

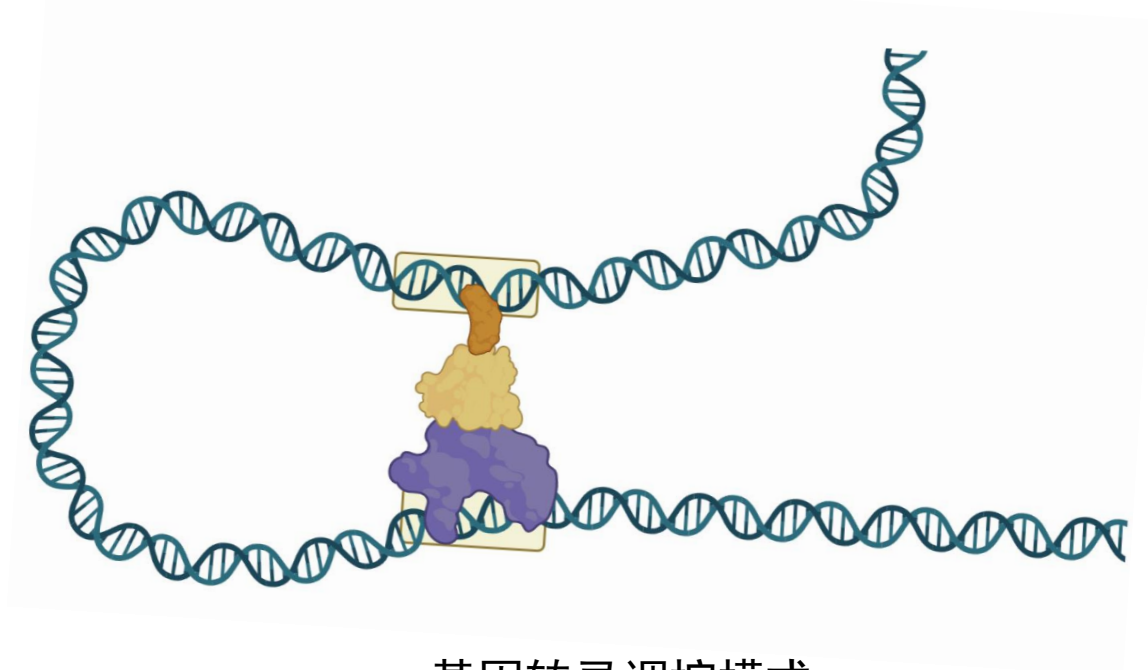
第八章 转录调控与表观遗传

张勇

同济大学

- 转录调控
 - 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
- 表观遗传
 - DNA甲基化组学数据分析
 - 组蛋白修饰组学数据分析
 - 三维基因组学数据分析

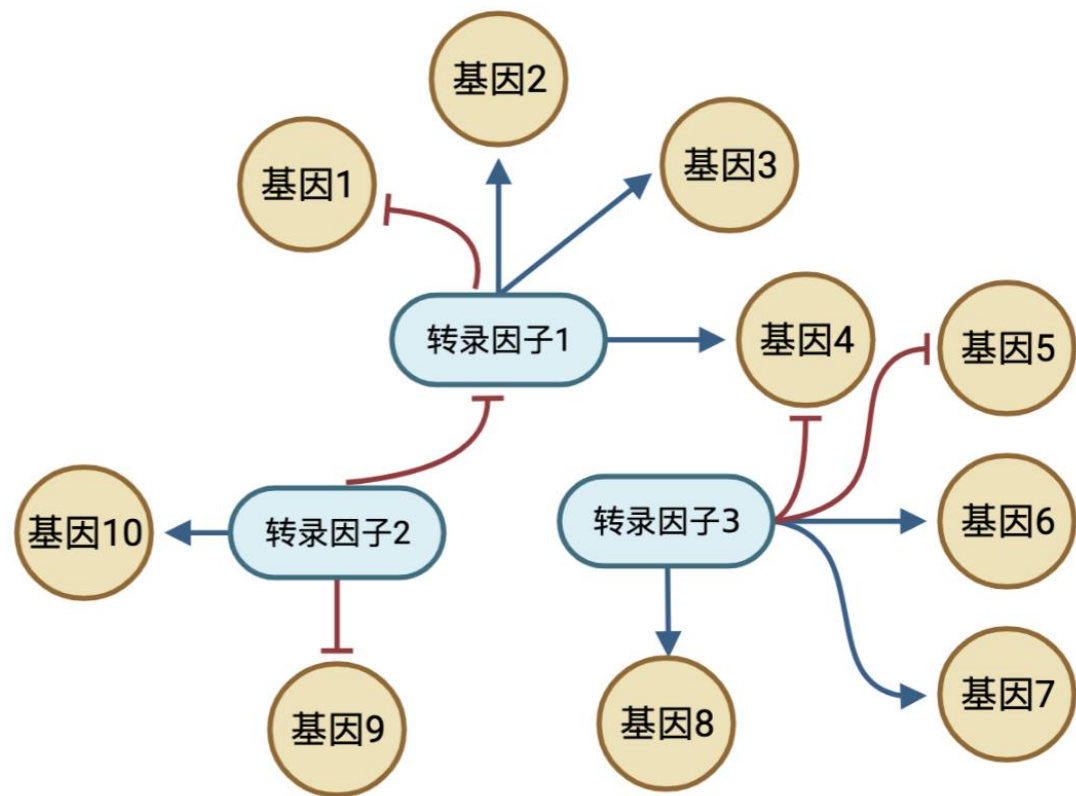
- 转录是基因表达过程的第一步，也是调控基因活性的核心步骤
- 转录因子 / 反式作用因子
- 启动子、增强子 / 顺式调控元件



基因转录调控模式

➤ 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

- ✓ 基于转录因子结合模体预测结合位点
- ✓ 基于ChIP-seq数据解析结合位点



基因转录调控网络

- 很多转录因子结合位点具有特定的DNA序列模式，称为转录因子结合模体（binding motif）
- 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	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

➤ IUPAC简并码

IUPAC简并码	碱基	IUPAC简并码	碱基
W	A 或 T	B	C、G 或 T
R	A 或 G	D	A、G 或 T
K	G 或 T	H	A、C 或 T
S	C 或 G	V	A、C 或 G
Y	C 或 T	N	A、C、G 或 T
M	A 或 C		

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

```

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



$$\begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 1 & 1 & 3 & 2 & \dots & 1 \\ 6 & 4 & 1 & 7 & \dots & 3 \\ 1 & 3 & 5 & 2 & \dots & 6 \\ 3 & 3 & 2 & 0 & \dots & 1 \end{bmatrix}$$



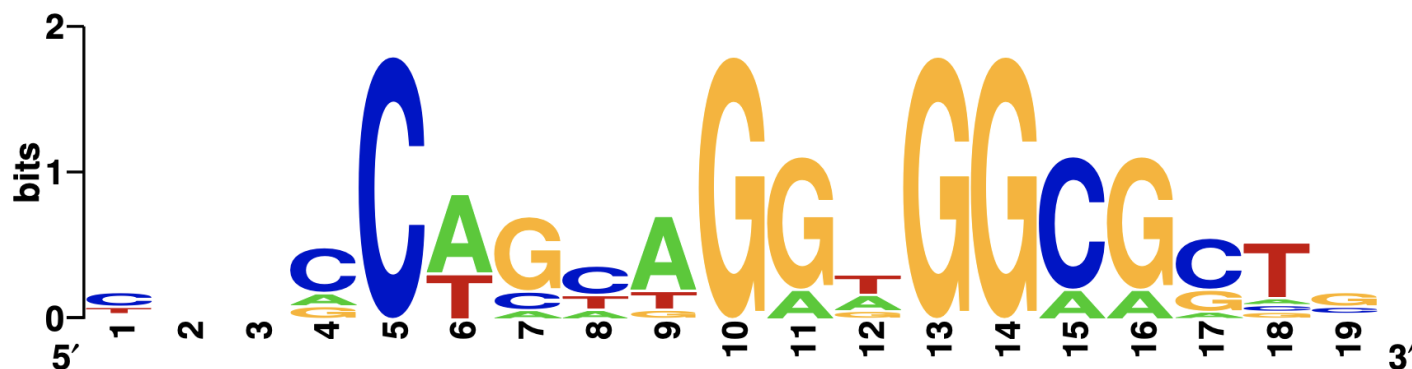
$$\begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 0.09 & 0.09 & 0.27 & 0.18 & \dots & 0.09 \\ 0.55 & 0.36 & 0.09 & 0.64 & \dots & 0.27 \\ 0.09 & 0.27 & 0.45 & 0.18 & \dots & 0.55 \\ 0.27 & 0.27 & 0.18 & 0.00 & \dots & 0.09 \end{bmatrix}$$

位置频率矩阵

➤ 表示方法3：序列标识图（sequence logo）

✓ 序列标识图第 j 位上某个碱基 i 的高度 $height_{i,j} = q_{i,j} \times R_j$

- 其中， $R_j = 2 - (H_j + e_n)$
- H_j 是位置 j 处的信息熵： $H_j = -\sum_{i=1}^4 q_{i,j} \times \log_2 q_{i,j}$
- e_n 是针对小样本的近似矫正： $e_n = \frac{1}{\ln 2} \times \frac{3}{2n}$

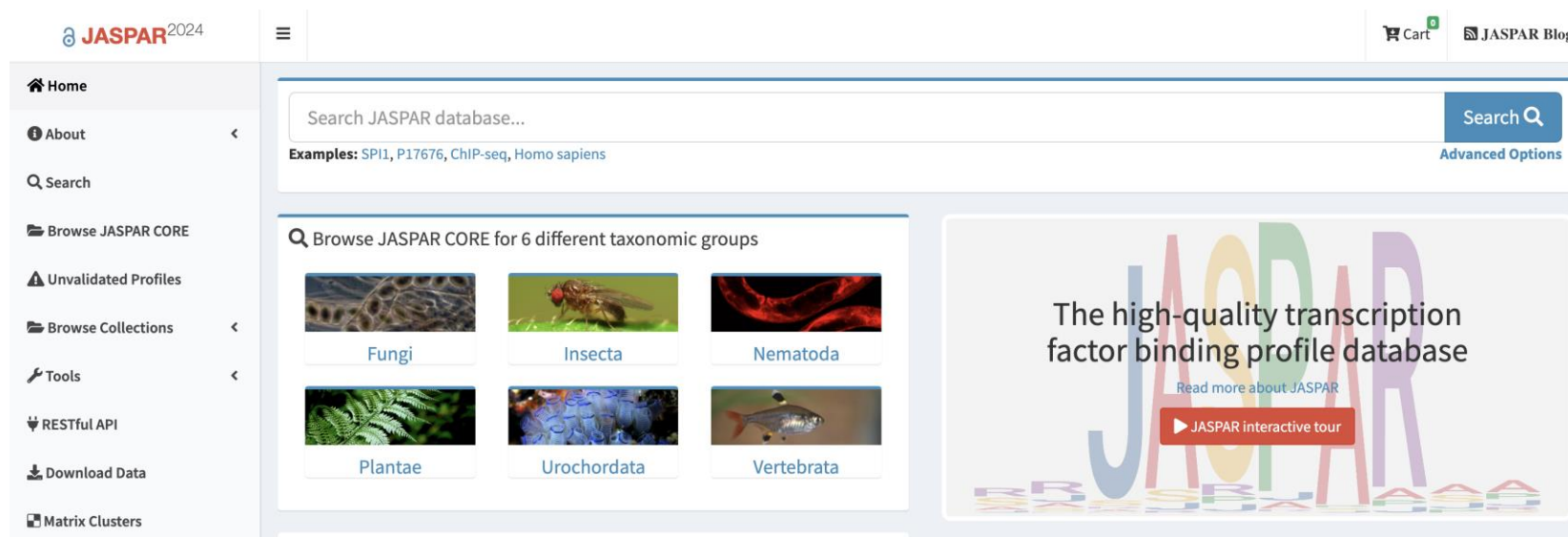


序列标识图

转录因子结合模体数据库

➤ JASPAR (<https://jaspar.elixir.no/>)

✓ 2004年发布第一版；目前为2024年更新版



JASPAR数据库主页

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。

位置频率矩阵

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

b_i 是碱基 i 在 DNA 序列中出现频率

位置权重矩阵

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

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

	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

➤ 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
- ✓ 基于共有序列的方法
- ✓ 基于位置频率矩阵的方法
 - 基于EM算法的识别方法
 - 基于吉布斯抽样法的识别方法

基于共有序列的结合模体从头发现

- 穷举所有可能的序列组合，得到具有统计显著性的短片段模式
 - ✓ 穷举策略的计算复杂度为 4^L ，不适用于片段长度 L 较大的情况
 - ✓ MobyDick方法在此基础上，应用启发式策略，只将序列中出现的片段作为候选序列，降低了计算量

基于EM算法的结合模体从头发现

➤ EM算法是一种迭代算法

- ✓ E-步骤（期望步骤）：观察数据和现有模型来估计参数，并用估计的参数值来计算似然函数的期望值
- ✓ M-步骤（最大化步骤）：寻找似然函数最大化时对应的参数
- ✓ EM算法可以保证在每次迭代之后似然函数增加

基于EM算法的结合模体从头发现

➤ E-步骤:

✓ 给定序列：CCGGCAGCGGGTGGCGCTG

✓ 假设转录因子结合模体的长度为9，位置频率矩阵为：

A	0.261	0.028	0.499	0.956	0.005	0.095	0.075	0.002	0.710
C	0.005	0.002	0.194	0.004	0.670	0.004	0.017	0.966	0.114
G	0.678	0.930	0.005	0.030	0.202	0.003	0.905	0.011	0.067
T	0.056	0.040	0.302	0.010	0.123	0.899	0.003	0.021	0.109

✓ 对给定序列中的每个长度为9的片段，计算似然比（以第3个片段为例）：

$$LR_3 = \frac{P(\text{GGCAGCGGG} \mid \theta)}{P(\text{GGCAGCGGG} \mid \theta_0)} \quad \begin{array}{l} \leftarrow \theta: \text{位置频率矩阵,} \\ \leftarrow \theta_0: \text{四种碱基在基因组上出现的概率, 此处假设均为0.25} \end{array}$$
$$= \frac{0.678 \times 0.930 \times 0.194 \times 0.956 \times 0.202 \times 0.004 \times 0.905 \times 0.011 \times 0.067}{0.25 \times 0.25 \times 0.25 \times 0.25 \times 0.25 \times 0.25 \times 0.25 \times 0.25 \times 0.25} = 0.0166$$

基于EM算法的结合模体从头发现

➤ M-步骤:

- ✓ 以E-步骤得到的似然比为权重，计算结合模体各个位置不同碱基得分，得到新的矩阵。
- ✓ 例如，第1、2、5、8个短片段的第一个位置均为C，那么在结合模体第一个位置上C的分数为：

$$T_1 = \frac{LR_1 + LR_2 + LR_5 + LR_8}{\sum_1^{11} LR_i}$$

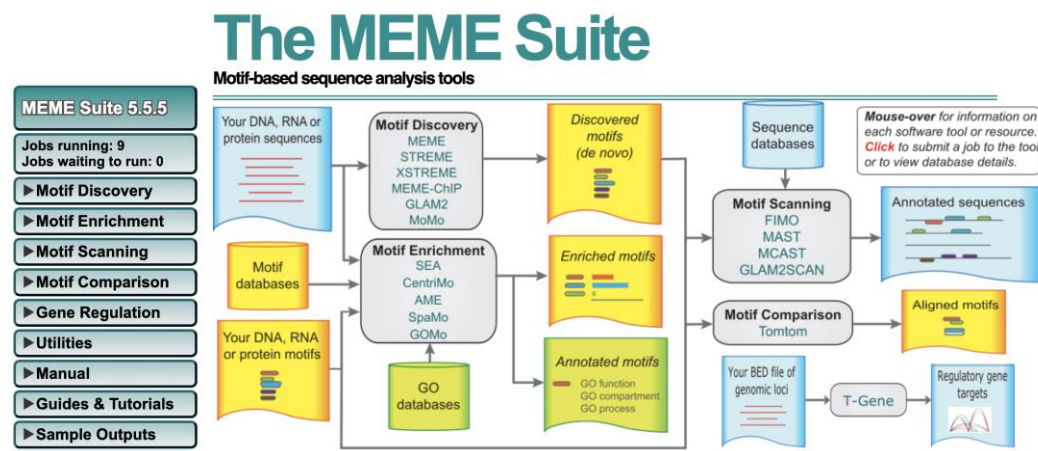
➤ 迭代

- ✓ 将M-步骤得到的新位置频率矩阵作为下一轮E-步骤的输入
- ✓ 最终位置频率矩阵趋于收敛，即为结合模体

基于EM算法的结合模体从头发现

➤ MEME方法改进了EM算法

- ✓ 遍历所有可能的起始矩阵，筛选出来具有统计学显著性的矩阵输入给EM算法，然后通过循环迭代得到最优解
- ✓ MEME系列软件（<https://meme-suite.org/meme>）包括网页版和本地版



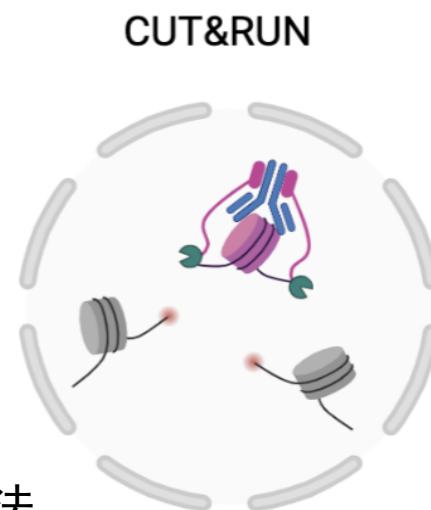
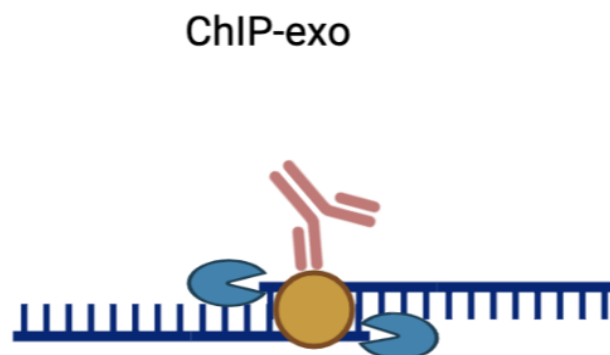
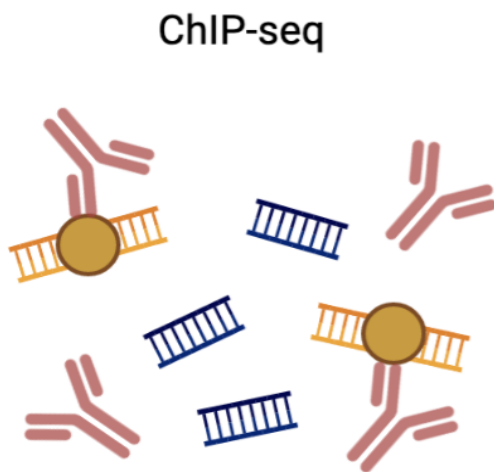
MEME系列软件网页版界面

基于吉布斯抽样法的结合模体从头发现

- 吉布斯抽样法通过随机采样，更新结合模体的位置频率矩阵和在各序列中匹配的位置：
 - ✓ 从多条DNA序列中选出一条序列 S_1 ，随机从剩余的每条DNA序列选取给定长度为 n 的片段，得到一个位置频率矩阵
 - ✓ 基于上述位置频率矩阵对序列 S_1 的每个长度为 n 的片段计算似然比
 - ✓ 按似然比从序列 S_1 中随机取一段长度为 n 的片段，更新位置频率矩阵
 - ✓ 进行多轮迭代，直至确定结合模体在序列 S_1 上的匹配位置
 - ✓ 将序列 S_1 放回，将序列 S_2 取出，进行同样的操作
 - ✓ 依次对每一条DNA序列进行如上操作，确定每条序列上结合模体匹配的位置，并得到结合模体

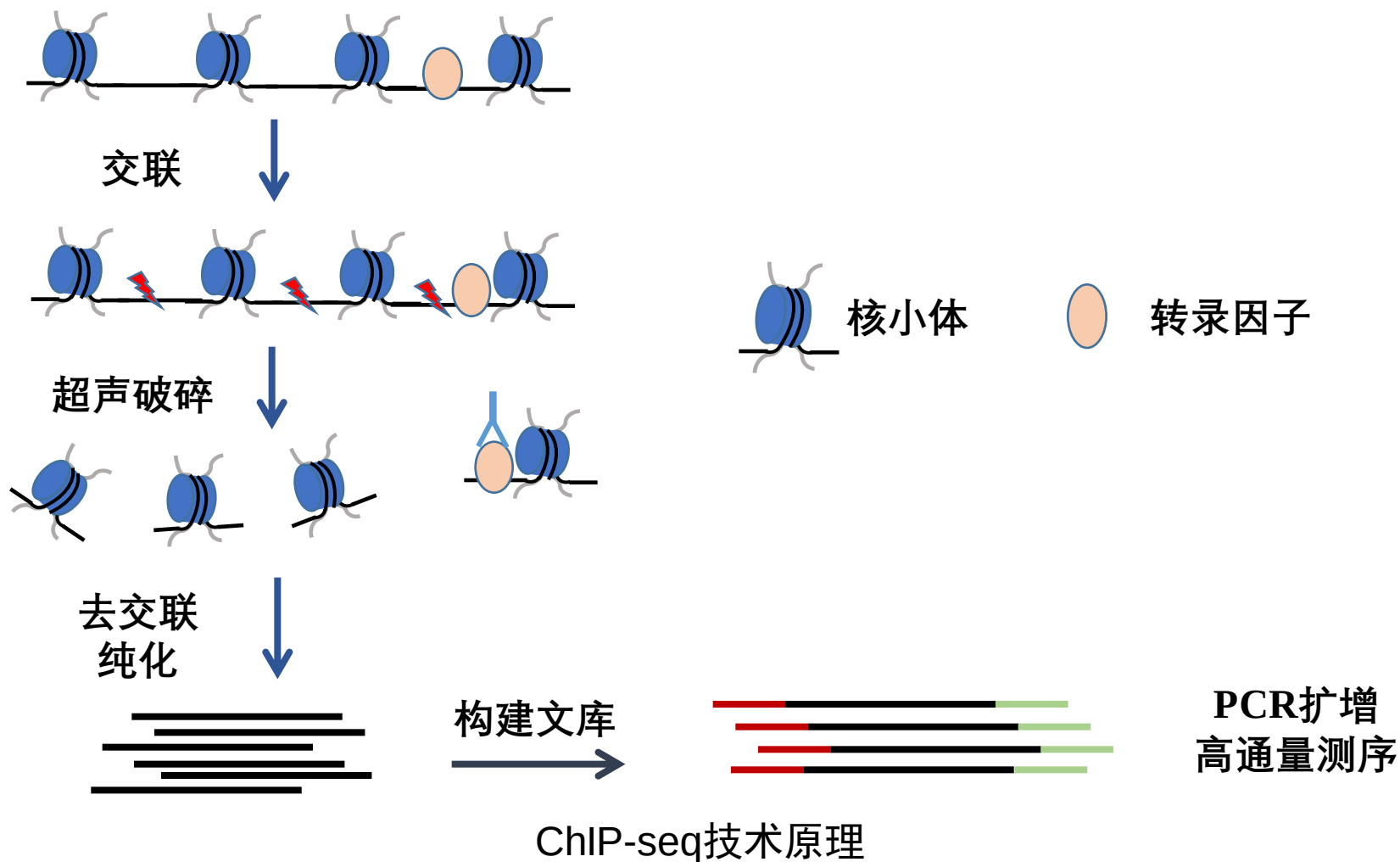
识别转录因子全基因组结合位点的技术

- ChIP-seq是最常用的识别转录因子全基因组结合位点的技术
- ChIP-exo具有分辨率高的优势
- CUT&RUN具有适用于稀缺样本的优势



识别转录因子全基因组结合位点的代表性方法

转录因子ChIP-seq技术原理



转录因子ChIP-seq实验设计

- 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
- 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测序读长
- 需要进行生物学重复和对照实验以增加可靠度

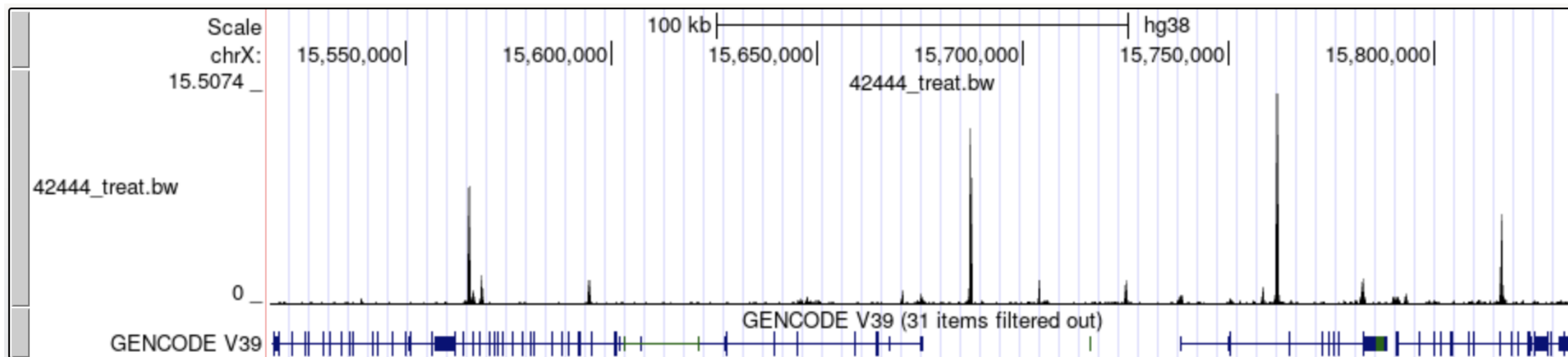
转录因子ChIP-seq数据质量控制

- 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

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



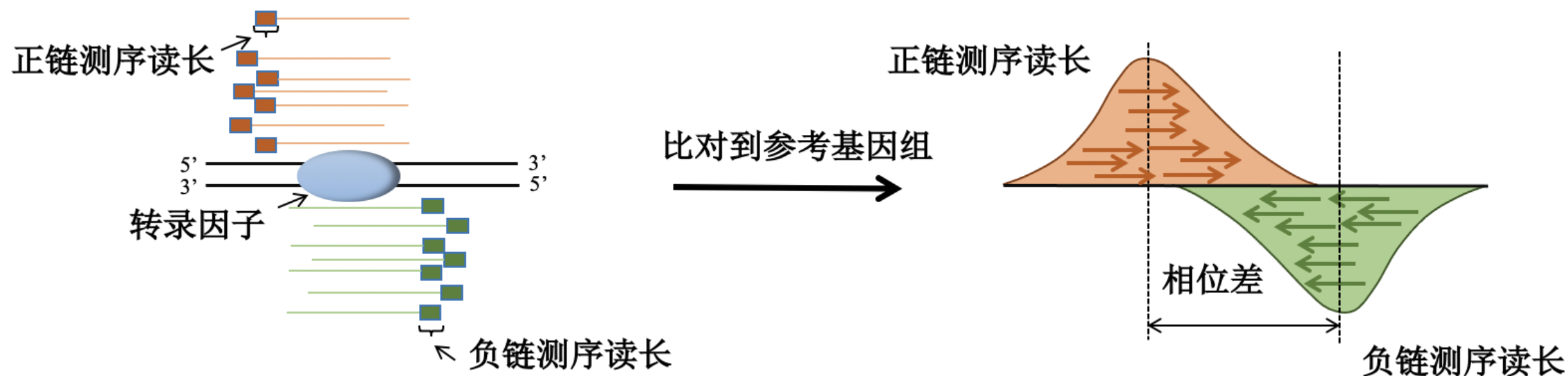
ChIP-seq信号可视化示例

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

信号峰层面质量控制

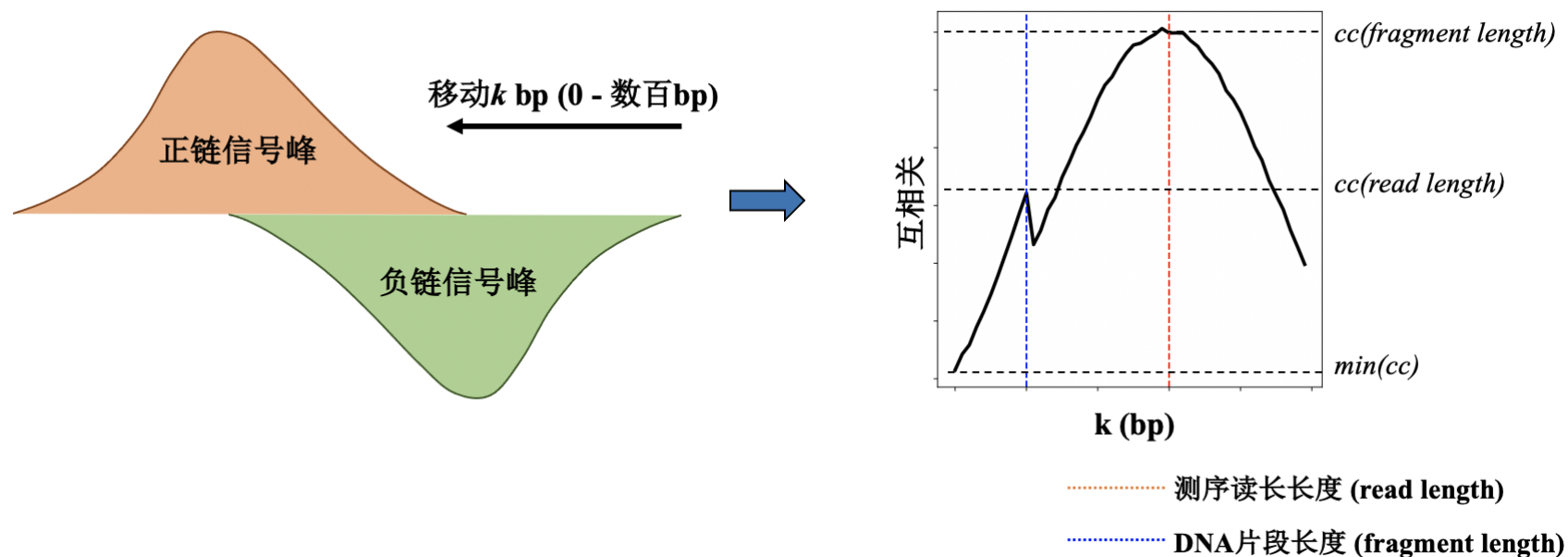
- 在转录因子ChIP-seq数据的信号峰区域，比对到正链的读长与比对到负链的读长之间会产生一个相位差



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



信号峰层面质量控制

➤ 互相关 (cross-correlation)

- ✓ 质量较好的ChIP-seq数据，上述曲线会出现两个峰值，对应的k分别为测序读长长度 (read length) 和ChIP实验得到的DNA片段长度 (fragment length)

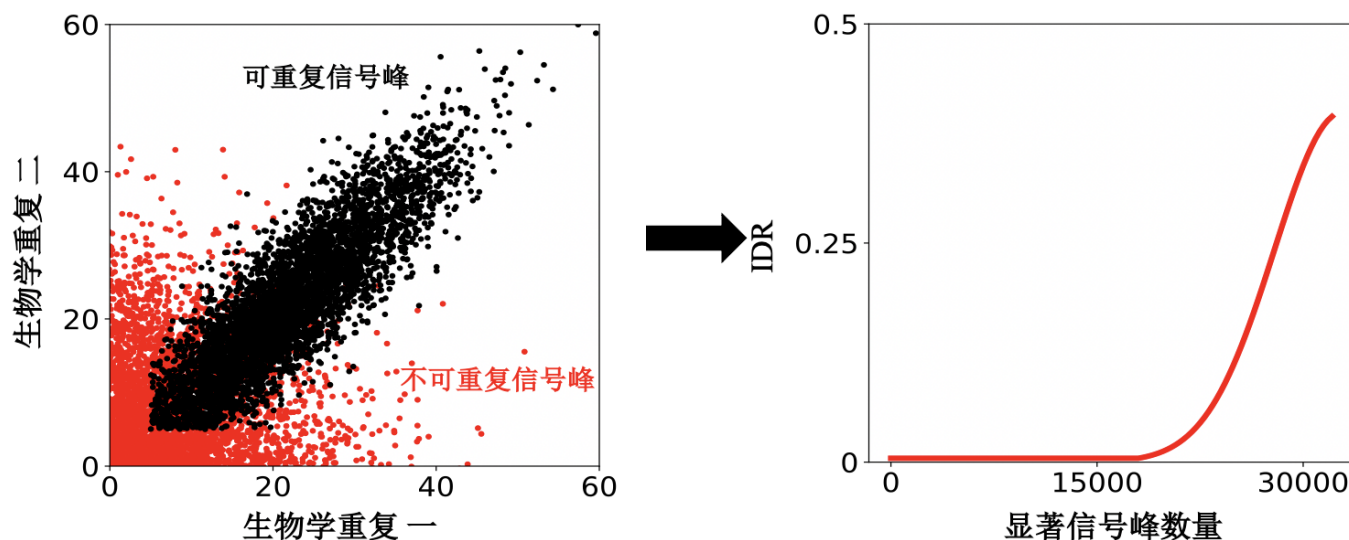
- ✓
$$NSC = \frac{CC(fragment\ length)}{min(cc)}$$

- ✓
$$RSC = \frac{CC(fragment\ length) - min(cc)}{CC(read\ length) - min(cc)}$$

- ✓ NSC、RSC可指征测序读长在信号峰的富集程度

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

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

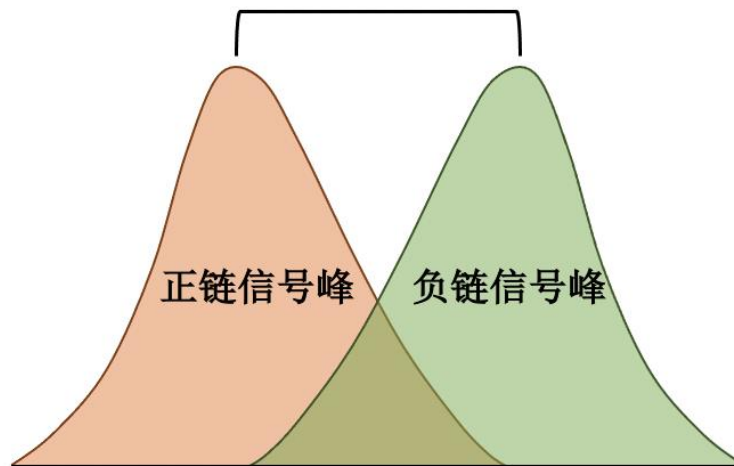
➤ DNA模体分析

- ✓ 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

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

第一步：估算相位差 d

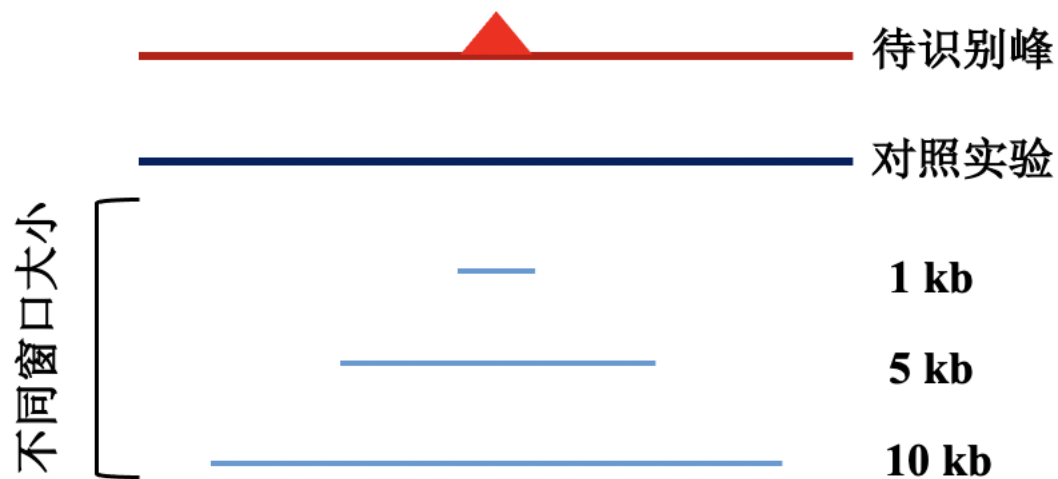


识别ChIP-seq数据信号峰

➤ 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

第二步：估算区部 λ 值



$$\lambda_{\text{local}} = \max(\lambda_{\text{BG}}, [\lambda_{1\text{k}}, \lambda_{5\text{k}}, \lambda_{10\text{k}}])$$

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

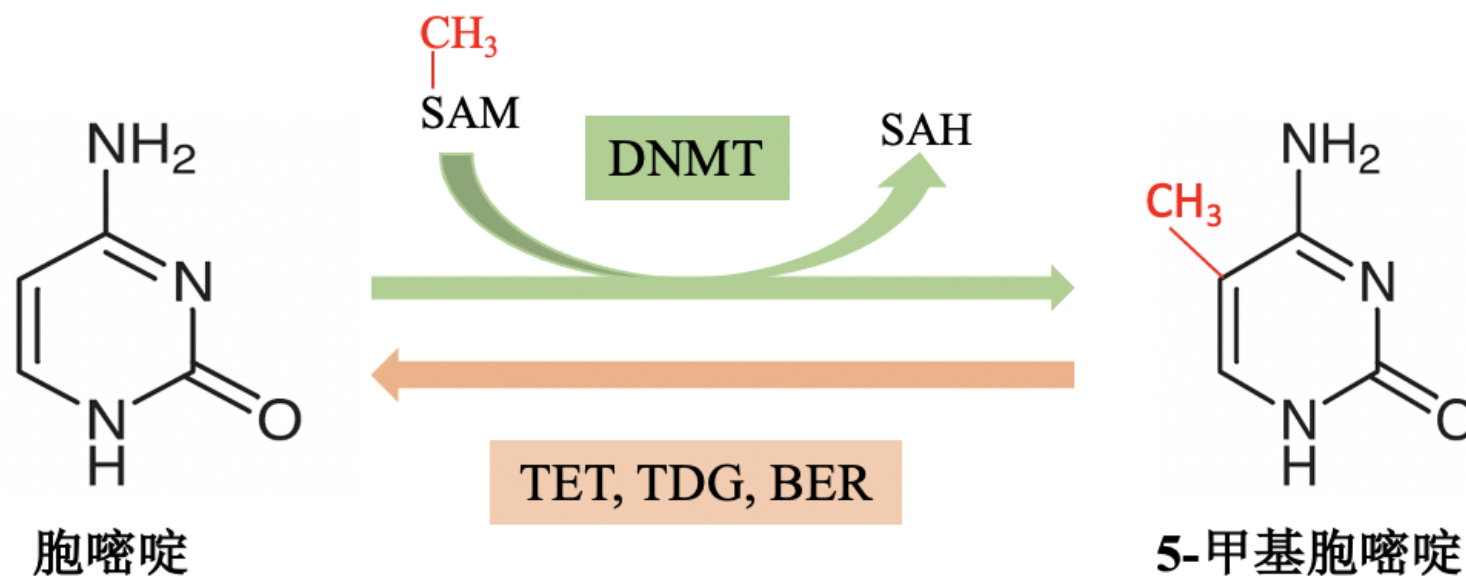
预测转录因子靶基因

- 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

- 转录调控
 - 转录因子结合模体表示方法
 - DNA共有序列、位置频率矩阵、序列标识图
 - 转录因子结合模体从头发现
 - 基于共有序列、基于EM算法、基于吉布斯抽样法
 - 转录因子ChIP-seq数据分析
 - 技术原理、质量控制、分析要点及工具

- 表观遗传组学：在全基因组水平上解析染色质修饰与结构对染色质环境及基因表达调控功能的影响
 - ✓ 表观遗传类型多样
 - ✓ 组学技术原理各异
 - ✓ 分析方法工具不同
- 组学技术决定数据特点，数据特点决定分析思路

DNA甲基化



DNA甲基化组学技术

➤ 亚硫酸盐转化方法

- ✓ WGBS是应用最为广泛的DNA甲基化组学技术

➤ 限制性内切酶方法

- ✓ 利用对DNA甲基化状态敏感的限制性内切酶切割DNA
- ✓ MRE-seq、McrBC-seq、HELP-seq、Methyl-seq等

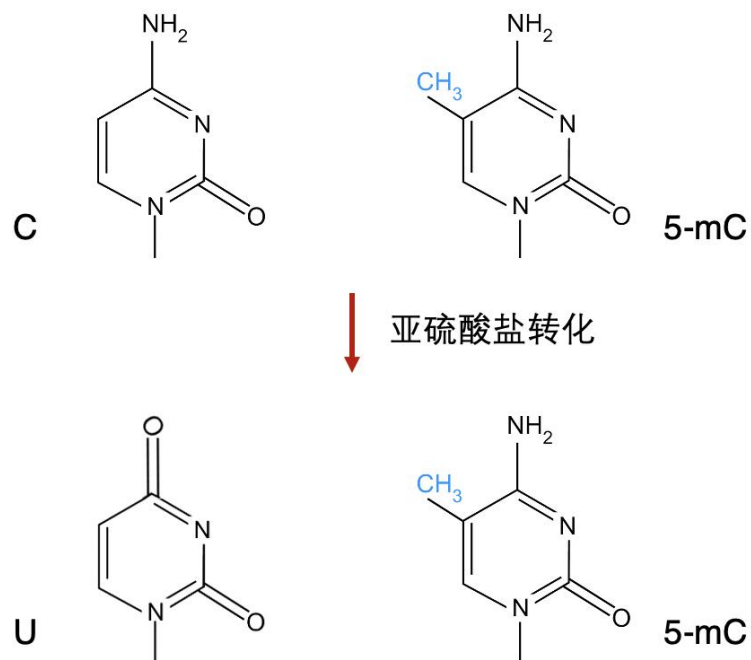
➤ 亲和纯化方法

- ✓ 利用特异性结合DNA甲基化的蛋白进行DNA富集操作
- ✓ MeDIP-seq、MBD-seq等

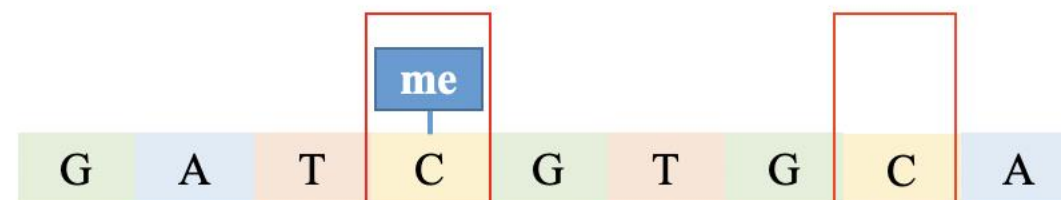
➤ 纳米孔测序方法

- ✓ Oxford Nanopore Technologies (ONT) 技术通过电流强度差异检测DNA甲基化

全基因组亚硫酸盐测序 (WGBS)



DNA 序列



亚硫酸盐转化



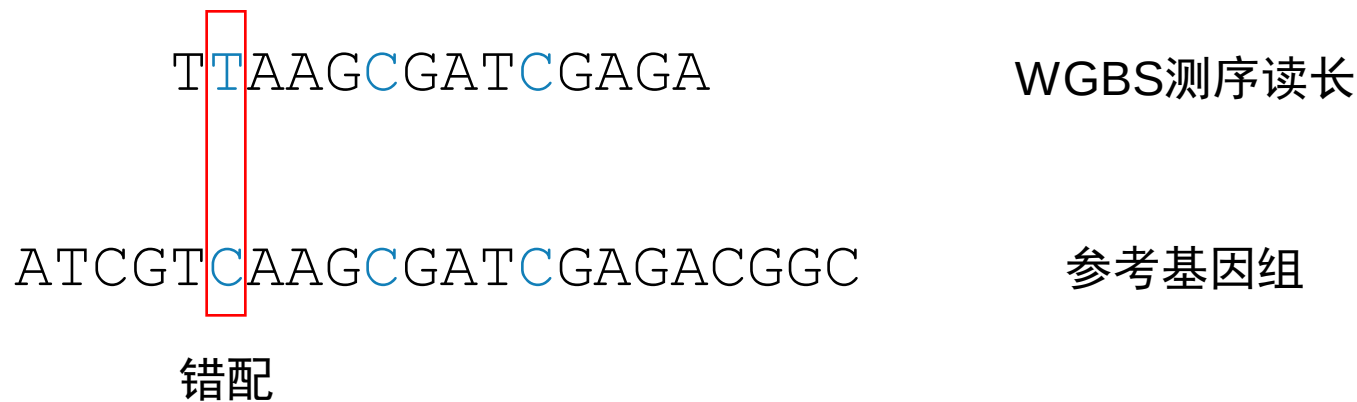
测序结果



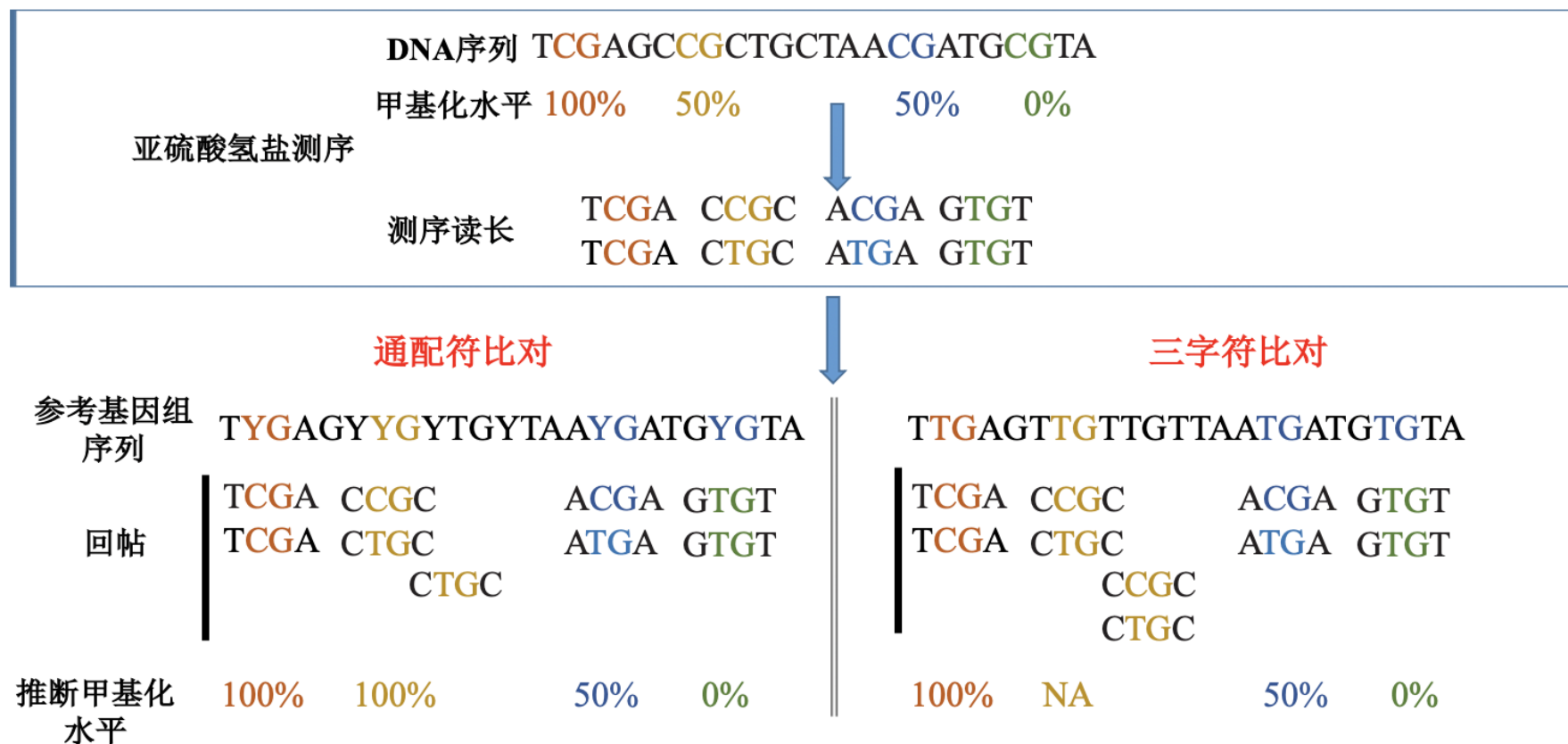
WGBS数据比对

➤ 数据比对困难：

- ✓ 亚硫酸盐转化将未甲基化的C转化为U，并在文库构建过程中转化为T
- ✓ 测序读长中的T与参考基因组中的C错配



WGBS数据比对



WGBS数据比对

通配符比对策略

- 用通配符Y在参考基因组上代替C
 - ✓ 能够比对更多测序读长
 - ✓ 部分位点上推断的甲基化水平较实际值偏高
 - ✓ 代表性算法：BSMAP

三字符比对策略

- 将参考基因组上所有C替换为T
 - ✓ 比对成功率偏低
 - ✓ 不会给甲基化水平推断带来偏差
 - ✓ 代表性算法：Bismark

DNA甲基化水平推断

- 一个CpG位点的DNA甲基化水平：该位点在细胞群体中发生DNA甲基化的比例。

参考基因组序列	C	A	T	G	C	A	C	G	G	C	T	A	T
测序读长			T	G	T	A	C	G	G	T	T		
			T	G	T	A	T	G	G	T	T		
			T	G	T	A	C	G	G	C	T		
			T	G	C	A	C	G	G	C	T		
			T	G	T	A	C	G	G	T	T		
					$1/5 = 0.2$		$4/5 = 0.8$			$2/5 = 0.4$			
					甲基化水平0.2		甲基化水平0.8			甲基化水平0.4			

DNA甲基化水平推断

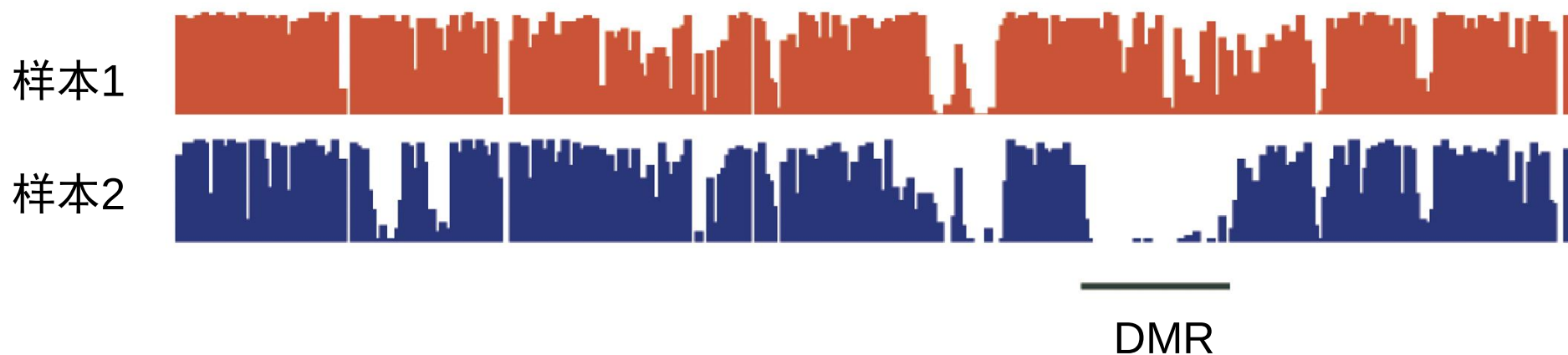
➤ 影响DNA甲基化水平推断因素

- ✓ DNA片段末端修复
- ✓ 亚硫酸盐转化效率
- ✓ DNA文库测通
- ✓ 3'端测序质量下降
- ✓ 文库的不均衡扩增

差异甲基化区域

➤ DMR (differentially methylated region)

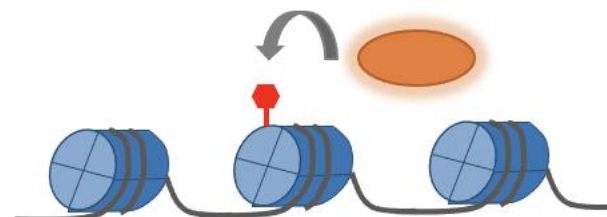
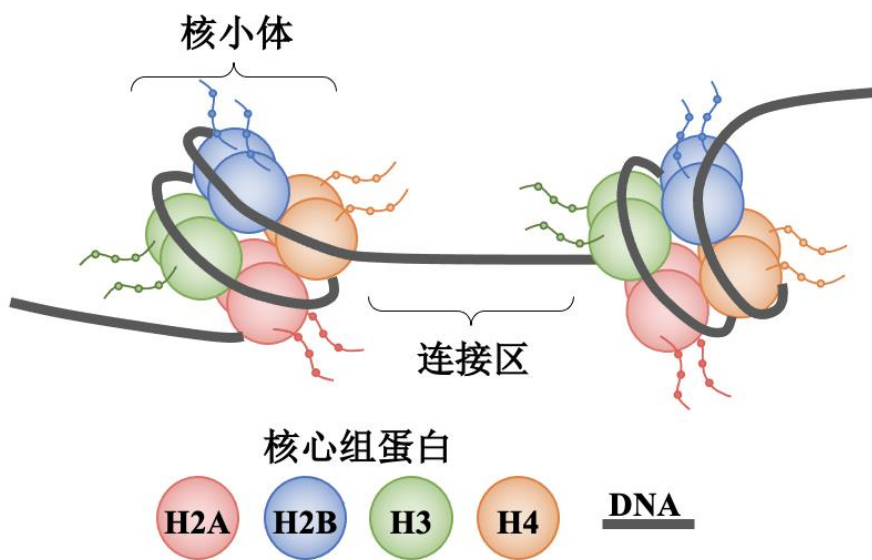
- ✓ DMR与样本间转录调控的差异相关
- ✓ 疾病样本与正常样本间存在大量DMR



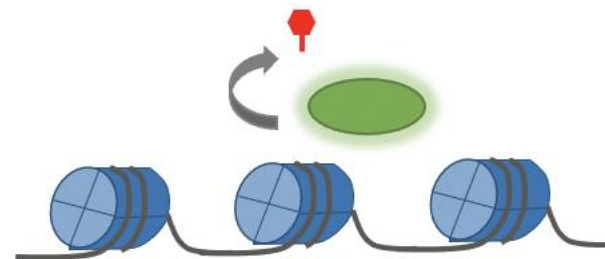
鉴定差异甲基化区域

- 以CpG二核苷酸为单位
 - ✓ 通过统计检验，识别出样本间统计显著的差异甲基化CpG（DMC）
 - ✓ 应用统计方法，将相邻的DMC归纳为DMR
- 以固定长度的窗口在基因组上滑动
 - ✓ 每个窗口内的 CpG一起用于评估样本间差异的统计显著性
 - ✓ 将相邻的样本间统计显著的差异甲基化窗口连接为DMR

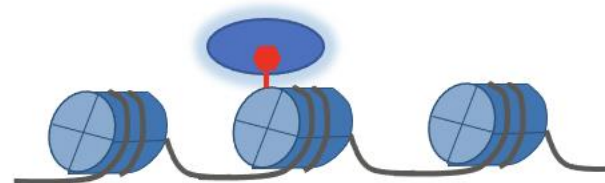
组蛋白修饰



写入器



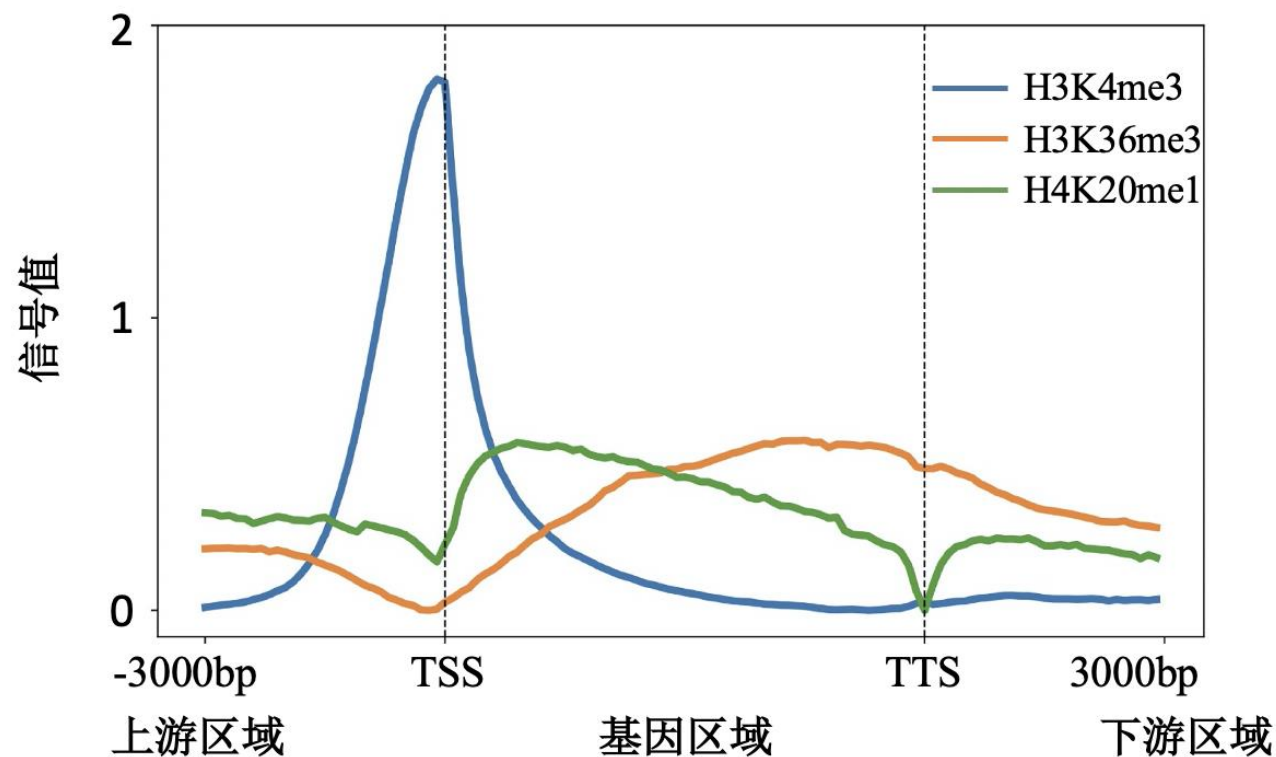
擦除器



读取器

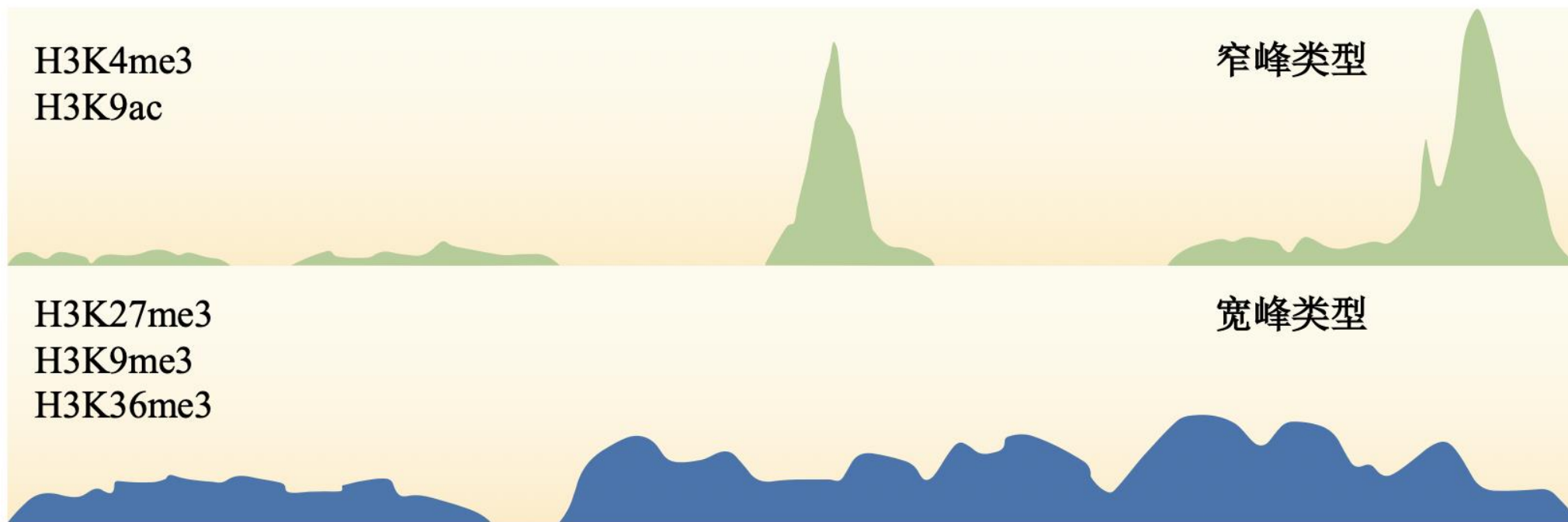
注释层面的组蛋白修饰数据质量控制

- 一些组蛋白修饰在基因组特定区域具有特定的分布特征



组蛋白修饰分布特征

识别组蛋白修饰富集区域

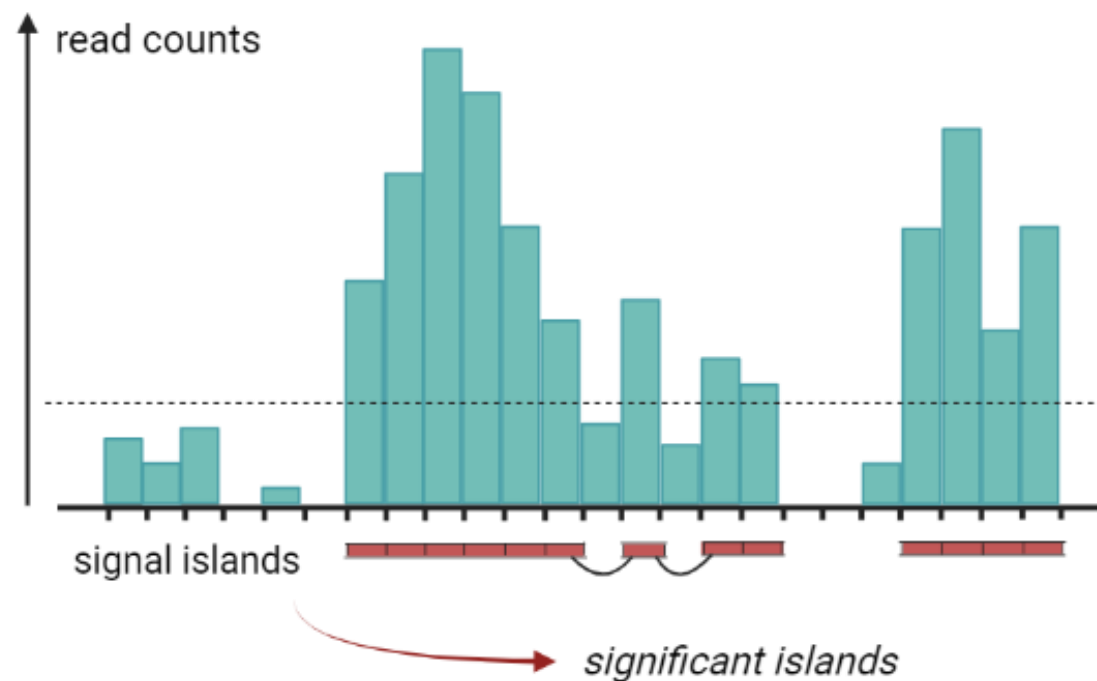


➤ 识别窄峰类型的组蛋白修饰富集区域的软件：MACS

识别组蛋白修饰富集区域

➤ 识别宽峰类型的组蛋白修饰富集区域的软件：

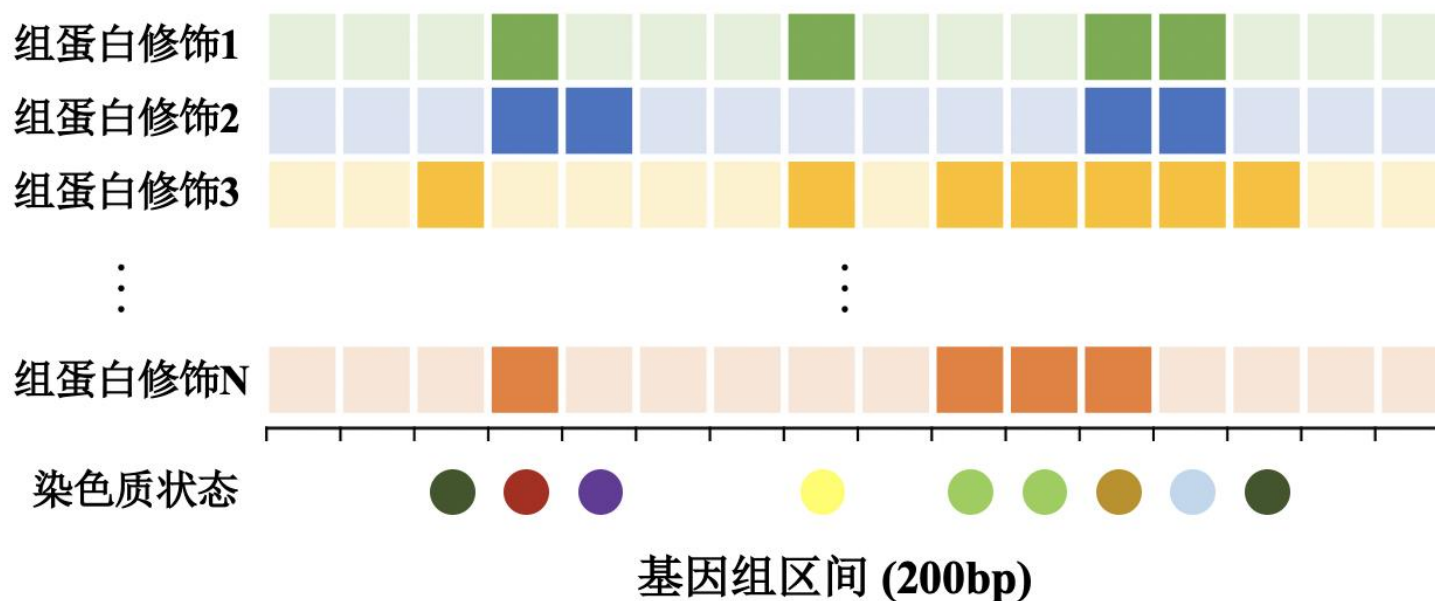
✓ MACS、SICER、SPP等



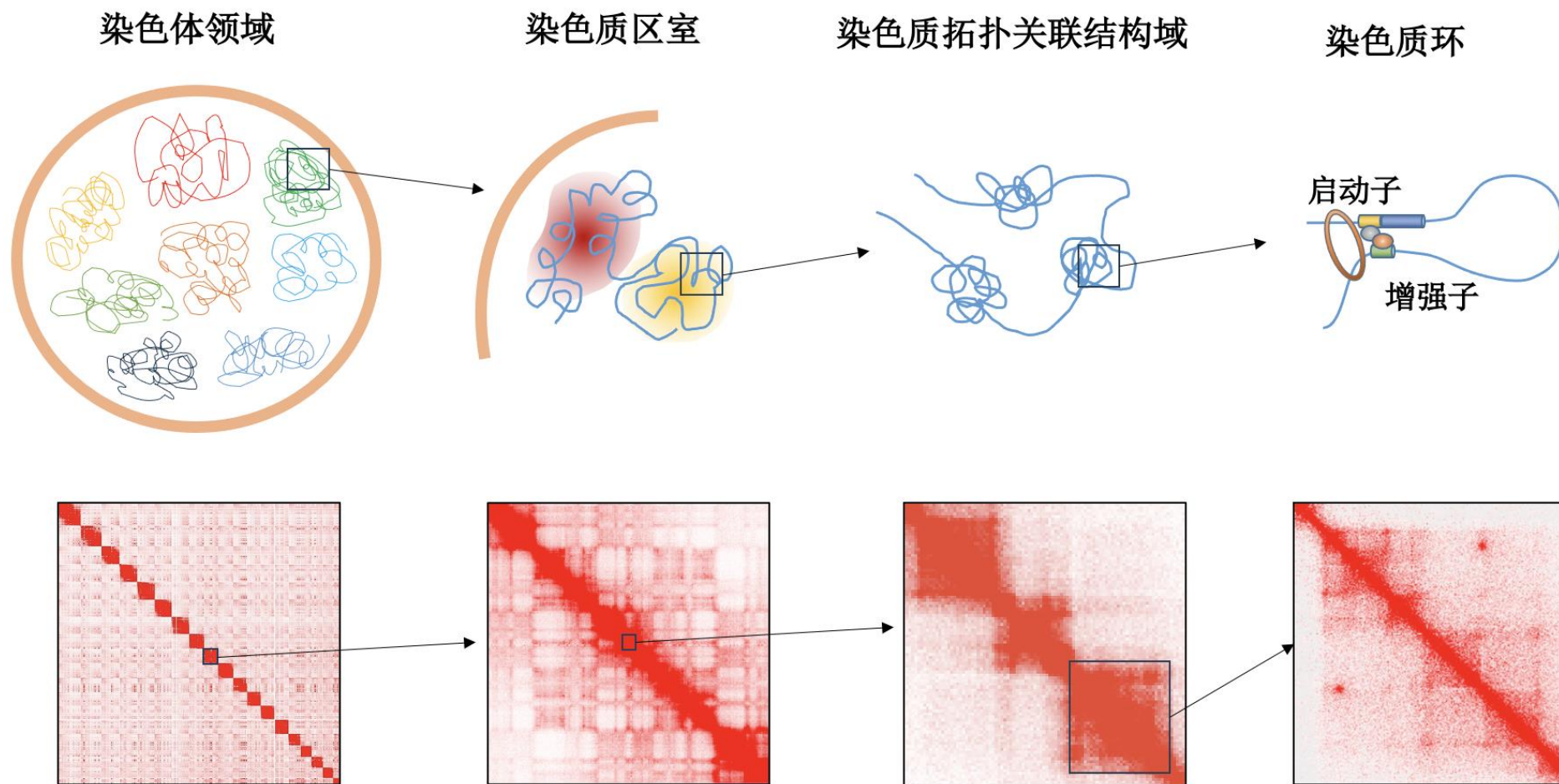
宽峰模式组蛋白修饰ChIP-seq数据信号峰识别原理

基于组蛋白修饰图谱推断染色质状态

- 组蛋白修饰之间具有关联性和冗余性
- 多种组蛋白修饰共同决定染色质状态

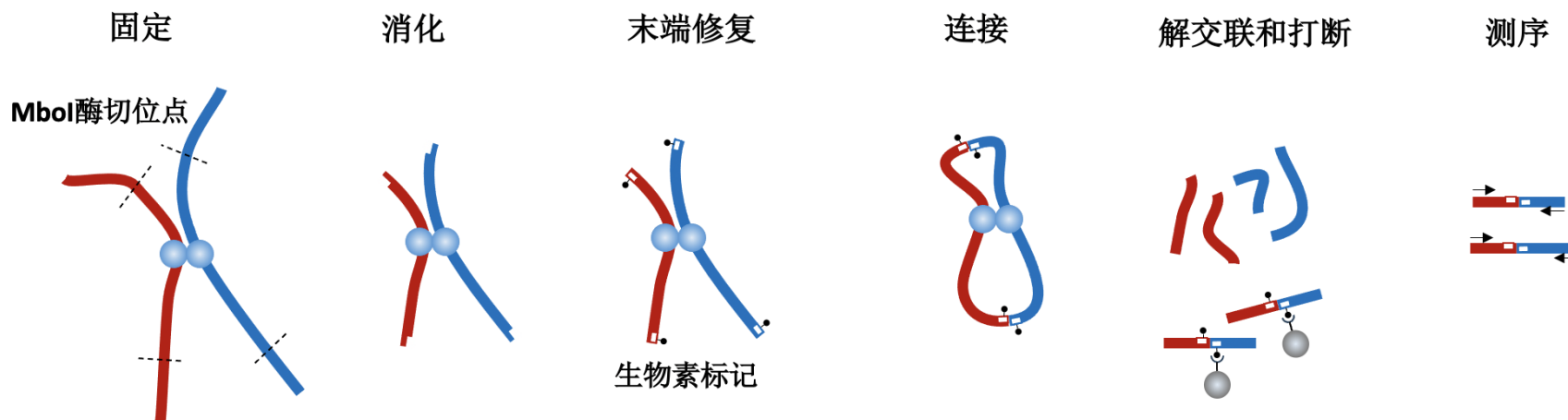


染色质三维结构的不同层级

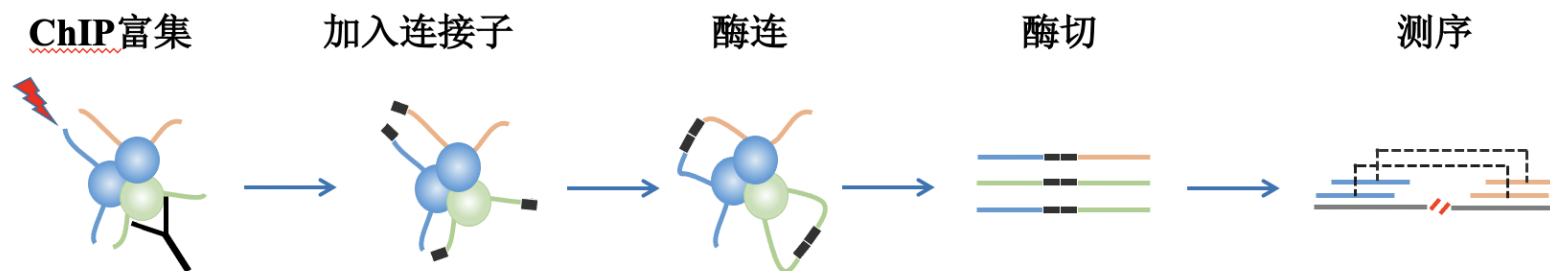


三维基因组学数据类型

Hi-C



ChIA-PET



三维基因组学技术特点

➤ Hi-C技术

- ✓ 理论上，可以获得三维基因组全部层级的结构
- ✓ 实际应用中，分辨率依赖于限制性内切酶及测序深度
- ✓ 实验操作简便，细胞用量少

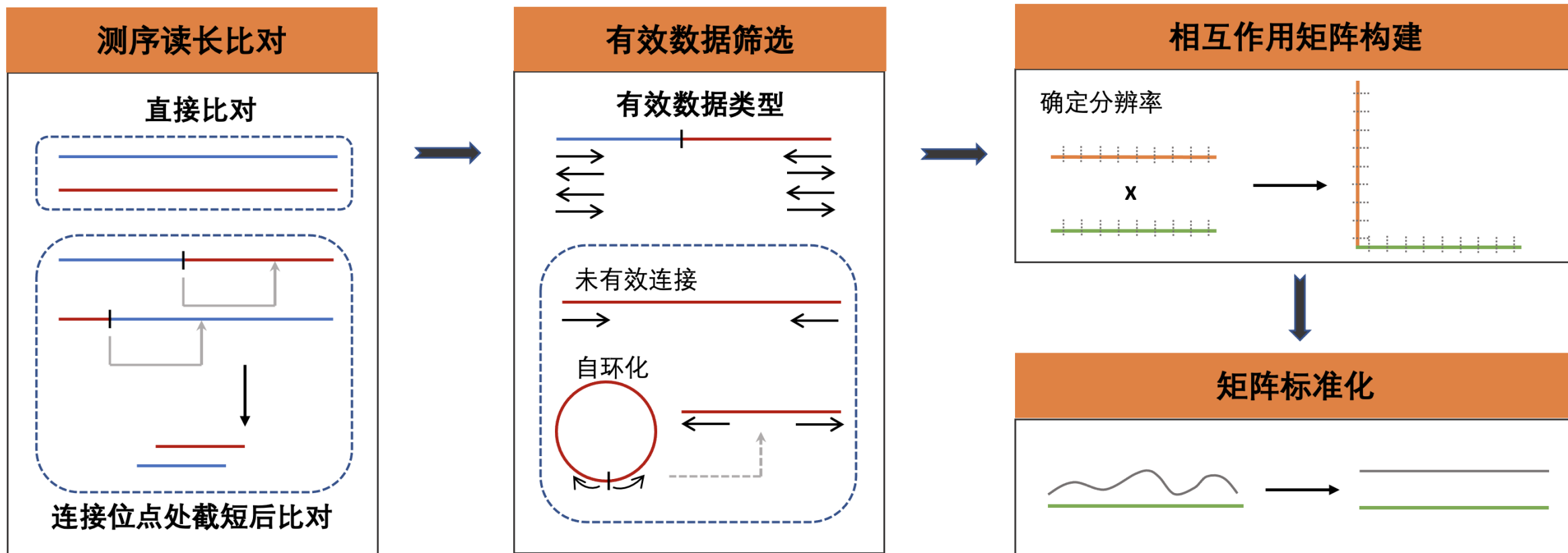
➤ ChIA-PET技术

- ✓ 目的是识别特定转录因子或组蛋白修饰所介导或参与的染色质环
- ✓ 分辨率高，染色质相互作用可解释性强
- ✓ 实验难度高，细胞用量多

Hi-C数据分析要点

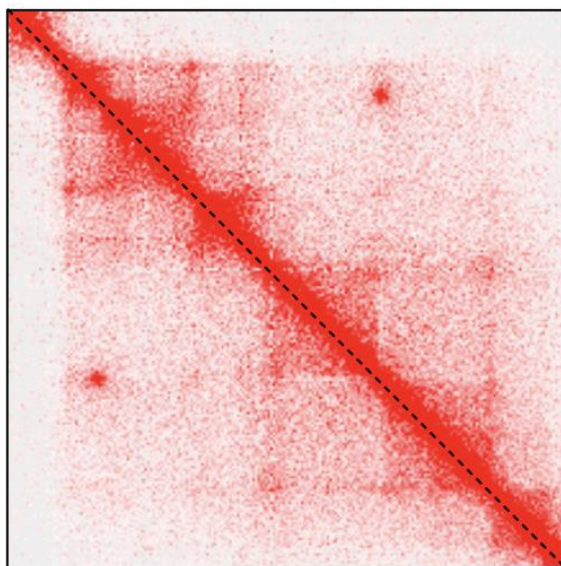
- 产生染色质相互作用矩阵
- Hi-C数据可视化
- 识别染色质区室
- 识别拓扑相关结构域
- 识别染色质环

产生染色质相互作用矩阵

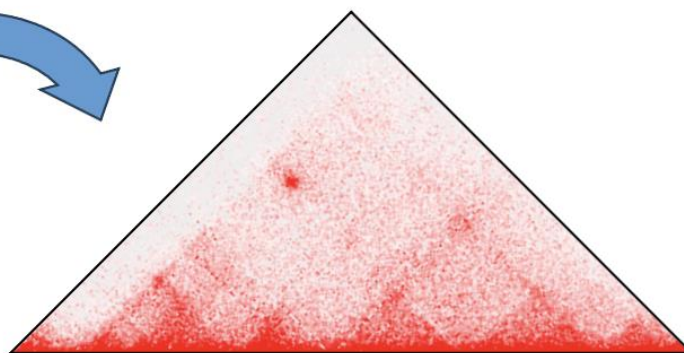


产生染色质相互作用矩阵的步骤

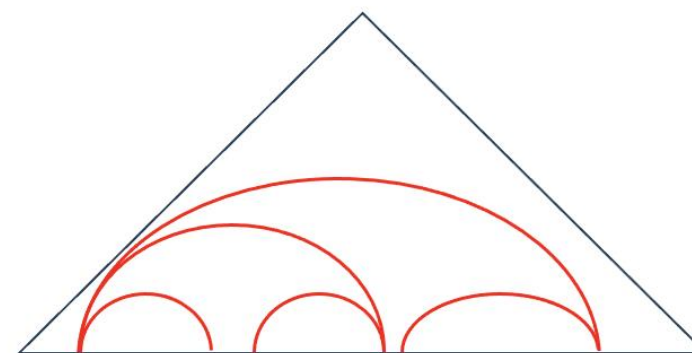
Hi-C数据可视化



正方形形式



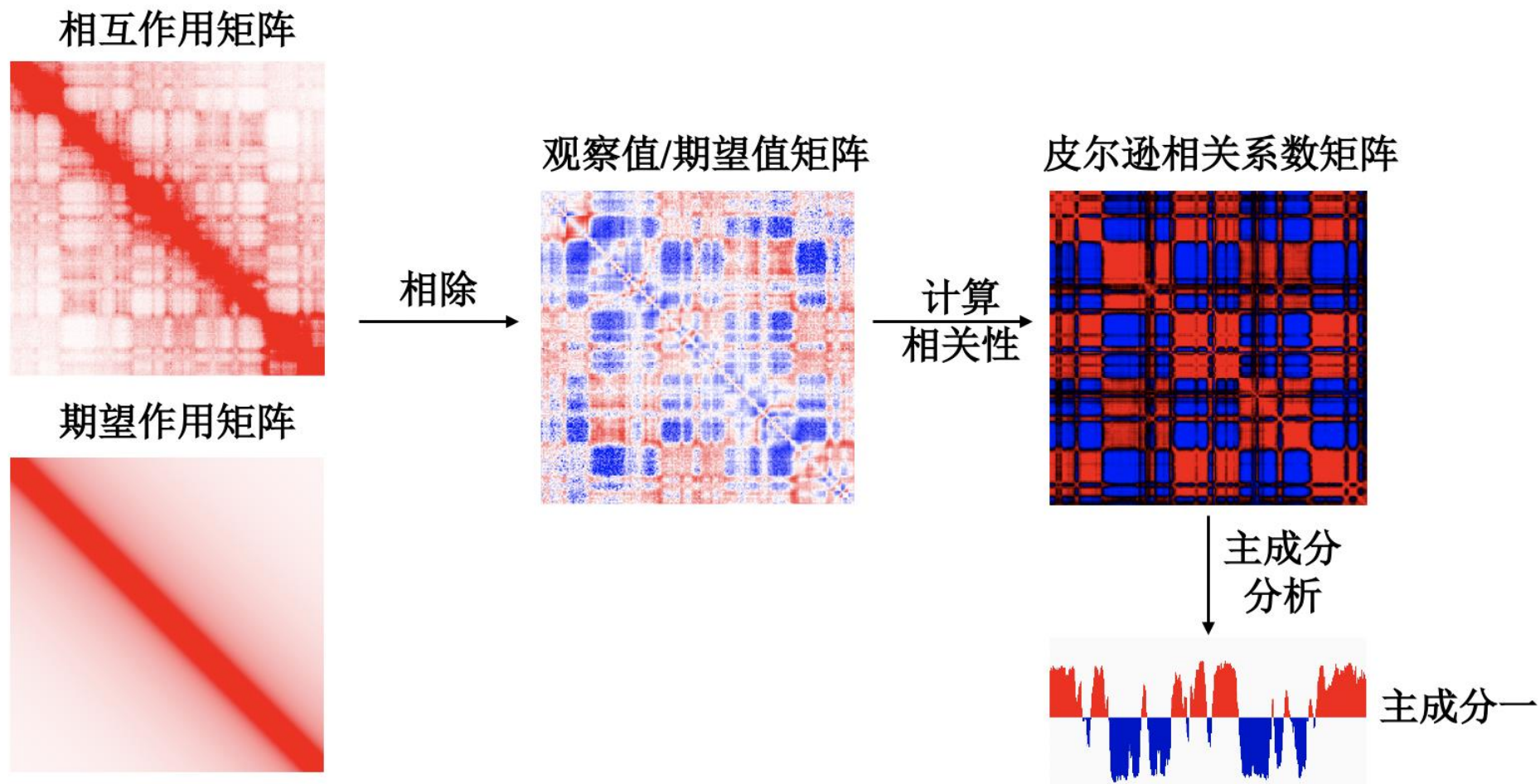
三角形形式



弧形形式

Hi-C数据的三种展示方式

识别染色质区室

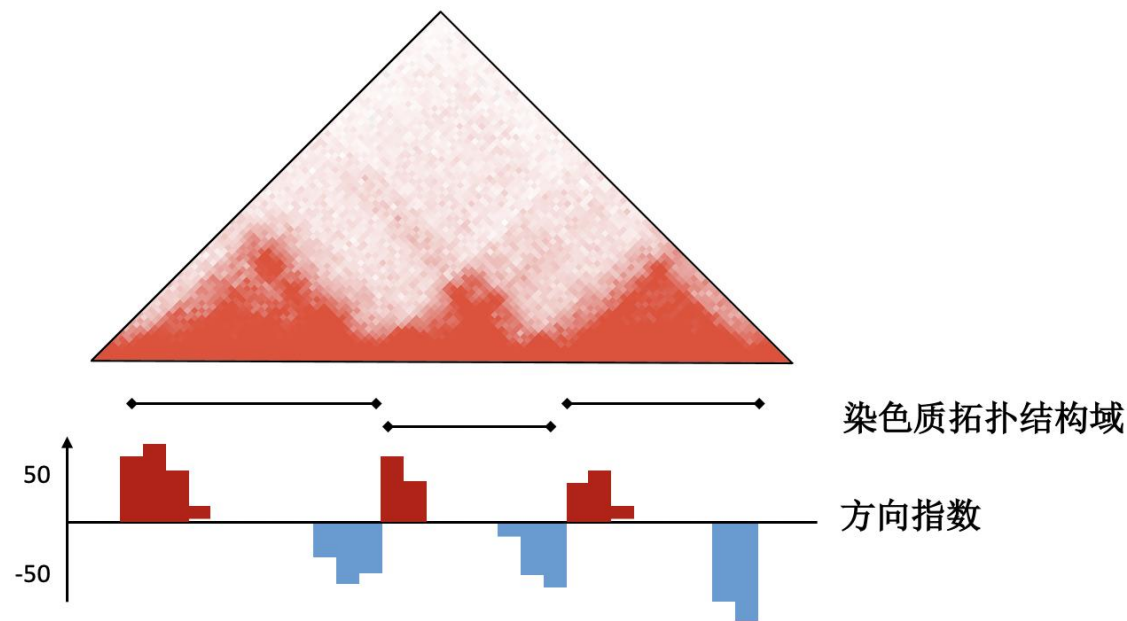


识别拓扑相关结构域

- 拓扑相关结构域内部的相互作用频率远大于之间的相互作用频率
- 方向指数（directionality index, DI）

$$DI = \left(\frac{B-A}{|B-A|} \right) \left(\frac{(A-E)^2}{E} + \frac{(B-E)^2}{E} \right)$$

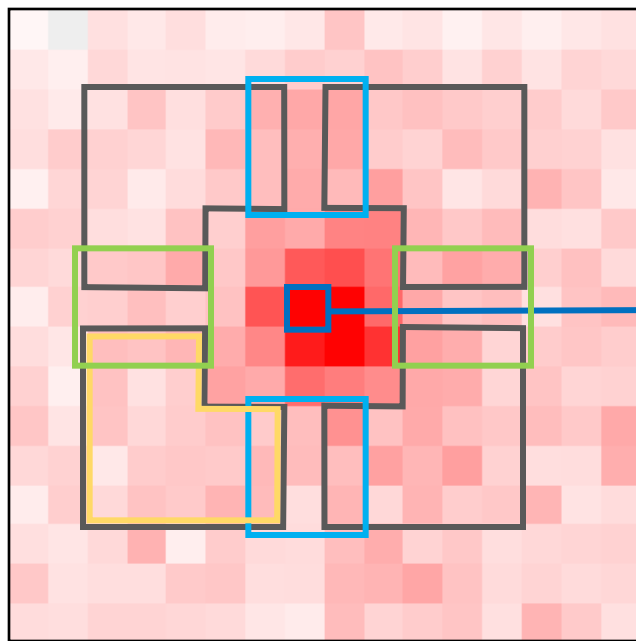
- ✓ 当方向指数值出现正负跳转时，该区间可能是拓扑相关结构域的边界



方向指数识别拓扑相关结构域原理

识别染色质环

- 染色质环的两端相对周围区域而言具有显著高的相互作用频率



背景值 (左下) = 35
背景值 (纵轴) = 32
背景值 (横轴) = 54
背景值 (四周) = 46
峰值 (中心) = 325

- 转录调控
 - 主要表观遗传信息
 - DNA甲基化、组蛋白修饰、三维基因组
 - 常用组学数据类型
 - WGBS、ChIP-seq、Hi-C
 - 数据分析要点及工具
 - WGBS数据、ChIP-seq数据、Hi-C数据

1. **版权声明：**本PPT及其所有内容（以下简称“本PPT”）仅用于教育和教学用途，版权归属于本PPT作者。
2. **使用要求：**任何使用本PPT的行为均须遵守以下条件：
 - 1) **致谢和标注：**若部分或全部使用本PPT的内容，请在使用内容的适当位置标注出处，并致谢本PPT作者。
 - 2) **修改和再分发：**未经作者书面许可，不得对本PPT进行修改或再分发。
3. **禁止商业化使用：**严禁将本PPT用于任何形式的商业化用途，包括但不限于：
 - 1) 通过网络或其他途径进行付费使用或分发；
 - 2) 在商业培训、广告或其他商业活动中使用本PPT的内容。
4. **法律责任：**任何违反上述条款的行为，作者保留追究法律责任的权利，包括但不限于：
 - 1) 要求停止侵权行为；
 - 2) 追究侵权使用者的经济赔偿责任。
5. **其他规定：**
 - 1) 本使用条款的解释权归本PPT作者所有。
 - 2) 作者保留随时更新本使用条款的权利，更新后的条款将即时生效。

谢谢大家!