Whole Genome Bisulfite Sequencing Analysis 报告

Part-1 数据预处理、比对和 call methylation

1.1 数据分析流程及对应文件

- Step-1：下机数据去接头及质控
  - 使用 fastqc 软件（version 0.11.9）对 raw data 进行数据质控分析（结果见目录 1_raw-data_QC）。
  - 使用 trim_galore (version 0.6.7)和 cutadapt（version 1.18）软件对 raw data 进行接头去除和低质量碱基修剪，得到 clean data（见 2_clean-data 目录中的以 fq.gz 为后缀的文件）。
  - 使用 fastqc 软件 (version 0.11.9) 对 clean data 进行数据质控分析 (结果见目录 2_clean-data_QC)。
- Step-2: Reads 比对、去重、call methylation 以及质控
  - 使用 bismark 软件(version 0.24.0)将来自 clean data 的双端测序 reads 比对到基因组上（参数：--score_min L,0,-0.6 -N 0 -L 20），得到 bam 文件。
  - 使用 bismark 软件里的 *deduplicate_bismark* 功能去除 bam 文件中以相同方向比对到相同位置的 reads，结果见 3_aligned_BISMARK 目录中的 SampleName_bismark_bt2_pe.deduplicated.bam 文件。可使用 SeqMonk 或 IGV 软件等对 bam 文件进行可视化。注：SampleName 为样本名，如 D1-2、D1-3 等。
  - 使用 bismark 软件里的 *bismark_methylation_extractor* 功能（参数：--no_overlap --comprehensive --gzip --CX --cytosine_report）对 CpG、CHG 和 CHH Context 进行 call methylation【注：依据 C 碱基的背景，将 C 碱基分为 CpG、CHG 和 CHH 三类，H 可以是 A，T 或 C】，见 3_aligned_BISMARK\methylation\SampleName 目录中的文件：
    - `CpG_context_SampleName_bismark_bt2_pe.deduplicated.txt.gz`
    - `CHG_context_SampleName_bismark_bt2_pe.deduplicated.txt.gz`
    - `CHH_context_SampleName_bismark_bt2_pe.deduplicated.txt.gz`

- 使用 bismark 软件里的 *bam2nuc* 功能生成核酸覆盖度的统计报告。使用 *bismark2report* 功能生成整合的 reads 比对、methylation extraction reports、去重、核酸覆盖度统计以及 M-bias 报告，网页报告见 3_aligned_BISMARK 目录中的 SampleName_bismark_bt2_PE_report.html：

  - D1-2_bismark_bt2_PE_report.html
  - D1-4_bismark_bt2_PE_report.html
  - D1-5_bismark_bt2_PE_report.html
  - D63-5_bismark_bt2_PE_report.html
  - D63-6_bismark_bt2_PE_report.html
  - D63-8_bismark_bt2_PE_report.html
  - D63-9_bismark_bt2_PE_report.html
  - D7-1_bismark_bt2_PE_report.html
  - D7-2_bismark_bt2_PE_report.html
  - D7-3_bismark_bt2_PE_report.html
  - D7-5_bismark_bt2_PE_report.html

  使用 *bismark2summary* 功能把所有样本生成一个汇总的网页报告，见 3_aligned_BISMARK 目录中 bismark_summary_report.html。

- 使用 bismark 软件里的 *bismark2bedGraph* 功能将 call methylation 的结果转为 bedGraph 格式，见 3_aligned_BISMARK\methylation\SampleName 目录中的 SampleName_bismark_bt2_pe.deduplicated.bedGraph.gz，可使用 SeqMonk 或 IGV 软件对 bedGraph 文件进行可视化，此外还会生成覆盖度文件，见 3_aligned_BISMARK\methylation\SampleName 目录中的 CpG.cov.gz.bismark.cov.gz、CHG.cov.gz.bismark.cov.gz 和 CHH.cov.gz.bismark.cov.gz 文件，文件内容形式为：

  ```
  <chromosome> <start position> <end position> <methylation percentage> <count methylated> <count unmethylated>
  ```

  这三种文件将会被用于下游分析，可使用 SeqMonk 可视化这些文件。

➢ Step-3: 汇总以上步骤的分析报告

- 使用 multiqc 软件（version 1.13）对以上步骤中的 trim_galore，fastqc，bismark 等软件产生的分析结果进行汇总展示，见目录 4_multiQC 中的 multiqc_report.html。

1.2 主要结果解读

这部分分析的简要统计结果，见 multiqc_report.html，基本统计信息如下：

General Statistics

Copy table | Configure Columns | Plot | Showing 11/11 rows and 11/15 columns.

Sample Name	% mCpG	% mCHG	% mCHH	M C's	C Coverage	% Dups	% Aligned	% BP Trimmed	% Dups	% GC	M Seqs
D1-2	60.9%	1.2%	1.4%	3 948.0	19.55X	19.7%	86.6%	1.1%	10.8%	21%	128.1
D1-4	61.7%	1.2%	1.4%	3 191.3	15.26X	31.8%	86.7%	0.8%	23.7%	21%	118.4
D1-5	60.2%	1.1%	1.4%	3 522.2	17.15X	23.9%	86.3%	1.0%	28.0%	21%	118.6
D63-5	61.6%	1.1%	1.3%	4 122.6	20.37X	18.2%	86.1%	1.2%	10.6%	22%	131.4
D63-6	62.0%	1.2%	1.4%	4 061.8	20.18X	15.5%	87.8%	1.1%	10.0%	21%	123.7
D63-8	62.2%	1.1%	1.4%	3 818.8	18.89X	20.4%	87.9%	1.1%	19.6%	21%	123.0
D63-9	62.7%	1.1%	1.4%	4 095.8	20.22X	15.4%	87.6%	1.3%	9.9%	21%	124.5
D7-1	60.8%	1.1%	1.4%	3 927.7	19.10X	25.9%	87.9%	1.2%	17.3%	21%	134.2
D7-2	60.9%	1.1%	1.4%	3 955.2	19.16X	24.8%	87.5%	1.2%	17.0%	21%	132.9
D7-3	58.9%	1.1%	1.3%	3 802.4	18.53X	16.8%	86.6%	1.4%	11.3%	22%	116.9
D7-5	61.4%	1.1%	1.4%	4 545.2	22.27X	19.6%	88.1%	1.0%	11.8%	21%	143.1

软件	缩写	全名
Bismark	% mCpG	CpG context 中 C 碱基的甲基化比例
Bismark	% mCHG	CHG context 中 C 碱基的甲基化比例
Bismark	% mCHH	CHH context 中 C 碱基的甲基化比例
Bismark	M C's	分析中覆盖到的 C 碱基位点数目（单位：百万）
Bismark	C Coverage	有效 reads 的测序深度
Bismark	% Dups	比对位置重复的 reads 的百分比
Bismark	% Aligned	比对率
Cutadapt	% BP Trimmed	序列修剪损失的碱基比例
FastQC	% Dups	重复 reads 的比例
FastQC	% GC	平均 GC 含量
FastQC	M Seqs	Clean data 中的 reads 数目（单位：百万对）

Part-2 下游分析

主要基于 msPIPE 分析流程。

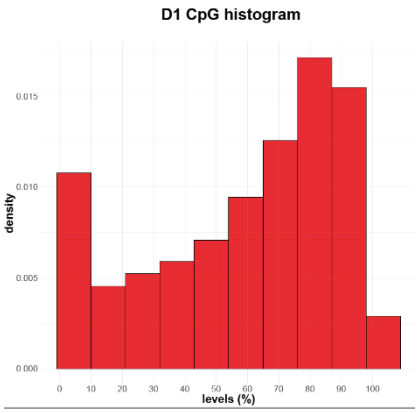
2.1 单样本水平分析及聚类和 PCA 分析

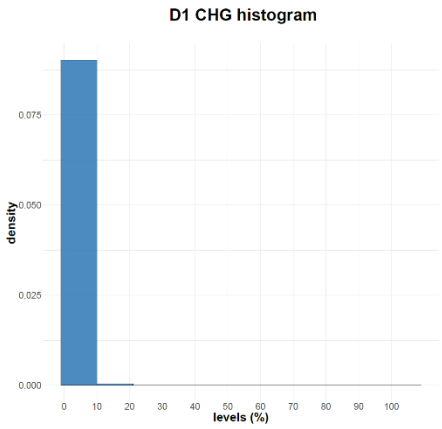
结果见网页 methylKit-Part1.html。

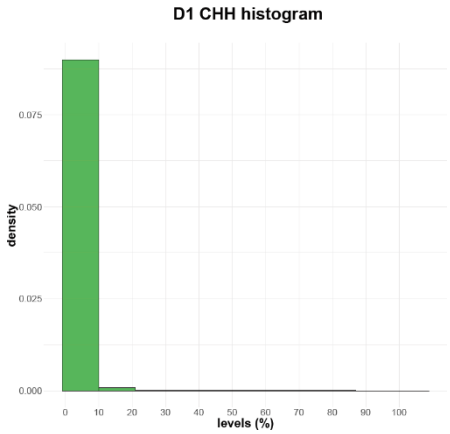
2.2 组水平分析

2.2.1 甲基化水平分布

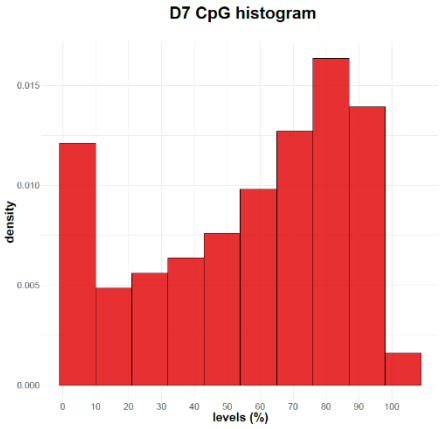
- **D1 组（CpG pdf; CHG pdf; CHH pdf）：**

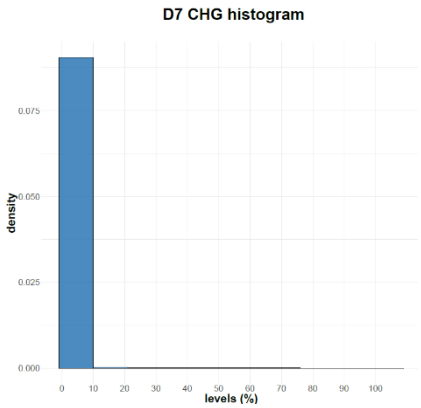


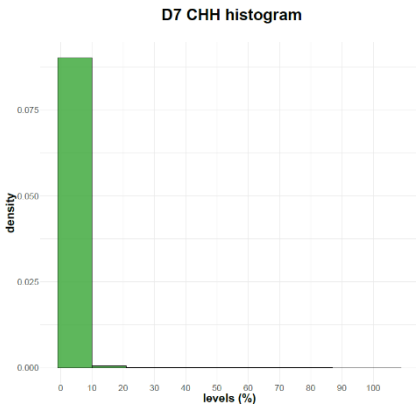




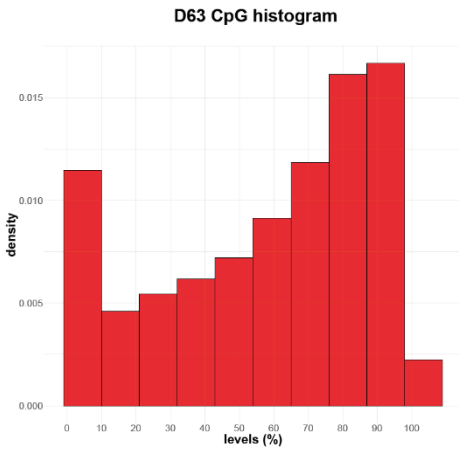
- **D7 组（CpG pdf; CHG pdf; CHH pdf）：**

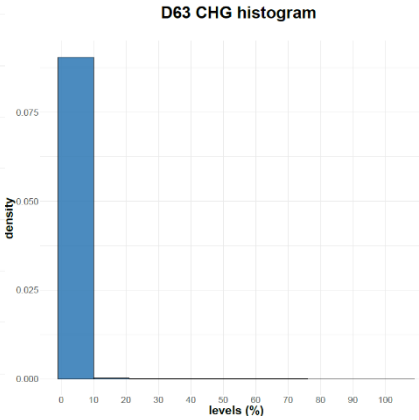


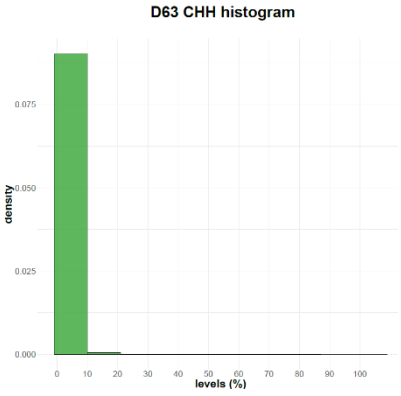




- **D63 组（CpG pdf; CHG pdf; CHH pdf）：**

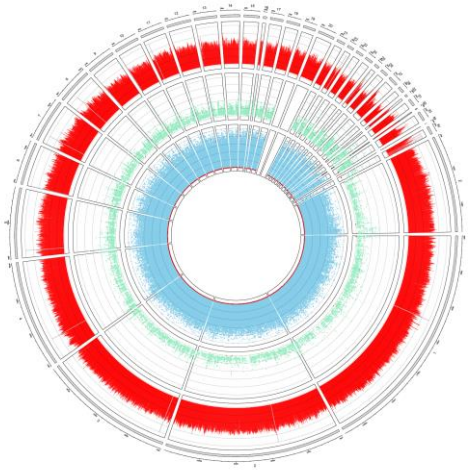




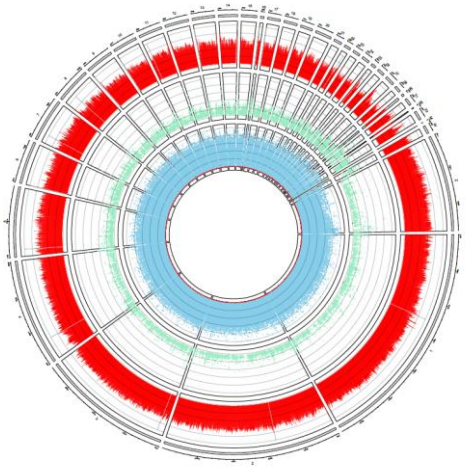


2.2.2 Circos plot 全基因组尺度甲基化谱（CpG/UMRs/LMRs）

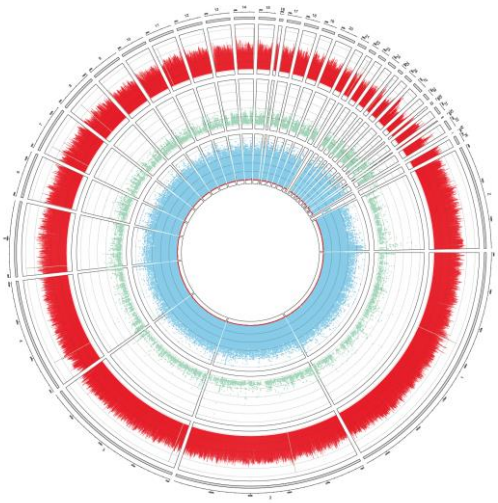
- D1 组(pdf)：



- D7 组(pdf)：



- D63 组(pdf)：

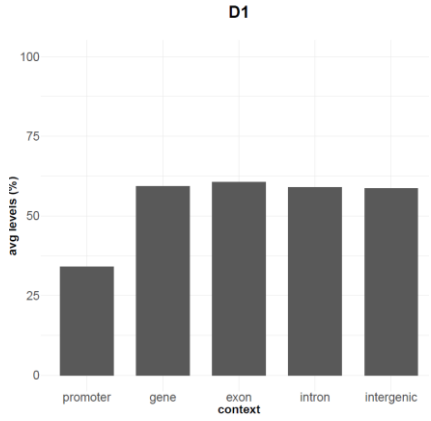


注：红色、浅绿色和浅蓝色分别表示 CpG、UMRs 和 LMRs 3 条 track。非甲基化区域（翻译可能不准确!）（hypomethylated regions，HMRs，由 R 包 MethylSeekR 预测）分为未甲基化区域（unmethylated regions， UMRs)和低甲基化区域（low methylated regions， LMRs）。红色柱状图是以 100kb 为 bin 的平均

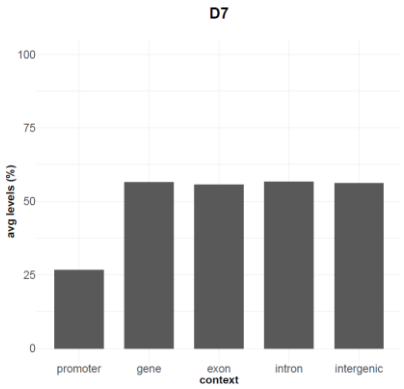
甲基化水平，灰色表示缺乏数据。图像高度表示每个区域的甲基化水平，UMR 或 LMR 以红点标注。图像较大，打开时可能会卡。

2.2.3 甲基化位点在基因组元件上的分布（The average CpG methylation levels in each genomic context）

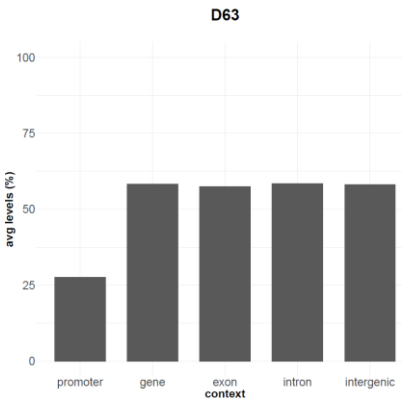
- D1 组(pdf)：



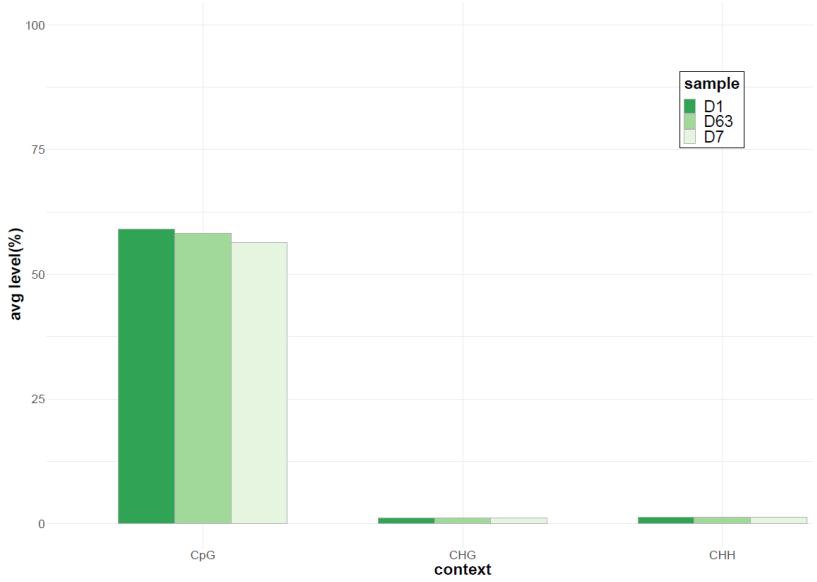
- D7 组(pdf)：



- D63 组(pdf)：



2.2.4 组间 CG、CHG 与 CHH 等的甲基化水平的比较(pdf file)



2.3 msPIPE 输出结果解读

Analysis 目录：

- avg_methlevel.pdf : CpG, CHG, and CHH context 的平均甲基化水平柱状图
- annotations : genes, exons, introns, promoters, intergenic regions 等区域的 bed 格式文件
- D1（D7 或 D63）：各组的甲基化分析结果
  - Average_methyl_lv.txt：每个基因及其 promoter 的平均甲基化水平
  - Avg_Genomic_Context_CpG.txt：每个基因组元件的平均甲基化水平(gene, exon, intron, promoter, and intergenic)
  - CXX_methylCalls.bed :每个 CX context (CXX is one of CpG, CHG, and CHH)的所有甲基化位点
  - AroundTSS/meth_lv_D1.txt：每个基因的 TSS (+/- 1500 bp)区域的滑动窗甲基化水平（bin 大小为 500bp，Step 大小为 100bp）
  - *MethylSeekR*：MethylSeekR 包的运行结果，主要用于鉴定 UMRs 和 LMRs。
  - UMR-Promoter.cnt.bed：每个 promoter 区域的 UMRs 数目（promoter 定义：基因上游的 1Kp 区域 ）
  - UMR-Promoter.pos.bed：每个 promoter 区域 UMRs 的基因组坐标
  - Circos.CpG_UMRs_LMRs.pdf：全基因组尺度的甲基化水平环状图（详见前面的介绍）
  - Genomic_Context_CpG.pdf：每个基因组元件的平均甲基化水平柱状图 (gene, exon, intron, promoter, and intergenic)
  - hist_sample1_CXX.pdf : CX context 的甲基化水平分布直方图 (CXX is one of CpG, CHG, and CHH)

Anslysis 目录中 DMR 目录为两两比较的分析结果，主要基于 R 包 methylkit

- D1.D7（D1.D63 或 D7.D93）：DMC/DMR 的分析结果目录, D1 为 control，D7 为 case，对于差异基因的设定尝试了 q0.5 和 q0.01 两个参数，q0.5 参数得到的差异基因数目更多，以下以 q0.5 为例。
  - DMR_q0.5.bed：差异区域的详细统计分析结果
  - *methylkit*：methylKit 包的输出结果。
  - DMC_q0.5.bed：q-value 0.5 参数过滤后的 DMCs
  - hypoDMC_detailed_count_methyl.txt：每个 promoter 区域的非甲基化 DMCs 数目(methylation level case < control)
  - hyperDMC_detailed_count_methyl.txt：每个 promoter 区域的超甲基化 DMCs 数目((methylation level case > control)
  - intersection.DMC2Promoter.txt：基因和 DMCs 的对应关系
  - DMC_genelist.txt：promoter 区域有 DMCs 的基因 list
  - DMC_gene.GOresult.txt :使用 g:Profiler 包对基因 list（methylKit 包）做 GO 分析的输出结果

    注：这里所用的 GO 分析工具是 R 包 gprofiler2，有一个问题是图片不显示 Term 名！客户可考虑使用在线的 gprofiler 重做一下，在线分析结果中，鼠标滑过时会显示 Term 名。建议考虑使用别的在线分析网站，我所找到的适用该物种的网站有：
    - KOBAS: GO 和 KEGG；可出图

        - DAVID: GO、KEGG 等
        - GeneOntology: GO
    - DMC_gene.GOresult.pdf : GO 分析结果绘图

注：bed 格式文件可以用 SeqMonk 或 IGV 软件打开。

参考文献

1. Kim H, Sim M, Park N, et al. msPIPE: a pipeline for the analysis and visualization of whole-genome bisulfite sequencing data[J]. BMC bioinformatics, 2022, 23(1): 1-13.
2. Akalin A, Kormaksson M, Li S, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles[J]. Genome biology, 2012, 13(10): 1-9.
3. Burger L, Gaidatzis D, Schübeler D, et al. Identification of active regulatory regions from DNA methylation data[J]. Nucleic acids research, 2013, 41(16): e155-e155.
4. Krueger F, Andrews S R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications[J]. bioinformatics, 2011, 27(11): 1571-1572.