中国科学院北京基因组研究所（国家生物信息中心）
BEIJING INSTITUTE OF GENOMICS CHINESE ACADEMY OF SCIENCES / CHINA NATIONAL CENTER FOR BIOINFORMATION

2025

单细胞DNA甲基化分析简介及实操指导

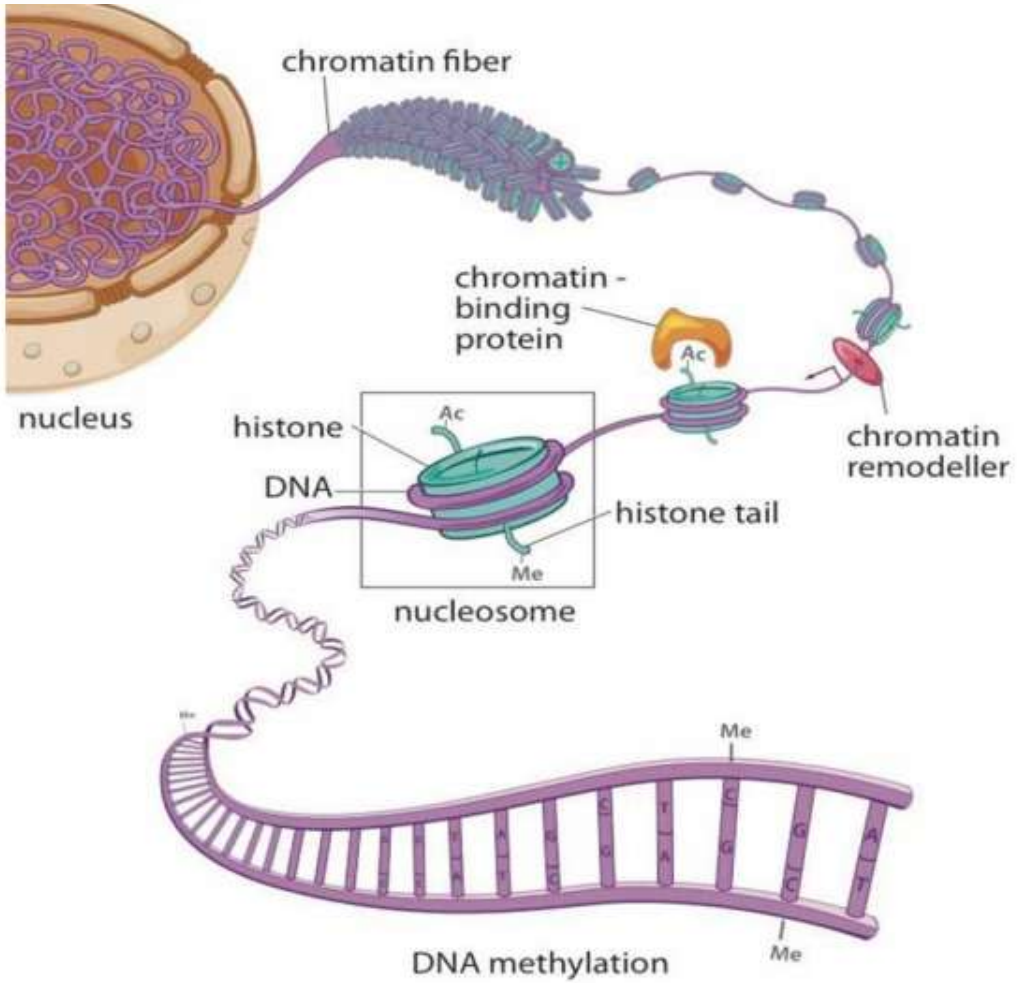
主讲人 宗文婷

时 间 2025.5.9

目 录

CONTENTS

Epigenetics



From The Cell Biology of Stem Cells (2010)

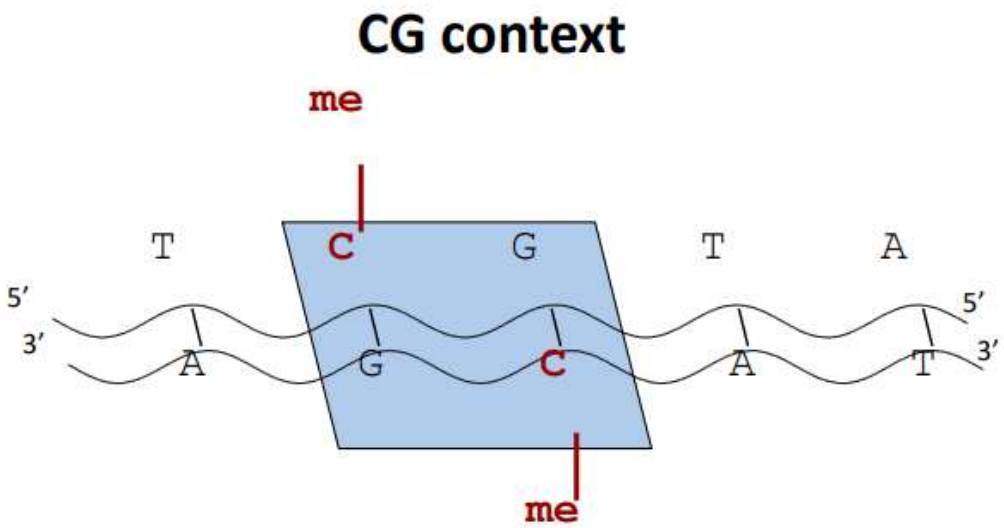
Studies changes in gene expression which are not encoded by the underlying DNA sequence

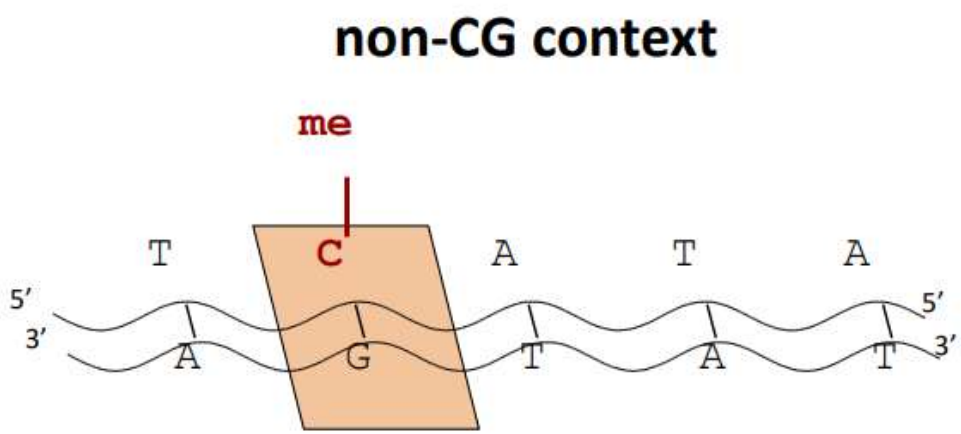
Chromatin

- histone modification
- non-coding RNAs
- higher order structure (accessibility/compaction)

DNA cytosine methylation

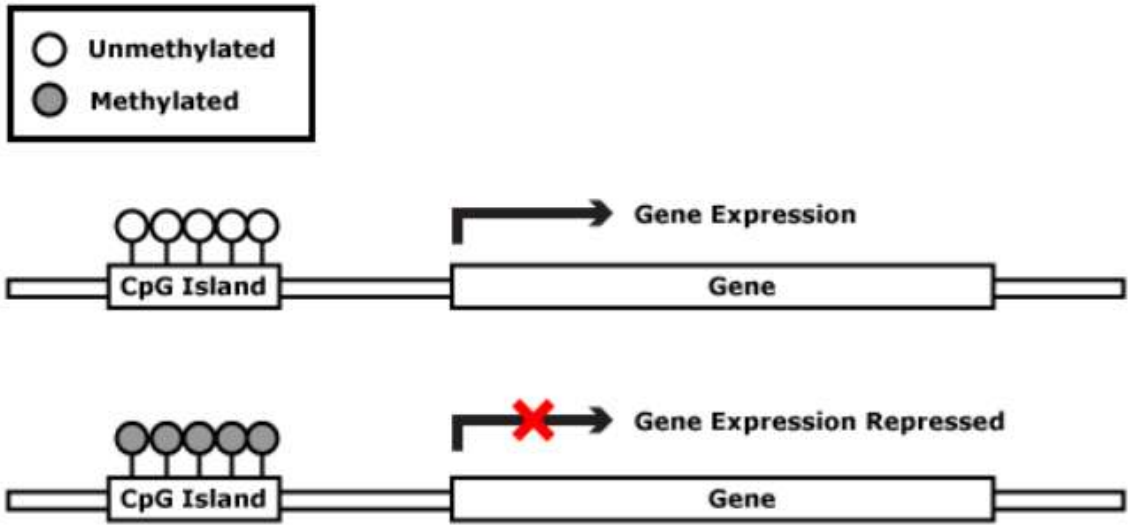
Context of DNA methylation





	Mammals	Plants
CG	present	present

Regulation by DNA methylation

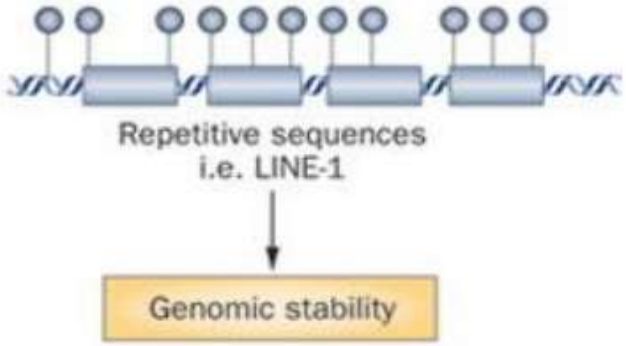


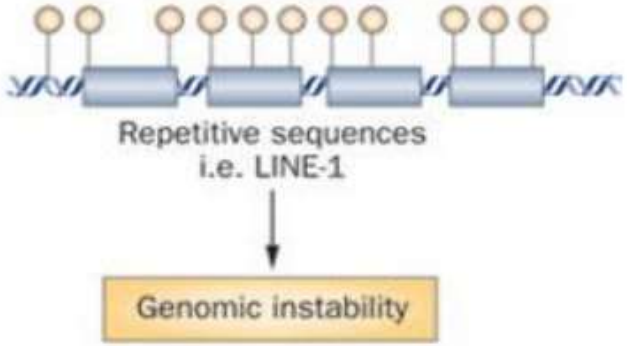
Silencing of gene expression

Tissue differentiation and embryonic development

Faults in correct DNA methylation may result in

- early development failure
- epigenetic syndromes
- cancer

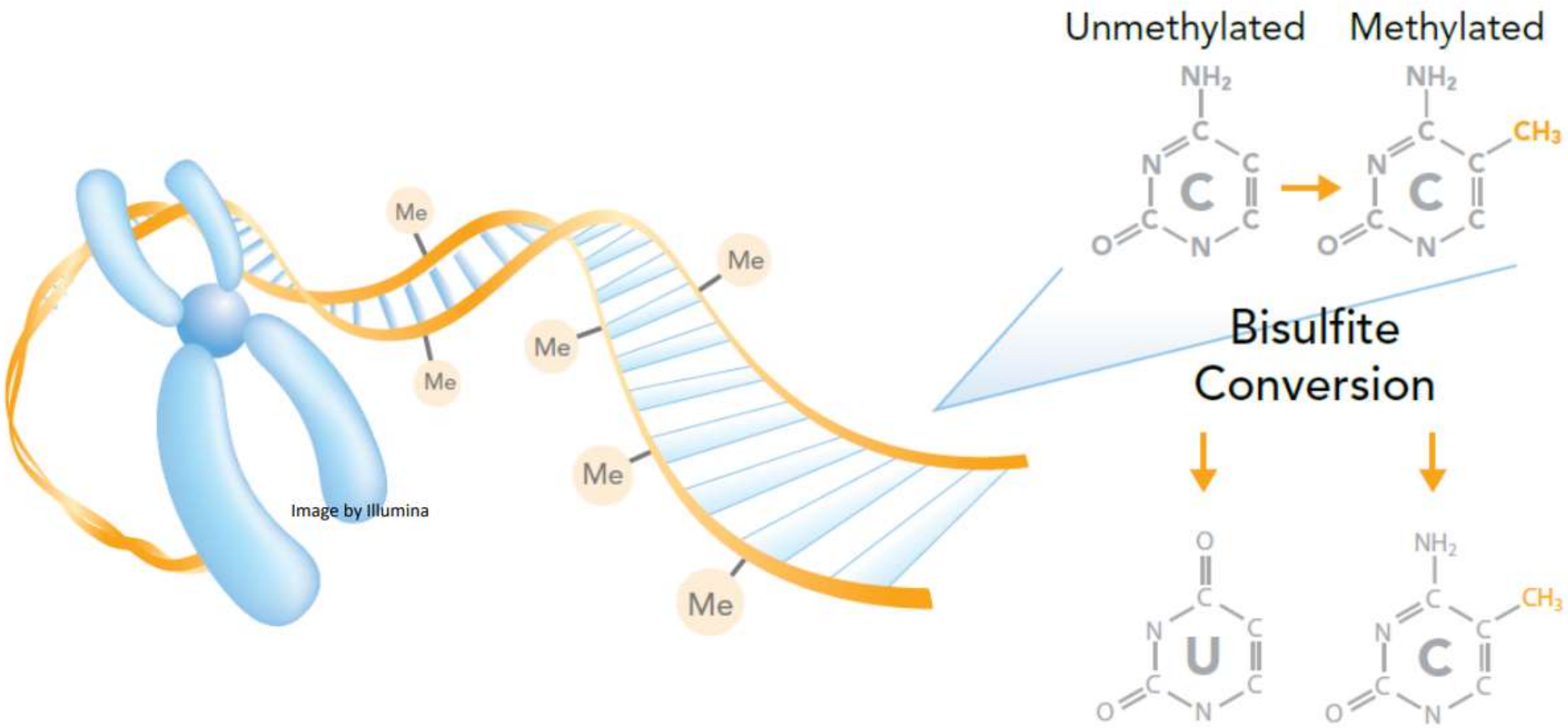




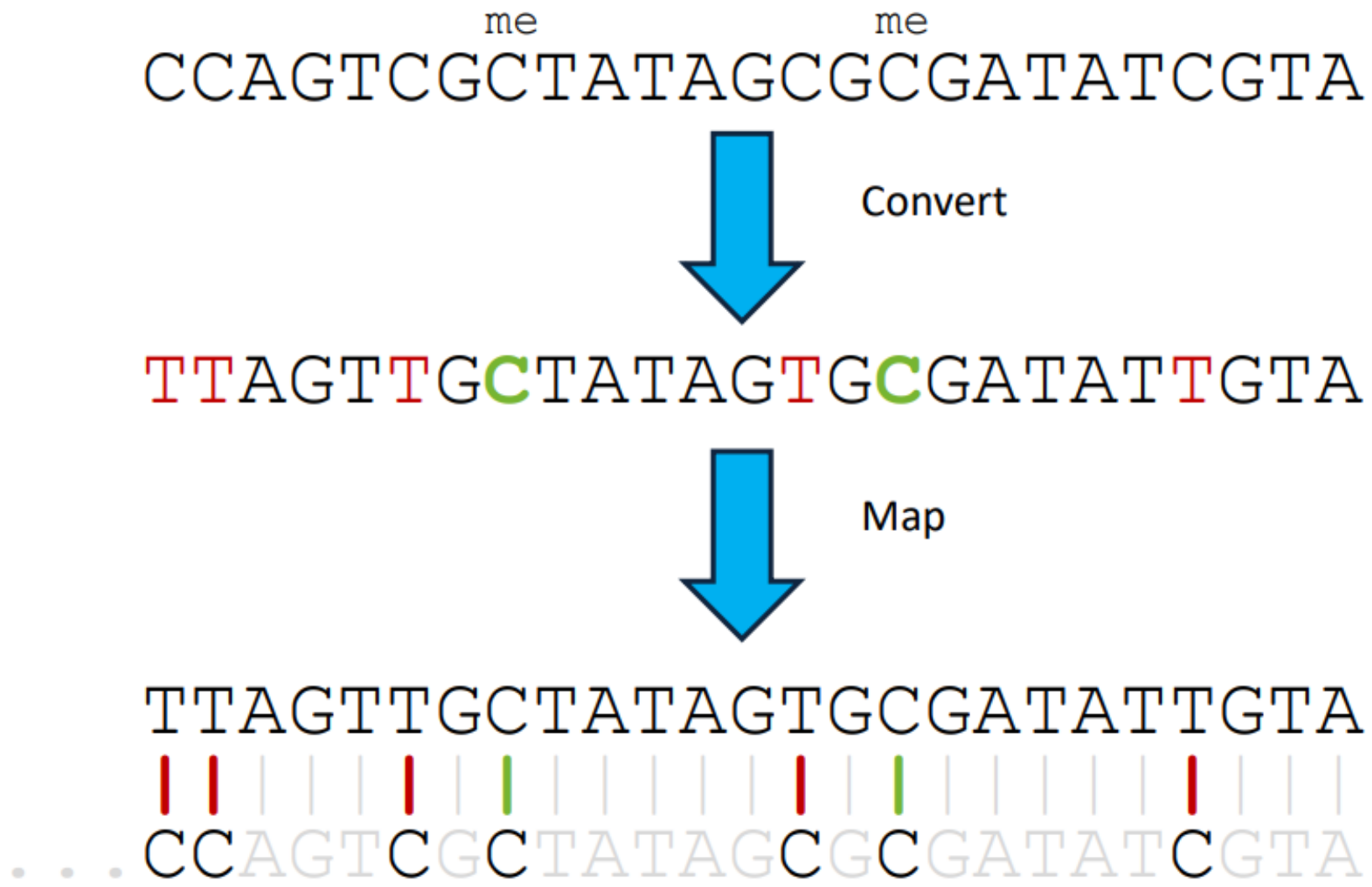
Repeat activity

Genomic stability

Measuring DNA methylation by Bisulfite-sequencing

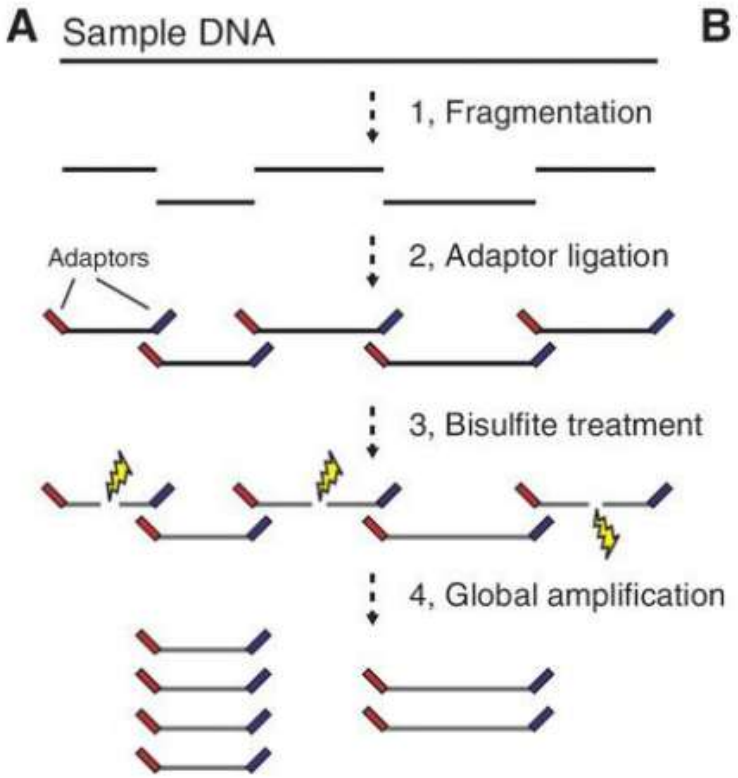


Measuring DNA methylation by Bisulfite-sequencing

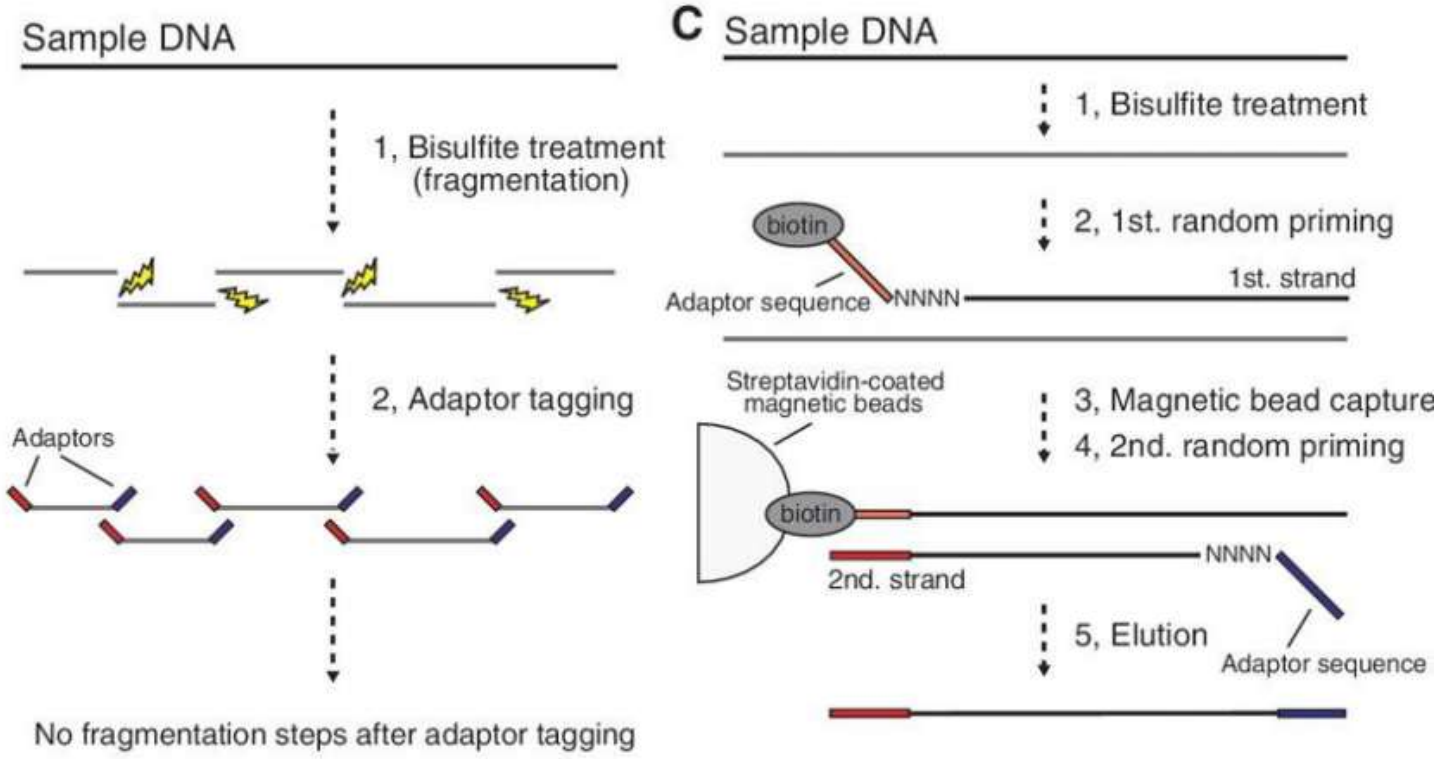


Bisulphite Library Preparation

Whole Genome Bisulphite Sequencing (WGBS)



Post Bisulphite Adapter Tagging (PBAT)



Enriched Bisulphite Sequencing

Reduced Representation (RRBS)

Cut with a restriction enzyme (MspI) whose recognition site contains a CpG

Size select for short fragments to enrich for CpG dense regions

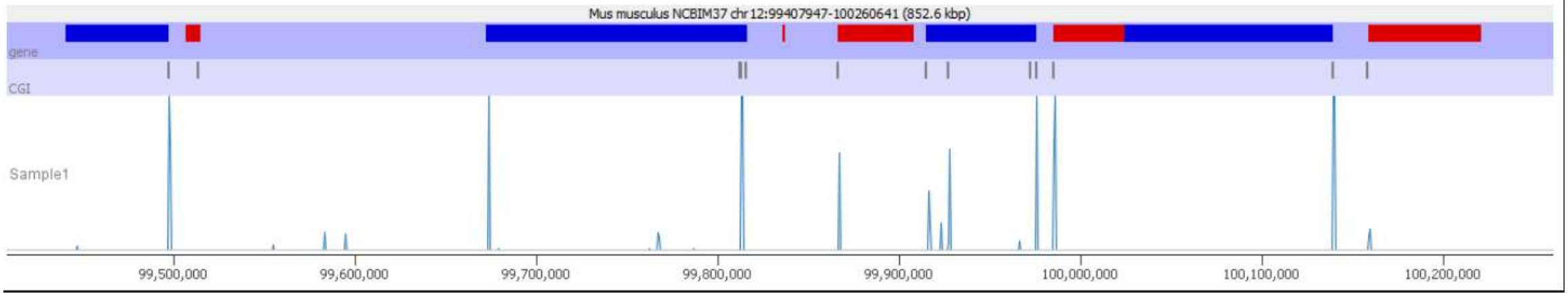
Enrichment Kits

Oligonucleotide hybridisation pull down systems

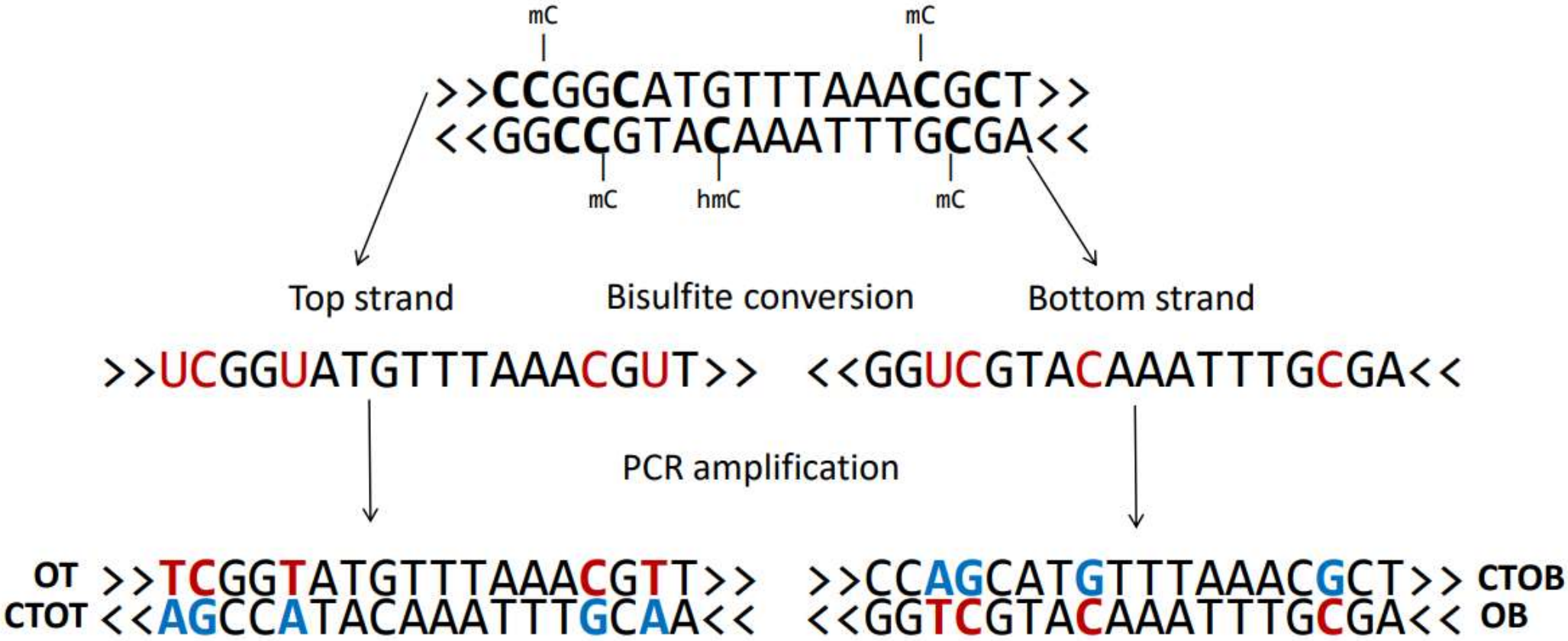
Can enrich for anything, special kits for methylation

CpG Islands and 'shores'

Relevant promoters and enhancers



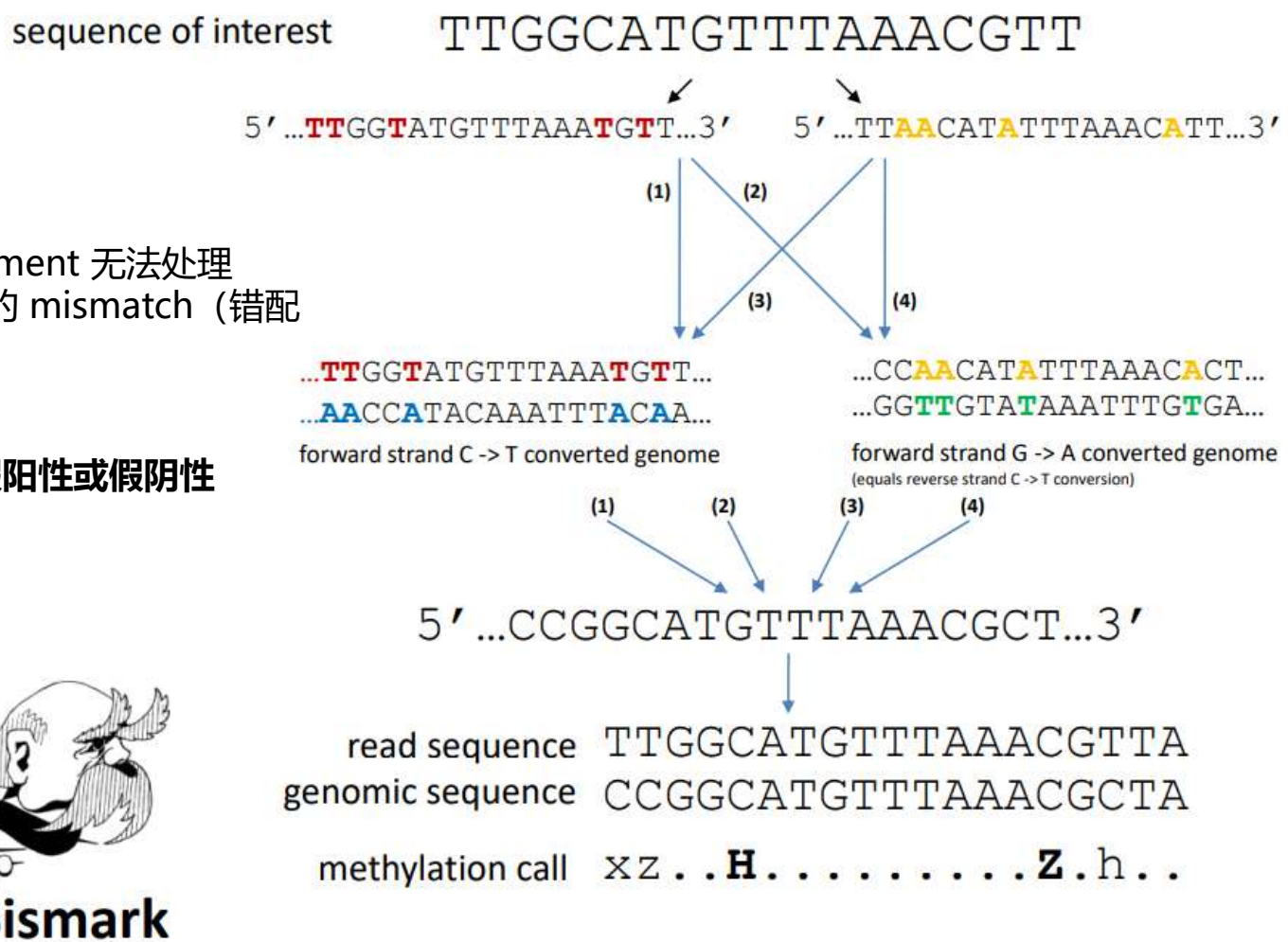
Measuring DNA methylation by Bisulfite-sequencing



- 2 different PCR products and 4 possible different sequence strands from one genomic locus
- each of these 4 sequence strands can theoretically exist in any possible conversion state

比对策略

3-letter alignment of Bisulfite-Seq reads



Fully bisulfite convert read
(as both forward and reverse strand)

Align to bisulfite converted genomes

3-letter genome

Read all 4 alignment outputs and extract the unmodified genomic sequence if the sequence could be mapped uniquely

Methylation Call

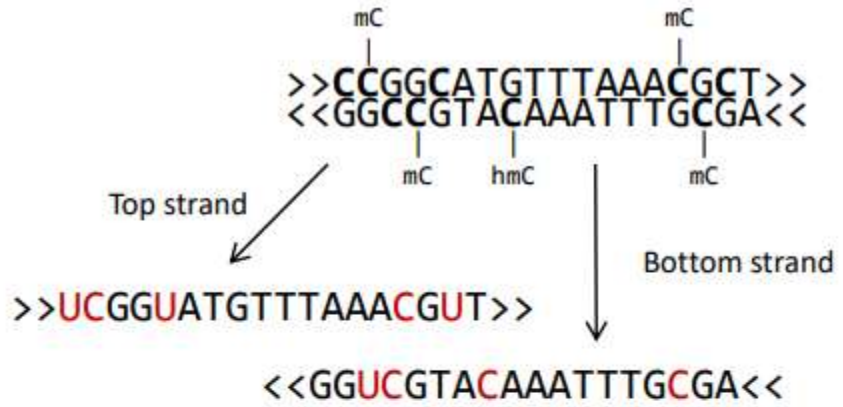
h unmethylated C in CHH context
H methylated C in CHH context
x unmethylated C in CHG context
X methylated C in CHG context
z unmethylated C in CpG context
Z methylated C in CpG context

标准的 read alignment 无法处理这么大量的 C→T 的 mismatch（错配容易：

- **比对失败**
- **误判为突变**
- **造成甲基化率的假阳性或假阴性**

Bismark

比对策略



- OB：原始下链（Original Bottom strand）
- CTOT：互补链对应的上链（Complementary to Original Top）
- CTOB：互补链对应的下链（Complementary to Original Bottom）
- OT：原始上链（Original Top strand）

1) Directional libraries

(vast majority of kits, also EpiGnome/Truseq)

OT >>TCGGTATGTTTAAACGTT>>
<<GGTCGTACAAATTTGCGA<< OB

2) PBAT libraries

CTOT <<AGCCATACAAATTTGCAA<<
>>CCAGCATGTTTAAACGCT>> CTOB

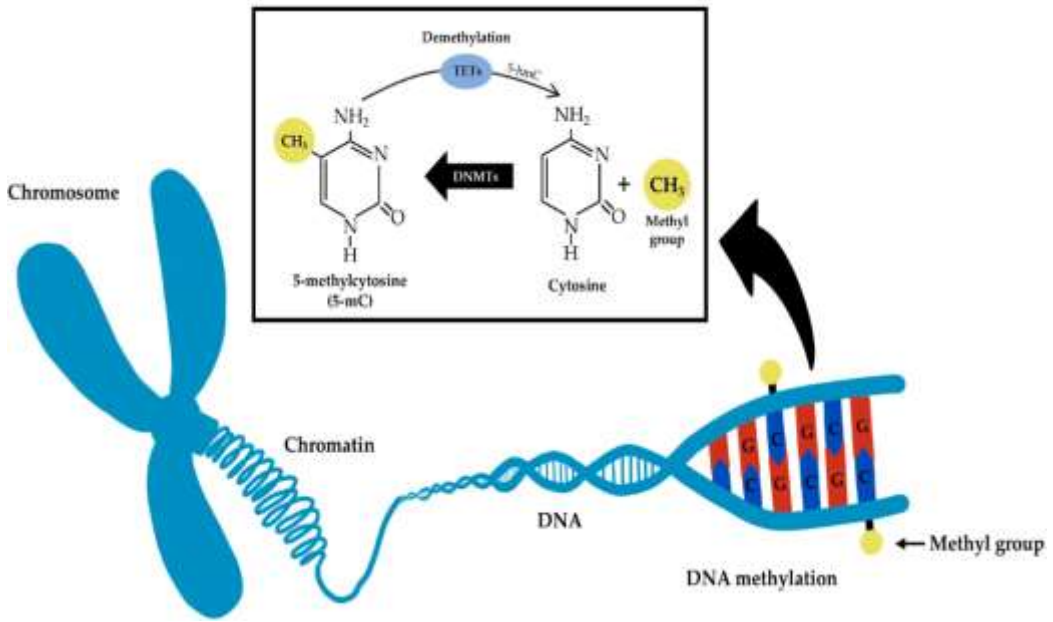
3) Non-directional libraries

(e.g. single-cell BS-Seq, Zymo Pico Methyl-Seq)

OT >>TCGGTATGTTTAAACGTT>>
CTOT <<AGCCATACAAATTTGCAA<<
>>CCAGCATGTTTAAACGCT>> CTOB
<<GGTCGTACAAATTTGCGA<< OB

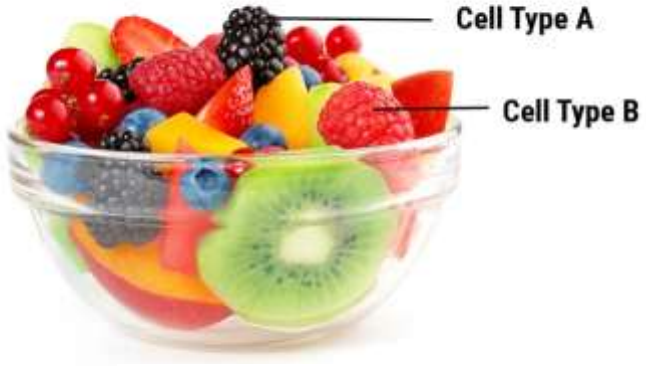
单细胞DNA甲基化

- DNA 甲基化 (DNA methylation)是最早被研究的重要的表观遗传修饰之一



- Bulk 测序结果是细胞整体的表征，一些细胞层面独有的低丰度信息会因为平均而丢失

Bulk测序



单细胞测序

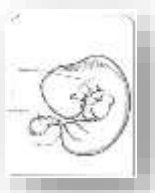
Valente et al., ***Nanomaterials***, 2023

单细胞DNA甲基化测序优势

- **高分辨率单细胞甲基化分析**：这种方法能够在单细胞水平上获取复杂的甲基化图谱。它有助于深入研究细胞异质性和动态变化，从而加深我们对细胞多样性和状态转变的理解。
- **揭示细胞状态转变和进化轨迹**：通过精确追踪单个细胞的甲基化动态，该方法揭示了细胞状态转变和发育轨迹的复杂机制。这些发现为深入了解细胞发育和进化的分子基础提供了宝贵的线索。
- **发掘单个细胞之间的功能差异**：通过对甲基化谱进行细致的比较分析，该方法阐明单个细胞之间的功能差异，从而阐明控制细胞功能和表观遗传调控的潜在机制。
- **最少的 DNA 输入要求**：该技术的特点是其对 DNA 输入的需求最少，仅需少量细胞即可达到超低检测阈值。

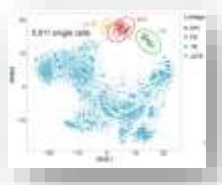
单细胞DNA甲基化应用

展示胚胎发育甲基化动态变化



研究癌症发生与转移过程

细胞衰老与凋亡



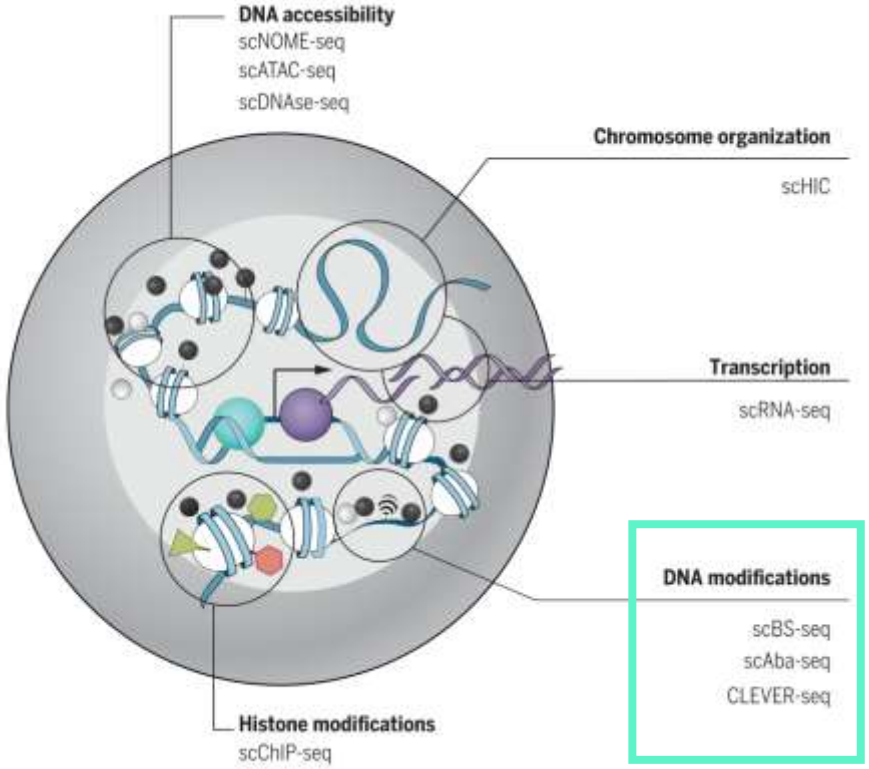
作为标记物分选细胞

干细胞分化

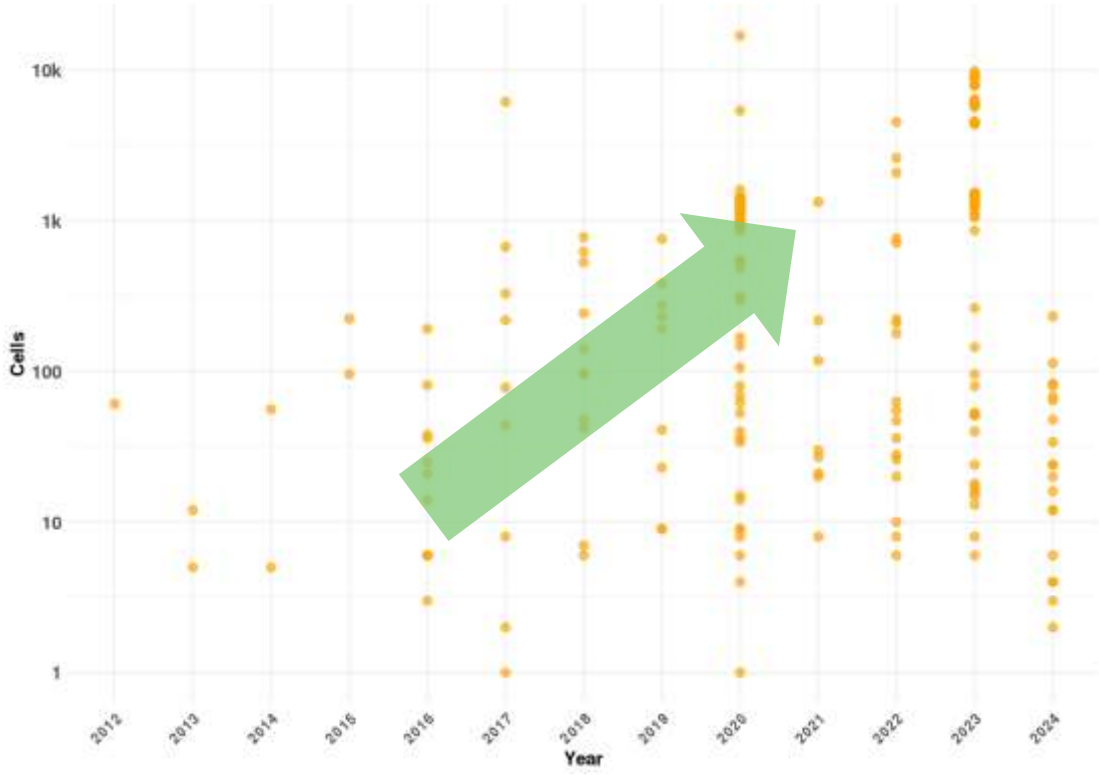
识别新神经元

单细胞DNA甲基化测序

- 单细胞DNA甲基化测序可以检测细胞水平单碱基的甲基化模式



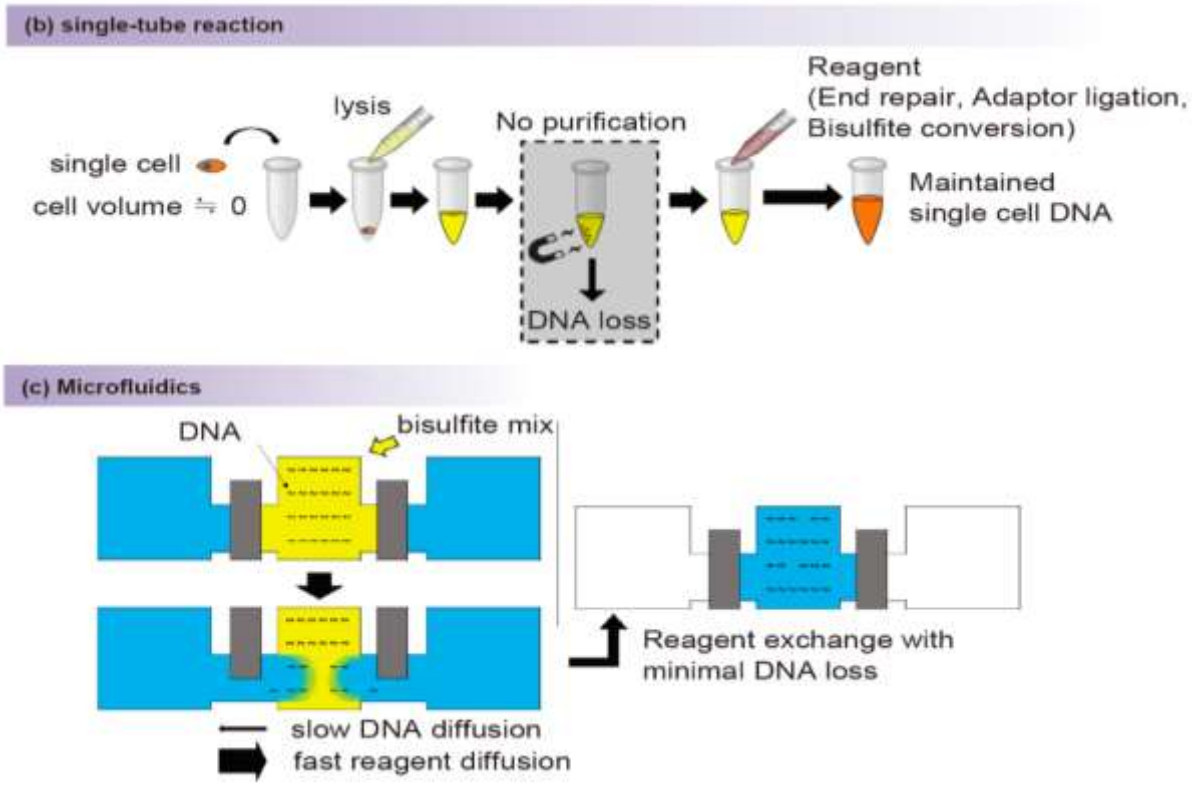
- 随着高通量测序技术的进步，单细胞DNA甲基化数据规模正在不断变大

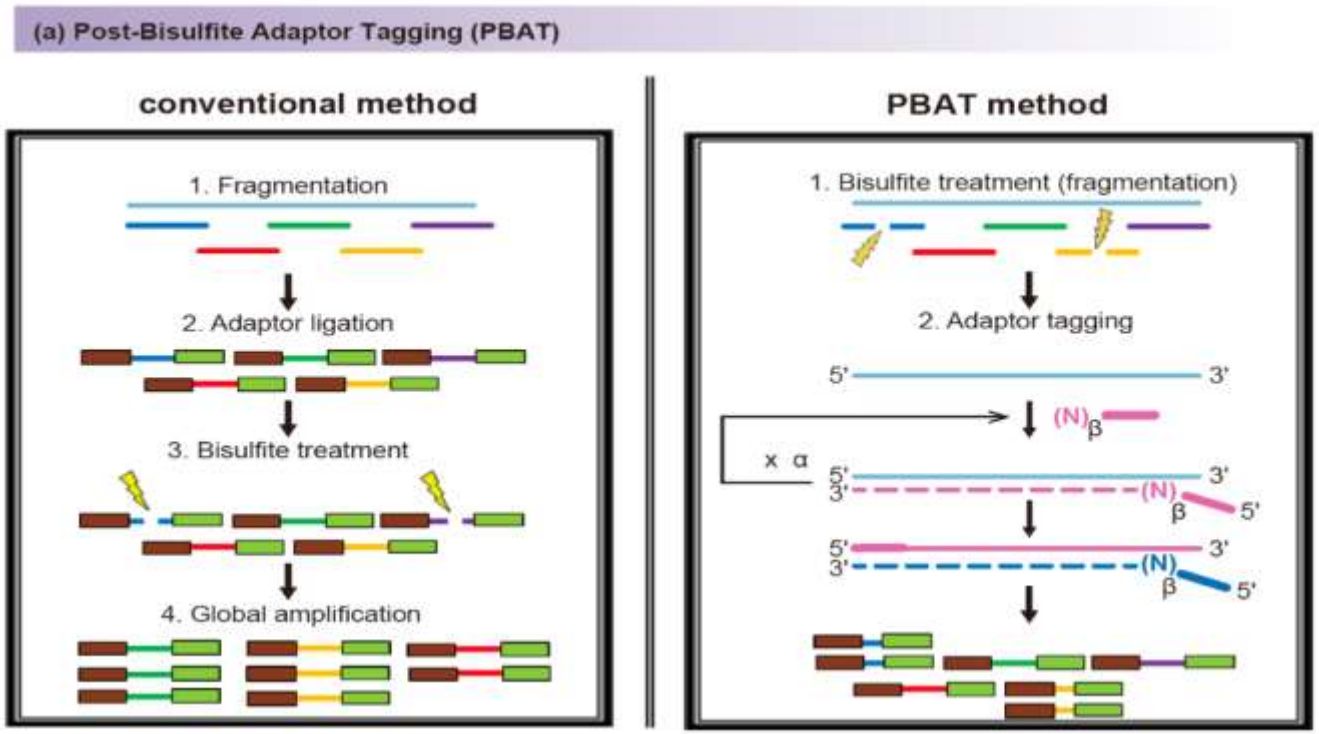


Kelsey G et al., ***Science***, 2017

单细胞DNA甲基化测序

- 二代测序是怎么检测细胞的甲基化状态的?





Isolation of single cell → Bisulfite conversion → Library Preparation → Sequencing → Data Analysis

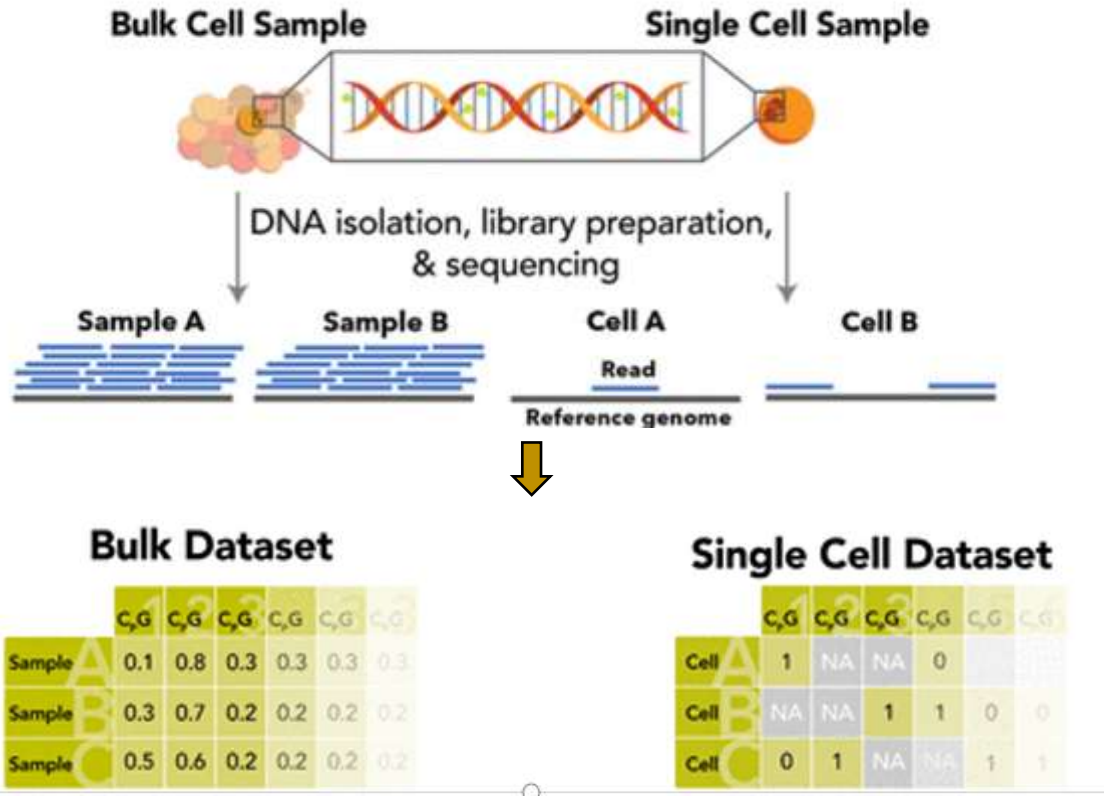
单细胞DNA甲基化测序

	基于限制性内切酶		基于PBAT		基于细胞标签	基于多组学		
方法	**scRRBS**	**scTAM**	**scBS-seq**	**snmc-seq2**	**sci-MET**	**scM&T-seq**	**scTrio-seq**	**scNMT-seq**
策略	单管酶促反应	基因特异性引物	亚硫酸盐处理后添加接头两轮随机启动	简并随机引物和适配酶进行接头序列整合	线性扩增组合索引带标签的PCR 富集	**转录组+甲基化组**	**转录组+甲基化组+拷贝数变异**	**转录组+甲基化组+染色体可及性**
通量	约100个细胞约一百万CpG位点	约1万个细胞650个CpG位点	96个细胞3.7M CpG位点	384个细胞30%-40%基因组	单孔96个细胞单次约1000个细胞0.05%-7%基因组	Smart-seq2scBS-seq	Smart-seqscRRBS	Smart-seq2NOMe-seq

单细胞DNA甲基化数据覆盖度低、所需计算量大

单细胞DNA甲基化数据具有独特性

与Bulk DNA甲基化测序数据不同：



噪声高,特征受到混杂因素影响

- 由于起始DNA数量少，以及测序文库制备的破坏性，单细胞数据覆盖度低且不均一

数据稀疏度高，需要处理缺失数据

- Bulk测序 产生从 0 到 1 的连续值矩阵，而单细胞样品主要表现出0和1的二值化现象，稀疏度高，维度极高

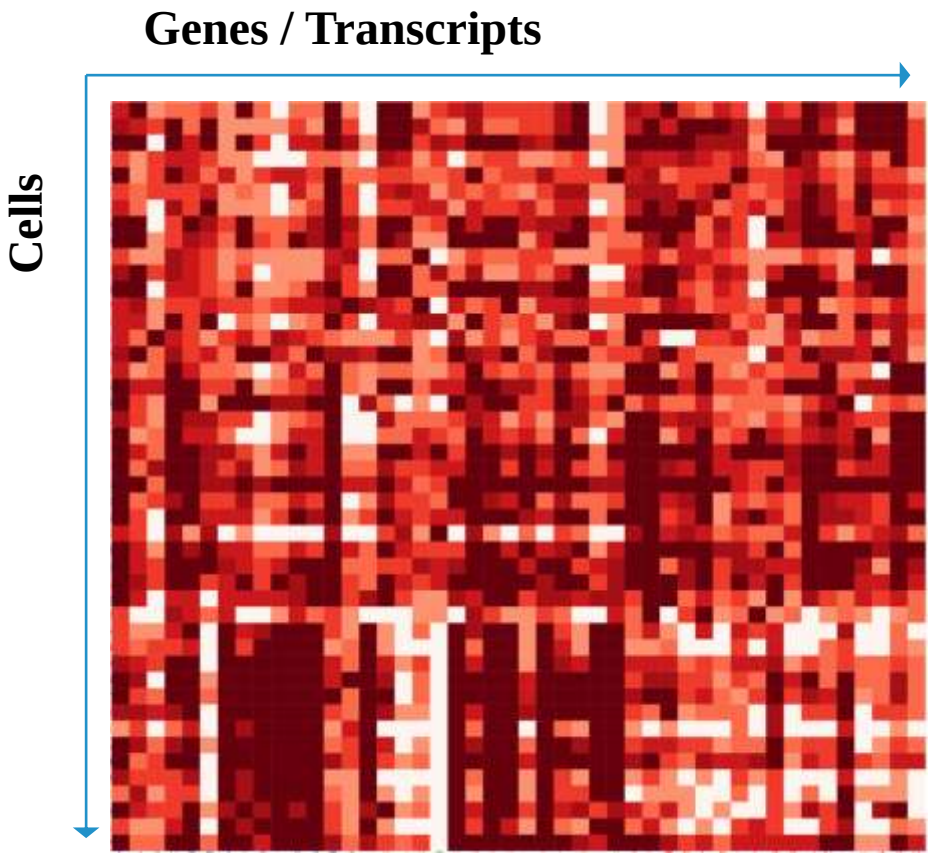
Alexandre T et al., ***Nature Aging***, 2021

单细胞DNA甲基化数据具有独特性

与scRNA-seq等其他单细胞组学数据不同：

数据维度更高，处理成本和复杂度更高

- scRNA-seq数据的特征空间是基因或者转录本，特征维度相对较低，约为10^2-10^4
- 单细胞DNA甲基化数据是胞嘧啶位点上的状态值，维度极度高，约10^4-10^7



单细胞DNA甲基化数据具有独特性

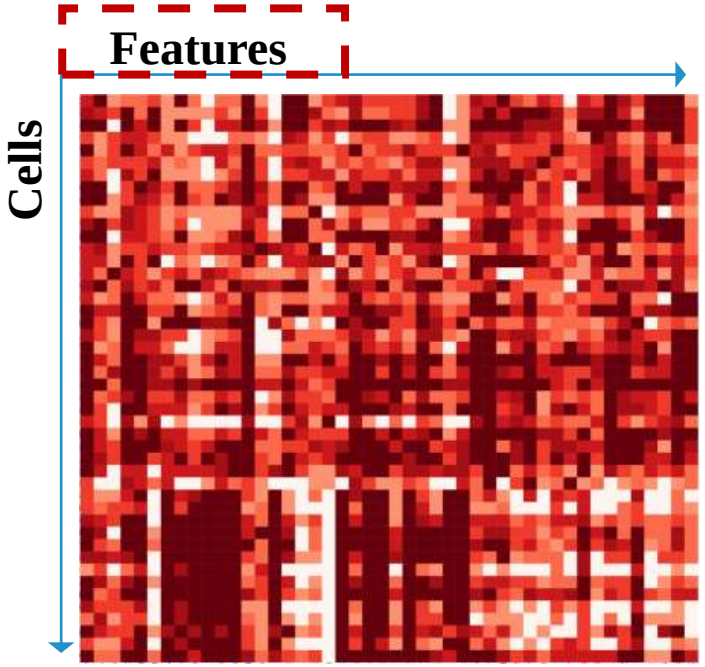
与scRNA-seq等其他单细胞组学数据不同：

windows

genome annotation

open chromatin

「硬分窗方法」

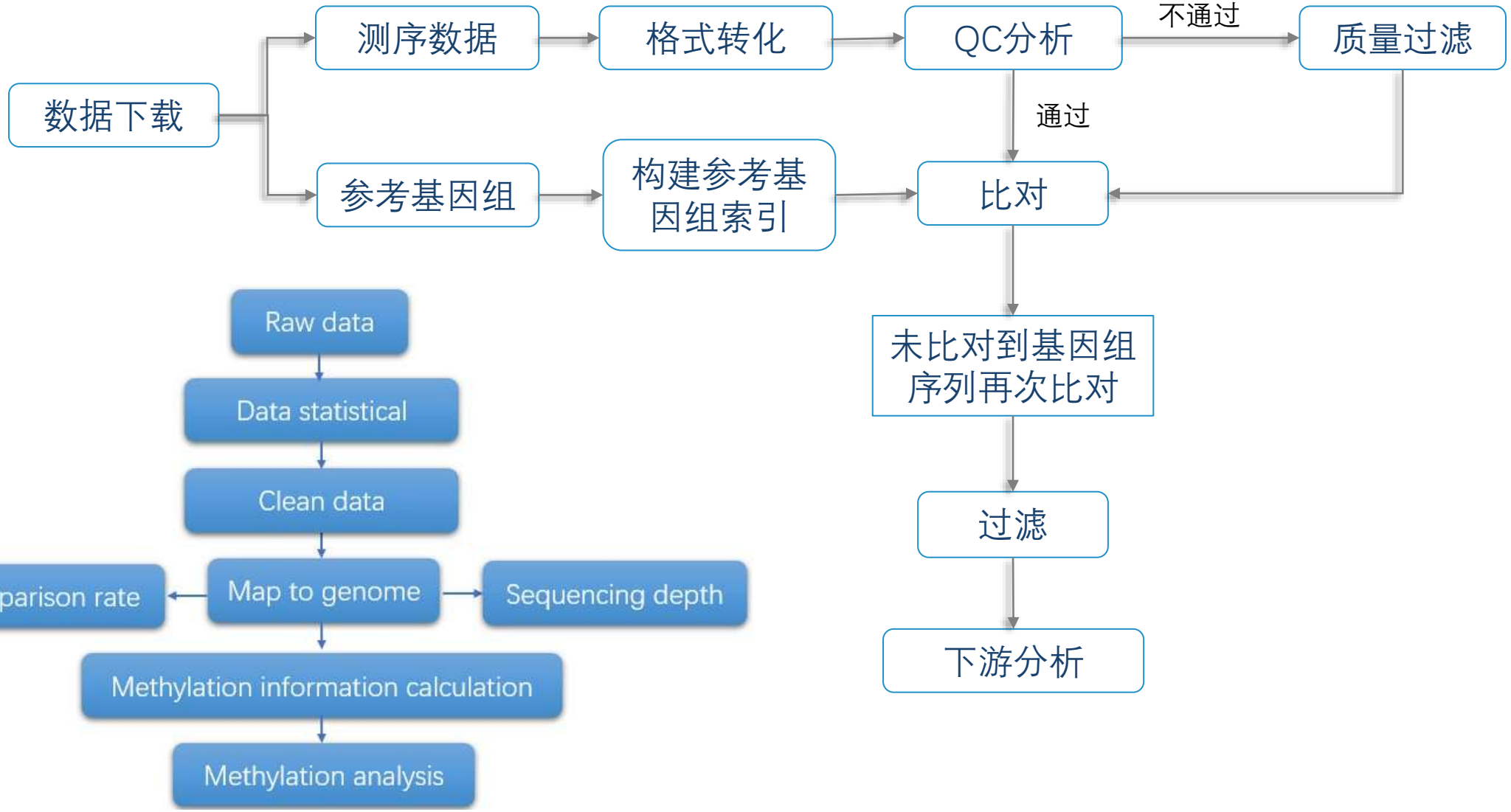


「从头识别特征」

De novo

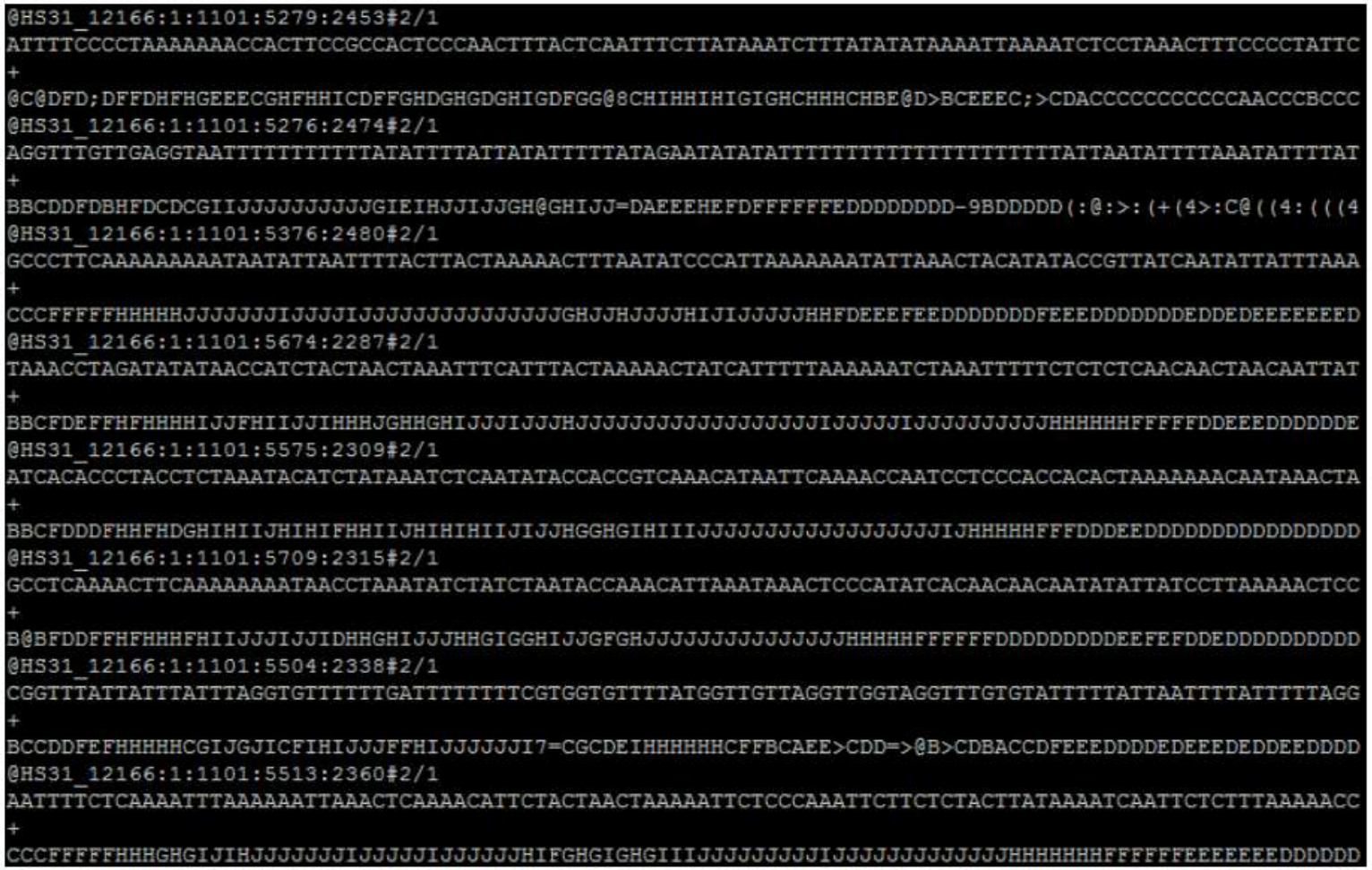
- 单细胞DNA甲基化数据的特征空间更难确定
- 能否通过数据驱动方式从头获取特征区域

单细胞DNA甲基化数据上游分析



原始数据（FastQ file）

- 与标准 DNA 文库相比，亚硫酸氢盐文库通常包含较短的片段，片段大小通常在 80 到 120bp 之间。



原始数据（FastQ file）

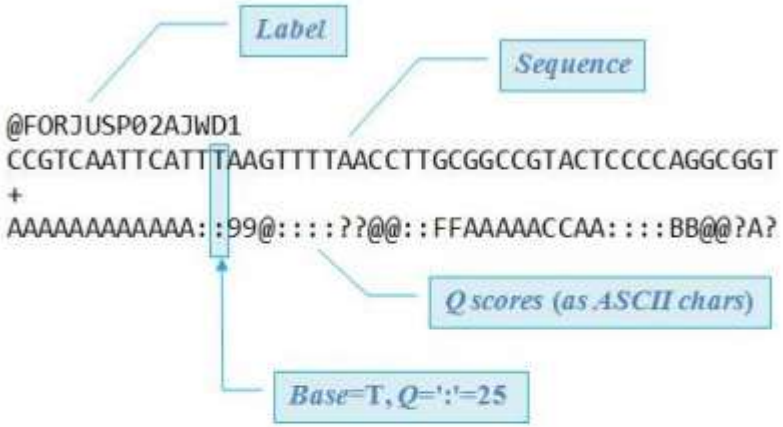


表 3.1.2 碱基质量值与碱基识别正确率的对应关系

	错误率	正确率
Q20	1%	99%
Q30	0.1%	99.9%
Q40	0.01%	99.99%

$$Q = -10log_{10}P_{error}$$

FASTQ文件每四行为一个单位，一个单位包含一条测序序列（read）的信息：

1. Read ID，是每一条read的唯一标识符。
2. Read序列。
3. 一般为一个“+”
4. 碱基质量值，与第二行一一对应，是对第二行每个碱基的测序准确性的描述。

质量控制

What dose QC tell you about your library

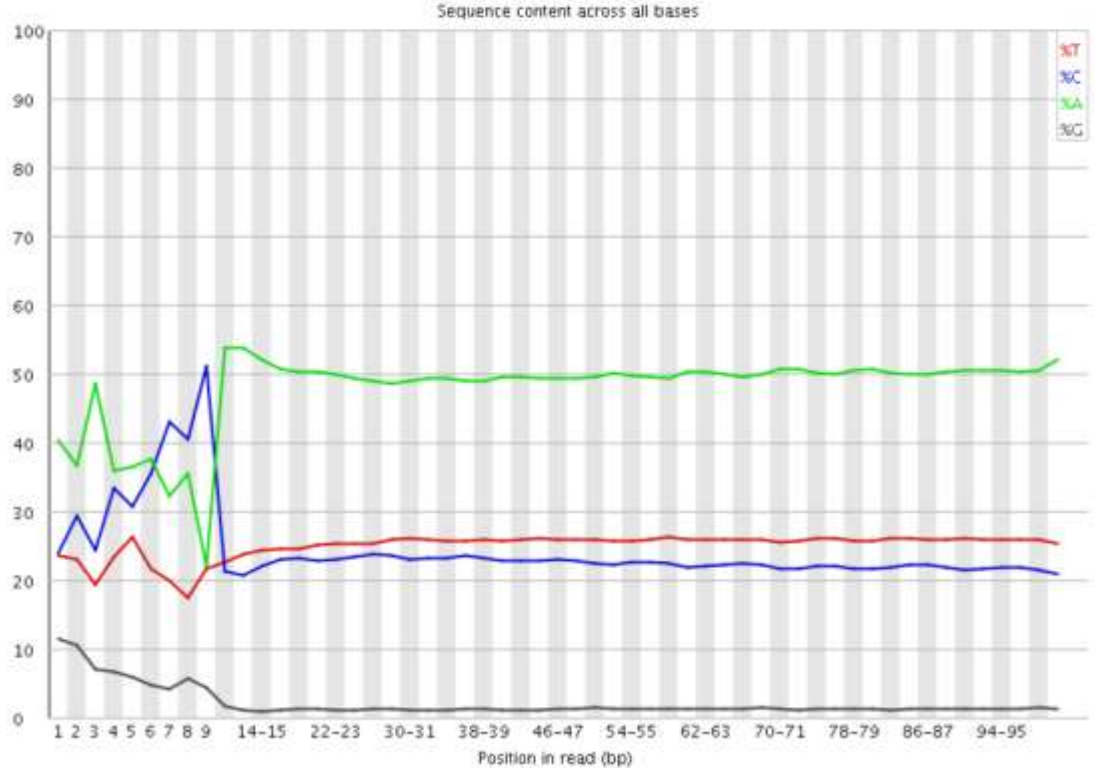
- # of sequences
- Basecall qualities
- Base composition
- Potential contaminants
- Expected duplication rate

Basic Statistics

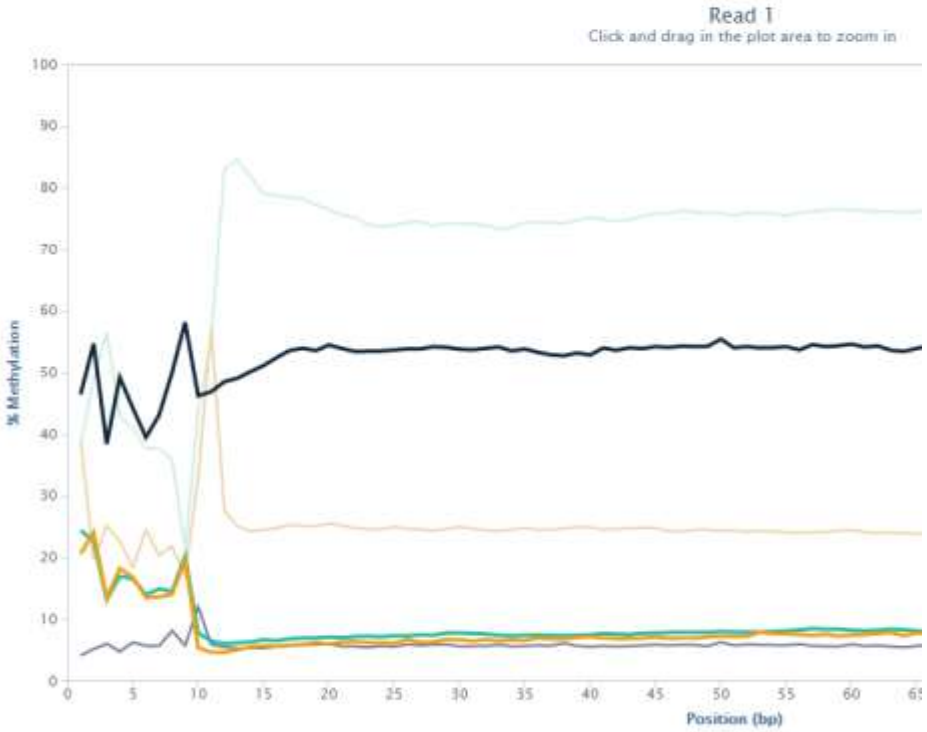
Measure	Value
Filename	s_4_1_sequence.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	35290120
Sequence length	40
%GC	46

质量控制

Per base sequence content

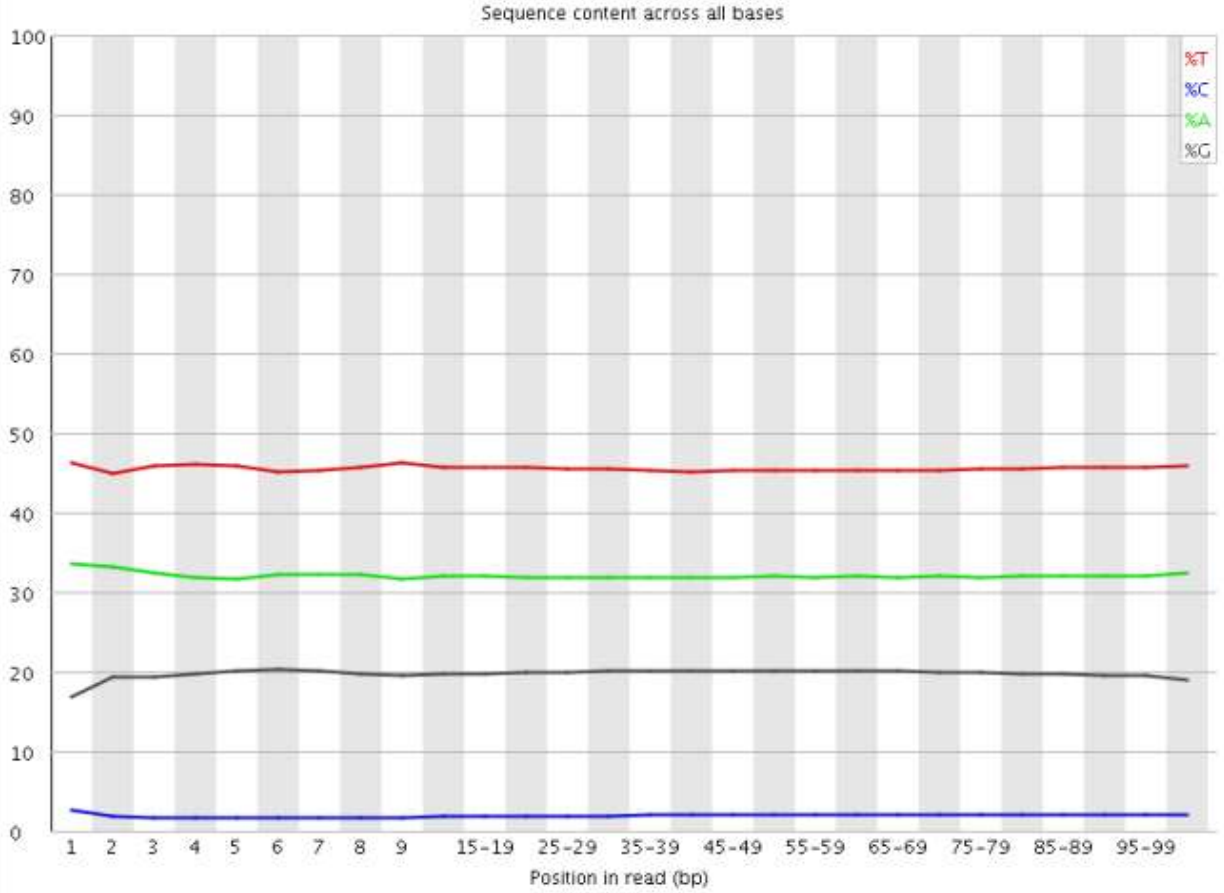


M-Bias Plot

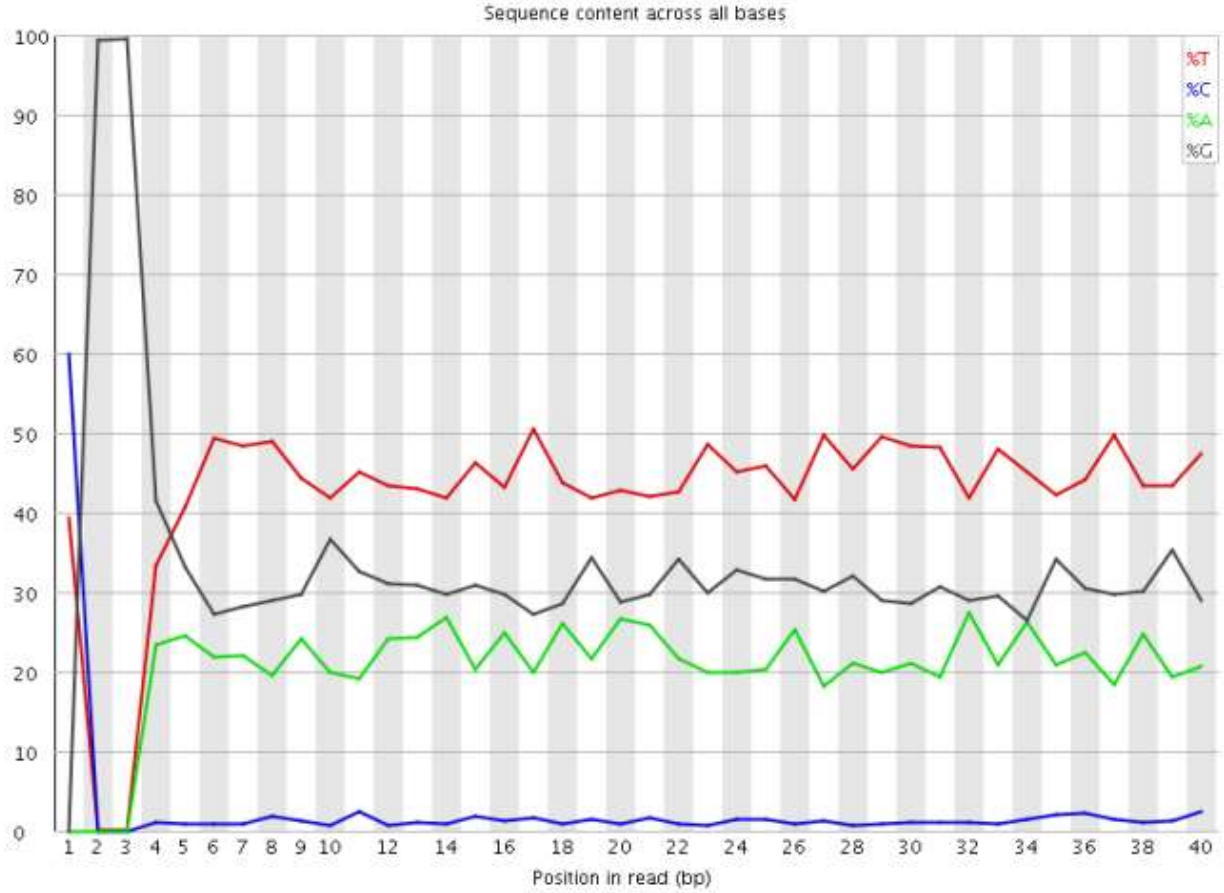


质量控制

WGSBS

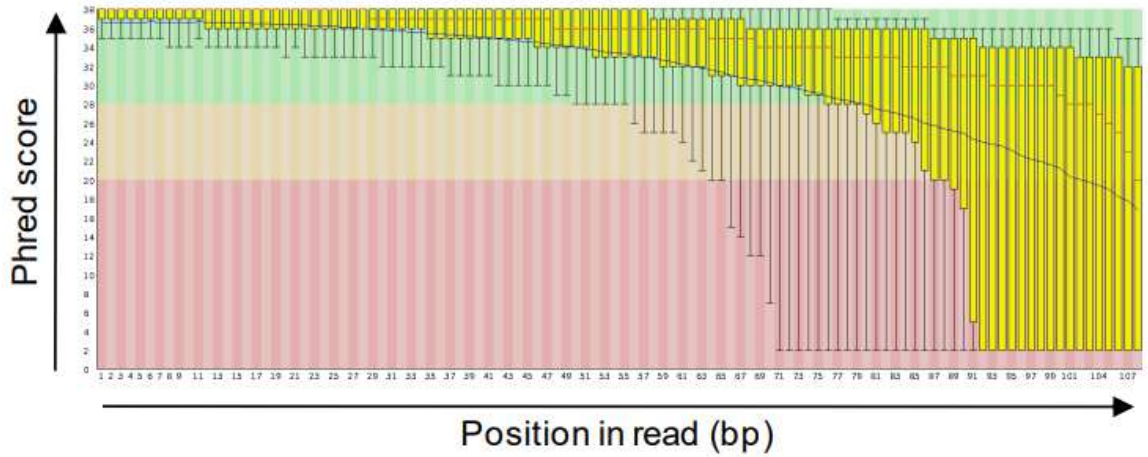


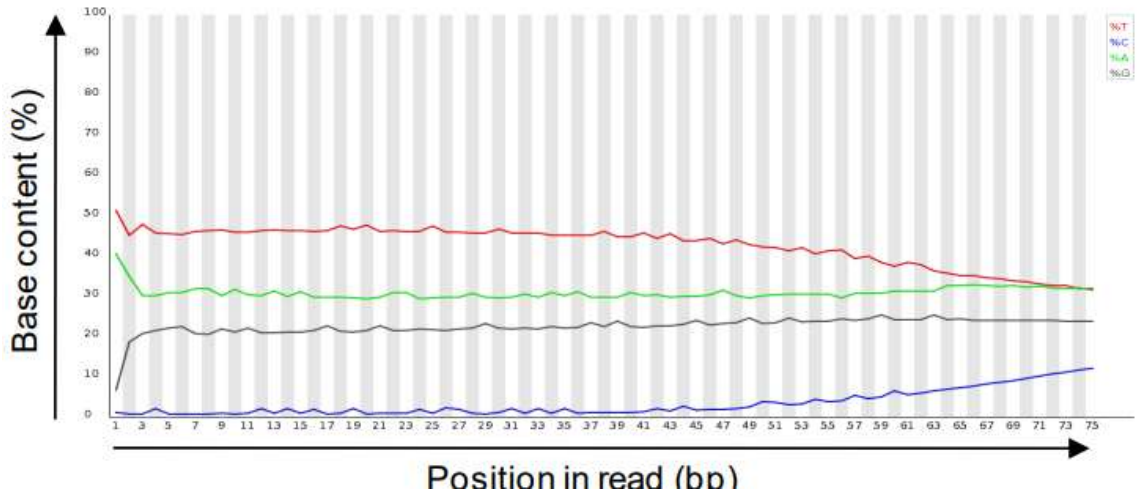
RRBS



质量控制

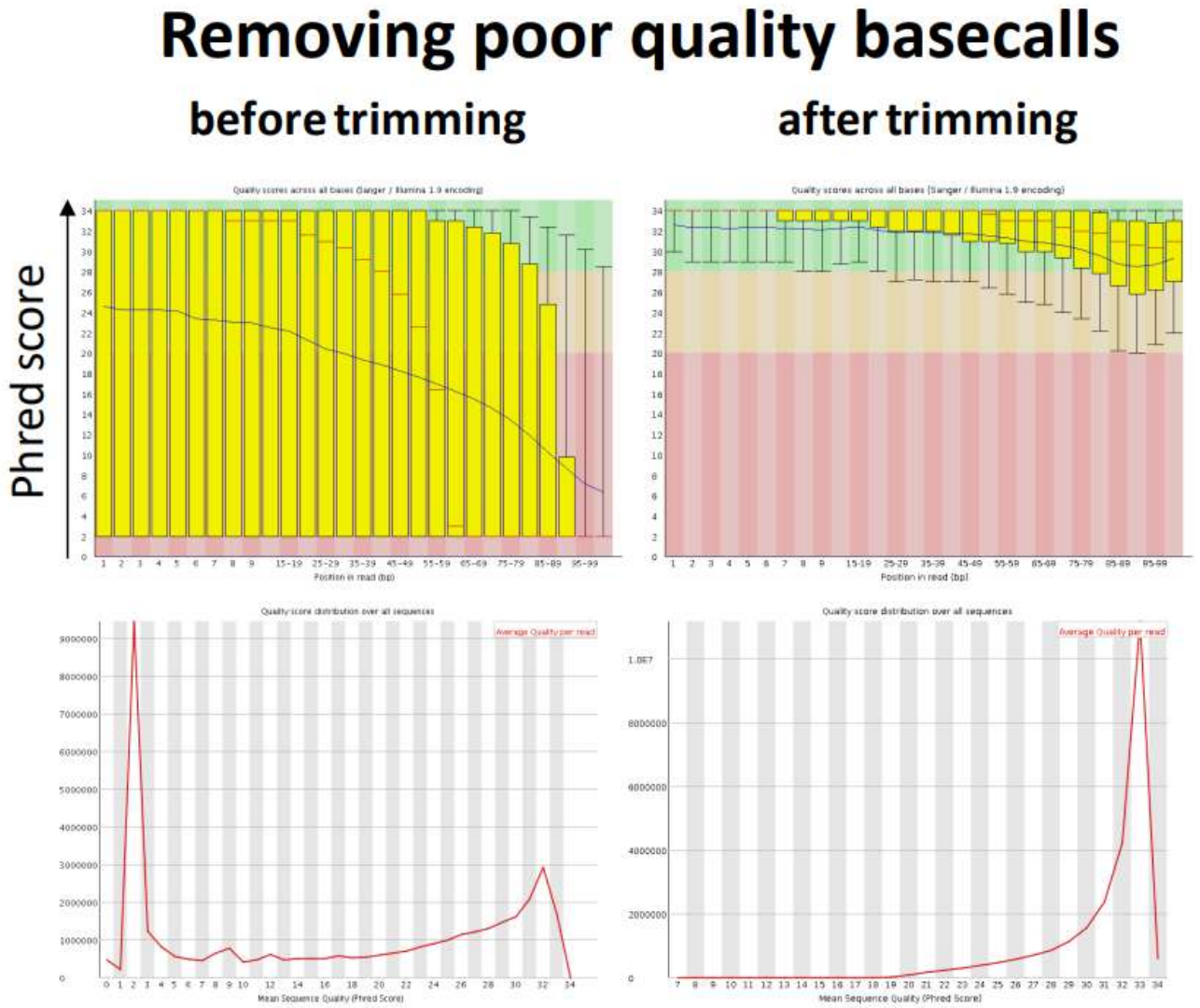
Common problems in BS-Seq



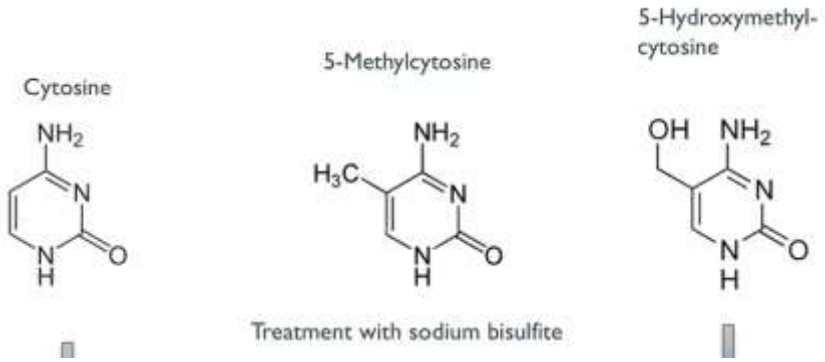


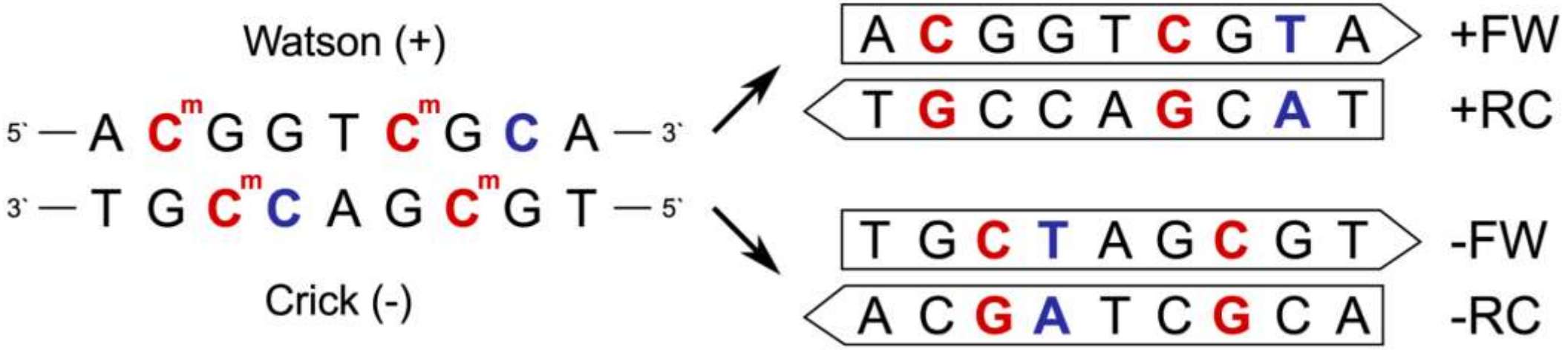
Not observed
in 'normal' libraries,
e.g. ChIP or RNA-Seq

质量控制

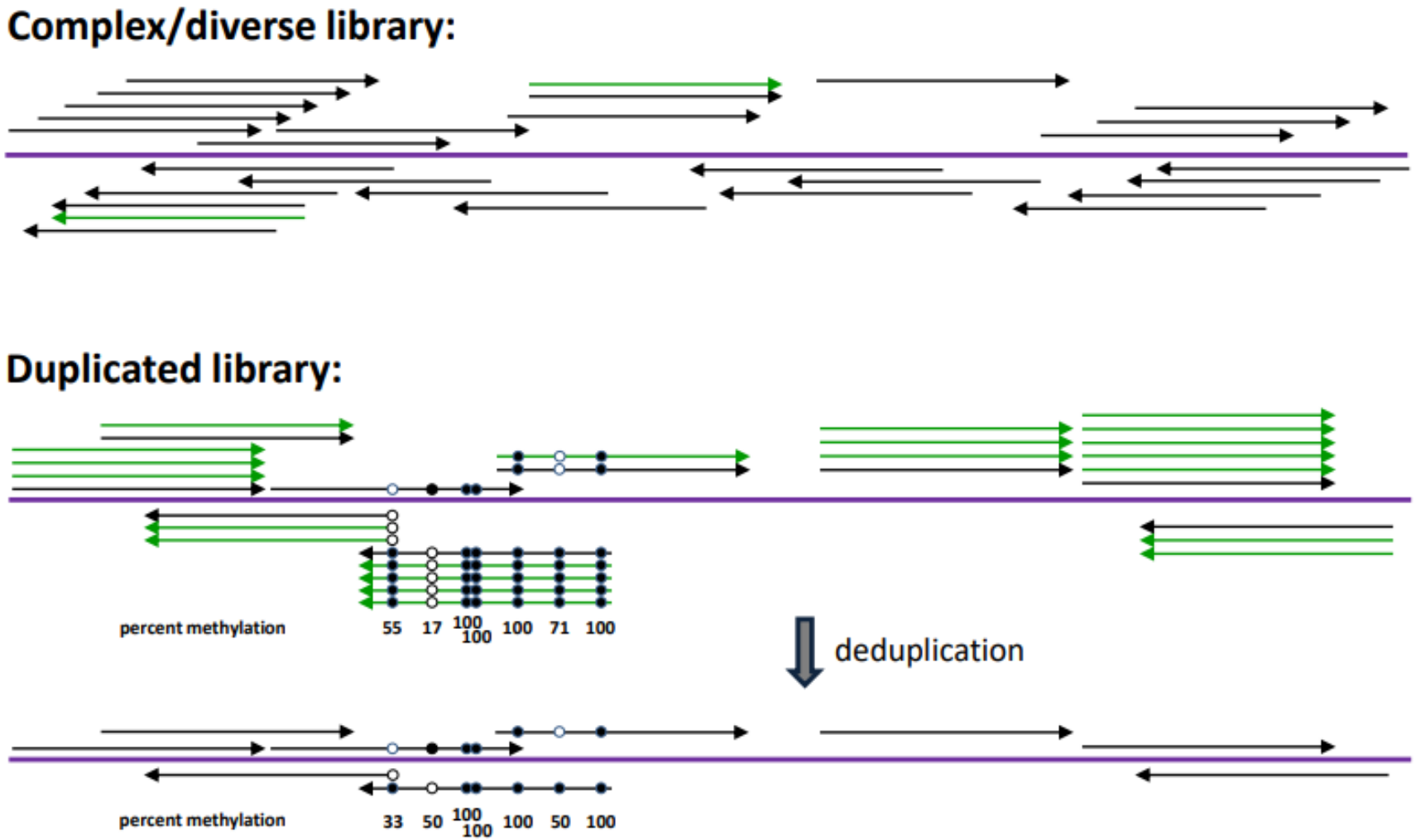

Removing poor quality basecalls
before trimming
after trimming
Phred score
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)
Quality score distribution over all sequences
Average Quality per read
Mean Sequence Quality (Phred Score)

单细胞DNA甲基化测序序列比对



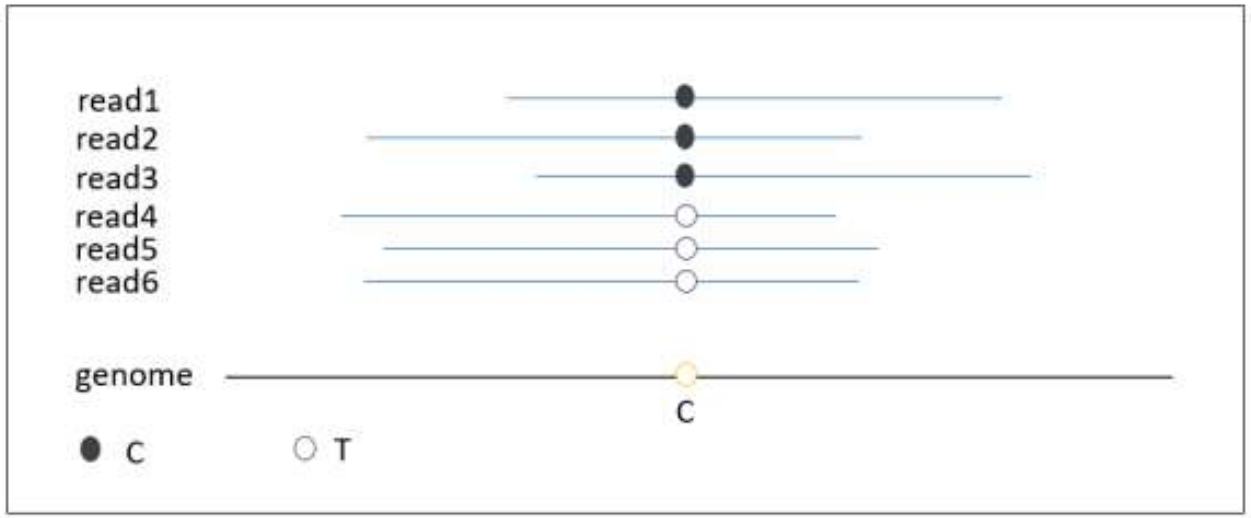


比对后去重


Complex/diverse library:
Duplicated library:
percent methylation
55 17 100 100 100 71 100
deduplication
percent methylation
33 50 100 100 100 50 100

怎么计算甲基化水平

WGBS样本



甲基化率 = C / （C + T）

甲基化率 = 3 / （3 + 3）= 0.5

Call Methylation 和 甲基化文件

染色体	正负链	基因组位置	状态	Context	甲基化水平	m_reads	total_reads
chr1	C	3000827	CG	CG	0.00	0	1
chr1	C	3001007	CG	CG	0.00	0	2
chr1	C	3001018	CG	CG	0.00	0	2
chr1	C	3001277	CG	CG	0.00	0	1
chr1	C	3001629	CG	CG	0.00	0	3
chr1	G	3005999	CG	CG	0.00	0	1
chr1	G	3006188	CG	CG	0.00	0	1
chr1	G	3014742	CG	CG	0.00	0	1
chr1	G	3014775	CG	CG	0.00	0	1
chr1	C	3016725	CG	CG	0.00	0	1
chr1	C	3016928	CG	CG	1.00	1	1
chr1	C	3017247	CG	CG	0.00	0	1
chr1	C	3017509	CG	CG	1.00	1	1
chr1	C	3017887	CG	CG	1.00	1	1
chr1	C	3020332	CG	CG	1.00	1	1
chr1	G	3020725	CG	CG	1.00	1	1
chr1	G	3020795	CG	CG	0.00	0	1
chr1	G	3020815	CG	CG	0.00	0	1
chr1	G	3021346	CG	CG	1.00	2	2
chr1	G	3022017	CG	CG	0.00	0	1
chr1	G	3022043	CG	CG	0.00	0	2

转化效率

- **胞嘧啶甲基化状态被称为甲基化，一般是至少三个因素的混合**:

  •真正的甲基化

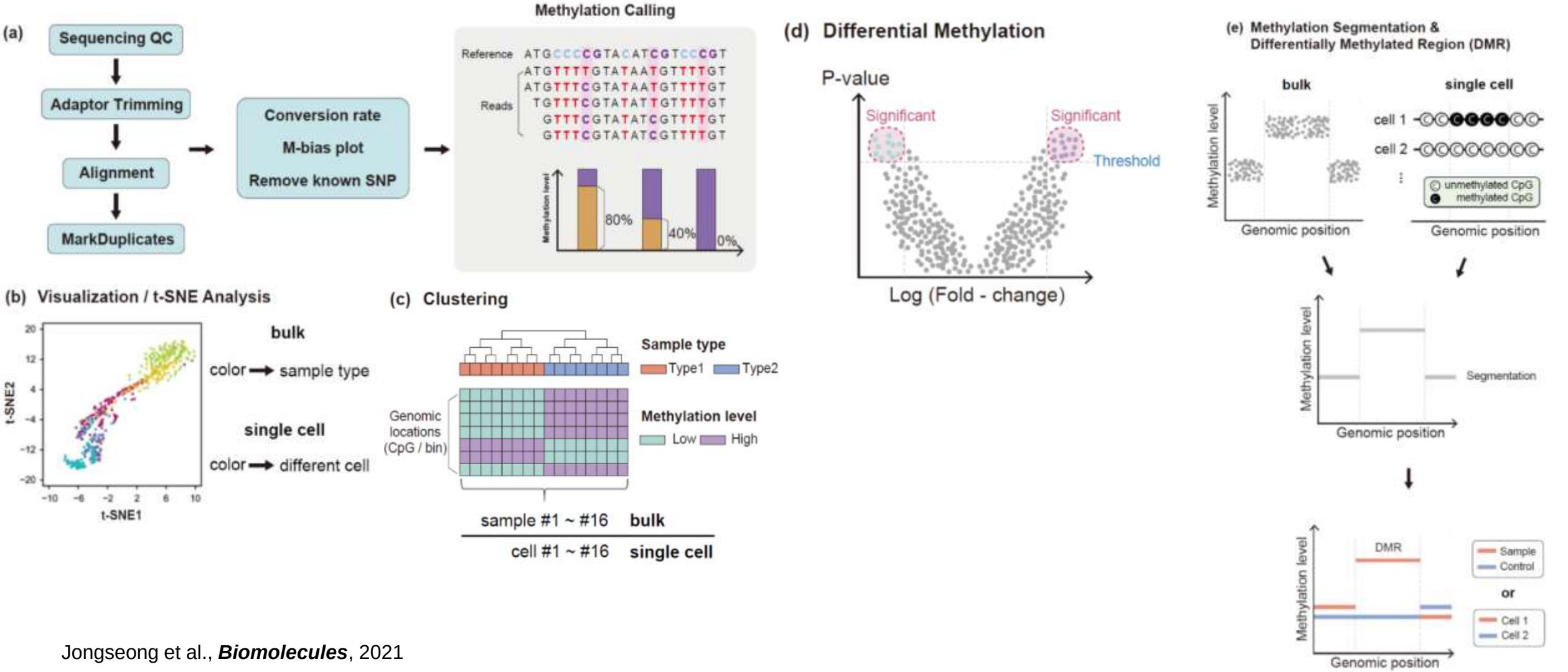
  •亚硫酸氢盐不转化

  •错误比对事件

- **如何判断亚硫酸氢盐转化率**:

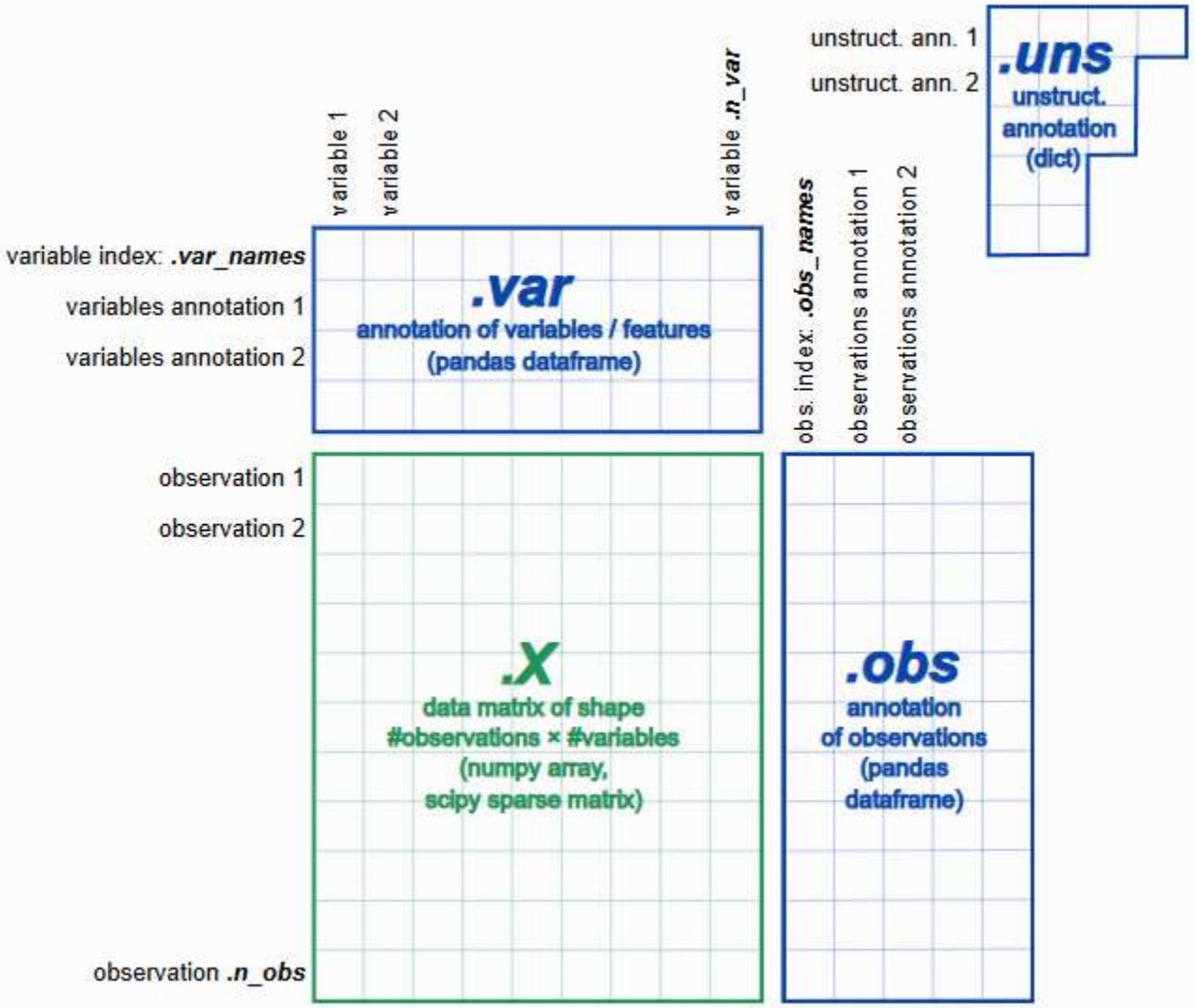
  - Spike-in（例如Lambda、phiX、M13）
  - 如果您没有任何spike-in对照, 可以简单地查看实验中观察到的最低甲基化水平。在哺乳动物系统中，非CG环境下的甲基化率通常非常低，因此仅查看非CG环境下的甲基化水平可能就足够了。例如，如果您在数百万的位点中看到非CG环境下的总体甲基化率为0.3%，那么亚硫酸氢盐转化效率至少为99.7%。
  - 找到甲基化程度最低的区域: 如CpG岛、线粒体（chrMT/chrM）、植物中与叶绿体序列比对的读段（叶绿体序列也不应该被甲基化）

单细胞DNA甲基化数据下游分析流程



Jongseong et al., ***Biomolecules***, 2021

使用Episcanpy进行下游分析

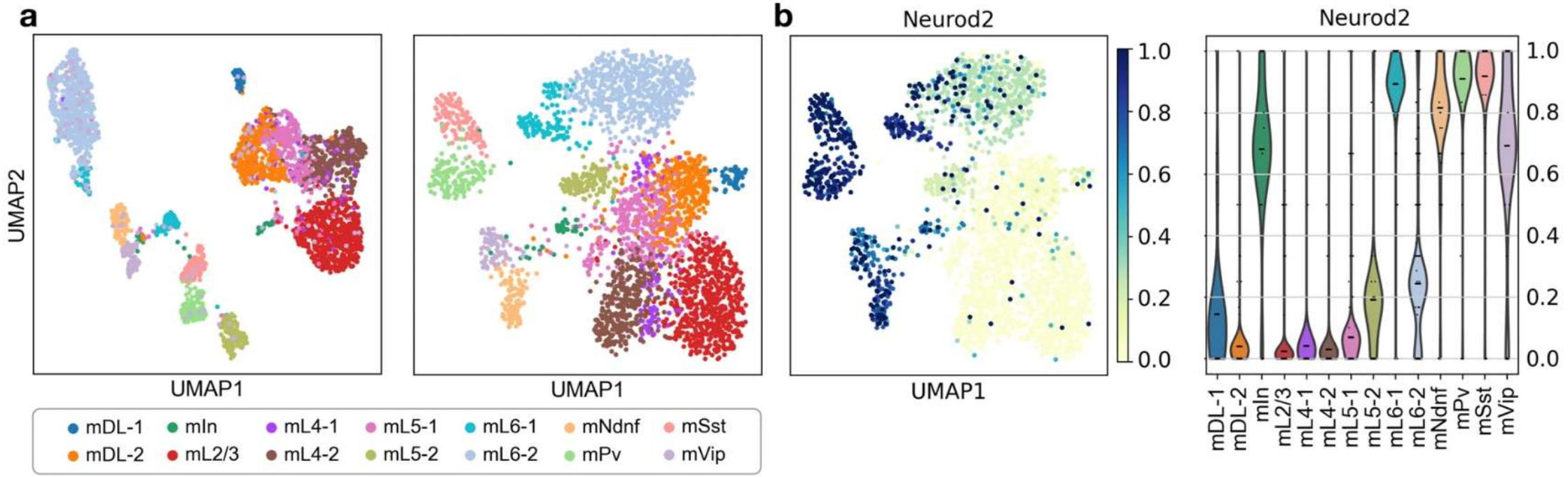


EpiScanpy – Epigenomics single cell analysis in python

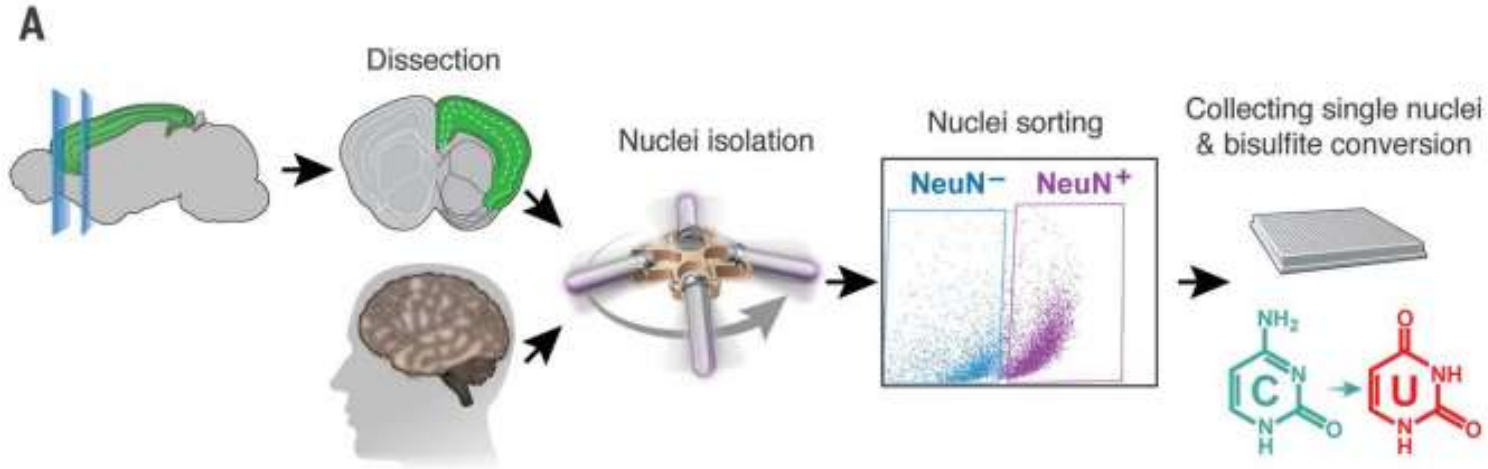
分析案例

Fig. 2: Clustering, visualisation and cell-type annotation for single-cell DNA methylation data and scATAC-seq data.

From: EpiScanpy: integrated single-cell epigenomic analysis



分析案例



NEURAL EPIGENOMICS

Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex

Chongyuan Luo,[1,2*] Christopher L. Keown,[3*] Laurie Kurihara,[4] Jingtian Zhou,[1,5] Yupeng He,[1,5] Junhao Li,[3] Rosa Castanon,[1] Jacinta Lucero,[6] Joseph R. Nery,[1] Justin P. Sandoval,[1] Brian Bui,[6] Terrence J. Sejnowski,[2,6,7] Timothy T. Harkins,[4] Eran A. Mukamel,[3†] M. Margarita Behrens,[6†] Joseph R. Ecker[1,2†]

- Single cell DNA methylation
- > 3000 cells from human/mouse frontal cortex

感谢观看

上机测试文档