计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(11):254−259 [doi: 10.15888/j.cnki.csa.008152]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

导向定位测序数据的甲基化序列比对算法优化①

刘梦雅[1,2], 徐　云[1,2]

[1](中国科学技术大学 计算机科学与技术学院, 合肥 230027)
[2](安徽省高性能计算重点实验室, 合肥 230026)
通讯作者: 徐　云, E-mail: xuyun@ustc.edu.cn

摘　要: 导向定位测序 (GPS) 是一种全基因组 DNA 甲基化检测的新测序技术, 产生的测序数据具有成本低、没有序列偏好等优势. 目前, 甲基化分析中最重要的一步是将其测序产生的序列比对到参考基因组上. 但是, 现有导向定位测序的方法使用 Smith-Waterman 进行局部序列比对, 时间消耗过大且容易对序列比对位置产生误判. 因此, 提出一种导向定位测序数据的改进比对算法, 该算法利用其双端测序的优势, 先用甲基化序列端数据进行序列比对, 对多位置匹配的序列再利用常规数据端数据进行比对位置确定. 实验结果表明: 本文方法和现有方法的准确率相当, 而具有更高的唯一比对比率, 时间性能有 3 倍以上的提升.

关键词: 甲基化; 导向定位测序; 亚硫酸氢盐测序; 序列比对; 相似性

引用格式: 刘梦雅,徐云.导向定位测序数据的甲基化序列比对算法优化.计算机系统应用,2021,30(11):254–259. http://www.c-s-a.org.cn/1003-3254/8152.html

Optimization of Methylation Sequences Alignment Algorithm Based on GPS Data

LIU Meng-Ya[1,2], XU Yun[1,2]

[1](School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China)
[2](Key Laboratory of High Performance Computing of Anhui Province, Hefei 230026, China)

Abstract: Guide Positioning Sequencing (GPS) is a novel method for genome-wide DNA methylation detection. The generated sequencing data has the advantages of low detection cost and no sequence preference. At present, the most important step in methylation analysis is to align the sequences to the reference genome. However, the existing method uses Smith-Waterman for local sequence alignment, which takes too much time and affects the mapping efficiency. Therefore, a new alignment algorithm for the GPS data is proposed. The algorithm uses the advantages of paired-end sequencing to determine the alignment positions. The methylation sequences are first aligned to the reference genome, and then corresponding regular sequences are used to determine the final positions. The experimental results show that compared with the existing method, the method presented in this paper has a high mapping efficiency with comparable accuracy and the time performance improved by more than 3 times.

Key words: methylation; Guide Positioning Sequencing (GPS); bisulfite sequencing; sequence alignment; similarity

DNA 甲基化是指在 DNA 甲基化转移酶的作用下, 将甲基化基团选择性的添加到胞嘧啶 (C) 的过程. 因此, 在人类基因组中一部分 C 被甲基化, 另一部分 C 未被甲基化, 在未改变基因序列的前提下, 控制基因表达[1,2]. 大量研究表明, 基因组中甲基化 C 的比例和所在区域, 能够为疾病的预测提供帮助, 同时也被证实在包括癌症在内的诸多疾病的治疗中发挥着重要的作用[3–6].

围绕全基因组甲基化的分析是近年研究的热点内

① 基金项目: 国家自然科学基金面上项目 (61672480)
Foundation item: General Program of National Natural Science Foundation of China (61672480)
收稿时间: 2021-01-26; 修改时间: 2021-02-24; 采用时间: 2021-03-03; csa 在线出版时间: 2021-10-22

容, 其中最重要的一步是将测序所产生的序列, 比对到参考基因组上, 获取整个基因组的甲基化状态[7]. 目前常用的测序技术是亚硫酸氢盐的全基因组甲基化测序, 但由于此类测序技术需要用亚硫酸氢盐对原始 DNA 片段进行预处理, 降低了序列的复杂性, 增加了后续比对的难度. 2019 年出现的导向定位测序数据很好地解决这一问题, 其利用双端测序的优势: 一端是亚硫酸氢盐处理后的序列 Read1, 另一端是原始序列 Read2, 没有经过处理的原始序列更容易确定在参考基因组上的位置, 通过双端测序序列的位置关系, 实现对甲基化的精确检测[8].

然而, 现有导向定位测序数据 (GPS) 的比对方法先确定原始序列 R2 的 20 个候选比对位置, 时间消耗大; 之后再用动态规划算法确定甲基化序列 Read1 的比对位置, 算法本身的时间成本高, 且需对多个候选位置进行动态规划验证. 同时, 根据 Read2 确定 Read1 的比对位置过于绝对, 可能会产生误判. 现有亚硫酸氢盐测序 (BS) 中的比对方法能将 70%–90% 的序列确定到唯一的位置, 比对的准确率高达 99%, GPS 数据的现有比对方法, 相比之下仍有较大改进空间[9].

因此, 本文提出一种新的导向定位测序数据的比对算法. 由于亚硫酸氢盐序列比对精度高达 99%, 对于能确定唯一位置的甲基化序列不再用常规序列进行定位, 保证高精度的同时节约了时间. 首先确定导向定位测序数据中的甲基化序列的候选比对位置; 然后根据甲基化序列和常规序列在参考基因组上对应的位置关系过滤偏离区域; 最后使用唯一比对序列的信息确定最佳比对位置. 充分利用辅助信息, 实现以时间高效的方式将更多的甲基化序列比对到参考基因组上.

1 相关工作

目前, 对 DNA 甲基化进行检测的金标准是亚硫酸氢盐测序的全基因组甲基化测序, 随着导向定位测序数据的出现, 在实现对全基因组甲基化位点高度覆盖的同时, 带来了新的研究问题. 接下来根据全基因组 DNA 甲基化测序数据的类型, 分别介绍数据的特点和相应比对方法, 分析其优缺点.

1.1 亚硫酸氢盐测序 (BS) 及其比对方法

亚硫酸氢盐测序技术通过对基因片段进行预处理, 使得甲基化的胞嘧啶 (C) 保持不变, 未发生甲基化的 C 先转换成尿嘧啶 (U), 再转换为胸腺嘧啶 (T), 如图 1 所示[10]. 因此, 在 DNA 甲基化序列比对的过程中, 序列中的 T 有可能比对到参考基因组上的 T 或 C, 但反之不行, 导致比对的难度增加[11]. 这是甲基化序列比对, 同常规 DNA 序列比对的不同之处. 测序得到的基因序列, 称为 BS-reads. 甲基化分析中很重要的一步就是将 BS-reads 比对到参考基因组上, 确定其位置.

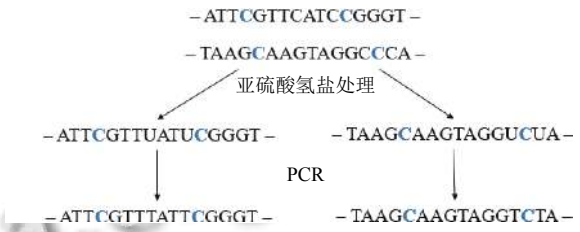

图 1 亚硫酸氢盐测序过程

现有亚硫酸氢盐测序的比对方法分成两种, 分别是基于三字符集和基于通配符的比对方法. 基于三字符集方法的特性, 是把 BS-reads 和参考基因组中的 C 都转化成 T, 将问题转化成常规的 DNA 序列比对, 在候选位置确定后, 再根据转化前的相似性对候选位置进行过滤, 代表方法有 Bismark[12], GEMBS[13], BS-Seeker3[14], BatMeth2[15]. 基于通配符方法的特性是 BS-reads 中的 C 转化成一个通配符, 同时允许通配符比对到参考基因组上的 C 和 T, 代表方法有 BSMAP[16] 和 RMAP[17].

随着越来越多的甲基化数据被测出, 这两类比对方法针对 BS-reads 不对称比对的特点, 适应序列长度短 (40 bp–400 bp)、数量多、规模大的特性, 实现将甲基化序列快速比对到参考基因组上, 使得全基因组甲基化分析成为可能. 但亚硫酸氢盐预处理将未发生甲基化的 C 转化成 T, 在大部分序列比对中, 字符集从 4 字符集 (A、T、C、G) 变成了 3 字符集 (A、T、G), 降低了序列的复杂性, 增加了 BS-reads 唯一比对位置确定的难度, 同时使参考基因组中重复区域的甲基化状态分析更为艰难.

1.2 导向定位测序 (GPS) 及其比对方法

导向定位测序是一种新的全基因组 DNA 甲基化检测的方法. 每条 DNA 链是由磷酸和脱氧核糖构成, 3’端和 5’端表示 DNA 链的两端, 其中连接磷酸基团的一端为 5’端, 另一端是 3’端. DNA 的复制方向是从 5’端到 3’端. 测序数据中 3’端的序列保持不变, 5’端的未甲基化的 C 转化成 T, 甲基化的 C 保持不变[8]. 获得的

两条 DNA 序列 (Read1 和 Read2), 其中 Read1 中未甲基化的 C 转化成 T, 和亚硫酸氢盐测序方法处理后的序列特性一致; Read2 是原始 DNA 序列, 更容易比对到参考基因组上, 如图 2 所示. 在 Read2 比对到参考基因组之后, Read1 比对到参考基因组的范围也相应确定. 其中 Read2 对 Read1 位置的确定起到定位作用, 为后续全基因组甲基化的分析奠定了基础.

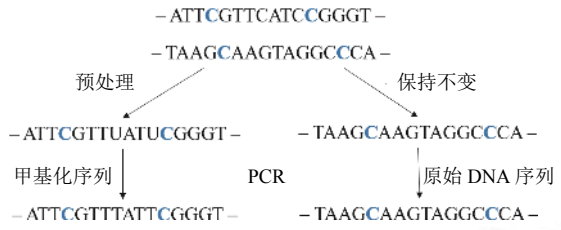

图 2 导向定位测序过程

现有方法调用 bowtie2[18] 将 Read2 比对到参考基因组上, 获取 Read2 在参考基因组上的 20 个候选比对位置. 由于 Illumina 测序原理可得, Read1 位于 Read2 下游的相反链上, 且由于 GPS 测序库中的碎片大小是 400 bp–500 bp, 可以确定 Read2 的比对范围. 通过使用 Smith-Waterman 算法[19], 将 Read1 比对到 Read2 所在参考基因组下游 1 kb 的反链上, 获取 Read1 和参考基因组上局部相似性最高的位置.

新型测序数据的出现, 实现以较低的覆盖率 (5X) 获得甲基化序列, 降低了甲基化测序的成本, 检测甲基化没有序列偏好. 同时, 比对过程中使用 Smith-Waterman 算法, 允许 Read1 中的 T 比对到参考基因组上的 T 或 C, 以解决甲基化序列比对中 C/T 的不对称性比对问题. 为受亚硫酸氢盐预处理影响较大的基因片段和部分物种, 提供了甲基化分析的新方法, 使得这部分序列甲基化信息的精确检测成为可能. 但使用 Smith-Waterman 算法对多个候选比对位置进行动态规划验证, 需要大量的时间. 且未考虑仅允许 Read1 中的 C 比对到参考基因组中的 C, 有可能导致 Read1 的错误比对, 从而影响到后续全基因组甲基化的分析. 且现有 GPS 数据的唯一比对比例为 79.8%–82.3%, 仍有提升的空间.

2 比对算法设计和优化

本文首先将导向定位数据中的甲基化序列比对到参考基因组上, 随后利用和常规序列间的位置关系对候选位置进行过滤, 最后对仍不能确定位置的甲基化序列, 利用唯一比对位置的信息进行定位, 该方法主要包括 4 个步骤: (1) 数据预处理; (2) 定位候选位置; (3) 过滤偏离区域; (4) 确定最佳位置.

2.1 数据预处理

由于 GPS 库的建立, 需要用到 T4 DNA 聚合酶处理基因片段, 从而保证 Read2 中的序列和原始 DNA 片段一致, 最后获取双端测序序列 (Read1 和 Read2). 但 T4 DNA 聚合酶可能产生处理不足或过度处理的现象, 直接影响获取数据的准确性, 影响比对的效率. 所以, 需要找到 Read1 和 Read2 处理的边界, 进而对数据进行预处理[8].

参考基因组中 CH 的甲基化水平较低, 若序列中出现 CH, 则说明酶处理充分. Read2 位于参考基因组的反链上, 根据碱基互补配对原理, 可知 CH 在 Read2 上的表现形式是 [A/G/T]G. 通过寻找 [A/G/T]G 确定酶处理边界, 对 Read2 进行预处理. 如图 3 所示, 最靠近右端, 且满足要求的处理边界是 TG. 确定处理边界后, 保留边界右边的序列作为处理后的 Read2 序列.

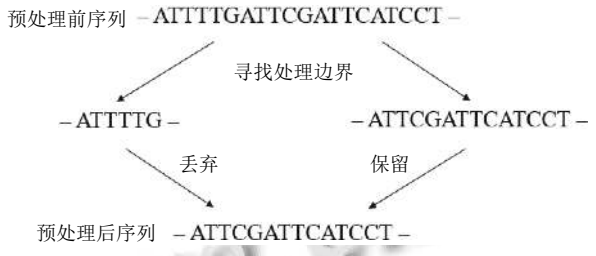

图 3 数据处理示意图

2.2 定位候选位置

本文直接将甲基化序列比对到参考基因组上. 一方面由于亚硫酸氢盐测序序列比对准确率较高, 另一方面易比对到多个位置的比例约为 20%–30%, 直接比对甲基化序列在保证准确率的同时, 减少了后续的计算成本. 本文使用基于三字符集方法和种子扩展策略的亚硫酸氢盐比对工具 BitmapperBS[9] 进行修改, 其包含高效的数据结构 FM-tree, 针对数据三字符集特性对传统 FM-index 索引进行优化, 能够获得高达 99.36% 的准确率.

首先 Read1 比对到参考基因组后, 分成两部分. 如图 4 所示, 将能够确定唯一位置的序列称为 Unique Reads; 比对到多个位置的序列称为 Multireads, 这部分序列比对到参考基因组的多个相似度较高的位置, 或者比对到了参考基因组的重复区域.

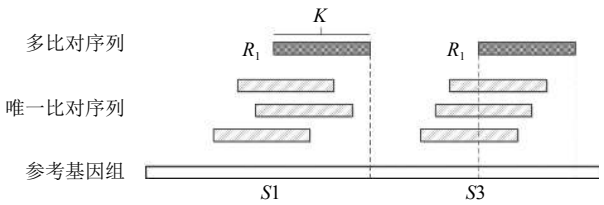

图 4 唯一比对序列和多比对序列举例

后续处理主要针对 Multireads, 找到其至多 20 个候选比对位置. 将 Multireads 符号化表示为集合 M, 设 R_1 为集合 M 中的一条序列, 候选比对位置的个数为 n, 其候选比对位置集合 $P(R_1)$ 表示为:

$$P(R_1)=\{p_1,p_2,\cdots,p_n|R_1\in M,n\le 20\} \quad (1)$$

2.3 过滤偏离区域

针对 GPS 数据中的测序原理可得, Read1 位于 Read2 下游的相反链上, 且距离相差不大于 1000 bp. 通过利用 Read2 的位置信息作为辅助信息, 对 Read1 的候选比对位置进行限制, 过滤位于偏离区域的候选位置.

设与 R_1 相对应的另一端序列是 R_2, 首先使用 bowtie2 将 R_2 比对到参考基因组上, 其候选比对的个数为 m, 得到候选比对集合:

$$P(R_2)=\{pos_1,pos_2,\cdots,pos_m|m\le 20\} \quad (2)$$

对 Read1 和 Read2 的候选位置进行两两比较, 过滤掉 Read1 候选比对集合中不能与 Read2 成对的位置. 如图 5 所示, R_1 的候选比对位置集合 $P(R_1)$ 中只有 p_1 和 p_3 存在与之相对应的 pos_1 和 pos_3, 所以对其余位置进行过滤, 此时 $P(R_1)=\{p_1,p_3\}$. 若此时 R_1 的候选比对位置个数为 1, 则转化为 Unique Reads, 否则其仍在 Multireads 的集合 M 中.

图 5 过滤偏离区域

2.4 确定最佳位置

过滤偏离区域后, 使用与 Multireads 重叠的 Unique Reads 信息, 以及对应参考基因组之间的信息, 逐个碱基计算相应位置的可能性, 最后对候选集合中每个位置得到一个总的得分, 确定最有可能的比对位置.

设甲基化序列 R_1 的长度为 K, 比对到候选比对位置的概率 S 为:

$$S=\prod_{k=1}^{K}s_k \quad (3)$$

其中, R_1 的第一个碱基比对到参考基因组对应位置的概率为 s_1, 依次类推得第 K 个碱基比对到参考基因组对应位置的概率为 s_K. 如图 6 所示, s_1–s_K 的计算使用工具 BAM-ABS[20], 该工具使用贝叶斯模型, 以 Multireads 和参考基因组之间的错配信息和对应甲基化区域信息; 以及重叠 Unique Reads 中获得的 SNP 和甲基化区域信息作为先验概率, 计算比对到每个位置的可能性. 最后选取候选比对集合中得分最高的位置为最佳比对位置.

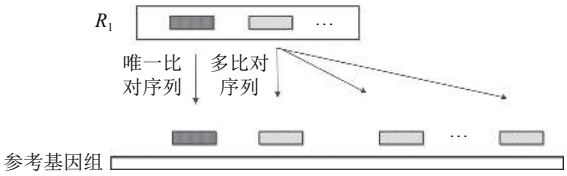

图 6 找最佳位置的过程

3 实验分析

3.1 数据集和实验环境

本文分别在真实和模拟数据集中对两种方法进行比较, 真实数据集来自 GSE92328, 在文献 [8] 中提出并被证实有利于甲基化信息的分析. 本文使用其中的 GPS 数据 SRR6443657 和 SRR6443658 进行后续实验. 模拟数据集使用模拟工具 ART 和 Astair 获得, 先用 ART 生成常规 DNA 数据, 再通过 Astair 对其中一条序列进行甲基化模拟. 在未特殊声明时, 本文使用工具的默认参数进行比较.

本文的实验平台包括 2 个 14 核 Intel Xeon Gold 5120 处理器和 512 GB 内存, 操作系统为 64 位的 Ubuntu 18.04.

3.2 评价指标

分别使用时间、唯一比对比率和准确率与现有方法进行比较. 其中时间包括数据预处理和得到最终结果的时间, 建索引的时间不包括在内, 因为索引只需建造一次, 在后续实验中通用.

(1) 唯一比对比率

该评价指标表示比对到唯一位置的甲基化序列占

全部甲基化序列的比例. 如式 (4) 所示, U 表示唯一比对序列集合, $n(U)$ 表示唯一比对序列集合中序列的条数, N 表示全部甲基化序列的条数.

$$Recall = \frac{n(U)}{N} \tag{4}$$

(2) 准确率

准确率这里表示唯一比对序列中, 比对到正确的位置所占的比例. 如式 (5) 所示, $n(R)$ 是唯一比对集合 U 中比对到正确位置的序列个数. 模拟数据集中序列在参考基因组上的位置是已知的, 当真实位置和比对结果相差 200 bp 以内, 则认为比对正确. 真实数据集中序列在参考基因组上的位置是未知的, 故不进行准确率的验证.

$$Accuracy = \frac{n(R)}{n(U)}, R \in U \tag{5}$$

3.3 实验结果

分别使用模拟数据集和真实数据集探究本文方法和现有方法[8] 的性能优劣. 模拟数据集和真实数据的序列长度均为 100 bp. 数据规模分别为甲基化序列 1w 条、5w 条、10w 条, 常规 DNA 序列 1w 条、5w 条、10w 条.

(1) 模拟数据集实验结果

如表 1 所示, 本文改进方法和现有方法相比, 准确率相差不大, 最多相差 0.7%. 而本文方法获得 3–30 倍时间性能的提升, 随着数据规模的增大, 对时间性能的提升越明显. 同时本文方法获得 6%–10% 唯一比对比率的提升, 将更多的序列比对到唯一位置, 有利于后续甲基化信息的分析. 因模拟数据集不能完全模拟真实数据中插入、删除, 以及发生测序错误、结构变异的情况, 更容易比对到参考基因组上, 唯一比对比率相比真实数据更高.

表 1 模拟数据集实验结果

数据集	比对工具	运行时间	准确率 (%)	唯一比对比率 (%)
1w	现有方法[8]	3′27.09″	99.86	80.35
	本文方法	49.66″	99.18	90.66
5w	现有方法[8]	15′8.65″	99.21	85.23
	本文方法	54.04″	99.08	91.77
10w	现有方法[8]	34′52.58″	99.70	80.05
	本文方法	55.47″	99.21	90.50

(2) 真实数据集实验结果

通过实验探究了不同方法在运行时间、唯一比对比率方面的优劣. 如表 2 所示, 在 SRR6443657 数据集中, GPS 方法的运行时间从 4 min 到 38 min, 受数据规模影响较大; 本文的改进方法在这 3 种数据规模下运行时间相差不大, 为 56–67 s, 小数据集时比现有方法快约 3 倍, 大数据集时比现有方法快 30 倍, 对大规模数据集的提升效果更为明显. 同时, GPS 比对方法的唯一比对比率为 79.32%–80.09%, 数据集规模对唯一比对比率的影响不大; 本文的改进方法唯一比对比率从 85.37% 到 89.32%, 比之前方法提升了 5%–10%, 且本文方法随着数据集规模越大, 唯一比对比率越来越大, 因获取比对到唯一位置的序列信息越多, 更容易比对到唯一位置. 第 2 个数据集整体结果和第 1 个数据集相似, 但唯一比对比率提升约为 2%–6%, 较上一个数据集提升不明显. 实验中发现部分甲基化序列未能找到与之配对的常规 DNA 序列, 使得该数据集比对难度增加.

表 2 真实数据集实验结果

数据集		现有方法[8]		本文方法	
		运行时间	唯一比对比率 (%)	运行时间 (s)	唯一比对比率 (%)
SRR6443657	1w	4′38.07″	80.09	56.13	85.87
	5w	24′39.40″	79.52	67.01	88.65
	10w	38′39.37″	79.32	62.09	89.32
SRR6443658	1w	5′42.05″	80.00	58.06	82.23
	5w	21′17.23″	79.64	59.33	84.73
	10w	43′17.3″	79.29	60.93	85.49

4 结论与展望

本文提出了一种高效的导向定位测序数据的比对算法, 首先对数据进行预处理, 将甲基化序列定位到参考基因组上; 再利用双端测序中两端序列的位置关系, 对甲基化序列的候选比对位置集合进行过滤; 最后通过比对到唯一位置的序列包含的信息, 找到最佳比对位置. 实验结果表明, 本文方法能够加速比对过程, 将更多的甲基化序列比对到唯一位置, 且对大规模数据

集的性能提升效果更为明显. 下一步的研究工作是提出启发式的算法, 探究影响准确率的因素, 在比对精度上取得更好的效果, 并探究比对性能的提升对后续甲基化信息的影响.

参考文献

1 Moore LD, Le T, Fan GP. DNA methylation and its basic function. Neuropsychopharmacology, 2013, 38(1): 23–38. [doi: 10.1038/npp.2012.112]

2 Adusumalli S, Omar MFM, Soong R, *et al*. Methodological aspects of whole-genome bisulfite sequencing analysis. Briefings in Bioinformatics, 2015, 16(3): 369–379. [doi: 10.1093/bib/bbu016]

3 Capper D, Jones DTW, Sill M, *et al*. DNA methylation-based classification of central nervous system tumours. Nature, 2018, 555(7697): 469–474. [doi: 10.1038/nature26000]

4 Soto J, Rodriguez-Antolin C, Vallespín E, *et al*. The impact of next-generation sequencing on the DNA methylation-based translational cancer research. Translational Research, 2016, 169: 1–18. [doi: 10.1016/j.trsl.2015.11.003]

5 Xu RH, Wei W, Krawczyk M, *et al*. Circulating tumour DNA methylation markers for diagnosis and prognosis of hepatocellular carcinoma. Nature Materials, 2017, 16(11): 1155–1161. [doi: 10.1038/nmat4997]

6 肖婷, 傅俊江. 肿瘤表观遗传标志物的研究进展及应用现状. 中国癌症防治杂志, 2019, 11(2): 93–98. [doi: 10.3969/j.issn.1674-5671.2019.02.02]

7 Zhang S, Qin CX, Cao GQ, *et al*. Genome-wide analysis of DNA methylation profiles in a senescence-accelerated mouse prone 8 brain using whole-genome bisulfite sequencing. Bioinformatics, 2017, 33(11): 1591–1595.

8 Li J, Li Y, Li W, *et al*. Guide Positioning Sequencing identifies aberrant DNA methylation patterns that alter cell identity and tumor-immune surveillance networks. Genome Research, 2019, 29(2): 270–280. [doi: 10.1101/gr.240606.118]

9 程昊宇. 面向大规模测序数据集的序列比对算法研究 [博士学位论文]. 合肥: 中国科学技术大学, 2019.

10 Prezza N, Vezzi F, Käller M, *et al*. Fast, accurate, and lightweight analysis of BS-treated reads with ERNE 2. BMC Bioinformatics, 2016, 17(S4): 69.

11 Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Briefings in Bioinformatics, 2010, 11(5): 473–483. [doi: 10.1093/bib/bbq015]

12 Krueger F, Andrews SR. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics, 2011, 27(11): 1571–1572. [doi: 10.1093/bioinformatics/btr167]

13 Merkel A, Fernández-Callejo M, Casals E, *et al*. gemBS: High throughput processing for DNA methylation data from bisulfite sequencing. Bioinformatics, 2019, 35(5): 737–742. [doi: 10.1093/bioinformatics/bty690]

14 Guo WL, Fiziev P, Yan WH, *et al*. BS-Seeker2: A versatile aligning pipeline for bisulfite sequencing data. BMC Genomics, 2013, 14(1): 774. [doi: 10.1186/1471-2164-14-774]

15 Lim JQ, Tennakoon C, Li GL, *et al*. BatMeth: Improved mapper for bisulfite sequencing reads on DNA methylation. Genome Biology, 2012, 13(10): R82. [doi: 10.1186/gb-2012-13-10-r82]

16 Xi YX, Li W. BSMAP: Whole genome bisulfite sequence MAPping program. BMC Bioinformatics, 2009, 10(1): 232. [doi: 10.1186/1471-2105-10-232]

17 Smith AD, Chung WY, Hodges E, *et al*. Updates to the RMAP short-read mapping software. Bioinformatics, 2009, 25(21): 2841–2842. [doi: 10.1093/bioinformatics/btp533]

18 Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods, 2012, 9(4): 357–359. [doi: 10.1038/nmeth.1923]

19 Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of Molecular Biology, 1981, 147(1): 195–197. [doi: 10.1016/0022-2836(81)90087-5]

20 Tran H, Wu XW, Tithi S, *et al*. A bayesian assignment method for ambiguous bisulfite short reads. PLoS One, 2016, 11(3): e0151826. [doi: 10.1371/journal.pone.0151826]