Whole Genome Bisulfite Sequencing Analysis

**Part-1 数据预处理、比对和call methylation**

1.1 数据基本处理及质控

对下机数据进行初步质控、去除接头和低质量碱基，得到clean data，对clean data进行二次质控。

软件：fastqc；trim\_galore（cutadapt）；multiqc

1.2 Reads 比对、去重及质控

将reads比对到基因组上(bisulfite-converted reference genome)，并去除duplicated reads。

软件：bismark（--score\_min L,0,-0.6 -N 0 -L 20）；multiqc

1.3 Methylation calling

分别对CpG、CHG和CHH Context进行call methylation，并对结果进行质控。

软件: Bismark（bismark\_methylation\_extractor， --no\_overlap --comprehensive --gzip --CX --cytosine\_report）

1.4 部分质控结果展示

**1.4.1 综合信息**

电脑萤幕画面

描述已自动生成

|  |  |  |
| --- | --- | --- |
| 软件 | 缩写 | 全名 |
| Bismark | % mCpG | % Cytosines methylated in CpG context |
| Bismark | % mCHG | % Cytosines methylated in CHG context |
| Bismark | % mCHH | % Cytosines methylated in CHH context |
| Bismark | M C's | Total number of C's analysed, in millions |
| Bismark | % Dups | Percent Duplicated Alignments |
| Bismark | M Unique | Deduplicated Alignments (millions) |
| Bismark | M Aligned | Total Aligned Sequences (millions) |
| Bismark | % Aligned | Percent Aligned Sequences |
| Cutadapt | % BP Trimmed | % Total Base Pairs trimmed |
| FastQC | % Dups | % Duplicate Reads |
| FastQC | % GC | Average % GC Content |
| FastQC | Read Length | Average Read Length (bp) |
| FastQC | % Failed | Percentage of modules failed in FastQC report (includes those not plotted here) |
| FastQC | M Seqs | Total Sequences (millions) |

**1.4.2 Bismark**

**比对率**

图表, 饼图

描述已自动生成

**C甲基化**

甲基化C碱基中CG, CHG 与CHH的甲基化数目及比例

图表

描述已自动生成

**Part-2 Methylation analysis and visualization**

**主要基于MethylKit和Bsseq两个不同的下游分析流程。**

2.1 单样本水平的甲基化谱

**2.1.1 C碱基的甲基化水平分布**

图表, 直方图

描述已自动生成

**2.1.2 C碱基甲基化reads的覆盖度分布**

图表, 直方图

描述已自动生成

2.2 Replicates merged methylation profiles

**2.2.1 Circos plot：全基因组甲基化谱**

图表, 雷达图

描述已自动生成

**2.2.2 甲基化位点在基因组元件上的分布**

图表, 条形图

描述已自动生成

2.3 Comparative analysis

**2.3.1 组间CG、CHG与CHH等的甲基化水平的比较**

图表, 条形图

描述已自动生成

**2.3.2 Correlation plot**

日历

中度可信度描述已自动生成

**2.3.3 Clustering samples**

图表, 箱线图

描述已自动生成

**2.3.4 PCA plot**

图表, 散点图

描述已自动生成

**2.3.5 DMR And DMC analysis**

Finding and annotating differentially methylated bases（DMCs） or regions（DMRs）

**差异碱基的区域分布**

图表, 饼图

描述已自动生成

**CpG island annotation 分布**

图表, 饼图

描述已自动生成

**Plot the DMRs**

图表, 直方图

描述已自动生成

**GO analysis of DMR related gene set**

Gene ontology analysis of DMR related gene set was conducted using the list of genes in the 1 Mbp region upstream and downstream of DMR.

日程表

描述已自动生成

**注意：**

1. 下游分析比较灵活多变，也可根据需求进行调整。
2. 推荐使用UCSC的参考基因组

部分结果的输出目录结构：

Output files and directories in Analysis

* avg\_methlevel.pdf : a bar plot of average methylation level for CpG, CHG, and CHH context
* annotations : a directory with information of genes, exons, introns, promoters, and intergenic regions in BED format files
* sample1 : a directory with all results of methylation analysis for sample1
  + Average\_methyl\_lv.txt : average methylation level for each gene and its promoter
  + Avg\_Genomic\_Context\_CpG.txt : average methylation level for each genomic context (gene, exon, intron, promoter, and intergenic)
  + CXX\_methylCalls.bed : all methylation calls for each CX context (CXX is one of CpG, CHG, and CHH)
  + AroundTSS/meth\_lv\_3M.txt : for each gene, average methylation levels in bins around TSS (+/- 1500 bp)
  + MethylSeekR : a directory with all results for running MethylSeekR
  + UMR-Promoter.cnt.bed : the number of UMRs in each promoter region
  + UMR-Promoter.pos.bed : the genomic coordinates of UMRs in each promoter region
  + Circos.CpG\_UMRs\_LMRs.pdf : a circos plot for methylation level in whole-genome scale
  + Genomic\_Context\_CpG.pdf : a bar plot for average methylation level of each genomic context (gene, exon, intron, promoter, and intergenic)
  + hist\_sample1\_CXX.pdf : the distribution of methylation in CX context (CXX is one of CpG, CHG, and CHH)

Examples of output files and directories in DMR for comparison pair sample1 and sample2

* sample1.sample2 : a directory with all results of DMC/DMR analysis, in this case sample1 will be treated as control and sample2 will be treated as case
  + DMR\_q0.5.bed : information of differentially methylated regions
  + methylkit : output of running methylKit
  + DMC\_q0.5.bed : filtered DMCs with q-value 0.5
  + hypoDMC\_detailed\_count\_methyl.txt : the number of hypomethylated DMCs in each promoter (methylation level case < control)
  + hyperDMC\_detailed\_count\_methyl.txt : the number of hypermethylated DMCs in each promoter (methylation level case > control)
  + intersection.DMC2Promoter.txt : a list of intersection between genes and DMCs
  + DMC\_genelist.txt : a list of genes with DMCs overlapped their promoter region
  + DMC\_gene.GOresult.txt : a text output of GO enrichment test for genes with DMCs from methylKit using g:Profiler
  + DMC\_gene.GOresult.pdf : Plots of GO enrichment test for genes with DMCs from methylKit using g:Profiler
  + DMR\_gene.GOresult.txt : a text output of GO enrichment test for genes with DMRs from BSmooth using g:Profiler
  + DMR\_gene.GOresult.pdf : Plots of GO enrichment test for genes with DMRs from BSmooth using g:Profiler