arXiv:2602.04554v1 [stat.AP] 4 Feb 2026

mmcmcBayes：用于检测差异甲基化区域的多阶段马尔可夫链蒙特卡洛框架 R 软件包

Zhexuan Yang[1], Duchwan Ryu[2], 和 Feng Luan[2]

[1]统计学系，宾夕法尼亚州立大学[2]统计学系与精算科学系，北伊利诺伊大学

摘要

识别差异甲基化区域是表观基因组关联研究中的一项重要任务，在这类研究中，差异信号通常出现在相邻 CpG 位点之间的组间差异中。许多现有方法通过汇总 CpG 级检验结果来检测差异甲基化区域，这可能会限制它们捕捉复杂区域甲基化模式的能力。本文介绍了 R 软件包 **mm-cmcBayes**，该软件包实现了一种用于差异甲基化区域区域级检测的多阶段马尔可夫链蒙特卡罗程序。该方法使用 alpha-偏斜广义正态分布对样本级区域甲基化汇总进行建模，并通过贝叶斯因子评估组间差异甲基化的证据。我们采用多阶段区域拆分策略，根据统计证据细化候选区域。本文介绍了其基础方法和软件实现，并通过模拟研究以及对 Illumina 450K 甲基化数据的应用展示其性能。**mmcmcBayes**

软件包为现有基于 CpG 的差异甲基化区域检测方法提供了一种实用的区域级替代方案，并包含用于汇总、比较和可视化已检测区域的辅助函数。

KEYWORDS：贝叶斯方法、差异甲基化区域、DNA甲基化、区域层面推断

1 引言

基因组通过对 DNA 化学结构的可遗传修饰（如特定碱基的甲基化）储存表观遗传信息。DNA 的一种化学修饰形式称为 DNA甲基化。DNA甲基化是一种基本的表观遗传机制，参与基因调控、细胞分化和疾病发生。在人类细胞中，约有百分之一的 DNA 碱基发生甲基化；在成熟体细胞组织中，甲基化主要发生于胞嘧啶 - 鸟嘌呤磷酸（CpG）位点。

早期分析侧重于单个 CpG位点，而如今人们已普遍认识到，具有生物学意义的信号往往出现在区域层面，因此，对差异甲基化区域（DMR）的识别日益受到关注。正如 Rakyan 等人（2011年）指出的，差异甲基化区域（DMR）定义为基因组中多个相邻 CpG位点呈现差异甲基化的区域，并被认为是一个功能区域，可能在调控基因转录水平方面发挥作用。

在包括唐氏综合征（Jin 等人，2013年）、阿尔茨海默病（Zhang 等人，2020）、类风湿关节炎（Shao 等人，2019）、糖尿病（Johnson 等人，2020；Dayeh 等人，2014年）、系统性红斑狼疮（Zhang 等人，2023）、自身免疫性甲状腺疾病（La-fontaine 等人，2024年；Absher 等人，2013年；Liu 等人，2013年）以及神经发育障碍等多种疾病相关研究中，均有差异甲基化区域的报道，

其中包括（Ladd-Acosta 等人，2014年）以及癌症（Irizarry 等人，2009年；Hansen 等，2011）。这些发现凸显了区域层面推断对于理解与疾病过程相关的表观遗传变异的重要性。然而，从统计学角度来看，由于甲基化数据的高维性、CpG位点之间的强空间相关性，以及不同样本间分布模式的异质性，差异甲基化区域的检测仍然面临挑战。

大多数现有的差异甲基化区域检测方法都是先在 CpG 水平进行推断，再汇总逐位点结果以界定区域。例如，**DMRcate** （Peters 等，2015）软件包对 CpG 水平检验统计量应用核平滑，而 **bumphunter** （Jaffe 等，2012）则识别平滑回归系数超过预定义阈值的连续基因组片段。同样，**DMRcaller** （Catoni 等，2018）、**DSS** （Feng 等，2014）和 **bsseq/B-Smooth** （Hansen 等，2014）依赖位点水平证据来构建区域。此外，若干广泛使用的 R和 Bioconductor 软件包为甲基化芯片分析提供了全面的工作流程，并纳入了区域查找或下游差异甲基化区域功能，包括 **minfi** （Aryee 等，2014）、**missMethyl** （Phipson 等，2016）和 **ChAMP** （Tian 等，2017）。测序甲基化分析由**methylKit**（Akalin 等，2012）等软件包支持，这些软件包同样根据位点水平统计量构建差异甲基化区域。在这些方法中，区域被视为由 CpG级推断派生的次级对象，而非主要推断单位。

除空间依赖性外，甲基化测量值在不同样本间通常还表现出异质性的分布模式，包括偏度和多峰性，尤其是在区域层面进行汇总时（Irizarry 等，2009；Teschendorff 等，2012）。这些特征反映了个体间潜在的生物异质性和变异性，并使主要依赖对称或基于均值的建模假设的推断程序变得复杂。

本文介绍了 R （R CoreTeam，2025）软件包 **mmcmcBayes** （Yang 等人，2025）。该软件包使用多阶段马尔可夫链蒙特卡罗算法检测差异甲基化区域。该软件包可通过 Comprehensive RArchive Network（CRAN）从 https://CRAN.R-project.org/package=mmcmcBayes 获取。**mmcmcBayes** 中实现的方法学基于汇总为 M-值的甲基化测量值运行，并分别对癌症组和正常组的区域级甲基化分布进行建模，从而允许样本间存在偏态和双峰模式。通过贝叶斯假设检验，使用贝叶斯因子评估差异甲基化的证据；贝叶斯因子提供了可解释的证据度量，同时避免依赖基于置换的 p-值。

mmcmcBayes 的一个显著特点是采用多阶段程序：从宽泛的基因组区域开始，并在统计证据支持时逐步将其细化为更局部的片段。该策略无需预设窗口大小，即可定位差异甲基化信号。该软件包允许用户指定关键调优参数，包括阶段数、区域拆分方式和贝叶斯因子阈值，从而适用于不同的数据类型和研究设计。

除了主要检测程序外，**mmcmcBayes** 还提供用于汇总、比较和可视化检测到的区域的辅助函数。函数summarize dmrs() 按染色体和阶段汇总已识别差异甲基化区域的信息，compare dmrs() 便于比较不同分析或参数设置的结果，而 plot dmr region() 则通过绘制癌症组和正常组在各CpG位点上的平均 M值，来可视化区域级甲基化模式。所提出的方法学适用于模拟生成和实验生成的甲基化数据，并专为基于芯片的测量值设计，例如 Illumina HumanMethylation 平台生成的测量值。

该　　本文其余部分安排如下。第2节介绍

支撑 **mmcmcBayes** 的建模框架和多阶段贝叶斯程序。第3节介绍该软件包的实现细节，包括数据结构和主要函数。第4节通过模拟研究以及真实甲基化数据的应用，展示 **mmcmcBayes** 软件包的使用方法。第5节对全文进行总结，并简要讨论。

2 模型

2.1 Alpha-skew GeneralizedNormalDistribution

CpG位点上的DNA甲基化水平通常用 β值汇总，定义为甲基化信号占甲基化与未甲基化信号强度总和的比例。 β值位于区间 $[0,1]$ 内，可自然地解释为甲基化比例。在本研究中，甲基化测量值在区域层面进行汇总。对于每个受试者 j、组 m、基因组片段k和阶段 ℓ，令 β^{ℓ}_{jml} 表示相应片段内所有CpG位点的 β值的均值。本文中，$m = 0$ 表示癌症组，$m = 1$ 表示对照组。

然而，由于 β值的取值范围有界，直接对其建模颇具挑战。按照Ryu等（2013）在甲基化分析中的标准做法，因此使用logit变换将 β值转换为 M值。具体而言，与 β^{ℓ}_{jmk} 对应的M值定义为

$$y^{\ell}_{jmk} = \log\left(\frac{\beta^{\ell}_{jmk} + c}{1 - \beta^{\ell}_{jmk} + c}\right), \tag{1}$$

其中， $c > 0$ 是引入的小偏移量，用于避免 β-值接近 0 或 1 时出现数值不稳定性。在实践中，我们在整个软件包中将 $c = 10^{-6}$ 设为该值；该值足够小，能够确保数值稳定性，同时不会实质性地影响转换后的值。尽管 M-值转换解决了其中一些

β-值的局限性，但所得分布仍可能呈现偏度和双峰性。

为适应偏斜和非对称分布，**mmcmcBayes** 包使用 Mahmoudi等人（2019）提出的 α偏斜广义正态（ASGN）分布对区域 M-值进行建模。ASGN分布通过引入偏度参数扩展了正态分布，使其能够呈现非对称行为，同时保留高斯核，因此非常适合对区域甲基化汇总进行建模。对于阶段 ℓ 、组 m 中给定的基因组片段 k ，假设区域 M-值 y_{jmk}^{ℓ} 相互独立地服从 ASGN分布，使得

$$y_{jmk}^{\ell} \overset{ind}{\sim} \mathrm{ASGN}(\alpha_{mk}^{\ell}, \nu_{mk}^{\ell}, \delta_{mk}^{2,\ell})$$

其概率密度函数为

$$f\left(y_{jmk}^{\ell} | \alpha_{mk}^{\ell}, \nu_{mk}^{\ell}, \delta_{mk}^{2\ell}\right) = \frac{\sqrt{2}\left\{(1-\alpha_{mk}^{\ell} y_{jmk}^{\ell})^2 + 1\right\}}{4\left\{\Gamma\left(\frac{3}{2}\right)\alpha_{mk}^{\ell} + \Gamma\left(\frac{1}{2}\right)\right\}} \exp\left\{-\frac{\left(y_{jmk}^{\ell} - \nu_{mk}^{\ell}\right)^2}{2\delta_{mk}^{2\ell}}\right\}, \quad (2)$$

其中， $\alpha_{mk}^{\ell} \in \mathbb{R}$ 控制偏度， $\nu_{mk}^{\ell} \in \mathbb{R}$是位置参数，而 $\delta^{2}_{,\ {}^{\ell}_{mk}} > 0$是尺度参数。在分析后续阶段的子片段时，也采用相同的建模假设。例如，对于阶段 $\ell+1$ 的子片段 k' ，

$$y_{jmk'}^{\ell+1} \overset{ind}{\sim} \mathrm{ASGN}(\alpha_{mk'}^{\ell+1}, \nu_{mk'}^{\ell+1}, \delta_{mk'}^{2,\ell+1}).$$

对于贝叶斯推断，我们采用以下先验分布：

$$\begin{aligned} \alpha_{mk}^{\ell} &\sim N(\mu_a, \sigma_a^2), \\ \nu_{mk}^{\ell} &\sim N(\mu_n, \sigma_n^2), \\ \delta_{mk}^{2,\ell} &\sim \mathrm{IG}(A_d, B_d), \end{aligned} \quad (3)$$

其中， N 和 IG 分别表示正态分布和逆伽马分布。 μ_a、 σ_a^2、 μ_n、 σ_n^2、 A_d 和 B_d 是相应先验分布的超参数。

在多阶段马尔可夫链蒙特卡罗框架中，先验信息会在各阶段间依次更新。具体而言，阶段 $\ell+1$ 的先验取自阶段 ℓ的后验均值。例如，阶段 $\ell+1$ 的偏度参数先验分布定义为

$$\alpha_{mk}^{\ell+1} \sim N(\hat{\alpha}_{mk}^{\ell}, 1),$$

其中，$\hat{\alpha}_{mk}^{\ell}$ 表示前一阶段 ℓ的后验均值。先验方差固定为 1，使前面阶段的信息能够提供适度指导，而不会主导对较小子区域的推断。

2.2 多阶段 MCMC 程序

基于第 2.1 节所述的区域级 M值 ASGN模型，我们提出了一种多阶段马尔可夫链蒙特卡罗（MCMC）程序，在进行区域层面推断的同时自适应地细化基因组区域。该方法从相对粗略的区域开始，仅在出现统计证据时才逐步将其划分。

假设该流程依次经过阶段 $\ell = 1$、2、.....。在阶段 ℓ，基因组区域被划分为 k 个片段。在阶段 ℓ，我们根据第 2.1 节所述，使用 ASGN分布对每个片段中癌症组和正常组的区域甲基化水平进行建模。例如，图 1 展示了如何实现多阶段结构。在初始阶段（$\ell = 1$），整个基因组区域被视为一个片段。对于每个片段，我们使用贝叶斯因子量化两组之间差异甲基化的证据。当某个片段的贝叶斯因子超过用户指定的阈值时，我们将该片段划分为若干子片段，并将其带入下一阶段（$\ell+1$）。当证据低于阈值时，则保留该片段，不再进一步分割。随后，在下一阶段对所有片段应用相同的流程。该流程

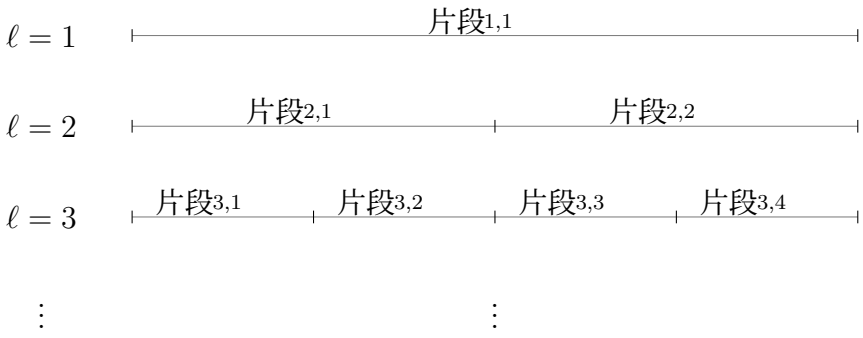

图 1：多阶段马尔可夫链蒙特卡洛程序

将持续进行，直到达到预先设定的最大阶段数，或没有片段满足分割准则为止。

2.3 模型比较

为正式比较癌症组和正常组的区域级差异，我们针对每个基因组片段考虑两个相互竞争的模型。在零假设 H_0 下，假设两组的区域甲基化值均来自同一个 α偏斜广义正态（ASGN）分布。在备择假设 H_1 下，则使用不同的 ASGN 分布对两组分别建模，从而允许其分布特征存在差异。

使用贝叶斯因子量化支持差异甲基化的证据。设$\boldsymbol{y}_{0k}^{\ell}$ 表示阶段 ℓ、片段 k 中癌症组的区域 M值集合， $\boldsymbol{y}_{1k}^{\ell}$ 表示正常组中相应的值。贝叶斯因子定义为

$$BF_k^{\ell} = \frac{\sum_j f\left(y_{j0k}^{\ell} \mid \hat{\alpha}_{0k}^{\ell}, \hat{\nu}_{0k}^{\ell}, \hat{\delta}_{0k}^{2,\ell}\right)}{\sum_j f\left(y_{j1k}^{\ell} \mid \hat{\alpha}_{1k}^{\ell}, \hat{\nu}_{1k}^{\ell}, \hat{\delta}_{1k}^{2,\ell}\right)},$$

其中， $f(\cdot)$ 表示第 2.1 节定义的 ASGN密度，帽号表示通过 MCMC 抽样获得的后验汇总估计值。较低的阈值会触发

更频繁的拆分，并倾向于生成更小的区域；较高的阈值则较少触发拆分，并倾向于生成更大、更粗略的区域。

3 mmcmcBayes 软件包

mmcmcBayes 可在 CRAN 上获取。我们可以使用以下代码进行安装并加载：

```
R> install.packages("mmcmcBayes")
R> library("mmcmcBayes")
```

3.1 数据结构

核心函数 `mmcmcBayes()` 需要两个数据框，分别对应两个比较组，通常称为 `cancer data` 和 `normal data`。这两个数据框必须包含相同的 CpG 位点，且行顺序相同。这一要求至关重要，因为候选区域是基于行索引而非基因组坐标匹配来定义和递归划分的。因此，分析前必须按基因组位置对输入数据排序，具体来说，要按染色体和物理位置排序。否则，输入数据中相邻的行可能并不对应生物学上连续的区域，从而导致结果无法解释。以下代码通过展示随软件包提供的示例数据集的前五列，说明输入数据的结构：

```
R> data(癌症_示例) R>
head(癌症_示例[,1:5])
```

	CpG_ID	染色体	M_样本_1	M_样本_2	M_样本_3
1	cg00000721	6	1.9781703	1.8551462	2.0203321
2	cg00002930	6	-2.3549588	-1.9455890	-2.0854731
3	cg00003181	6	0.5899144	1.3358503	1.0194316
4	cg00004061	6	1.6567481	0.8350491	0.3173226
5	cg00004562	6	1.1449919	1.1267618	1.0187141
6	cg00004608	6	-0.3366765	0.6255285	0.9428262

每一行 表示一个 CpG 位点的信息。前两列必须

分别为_ CpG ID和 Chromosome，用于标识 CpG位点及其染色体标签。无需额外的基因组注释。其余所有列均视为包含各个样本的数值型甲基化测量值，以 M值表示，定义见公式 1。对于给定区域， mmcmcBayes() 函数会自动计算该区域内所有CpG位点的样本平均 M值。这些区域层面汇总值作为模型中的观测数据，并会在多阶段程序中每次划分区域时重新计算。允许存在 CpG 水平的缺失值。计算区域级均值时，仅使用可用的测量值。

3.2 主要函数

mmcmcBayes 包的主要函数是 mmcmcBayes()，该函数通过递归区域拆分实现多阶段马尔可夫链蒙特卡罗（MCMC）过程，以检测差异甲基化区域（DMR），如第 2.2 节所述。函数的调用方式如下：

mmcmcBayes(癌症_数据, 正常_数据, 阶段 = 1, 最大_阶段 = 3, 数量_分割数 = 50, mcmc = NULL, 先验_癌症 = NULL, 先验_正常 = NULL, bf_阈值 = c(0.5, 0.8, 1.05))

- 参数 cancer data 和 normal data 是数据框，包含两个比较组的甲基化测量值。这些输入所需的结构详见第 3.1 节。

- 参数 stage 指定多阶段程序的起始阶段，通常保持默认值 1。递归拆分的最大阶段数由 max stages控制。当程序达到max stages，或没有子区域满足分割准则时，将不再

进行细分。根据第 4.1 节所述的模拟研究，默认值 `max stages = 3` 在假发现率（FDR）和检测精度之间实现了良好的平衡。

- 参数 `num splits` 控制区域分裂时创建的子区域数量。实际应用中，子区域的数量受区域内 CpG 位点数量的限制。增大 `num splits` 的值可以更精确地定位差异甲基化区域，但会增加计算成本。默认值num splits=50参考了第 4.1 节中的模拟结果，这些结果综合考量了错误发现率、精确度和计算时间。

- 用户可以根据自身偏好指定 MCMC 抽样设置。参数 `mcmc` 是一个包含 `nburn`、`niter`和 `thin`的列表。如果 `mcmc=NULL`，则会将默认设置 `list(nburn = 5000, niter = 10000, thin = 1)` 直接传递给函数 `asgn func()`，以获取 α-偏斜广义正态（ASGN）模型参数的后验均值。

- 参数 `priors cancer` 和 `priors normal` 允许用户为每个组指定公式 3 中定义的 ASGN 模型超参数。例如，当癌症组的 $\mu_a = 0$、$\mu_n = 2$和 $A_d = 1$ 分别为以下值时，相应的先验可以按如下方式指定：R> 先验_癌症 <- list(alpha=0, mu=2, sigma2=1)

其余超参数σ_a^2、σ_n^2和 B_d 在内部固定。如果参数设置为 `NULL`，则会根据当前阶段的观测数据自动构建弱信息先验。

- 各阶段的贝叶斯因子决策阈值通过 `bf threshold` 指定。如果贝叶斯因子超过阈值且当前阶段小于`max stages`，则会根据 `num splits` 将该区域划分为更小的子区域

并传递到下一阶段进行进一步评估。如果在最后阶段贝叶斯因子超过阈值，则该区域将作为检测到的差异甲基化区域报告，且不再进行进一步划分。如果 NULL，则将使用默认值 bf threshold <-list(0.5, 0.8, 1.05)。

3.3 次级函数

mmcmcBayes 包中的次级函数为 asgn func()。该函数用于在第 2.1 节所述的α偏斜广义正态（ASGN）分布下进行参数估计。函数的调用方式如下：

asgn_func(data, 先验 = NULL, MCMC = list(nburn = 5000,
niter = 10000, thin = 1), 随机种子 = NULL)

- 参数 data是一个单变量数值向量或单列矩阵。在该包中，此输入包含针对候选区域计算的逐样本区域平均 M 值。

- 参数 priors允许用户指定 ASGN 分布的超参数。其设置方式与第 3.2 节中介绍的相同。请注意，当从 mmcmcBayes()调用 asgn func()时，母区域的后验均值会被用作其子区域的先验。

- mcmc 参数遵循第3.2节中介绍的相同规范。

- 可选的 seed参数可用于设置随机种子，以确保结果可复现。

该函数返回 ASGN 分布的后验均值估计及相应的 95% 可信区间。尽管它在软件包中的主要作用是对区域甲基化汇总进行建模，但也可以独立使用，以拟合偏斜且可能呈双峰的数据。

3.4 其他函数

除了主要的检测函数外，该软件包还提供了若干辅助函数，用于汇总、比较和可视化检测到的差异甲基化区域（DMR）。表1概述了软件包中包含的这些辅助函数及其主要用途。

表1： mmcmcBayes 包提供的辅助函数。

函数	用途	描述
summarize_dmrs()	摘要	该函数报告检测到的区域总数、区域大小的汇总统计量（CpG Count）和决策值，以及按染色体统计的差异甲基化区域数量。如果存在 Stage 列，还会报告各阶段的数量。
compare_dmrs()	比较	识别两组结果之间重叠的差异甲基化区域这些结果通过不同的方法或参数设置获得。对于每个重叠区域，该函数都会报告一个重叠百分比。如果没有有效的，则返回 NULL 重叠被找到。
plot_dmr_region()	可视化	提供用于解释所检测到的可视化来自 mmcmcBayes() 函数的差异甲基化区域，通过图示-癌症与之间的甲基化差异每个区域内的正常组。

3.5 与其他软件包的比较

表 2 总结了**mmcmcBayes**与若干广泛使用的差异甲基化区域检测软件包在方法上的主要差异。现有软件包主要在 CpG 水平上运行，并依赖频率学派检验程序；而 mmcmcBayes 则在贝叶斯框架内直接在区域层面进行推断。这使mmcmcBayes 能够处理均值差异型 CpG级方法难以充分捕捉的偏斜或多峰甲基化模式。

表 2：**mmcmcBayes** 与其他现有软件包的比较。

特性	**mmcmcBayes**	**bumphunter**	**差异甲基化区域**	**DMRcaller**
主要推断层级	区域级	CpG级	CpG级	CpG级
统计框架	贝叶斯	频率主义	频率主义	频率主义
分布建模	Yes	No	No	No
处理偏度	Yes	No	No	No
处理双峰性	Yes	No	No	No
检验	贝叶斯因子	置换	FDR	FDR

总体而言，**mmcmcBayes** 通过关注分布差异而非仅关注均值变化，为现有CpG级方法提供了一种互补的区域级方法。

4 示例

我们首先利用模拟结果来说明 `mmcmcBaeyes()` 中默认值的选择。

接着，我们使用随软件包分发的演示示例展示该软件包的用法，随后将其应用于一个真实的 Illumina 450K 甲基化数据集。本节中的所有示例均使用 `R` version 4.5.1 在标准 macOS 平台上完成。请注意，在不同平台上获得的结果可能有所不同。

4.1 模拟

我们首先开展模拟研究，以确定**mmcmcBayes** 软件包所用默认值的合理性。我们使用来自6号染色体的真实甲基化数据作为基准，以保留Illumina 450K数据中常见的真实基因组结构、探针密度和空间相关模式。

为引入技术变异和生物变异，我们向所有样本添加均值为0、标准差为0.5的独立高斯噪声。在每次模拟重复中，向数据集中生成十个模拟的差异甲基化区域（DMR）。对于每个差异甲基化区域，相对于背景的平均甲基化偏移随机设为1。0或2。0，代表中等到强烈的生物信号。每个差异甲基化区域的长度从连续的10、20或50个CpG位点中随机抽取，并限制这些区域形成互不重叠的基因组区块。

为评估区域划分的影响，我们评估了六种划分配置（20、30、40、50、80 和 100 个分割数）。多阶段贝叶斯因子阈值固定为（ 0.5，0.8，1.05 ）。对于每种配置，我们记录了灵敏度、假发现率数量和总计算时间。

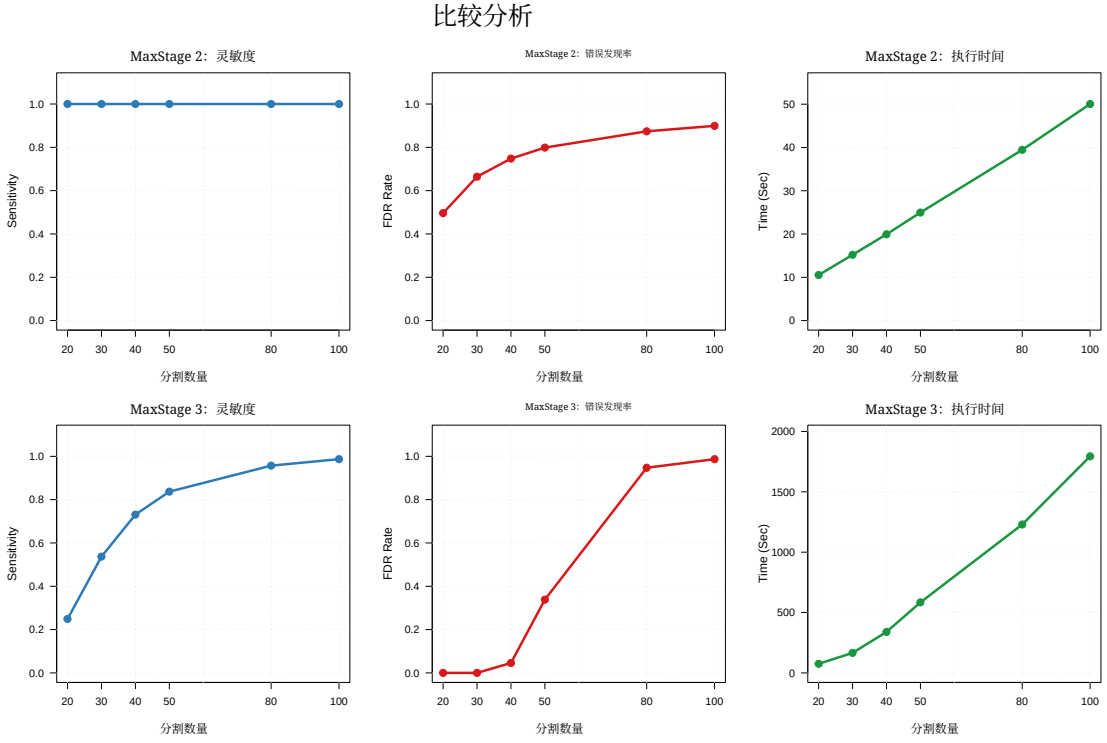

图2：MaxStage 2 与 MaxStage 3 的比较

图2展示了分割数对结果的影响。当 `max stage=2` 时，增加分割数并不能提高灵敏度，却会明显增加错误发现率（FDR）。当 `max stages=3` 时，灵敏度会随着分割数增加而提高，并在分割数约为 50 时趋于稳定。与此同时，FDR 在分割数达到约 50 之前一直较低，但使用更多分割时会迅速上升。超过这一点后，计算时间也会大幅增加。根据这些结果，我们将 `max stage = 3`、 `num splits=50` 和 `bf threshold=c(0.5,0.8,1.05)` 设为默认配置，因为它们能够在不产生过多假阳性或计算开销的情况下提供良好的灵敏度。需要注意的是，尽管该软件包本身按顺序运行，用户仍可使用标准的 `R`工具并行处理不同染色体上的独立分析。

4.2 演示数据

在本小节中，我们使用包中包含的演示数据集 `cancer demo` 和 `normal demo` ，展示 `mmcmcBayes()` 的基本工作流程。这些数据集分别包含癌症样本和正常样本的甲基化 M值。两个数据集均取自一个 Illumina 450K 数据集的 6号染色体上前 5,000 个 CpG位点，并作为精简示例供演示和测试使用。以下代码用于读取这些数据集。

```
R> library("mmcmcBayes") R> data(cancer_demo,
package = "mmcmcBayes") R> head(cancer_demo
[,1:5])
```

	CpG_编号	染色体	M_样本_1	M_样本_2	M_样本_3
1	cg00000721	6	1.9781703	1.8551462	2.0203321
2	cg00002930	6	-2.3549588	-1.9455890	-2.0854731
3	cg00003181	6	0.5899144	1.3358503	1.0194316
4	cg00004061	6	1.6567481	0.8350491	0.3173226
5	cg00004562	6	1.1449919	1.1267618	1.0187141
6	cg00004608	6	-0.3366765	0.6255285	0.9428262

R> 数据（正常_演示，程序包 = "mmcmcBayes"）

R> 头部（正常_示例[,1:5]）

	CpG_编号	染色体	M_样本_1	M_样本_2	M_样本_3
1	cg00000721	6	1.9716225	1.93061870	2.2024953
2	cg00002930	6	-2.0804008	-2.42836314	-2.2345472
3	cg00003181	6	0.8372380	0.66398628	0.6713868
4	cg00004061	6	0.7094518	0.38711073	0.4195223
5	cg00004562	6	1.3338421	1.19899281	1.2833079
6	cg00004608	6	-0.1988843	-0.08407308	-0.1059526

每个数据集均存储为一个数据框，其中行表示 CpG位点，列表示样本甲基化值。两个数据集具有相同的 CpG 排序和基因组注释，因此可直接作为配对输入使用。以下代码用于运行 `mmcmcBayes()` 函数。为简单起见，预烧期设为 1,000 次迭代，随后保存 2,000 个样本，不进行抽稀。

```
R> set.seed(2021) R> mcmc <-列表(nburn=1000, niter=2000,
thin=1) R> prior_c <-列表(alpha=0, mu=2, sigma2=0.1) R>
prior_n <-列表(alpha=0, mu=1, sigma2=0.1) R> rst <-
mmcmcBayes(cancer_data = cancer_demo,
+                          normal_data =normal_demo,+ 阶段
 = 1, max_stages = 2,+                              num_分割数
 = 5, mcmc =mcmc,+                                priors_cancer =
先验_c,+                              priors_normal =先验_n,
+                          bf_thresholds =c(0.5, 0.8))
```

参数 `prior cancer` 和 `prior normal` 表示公式 3 中定义的超参数。如果 `prior=NULL`，函数会根据观测数据自动构建弱信息先验。

返回对象 `rst` 的输出如下。

R> print(rst)

	染色体	起始_CpG	结束_CpG	CpG_数量	决策_值	阶段
1	6	cg00000721	cg00613945	1000	9.755850	2
2	6	cg00614280	cg01300096	1000	9.775594	2
3	6	cg01300341	cg01971547	1000	9.732180	2
4	6	cg01971789	cg02683846	1000	9.692494	2
5	6	cg02683893	cg03338154	1000	9.792246	2

上表总结了该函数检测到的差异甲基化区域（DMR）。每行对应一个检测到的区域。

- `Chromosome` 列表示该区域所在的染色体。在此示例中，所有检测到的区域均位于6号染色体上。

- `Start CpG` 列标记检测到的区域的起点。

- `End CpG` 列标记检测到的区域的终点。

- `CpG Count` 列显示该区域包含的CpG位点数量。

- `Decision Value` 列汇总了针对该区域计算的贝叶斯因子。数值越大，差异甲基化的证据越强。

- `Stage` 列表示该区域是在第几个阶段被识别的。例如，值为2表示该区域在第二阶段的拆分和检验过程中满足了决策标准。

我们可以看到，在使用 `max stage=2` 和 `num splits=5`的情况下，演示数据集中的 5,000 个 CpG 位点会被划分为每块约 1,000 个 CpG 位点的粗略区块进行检查。输出中列出的区域是通过第二阶段贝叶斯因子阈值 0.8 的区域，其中 `Decision Value` 报告了相应的贝叶斯因子。尽管这些贝叶斯因子更大并且超过了第二阶段阈值，但由于我们将 `max stages=2` 设为 2，阶段 2 是允许的最大阶段，因此不会进一步拆分。

4.3 Illumina 450K数据

在本小节中，我们演示如何在真实的甲基化数据集上使用**mmcmcBayes** 软件包。该数据集来自一项针对肺癌的 Illumina HumanMethylation450 BeadChip (450K) 研究，其中包含以下数据的甲基化测量值：

来自癌症和正常样本、范围限定在6号染色体上的样本，且以 M值的形式呈现。以下代码可用于从 GitHub 仓库下载该数据集。

```
R> c<-"https://raw.githubusercontent.com/zyang1919/mmcmcBayes/main/chr6_癌症.RData" R> n<-"https://raw.githubusercontent.com/zyang1919/mmcmcBayes/main/chr6_正常.RData"

R>download.file(c, destfile = "chr6_cancer.RData", mode = "wb", quiet = TRUE) R> download.file(n, destfile = "chr6_正常.RData", mode = "wb", quiet = TRUE)

R> load("chr6_cancer.RData") R> load("chr6_normal.RData")

R> dim(chr6_癌症)

[1] 36438    21

R> dim(chr6_正常)

[1] 36438    21
```

输出显示，两个数据集均包含 36,438 个 CpG位点和 21 列。前两列存储基因组信息，即 `CpG site` 和 `Chromosome`，如第4节 4.2 中所示；其余 19 列对应样本甲基化 M值。癌症数据集和正常数据集的维度相同，且 CpG 排序一致，因此无需额外调整即可直接将它们用作 `mmcmcBayes()`的输入。

以下代码将 `mmcmcBayes()` 函数应用于已加载的癌症和正常数据集。此处使用三阶段程序，分割数为 50，且各阶段的贝叶斯因子阈值逐步提高；其他设置均保留默认值。结果对象存储为 `rst 450k`。

```
R> library("mmcmcBayes") R> set.seed(2021) R> rst_450
k <-mmcmcBayes(癌症_数据 = chr6_癌症,
+                              正常_数据 =chr6_正常,
+                              阶段 = 1, max_阶段 = 3,
```

```
+                                num_分割数  = 50,
+                                bf_阈值  = c(0.5, 0.8, 1.05))
```

将**mm-cmcBayes** 包中的 `summarize dmrs()` 函数应用于返回对象 `rst 450k` 后得到的输出如下。

```
R> summarize_差异甲基化区域(rst_450k)

差异甲基化区域结果摘要（mmcmcBayes）

差异甲基化区域数量: 1514

区域大小（CpG_数量）：最小值 q1 中位数 均值 q3
最大值 14.000 14.000 15.00014.565 15.000 15.000

决策_值:
    最小值 q1 中位数 均值 q3 最大值 1.050 1.089 1.143
    1.184 1.242 1.962

按染色体统计的差异甲基化区
域：染色体 n_差异甲基化区域
1 6 1514

按阶段统计的差异甲基
化区域：阶段 n_差异甲
基化区域 1 3 1514
```

我们可以看到，总共检测到 1,514 个差异甲基化区域。大多数检测到的区域都相对较小，每个区域的 CpG 位点数为 14 至 15 个，中位数为 15 个 CpG。这表明，多阶段程序在最终阶段将候选区域细化为狭窄的片段。

`Decision Value`列对应于每个区域的贝叶斯因子，显示所有值均超过1.05的最终阶段阈值，中位数约为1.14。这表明每个报告的区域均满足多阶段程序最严格的决策标准。

所有检测到的区域均位于6号染色体上，这与分析仅限于该染色体的设定一致。此外，所有差异甲基化区域均在

阶段3被识别出来，这是该流程的最终阶段。这证实了该流程首先评估较大的区域，然后逐步将其细分为更小的区域。

以下 mmcmcBayes 软件包中的代码用于绘制在6号染色体上检测到的差异甲基化区域。

```
R> pdf("6号染色体 差异甲基化区域4 至 9.pdf", width = 11, height = 11)

R> par(mfrow =c(3, 2), mar = c(2.5, 2.5, 2, 1),+ cex = 0.85, cex.axis = 0.75,
cex.lab = 0.85, cex.main = 0.9)

R>绘制_差异甲基化区域_区域(rst_450k, chr6_癌症, chr6_正常, 差异甲基化区域_索引 = 4) R>绘制_差异甲基化区域_区域(rst_450k, chr6_癌症, chr6_正常, 差异甲基化区域_索引 = 5) R>绘制_差异甲基化区域_区域(rst_450k, chr6_癌症, chr6_正常, 差异甲基化区域_索引 = 6) R>绘制_差异甲基化区域_区域(rst_450k, chr6_癌症, chr6_正常, 差异甲基化区域_索引 = 7) R>绘制_差异甲基化区域_区域(rst_450k, chr6_癌症, chr6_正常, 差异甲基化区域_索引 = 8) R>绘制_差异甲基化区域_区域(rst_450k, chr6_癌症, chr6_正常, 差异甲基化区域_索引 = 9)
```

图3绘制了使用plot dmr region() **mmcmcBayes** 软件包中的函数在6号染色体上检测到的几个代表性差异甲基化区域。每个面板显示了检测到的区域内各CpG位点上癌症样本和正常样本的平均 M值。两组之间的分离程度各不相同，反映了区域甲基化模式的异质性。

需要注意的是， plot dmr regions() 是一个可视化工具，用于评估由 mmcmcBayes()检测到的差异甲基化区域。该函数需要 mmcmcBayes()返回的差异甲基化区域结果对象、原始癌症和正常数据集，以及一个用于指定要可视化的检测区域的索引。

5 结论

在本文中，我们介绍了用于检测差异甲基化区域（DMR）的 R 包**mmcmcBayes** ，该方法采用多阶段马尔可夫链蒙特卡罗（MCMC）程序。所提出的方法通过对样本级区域甲基化汇总进行建模，并比较癌症组和

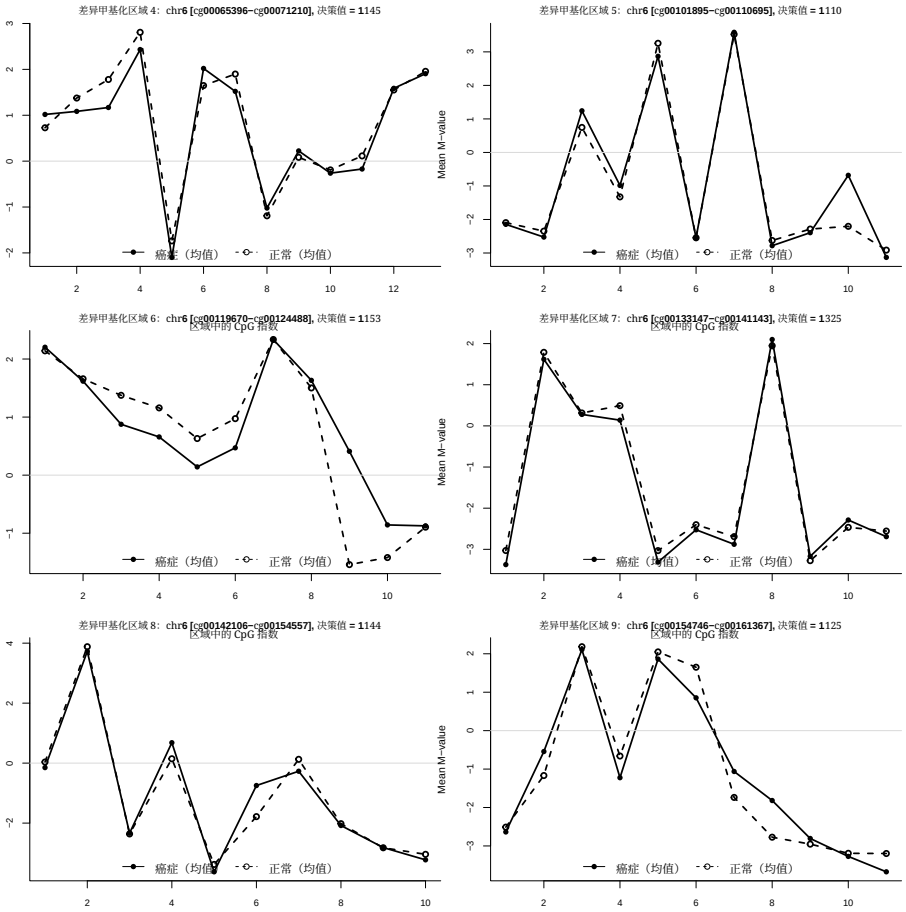

图3：在肺癌 Illumina 450K 数据集中检测到的 6号染色体上的差异甲基化区域示例。每个面板显示了癌症和正常样本在检测到的区域内各 CpG位点的平均 M 值。

正常组，使用贝叶斯因子直接在区域层面进行推断。**mmcmcBayes** 并不将区域定义为CpG级结果的聚合，而是将区域视为主要推断单位，并通过一系列基于证据的划分来细化其边界。

该方法的一个关键特点是采用多阶段马尔可夫链蒙特卡罗方案：从较粗略的基因组片段开始，只有在统计证据支持时，才逐步聚焦于更小的区域。这种结构为在差异信号的定位精度与计算成本之间取得平衡提供了一种自然方式。此外，使用 α 偏斜广义正态分布对区域甲基化汇总进行建模

使该方法能够处理实践中常见的偏斜或不对称甲基化模式，而许多现有方法并未明确考虑这些模式。

模拟研究用于考察关键调优参数（如阶段数和区域划分数）的影响，并为软件包提供的默认设置提供依据。对示例数据集和真实的 Illumina 450K 肺癌数据集的应用，展示了完整的分析流程，并说明了如何将该方法应用于实验生成的甲基化数据。这些示例共同表明，**mmcmcBayes** 能够识别差异甲基化的局部区域，同时在中等规模数据集上仍具有实用性。**mmcmcBayes**软件包提供了所提出方法学的可复现实现，以及用于汇总检测到的区域、比较不同分析结果和可视化区域级甲基化模式的工具。

致谢

本文基于第一作者的博士论文（Yang，2025）《贝叶斯函数型数据分析及其应用》的第3章。作者谨此感谢审稿人提出的宝贵意见，并感谢所有报告错误和提出建议的用户。

参考文献

Absher, D. M., Li, X., Waite, L. L., Gibson, A., Roberts, K., Edberg, J., Kimberly, R. P., 与 Alarc´on-Riquelme, M. E. (2013年)，“系统性红斑狼疮的全基因组DNA甲基化分析揭示干扰素基因持续低甲基化及CpG岛组成变化”，*PLoS Genetics*，9，e1003678。

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Mel- nick, A., 与 Mason, C. E. (2012)，“methylKit：用于分析全基因组 DNA 甲基化谱的综合性R软件包”，《基因组生物学》，13，R87。

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., and Irizarry, R. A. (2014年), “Minfi：一款灵活且全面的 Bioconductor 软件包，用于分析 Infinium DNA 甲基化微阵列”，《生物信息学》，30, 1363–1369。

Catoni, M., Tsang, J. M. F., Greco, A. P., and Zabet, N. R. (2018), “DMRcaller：一款多用途的 R/Bioconductor 软件包，用于检测和可视化 CpG 和非 CpG 背景下的差异甲基化区域”，《核酸研究》, 46, e114。

Dayeh, T., Volkov, P., Salo, S., Hall, E., Nilsson, E., Olsson, A. H., Kirkpatrick, C. L., Wollheim, C. B., Eliasson, L., R¨onn, T., and Ling, C. (2014年), “对 2 型糖尿病供者和非糖尿病供者的人胰岛进行全基因组 DNA甲基化分析，发现了影响胰岛素分泌的候选基因”，*PLoSGenetics*, 10, e1004160。

Feng, H.、Conneely, K. N. 和 Wu, H.（2014年），“一种用于从单核苷酸分辨率测序数据中检测差异甲基化位点的贝叶斯层次模型”，《核酸研究》，42，e69。

Hansen, K. D.、Langmead, B. 和 Irizarry, R. A.（2014年），“BSmooth：从全基因组亚硫酸氢盐测序读段到差异甲基化区域”，《基因组生物学》，15，R83。

Hansen, K. D., Timp, W., Bravo, H. C., Sabunciyan, S., Langmead, B., McDonald, O. G., Wen, B., Wu, H., Liu, Y., Diep, D., Briem, E., Feinberg, A.P. 和Irizarry,

R. A. (2011年), “癌症类型中表观遗传结构域的甲基化变异增加”, 《自然遗传学》, 43, 768–775。

Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., Ji, H., Potash, J. B., Sabunciyan, S. 和 Feinberg, A. P.（2009年）, “人类结肠癌甲基组显示出保守的组织特异性 CpG 岛岸区域相似的低甲基化和高甲基化”, 《自然遗传学》, 41, 178–186。

Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P. 和 Irizarry, R. A.（2012）, “在表观遗传流行病学研究中通过峰值搜寻识别差异甲基化区域”, 《国际流行病学杂志》, 41, 200–209。

Jin, S., Lee, Y. K., Lim, Y. C., Zheng, Z., Lin, X. M., Ng, D. P. Y., Holbrook, J.D., Law, H. Y., Kwek, K. Y. C., Yeo, G. S. H., and Ding, C. (2013年), “唐氏综合征胎盘中的全局 DNA 高甲基化”, 《公共科学图书馆·综合》, 9, e66578。

Johnson, R. K., Vanderlinden, L. A., Dong, F., Carry, P. M., Seifert, J., Waugh, K., Shorrosh, H., Fingerlin, T., Hrohnert, B. I., Yang, I. V., Kechris, K., Rewers, M., and Norris, J. M.(2020), “纵向 DNA 甲基化差异先于 1 型糖尿病”, 《科学报告》, 10, 3721。

Ladd-Acosta, C., Hansen, K. D., Briem, E., Fallin, M. D., Kaufmann, W. E., 和 Feinberg, A. P.（2014年）, 《自闭症多个脑区的常见 DNA 甲基化改变》, 《分子精神病学》, 19, 862–871。

Lafontaine, N., Shore, C. J., Campbell, P.J., Mullin, B.H., Brown, S.J., Panicker, V., Dudbridge, F., Brix, T. H., Hegedus, L., Wilson, S. G., Bell, J. T., and Walsh, J. P. (2024年), “全表观基因组关联研究显示自身免疫性甲状腺疾病中的差异性 DNA 甲基化

MDC1、KLF9 和 CUTA，”《临床内分泌与代谢杂志》，109，992–999。

Liu, Y., Aryee, M. J., Padyukov, L., Fallin, M. D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M., Shchetynsky, K., Scheynius, A., Klareskog, L., Ekstr¨om, T. J., 和 Feinberg, A. P.（2013年），“全表观基因组关联数据表明，DNA甲基化是类风湿关节炎遗传风险的中介，”《自然·生物技术》，31，142–147。

Mahmoudi, E., Jafari, H., 和 Meshkat, R. S.（2019），“Alpha偏斜广义正态分布及其应用，”《应用与应用数学：国际期刊》（*AAM*），14，784–804。

Peters, T. J., Buckley, M. J., Statham, A. L., Pidsley, R., Samaras, K., Lord, R. V., Clark, S. J., 和 Molloy, P. L.（2015），“人类基因组中差异甲基化区域的从头鉴定，”《表观遗传学与染色质》，8，6。

Phipson, B., Maksimovic, J. 和 Oshlack, A.（2016），“missMethyl：一个用于分析 Illumina HumanMethylation450平台数据的 R 软件包”，《生物信息学》，32，286–288。

R 核心团队（2025），*R*：一种用于统计计算的语言和环境，R 统计计算基金会，维也纳，奥地利。

Rakyan, V. K., Down, T. A., Balding, D. J. 和 Beck, S.（2011），“常见人类疾病的表观基因组范围关联研究”，《自然综述：遗传学》，12，529–541。

Ryu, D., Xu, H., George, V., Su, S., Wang, X., 和 Podolsky, R. H.（2013年），《量化并标准化 Illumina 芯片上的甲基化水平》，《生物测定学与生物统计学》，4，164。

Shao, X., Hudson, M., Colmegna, I., Greenwood, C.M.T., Fritzler, M.J., Awadalla, P., Pastinen, T., 和 Bernatsky, S.（2019），“利用甲基化组捕获测序鉴定 ACPA阳性无症状个体中与类风湿关节炎相关的 DNA甲基化变化”，临床表观遗传学，11，110。

Teschendorff, A.E., Jones, A., Fiegl, H., Sargent, A., Zhuang, J. J., Kitchener, H. C., 和 Widschwendter, M. (2012), “正常细胞学细胞中的表观遗传变异性与未来发生形态学转化的风险相关”，基因组医学, 4, 24.

Tian, Y., Morris, T. J., Webster, A. P., Yang, Z., Beck, S., Feber, A., 和 Teschen- dorff, A. E. (2017), “ChAMP：面向 Illumina BeadChips 的更新版甲基化分析流程”，《生物信息学》, 33,3982–3984.

Yang, Z. (2025), “贝叶斯函数型数据分析及其应用”，博士论文，北伊利诺伊大学。

Yang, Z., Ryu, D., 和 Luan, F. (2025), *mmcmcBayes*：用于检测差异甲基化区域的多阶段*MCMC*方法, R软件包版本 0.2.0.

Zhang, L., Silva, T. C., Young, J. I., Gomez, L., Schmidt, M. A., Hamilton-Nelson, K. L., Kunkle, B. W., Chen, X. 和 Martin, E. R.（2020），“前额叶皮层DNA甲基化差异的全表观基因组荟萃分析揭示了阿尔茨海默病中的免疫过程”，自然通讯，11，6114。

Zhang, W., Liang, G., Zhou, H., Zeng, X., Zhang, Z., Xu, X. 和 Lai, K.（2023），“通过对基因表达和甲基化数据进行整合分析识别系统性红斑狼疮的潜在生物标志物”，临床风湿病学，42，1423–1433。