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

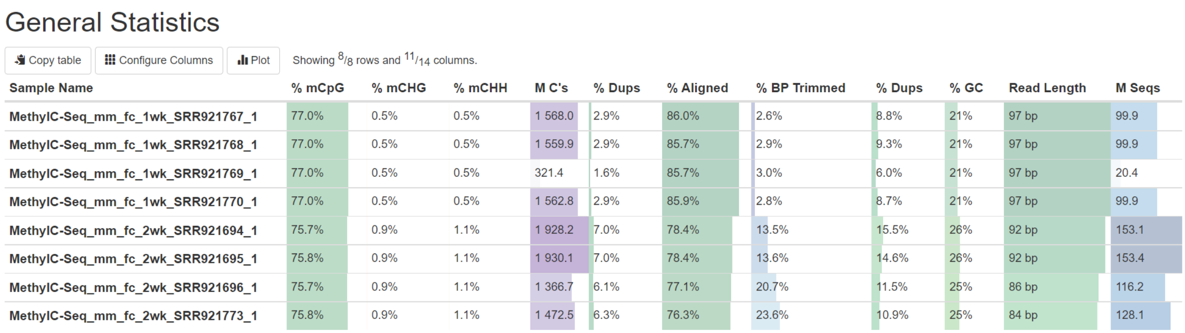
1.3 Methylation calling

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



General Statistics

Showing 8/8 rows and 11/14 columns.

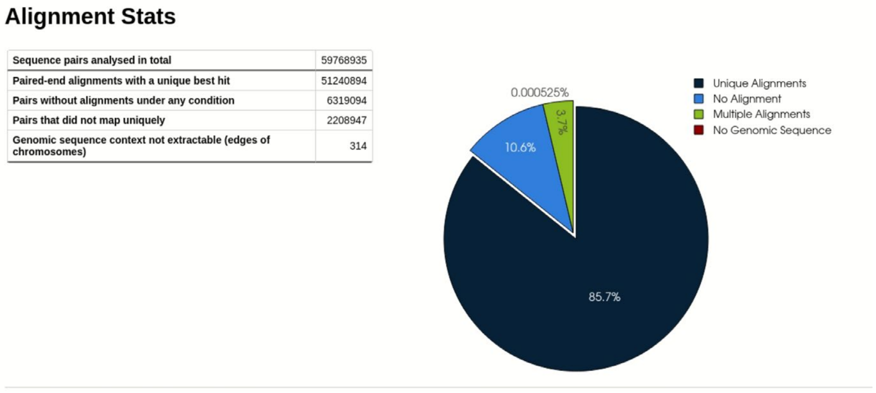
Sample Name	% mCpG	% mCHG	% mCHH	M C's	% Dups	% Aligned	% BP Trimmed	% Dups	% GC	Read Length	M Seqs
MethylC-Seq_mm_fc_1wk_SRR921767_1	77.0%	0.5%	0.5%	1 568.0	2.9%	86.0%	2.6%	8.8%	21%	97 bp	99.9
MethylC-Seq_mm_fc_1wk_SRR921768_1	77.0%	0.5%	0.5%	1 559.9	2.9%	85.7%	2.9%	9.3%	21%	97 bp	99.9
MethylC-Seq_mm_fc_1wk_SRR921769_1	77.0%	0.5%	0.5%	321.4	1.6%	85.7%	3.0%	6.0%	21%	97 bp	20.4
MethylC-Seq_mm_fc_1wk_SRR921770_1	77.0%	0.5%	0.5%	1 562.8	2.9%	85.9%	2.8%	8.7%	21%	97 bp	99.9
MethylC-Seq_mm_fc_2wk_SRR921694_1	75.7%	0.9%	1.1%	1 928.2	7.0%	78.4%	13.5%	15.5%	26%	92 bp	153.1
MethylC-Seq_mm_fc_2wk_SRR921695_1	75.8%	0.9%	1.1%	1 930.1	7.0%	78.4%	13.6%	14.6%	26%	92 bp	153.4
MethylC-Seq_mm_fc_2wk_SRR921696_1	75.7%	0.9%	1.1%	1 366.7	6.1%	77.1%	20.7%	11.5%	25%	86 bp	116.2
MethylC-Seq_mm_fc_2wk_SRR921773_1	75.8%	0.9%	1.1%	1 472.5	6.3%	76.3%	23.6%	10.9%	25%	84 bp	128.1

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

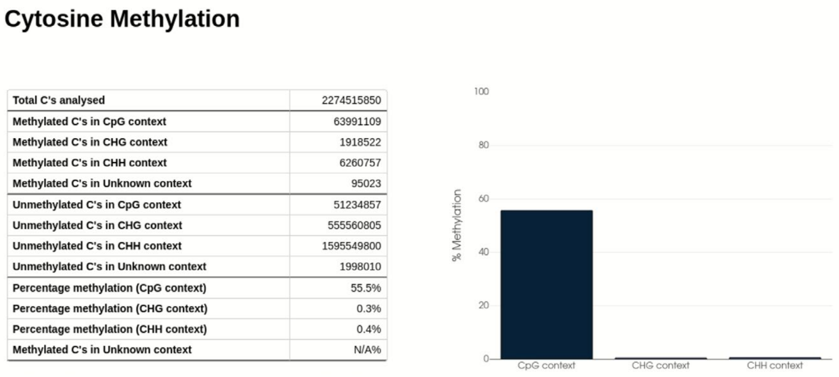


Alignment Stats

Sequence pairs analysed in total	59768935
Paired-end alignments with a unique best hit	51240894
Pairs without alignments under any condition	6319094
Pairs that did not map uniquely	2208947
Genomic sequence context not extractable (edges of chromosomes)	314

C 甲基化

甲基化 C 碱基中 CG, CHG 与 CHH 的甲基化数目及比例



Cytosine Methylation

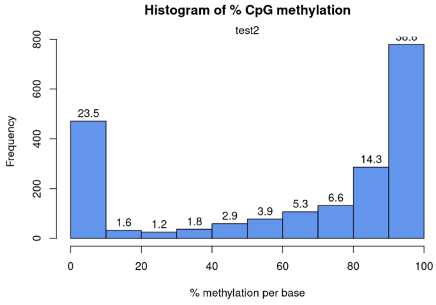
Total C's analysed	2274515850
Methylated C's in CpG context	63991109
Methylated C's in CHG context	1918522
Methylated C's in CHH context	6260757
Methylated C's in Unknown context	95023
Unmethylated C's in CpG context	51234857
Unmethylated C's in CHG context	555560805
Unmethylated C's in CHH context	1595549800
Unmethylated C's in Unknown context	1998010
Percentage methylation (CpG context)	55.5%
Percentage methylation (CHG context)	0.3%
Percentage methylation (CHH context)	0.4%
Methylated C's in Unknown context	N/A%

Part-2 Methylation analysis and visualization

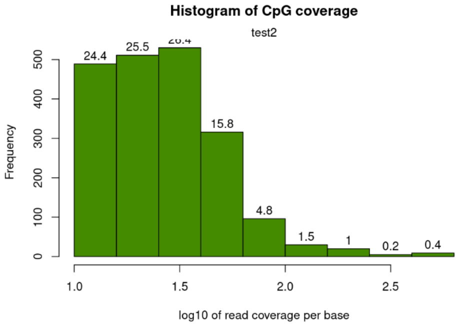
主要基于 MethylKit 和 Bsseq 两个不同的下游分析流程。

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

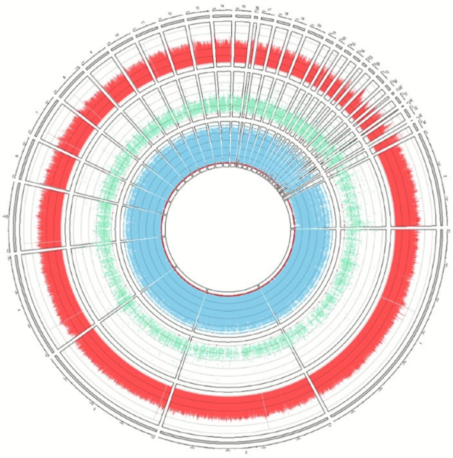


2.1.2 C 碱基甲基化 reads 的覆盖度分布

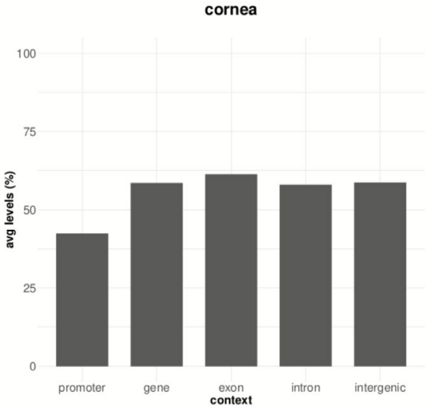


2.2 Replicates merged methylation profiles

2.2.1 Circos plot：全基因组甲基化谱

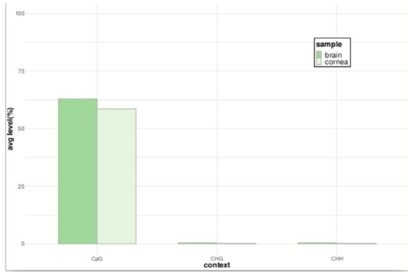


2.2.2 甲基化位点在基因组元件上的分布

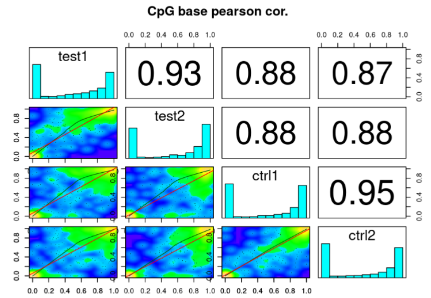


2.3 Comparative analysis

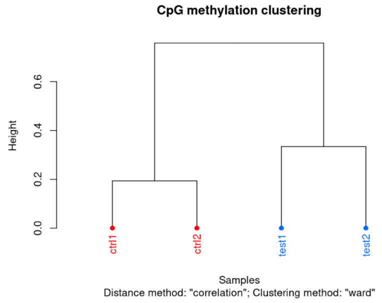
2.3.1 组间 CG、CHG 与 CHH 等的甲基化水平的比较



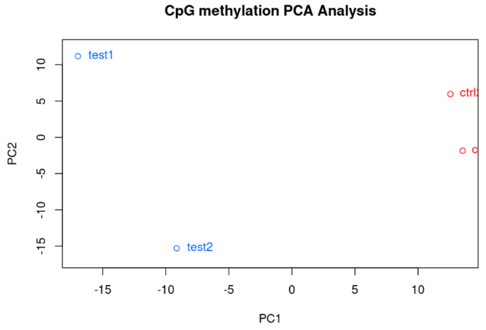
2.3.2 Correlation plot



2.3.3 Clustering samples



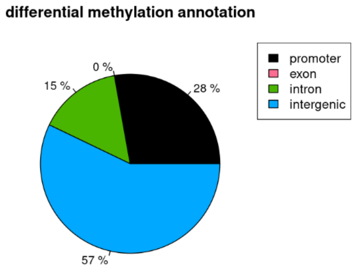
2.3.4 PCA plot



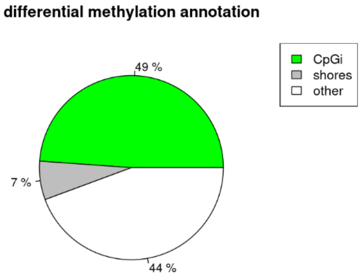
2.3.5 DMR And DMC analysis

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

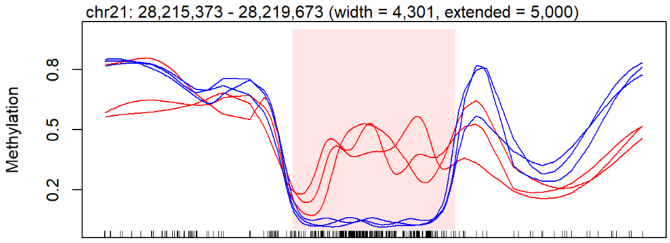
差异碱基的区域分布



CpG island annotation 分布

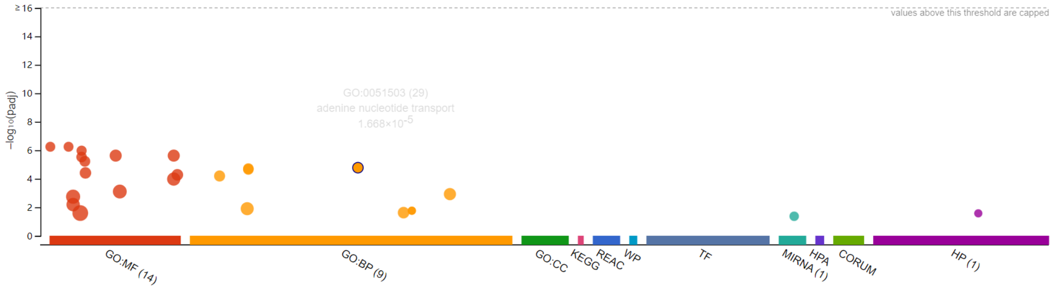


Plot the DMRs



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.



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