Whole Genome Bisulfite Sequencing Analysis报告

**Part-1 数据预处理、比对和call methylation**

* 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](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)软件（version 0.11.9）对[raw data](1_raw-data)进行数据质控分析（结果见目录<1_raw-data_QC>）。
  + 使用[trim\_galore](https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) (version 0.6.7)和[cutadapt](https://cutadapt.readthedocs.io/en/stable/)（version 1.18）软件对[raw data](1_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](https://www.bioinformatics.babraham.ac.uk/projects/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](https://www.bioinformatics.babraham.ac.uk/projects/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](https://www.bioinformatics.babraham.ac.uk/projects/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](https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/)或[IGV](https://software.broadinstitute.org/software/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](https://www.bioinformatics.babraham.ac.uk/projects/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](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](https://www.bioinformatics.babraham.ac.uk/projects/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](3_aligned_BISMARK/D1-2_bismark_bt2_PE_report.html)
* [D1-4\_bismark\_bt2\_PE\_report.html](3_aligned_BISMARK/D1-4_bismark_bt2_PE_report.html)
* [D1-5\_bismark\_bt2\_PE\_report.html](3_aligned_BISMARK/D1-5_bismark_bt2_PE_report.html)
* [D63-5\_bismark\_bt2\_PE\_report.html](3_aligned_BISMARK/D63-5_bismark_bt2_PE_report.html)
* [D63-6\_bismark\_bt2\_PE\_report.html](3_aligned_BISMARK/D63-6_bismark_bt2_PE_report.html)
* [D63-8\_bismark\_bt2\_PE\_report.html](3_aligned_BISMARK/D63-8_bismark_bt2_PE_report.html)
* [D63-9\_bismark\_bt2\_PE\_report.html](3_aligned_BISMARK/D63-9_bismark_bt2_PE_report.html)
* [D7-1\_bismark\_bt2\_PE\_report.html](3_aligned_BISMARK/D7-1_bismark_bt2_PE_report.html)
* [D7-2\_bismark\_bt2\_PE\_report.html](3_aligned_BISMARK/D7-2_bismark_bt2_PE_report.html)
* [D7-3\_bismark\_bt2\_PE\_report.html](3_aligned_BISMARK/D7-3_bismark_bt2_PE_report.html)
* [D7-5\_bismark\_bt2\_PE\_report.html](3_aligned_BISMARK/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](file:///D:\其它\WGBS\3_aligned_BISMARK)目录中[bismark\_summary\_report.html](3_aligned_BISMARK/bismark_summary_report.html)。

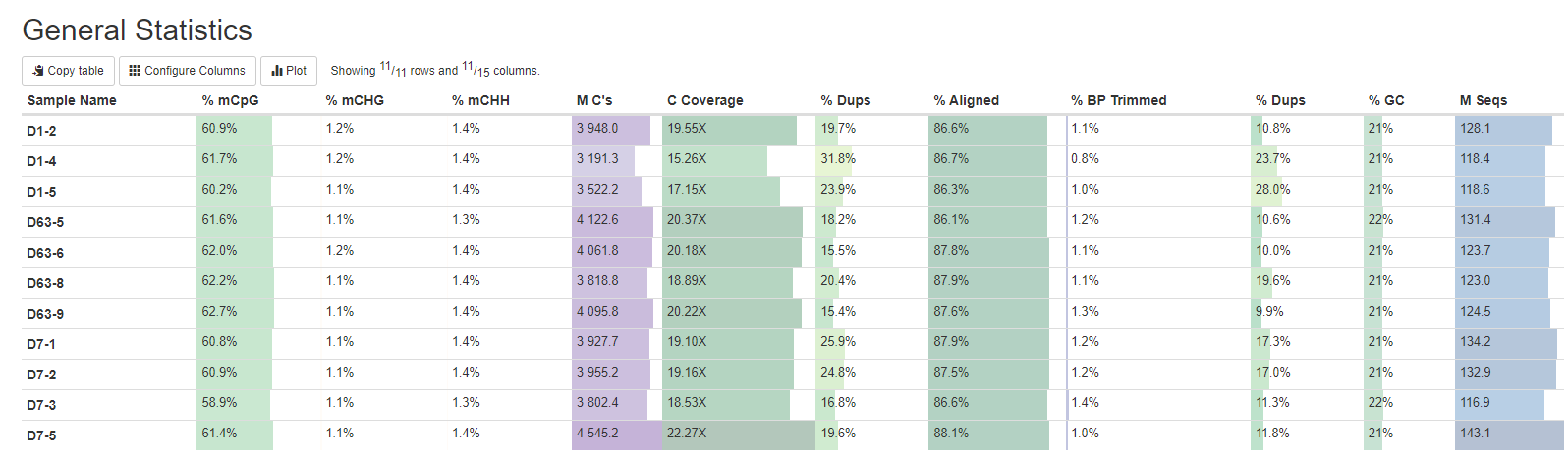
* + 使用[bismark](https://www.bioinformatics.babraham.ac.uk/projects/bismark/)软件里的*bismark2bedGraph*功能将call methylation的结果转为bedGraph格式，见[3\_aligned\_BISMARK\methylation](3_aligned_BISMARK/methylation)\SampleName目录中的SampleName\_bismark\_bt2\_pe.deduplicated.bedGraph.gz，可使用[SeqMonk](https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/)或[IGV](https://software.broadinstitute.org/software/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](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](https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/)可视化这些文件。

* Step-3: 汇总以上步骤的分析报告
  + 使用[multiqc](https://multiqc.info/)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](4_multiQC/multiqc_report.html)。
  1. 主要结果解读

这部分分析的简要统计结果，见[multiqc\_report.html](4_multiQC/multiqc_report.html)，基本统计信息如下：



|  |  |  |
| --- | --- | --- |
| 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**

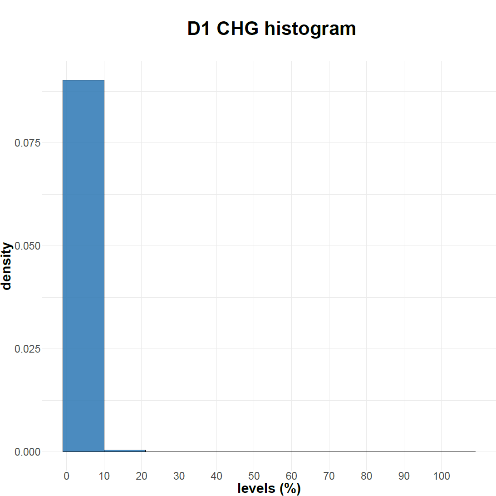
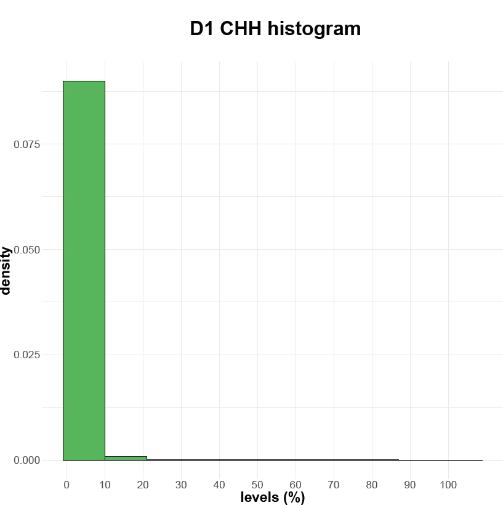
**主要基于**[msPIPE](https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04925-2)**分析流程。**

2.1 单样本水平分析及聚类和PCA分析

**结果见网页**[methylKit-Part1.html](6_supplement/methyKit-part/methylKit-Part1.html)**。**

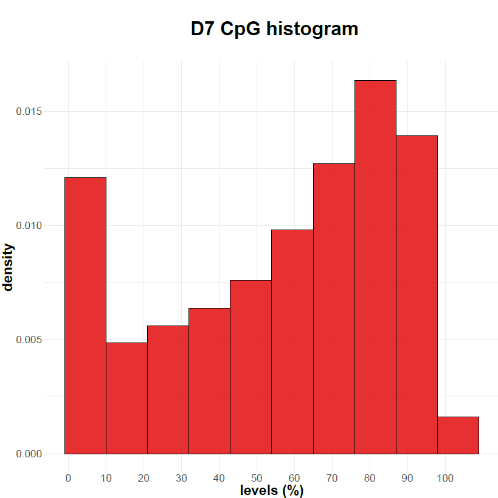
2.2 组水平分析

2.2.1甲基化水平分布

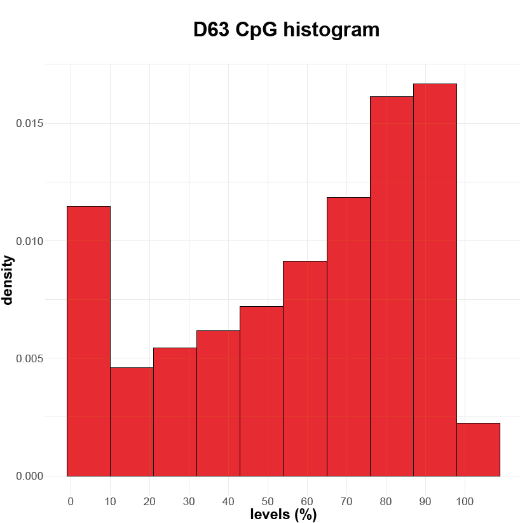
* 图表, 直方图

  描述已自动生成**D1 组（**[**CpG pdf**](5_msPIPE/Analysis/D1/hist_D1_CpG.pdf)**;** [**CHG pdf**](5_msPIPE/Analysis/D1/hist_D1_CHG.pdf)**;** [**CHH pdf**](5_msPIPE/Analysis/D1/hist_D1_CHH.pdf)**）：**
* 图表

  描述已自动生成图表, 条形图

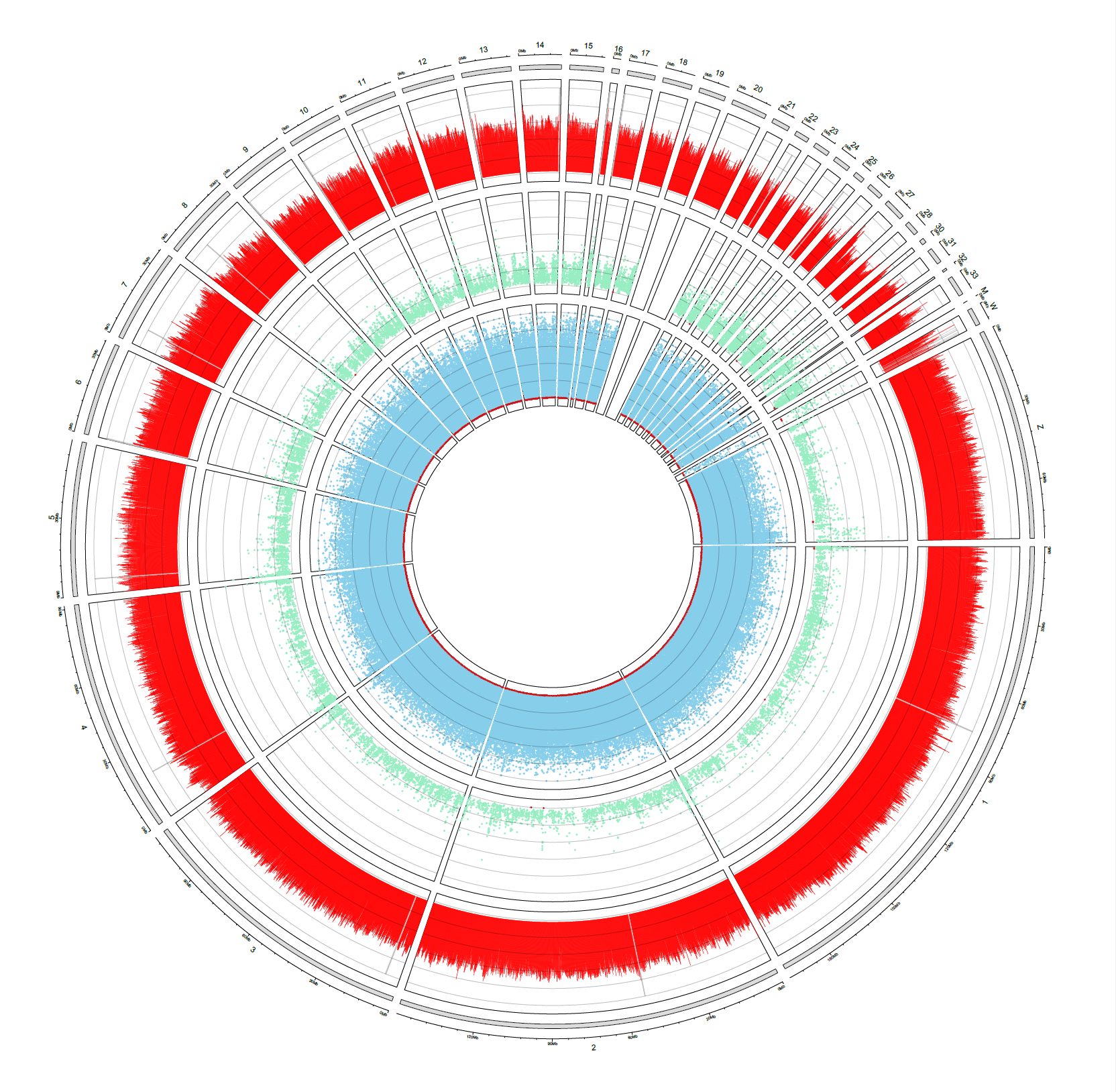
  描述已自动生成**D7 组（[CpG pdf](5_msPIPE/Analysis/D7/hist_D7_CpG.pdf);** [**CHG pdf**](5_msPIPE/Analysis/D7/hist_D7_CHG.pdf)**;** [**CHH pdf**](5_msPIPE/Analysis/D7/hist_D7_CHH.pdf)**）：**
* 图表

  描述已自动生成图表, 条形图

  描述已自动生成**D63 组（[CpG pdf](5_msPIPE/Analysis/D63/hist_D63_CpG.pdf);** [**CHG pdf**](5_msPIPE/Analysis/D63/hist_D63_CHG.pdf)**;** [**CHH pdf**](5_msPIPE/Analysis/D63/hist_D63_CHH.pdf)**）：**

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

* **D1 组(**[**pdf**](5_msPIPE/Analysis/D1/Circos_D1.CpG_UMRs_LMRs.pdf)**)：**

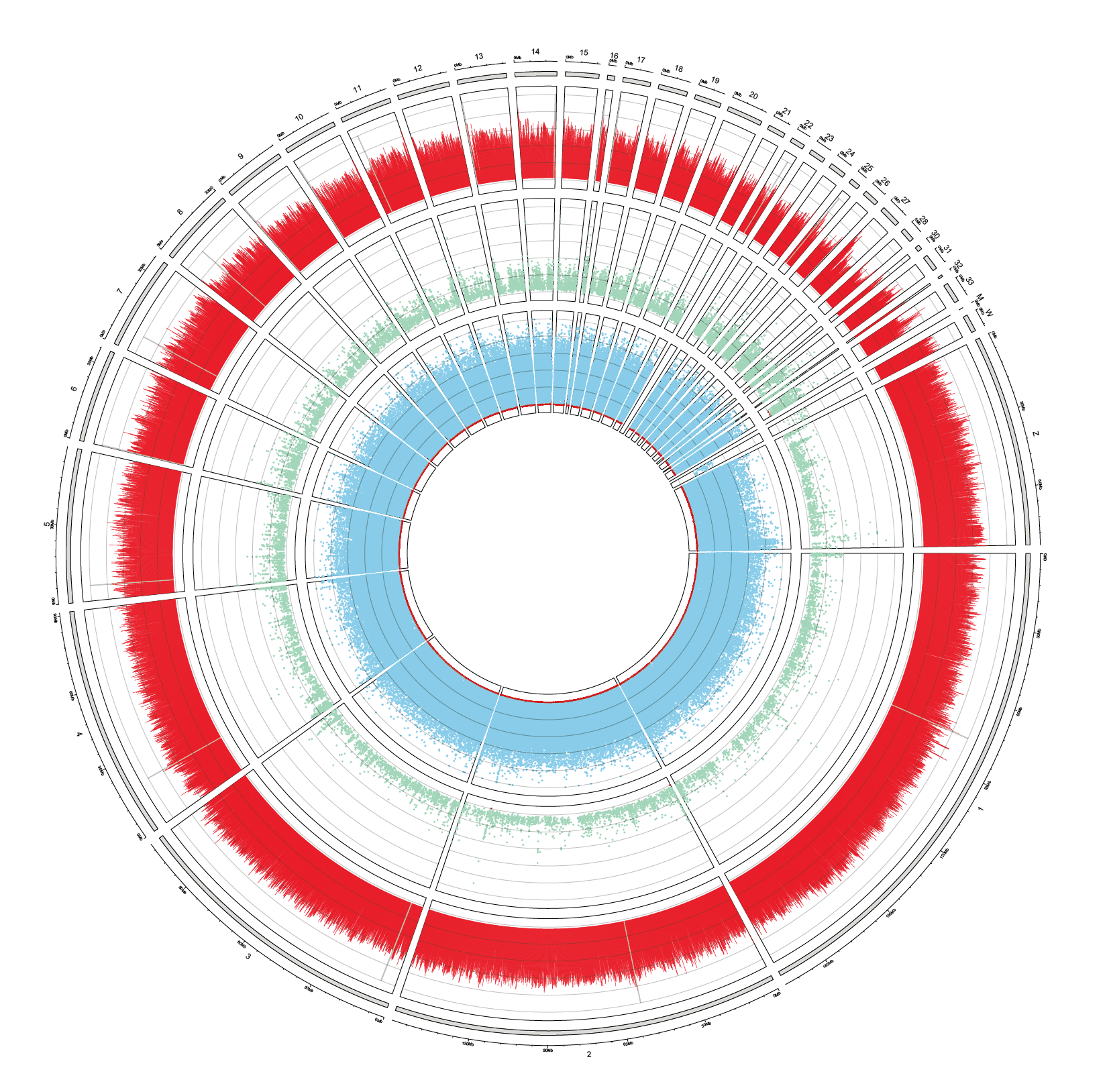


* **D7 组(**[**pdf**](5_msPIPE/Analysis/D7/Circos_D7.CpG_UMRs_LMRs.pdf)**):**

图表, 雷达图

描述已自动生成

* **D63 组(**[**pdf**](5_msPIPE/Analysis/D63/Circos_D63.CpG_UMRs_LMRs.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**](5_msPIPE/Analysis/D1/Genomic_Context_CpG.pdf)**)：**

图表, 条形图

描述已自动生成

* **D7 组(**[**pdf**](5_msPIPE/Analysis/D7/Genomic_Context_CpG.pdf)**)：**

图表, 条形图

描述已自动生成

* **D63 组(**[**pdf**](5_msPIPE/Analysis/D63/Genomic_Context_CpG.pdf)**)：**

图表, 条形图

描述已自动生成

**2.2.4 组间CG、CHG与CHH等的甲基化水平的比较(**[**pdf file**](5_msPIPE/Analysis/avg_methlevel.pdf)**)**

图表, 条形图

描述已自动生成

2.3 [msPIPE输出结果解读](5_msPIPE)

[Analysis](5_msPIPE/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](https://bioconductor.org/packages/release/bioc/vignettes/MethylSeekR/inst/doc/MethylSeekR.pdf)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](5_msPIPE/Analysis/DMR)目录为两两比较的分析结果，主要基于R包[methylkit](https://www.bioconductor.org/packages/release/bioc/vignettes/methylKit/inst/doc/methylKit.html)

* 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](https://biit.cs.ut.ee/gprofiler/gost)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](http://kobas.cbi.pku.edu.cn/genelist/): GO和KEGG；可出图
* [DAVID](https://david.ncifcrf.gov/tools.jsp): GO、KEGG等
* [GeneOntology](http://geneontology.org/): GO
  + DMC\_gene.GOresult.pdf : GO分析结果绘图

注：bed格式文件可以用[SeqMonk](https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/)或[IGV](https://software.broadinstitute.org/software/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.