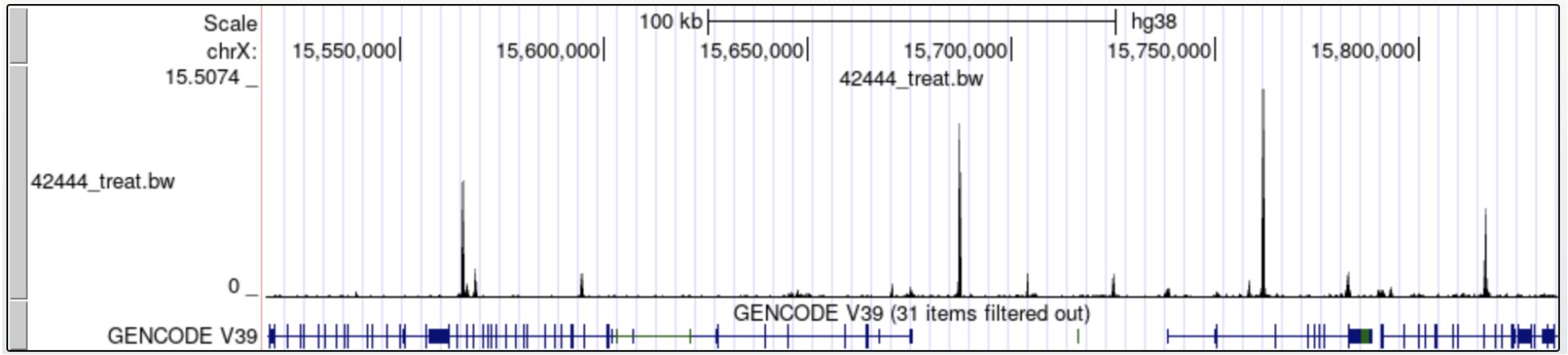
3.2 预测转录因子靶基因

信号峰层面质量控制

- ChIP-seq信号可视化展示

  将ChIP-seq数据转换为bigWig或bedGraph格式

  对样品背景噪音的高低进行直观判断



ChIP-seq信号可视化示例

03. 转录因子相关测序数据分析

3.2 预测转录因子靶基因

信号峰层面质量控制

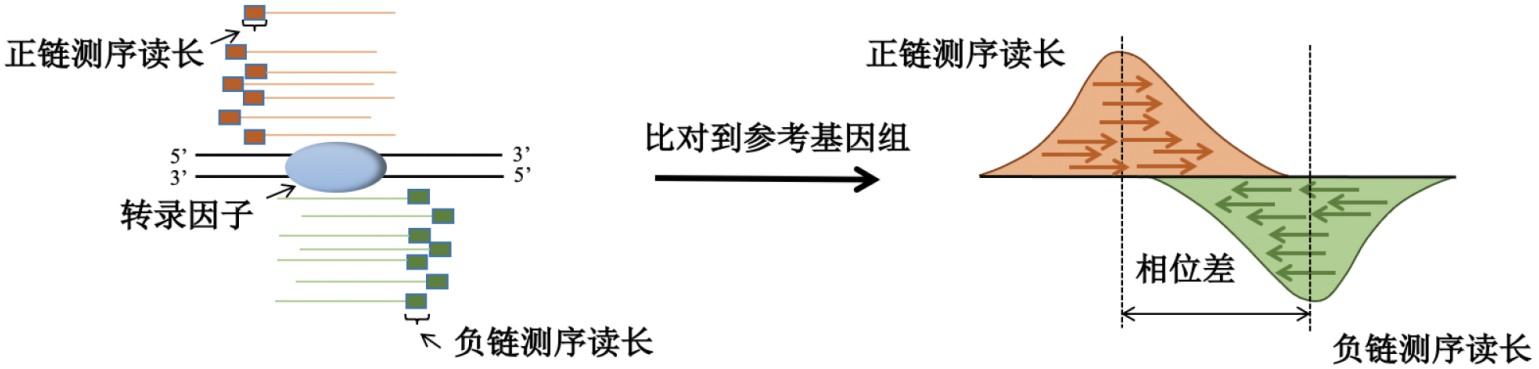
- 测序读长位于信号峰的比例（fraction of reads in peaks，FRiP）
  - ✓ 可以直观地指征测序读长在信号峰的富集程度
    - 质量较好的ChIP-seq数据具有较低的背景噪音，FRiP值较高
    - 质量较差的ChIP-seq数据FRiP值较低
  - ✓ 局限性
    - FRiP值的大小与信号峰的数量正相关，改变信号峰识别阈值会改变FRiP值
    - 由于抗体不同、结合位点数量不同，FRiP值在不同转录因子间通常不具备可比性

03. 转录因子相关测序数据分析

3.2 预测转录因子靶基因

信号峰层面质量控制

- 在转录因子ChIP-seq数据的信号峰区域，比对到正链的读长与比对到负链的读长之间会产生一个相位差

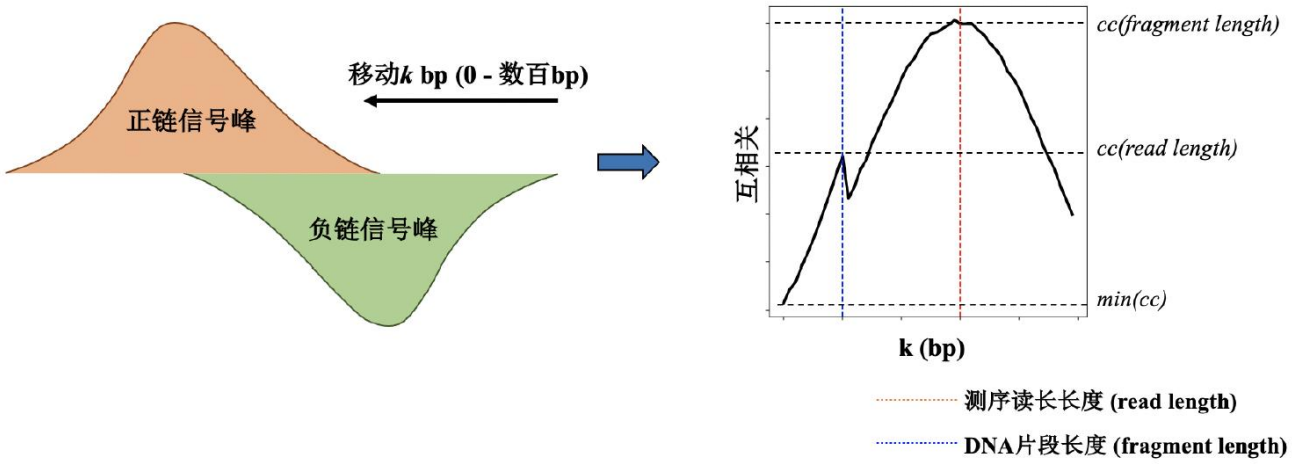


03. 转录因子相关测序数据分析

3.2 预测转录因子靶基因

信号峰层面质量控制

- 互相关（cross-correlation）
  - ✓ 负链读长向其3’端方向移动k bp（k的取值范围从1至数百），每次移动后计算正链和负链信号的皮尔森相关系数，得到相关系数随k变化的曲线

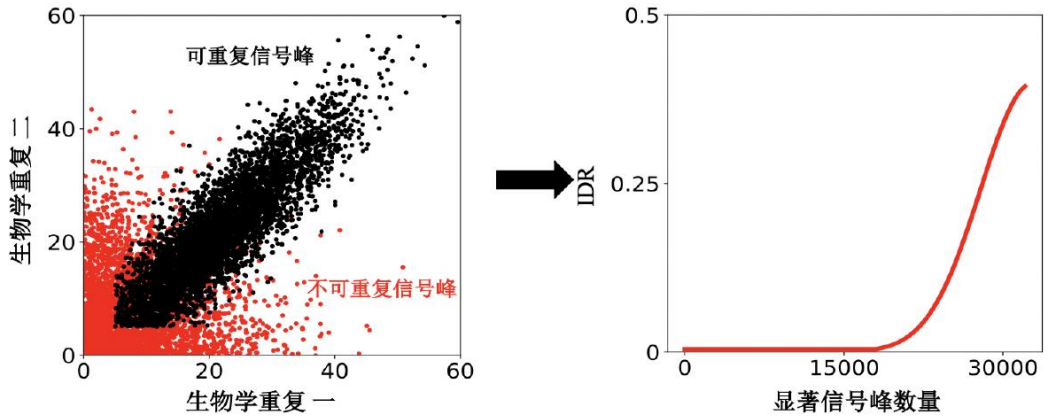


03. 转录因子相关测序数据分析

3.2 预测转录因子靶基因

信号峰层面质量控制

- 不可重复发现率（irreproducible discovery rate，IDR）
  - ✓ 用于衡量ChIP-seq信号峰在生物学重复之间的可重复性
  - ✓ 可以作为识别信号峰的阈值



不可重复发现率原理示意图

03. 转录因子相关测序数据分析

3.2 预测转录因子靶基因

注释层面的质量控制

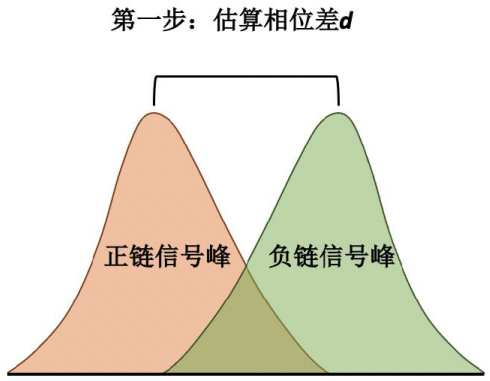
- 基因组分布特征
  - ✓ 高质量的转录因子ChIP-seq数据，信号峰富集于启动子和增强子区域
- 序列保守性分析
  - ✓ 转录因子结合位点倾向于在进化中保守
  - ✓ 高质量的转录因子ChIP-seq数据，信号峰顶点倾向具有更高的序列保守性
- DNA模体分析
  - ✓ 高质量的转录因子ChIP-seq数据，结合模体出现频率高，且倾向位于信号峰顶点

03. 转录因子相关测序数据分析

3.2 预测转录因子靶基因

识别ChIP-seq数据信号峰

- 常用方法包括MACS、CisGenome、SISSRs等
- MACS方法包括两个步骤：
  - ✓ 估计比对到正链的读长与比对到负链的读长之间的相位差*d*，以提高信号峰分辨率




```
bowtie2 -p 6 -x /path/to/index/genome_index \
-1 sample.R1.clean.fq.gz -2 sample.R2.clean.fq.gz \
-S sample.sam --local
```

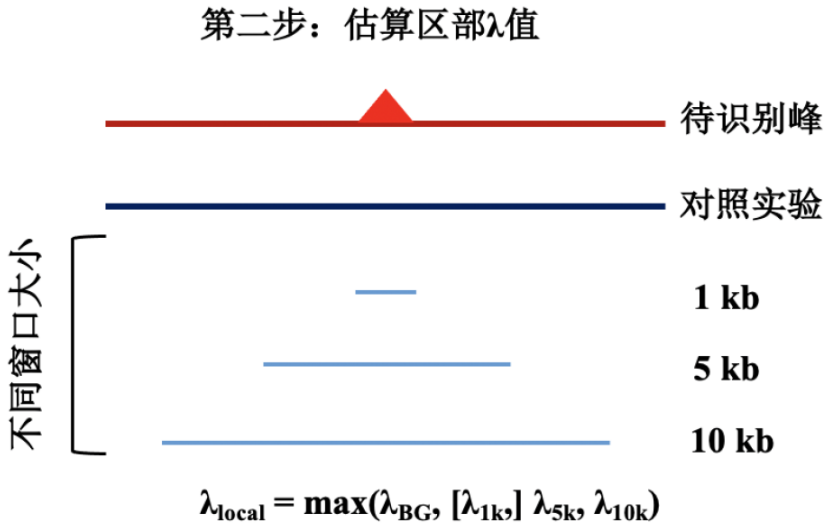
```
macs3 callpeak -t sample.bam -c control.bam -f BAMPE -n
sample_name -g hs -B -q 0.01
```

03. 转录因子相关测序数据分析

3.2 预测转录因子靶基因

识别ChIP-seq数据信号峰

- MACS方法包括两个步骤：
  - ✓应用动态泊松分布计算信号峰的统计显著性，以降低信号峰识别的假阳性率



03. 转录因子相关测序数据分析

3.2 预测转录因子靶基因

识别ChIP-seq数据差异信号峰

- 定性方法
  - ✓ 两套ChIP-seq数据分别用两个阈值识别信号峰
  - ✓ 差异信号峰：在一套数据中用较严格的阈值可以识别，在另一套数据中用较宽松的阈值仍不能识别的信号峰
- 定量方法
  - ✓ 在信号峰分别计算两套ChIP-seq数据的测序读长数目
  - ✓ 通过统计推断精确识别差异信号峰

```
library(DiffBind)
dba <- dba(sampleSheet = samples)
dba <- dba.count(dba)
dba <- dba.contrast(dba, categories=DBA_CONDITION)
dba <- dba.analyze(dba)
```

03. 转录因子相关测序数据分析

3.2 预测转录因子靶基因

预测转录因子靶基因

- 基于ChIP-seq信号峰与转录起始位点距离进行预测
  - ✓ 第一步：将信号峰与基因进行关联
  - ✓ 第二步：对信号峰设置权重，并对打分进行整合
- 联合使用转录组数据进行预测
  - ✓ 如果基因A的周围有转录因子B的结合位点，并且在敲除、敲降或过表达B时，A的转录水平发生了显著的变化，那么A很可能是B的靶基因

Epigenetic Sequencing

Epigenetic Sequencing

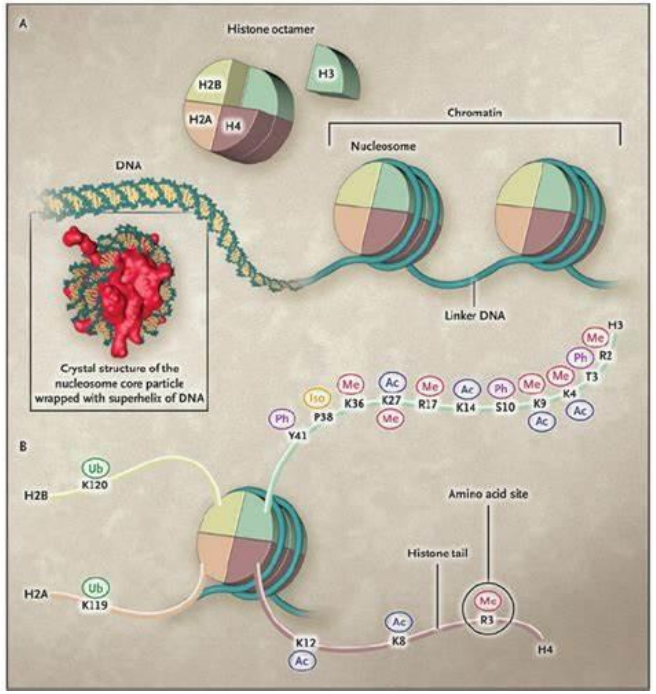
1. DNA-methylation (WGBS, RRBS, MeDIP-seq, single-cell WGBS)
2. Histone-modification

   (ChIP-seq, CUT&RUN, CUT&Tag, single-cell CUT&Tag)
3. Chromatin accessibility (ATAC-Seq, DNase-seq, NOMe-seq, single-cell ATAC-seq)
4. Chromatin Structure (3C-seq, Hi-C-seq, single-cell Hi-C)

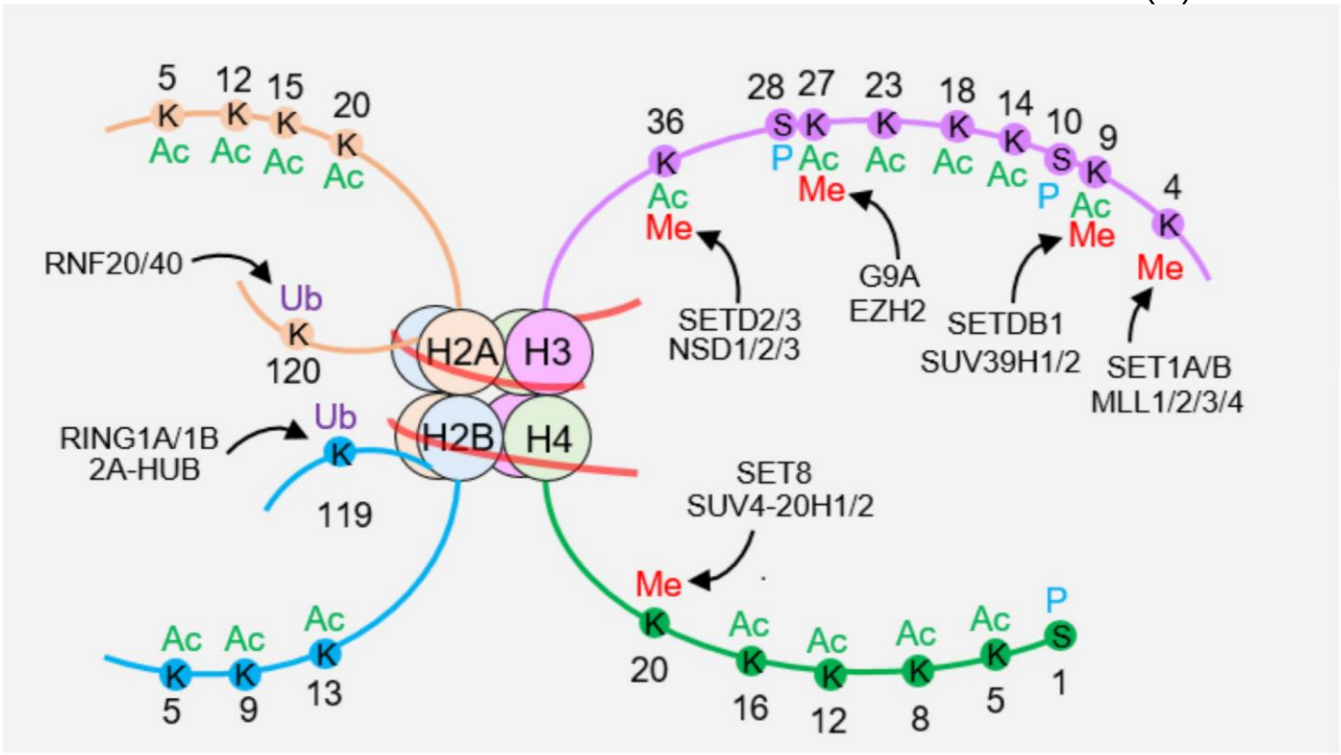
02 Histone Modification

Histone Modifications

Histone modifications, such as methylation, acetylation, phosphorylation, and ubiquitination, regulate gene expression and chromatin structure.



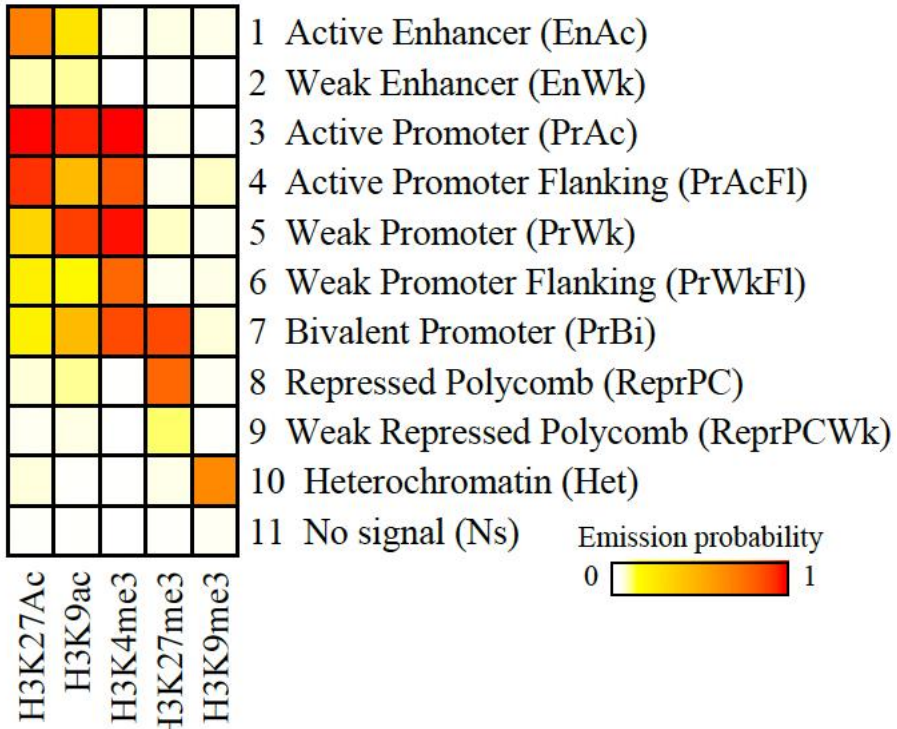
Lysine (K)
Serine (S)

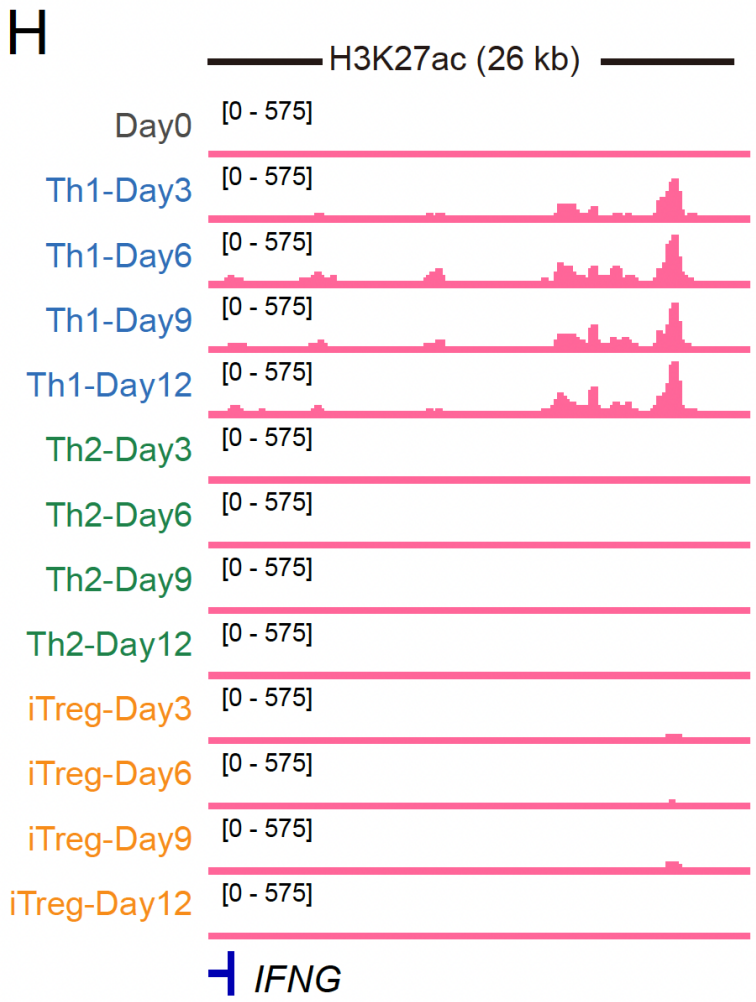


01. 基因顺式作用元件

1.2 真核生物的特异DNA序列

增强子 （enhancer）





02 Histone Modification

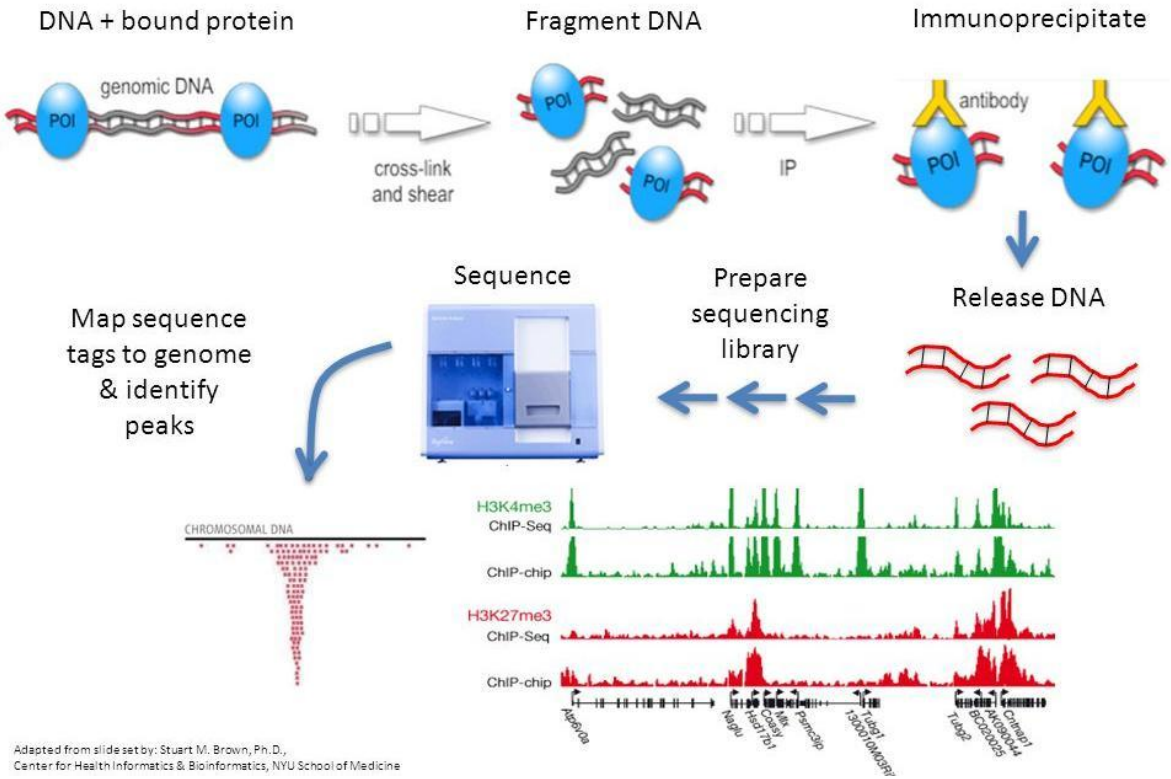
Histone Modifications

1. **ChIP-seq** (Chromatin Immunoprecipitation Sequencing)
2. **CUT&RUN** (Cleavage Under Targets and Release Using Nuclease)
3. **CUT&TAG** (Cleavage Under Targets and Tagmentation)

02 Histone Modification

2.1. ChIP-seq

Chromatin Immunoprecipitation Sequencing (ChIP-seq) is a powerful technique used to study protein-DNA interactions on a genome-wide scale. It combines chromatin immunoprecipitation (ChIP) with next-generation sequencing (NGS) to identify binding sites of DNA-associated proteins, such as transcription factors, histones, or other chromatin-modifying enzymes.



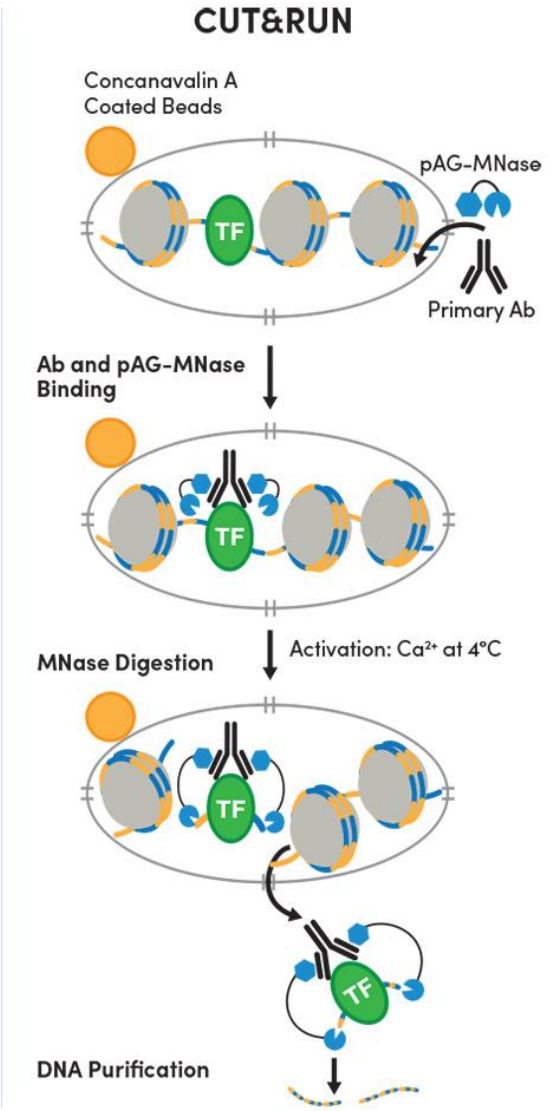
Workflow:

1.Chromatin crosslinking & shearing (via sonication or MNase digestion).
2.Immunoprecipitation (IP) with modification-specific antibodies.
3.Library preparation & sequencing (typically Illumina short reads).

02 Histone Modification

2.2. CUT&RUN

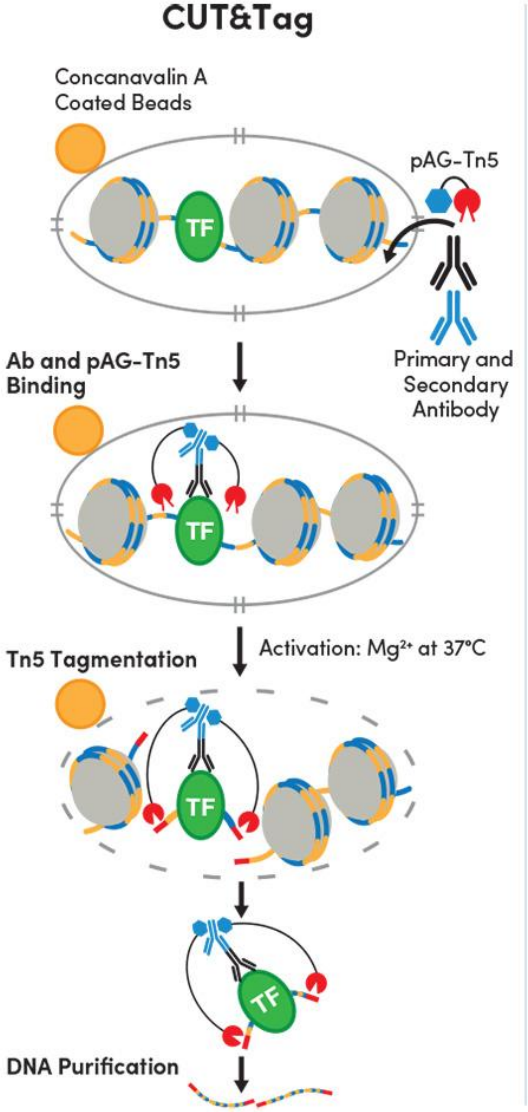
CUT&RUN (Cleavage Under Targets & Release Using Nuclease) : is an innovative, high-resolution method for mapping protein-DNA interactions in the genome. It is more efficient and requires fewer cells than traditional **ChIP-seq**, making it ideal for studying rare cell populations or low-abundance proteins.



02 Histone Modification

2.3. CUT&Tag

CUT&Tag (Cleavage Under Targets & Tagmentation) is an advanced epigenomic profiling technique that builds upon CUT&RUN, offering even higher sensitivity and scalability for mapping protein-DNA interactions. Instead of using Micrococcal Nuclease (MNase) like CUT&RUN, CUT&TAG employs a fusion protein of Protein A-Tn5 transposase (pA-Tn5) to simultaneously cleave and tag DNA at protein-binding sites.



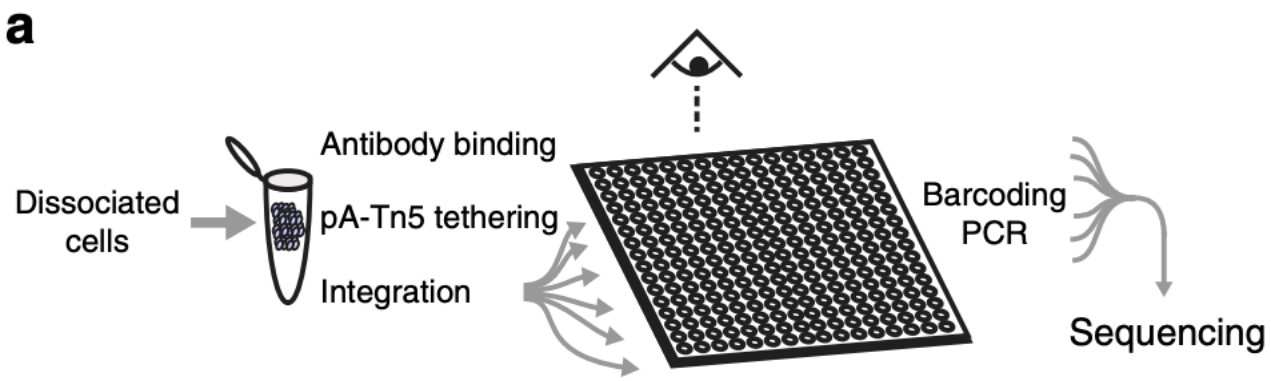
02 Histone Modification

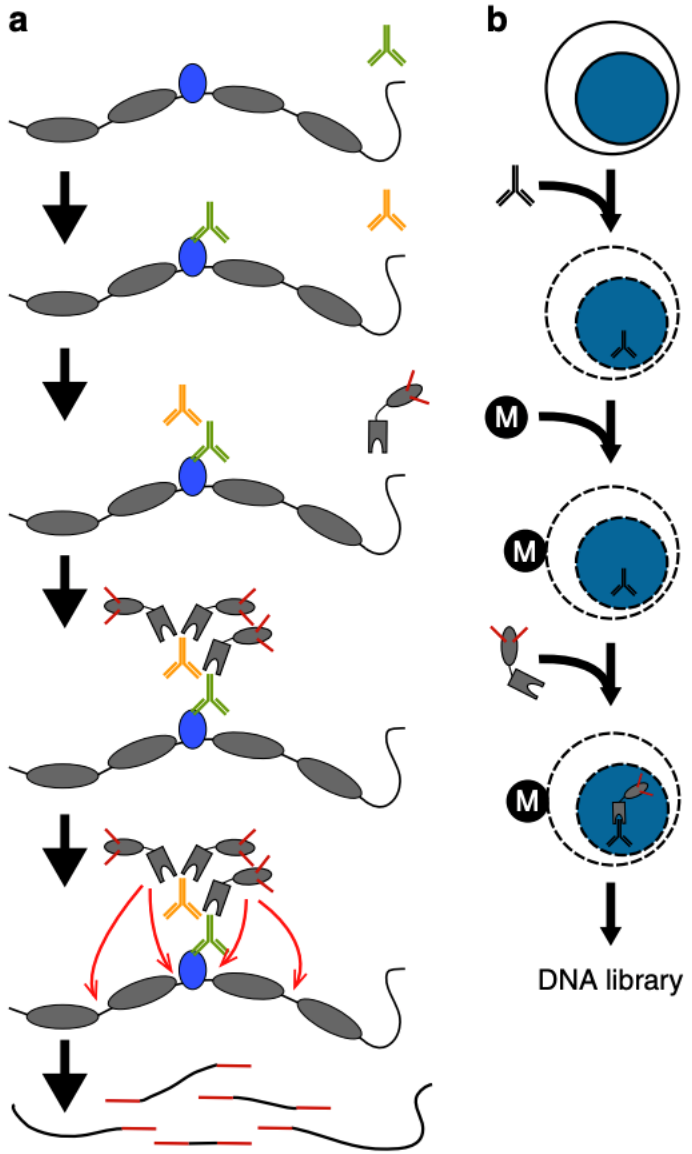
2.3. CUT&Tag

Method	Input (Cells/DNA)	Resolution	Advantages	Disadvantages
ChIP-seq	High (~1M cells)	~200 bp	Well-established, widely used	High background noise, requires crosslinking
CUT&RUN	Low (~100-1000 cells)	~300 bp	High signal-to-noise, no crosslinking	
CUT&TAG	Ultra-low (~10-100 cells)	~100 bp	Direct tagmentation, high resolution	Tn5 bias

02 Histone Modification

2.4. single cell CUT&Tag





02 Histone Modification

2.5. Data analyses

FASTQ → 质控 (FastQC, Trim Galore)
↓
比对到参考基因组 (bowtie2)
↓
去除PCR duplicate （samtools markdup)
↓
提取修饰peak (MACS2)

质控

```
fastqc sample_R1.fastq.gz sample_R2.fastq.gz -o qc/
multiqc qc/ -o qc_report/
```

测序质量、接头污染、序列重复水平、GC 偏差、过低质量的 bases。

02 Histone Modification

2.5. Data analyses

FASTQ → 质控 (FastQC, Trim Galore)
↓
比对到参考基因组 (bowtie2)
↓
去除PCR duplicate （samtools markdup)
↓
提取修饰peak (MACS2)

比对到参考基因组

```
bowtie2 -x hg38_index -1 trimmed_R1.fq.gz -2
trimmed_R2.fq.gz -S sample.sam --threads 8

samtools view -bS sample.sam > sample.bam
samtools sort -o sample.sorted.bam sample.bam
samtools index sample.sorted.bam
```

02 Histone Modification

2.5. Data analyses

FASTQ → 质控 (FastQC, Trim Galore)
↓
比对到参考基因组 (bowtie2)
↓
去除PCR duplicate （samtools markdup)
↓
提取修饰peak (MACS2)

去除PCR duplicate

```
picard MarkDuplicates I=sample.sorted.mapq30.bam
O=sample.dedup.bam M=dup_metrics.txt REMOVE_DUPLICATES=true
# 去除 chrM
samtools idxstats sample.dedup.bam | cut -f1 | grep -v chrM | xargs
samtools view -b sample.dedup.bam > sample.filtered.bam
# 排除 blacklist
bedtools intersect -v -abam sample.filtered.bam -b hg38_blacklist.bed >
sample.clean.bam
samtools index sample.clean.bam
```

02 Histone Modification

2.5. Data analyses

FASTQ → 质控 (FastQC, Trim Galore)
↓
比对到参考基因组 (bowtie2)
↓
去除PCR duplicate （samtools markdup)
↓
提取修饰peak (MACS2)

Peak calling

窄峰（H3K4me3、H3K27ac 等）：MACS2 callpeak --nomodel --extsize/--shift 或默认模式。

宽峰（H3K27me3、H3K36me3）：MACS2 --broad、SICER、epic2

```
macs2 callpeak -t ChIP.bam -c Input.bam -f BAM -g hs -n sample_H3K27ac -q 0.01 --outdir macs2_out
```

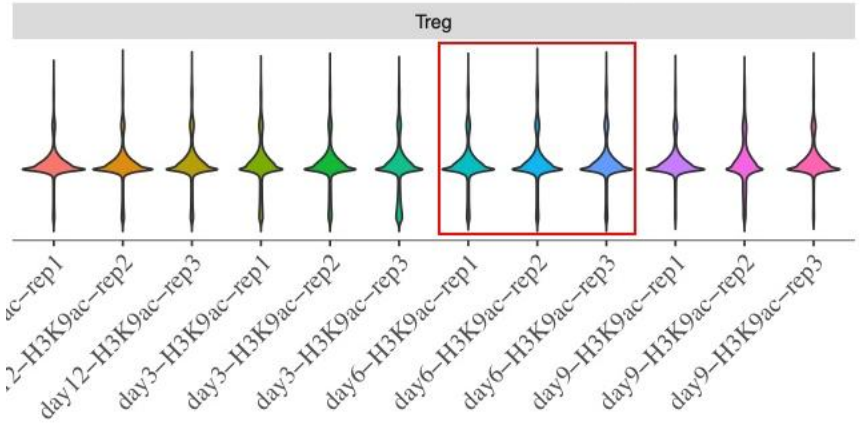
```
macs2 callpeak -t ChIP.bam -c Input.bam -f BAM -g hs --broad -n sample_H3K27me3 --broad-cutoff 0.1
```

02 Histone Modification

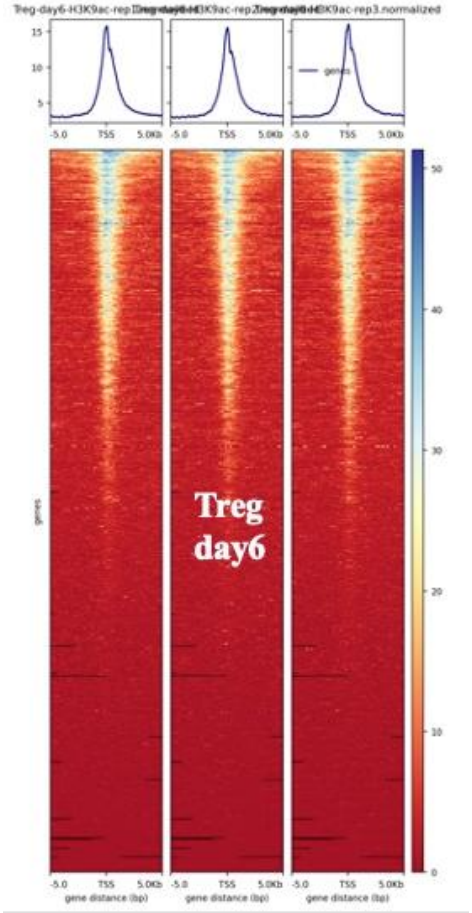
2.5. Data analyses

质控

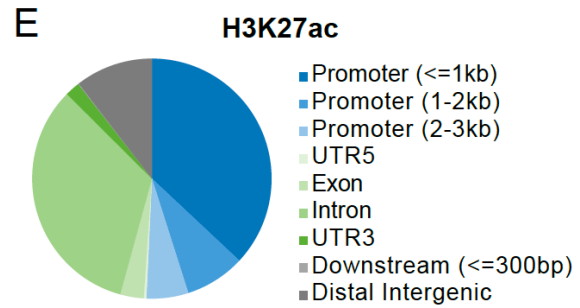
Peak 长度



Peak 分布



Peak 分布

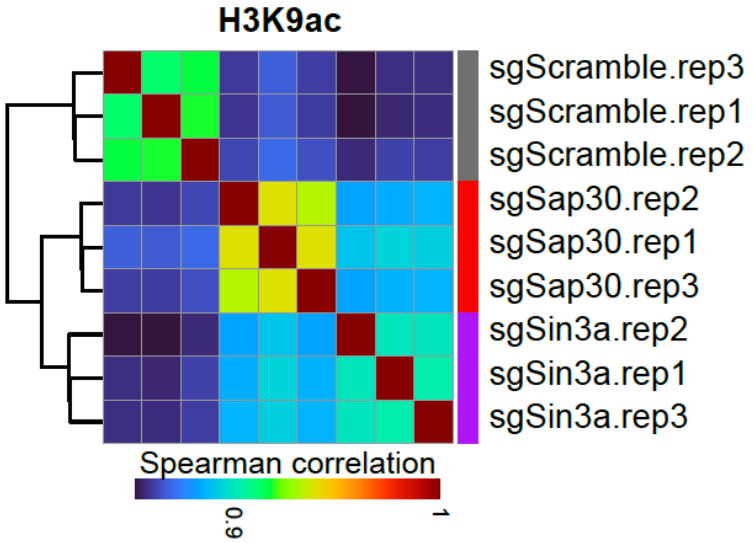


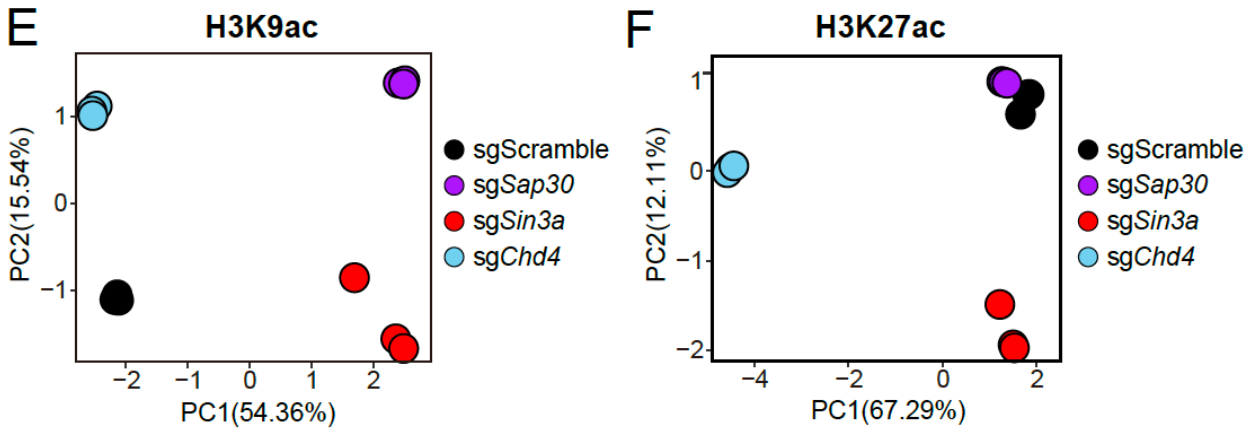
02 Histone Modification

2.5. Data analyses

样品聚类

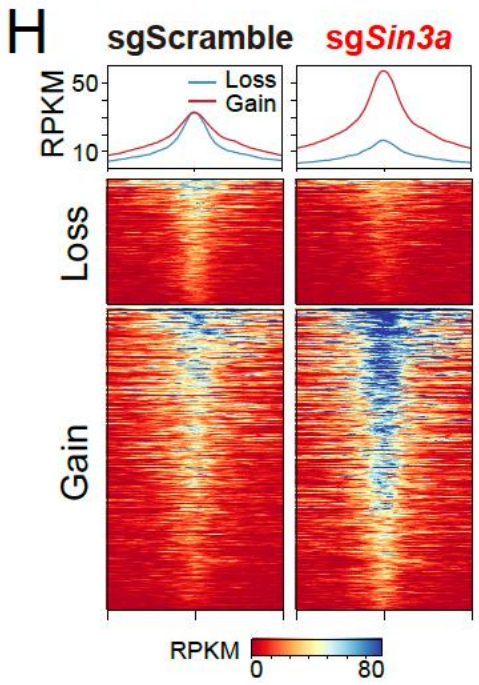
按照peak 或 promoter 修饰程度均可聚类



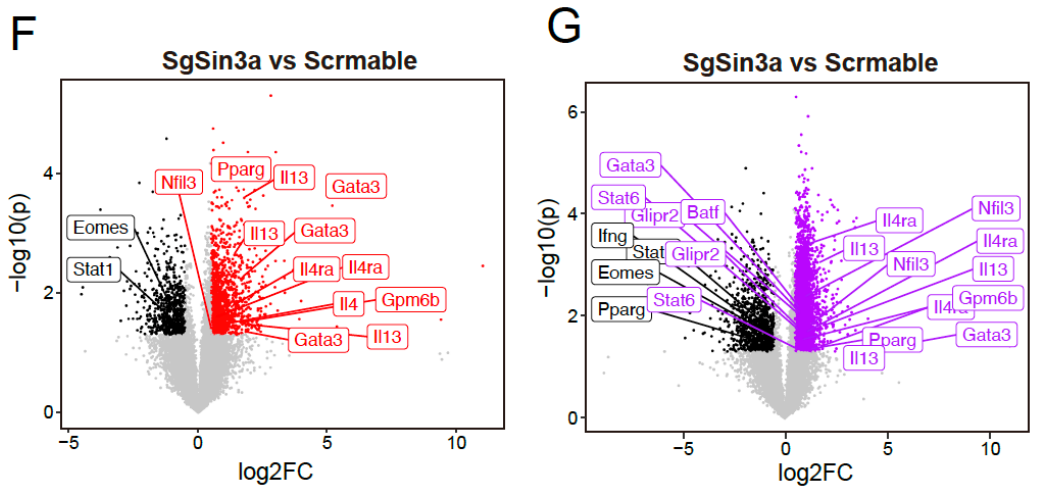


2.5. Data analyses

差异peak



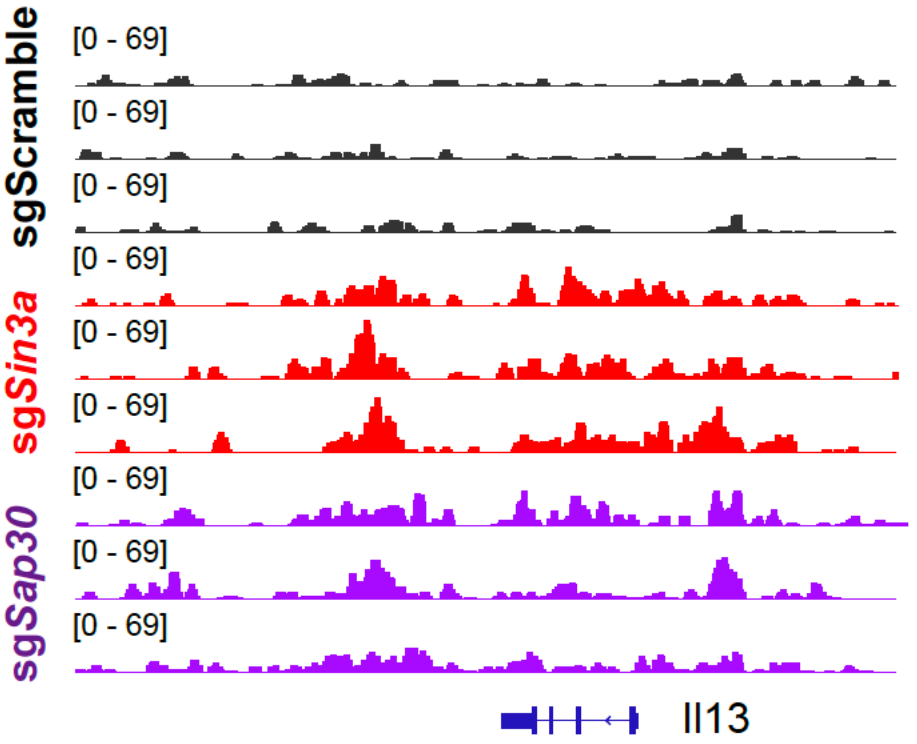
差异peak注释



02 Histone Modification

2.5. Data analyses

IGV 可视化



Thank you

Yu Hou Zhejiang University 2025