SMAP Further Analysis Report

目录

分析结果

1 数据基本处理与质控

所有样品的 WGBS 测序数据，将下机数据进行过滤，包括去污染，去测序接头和低质量碱基比例过高的 reads，得到 clean data。表 1-1 中列出了所有样本的数据产出概况。图 1.1 显示的是样品 Simulate01 的测序碱基含量分布，图 1.2 显示的是样品 Simulate01 的碱基测序质量分布情况。其余样品的测序碱基含量分布图与碱基测序质量分布情况图可在路径 /output/sample_name/sample_type/01.Data_Summary_and_QC 下查询。

表 1-1：数据基本处理与质控

Sample Name	Sample Type	Total read	Total base	Clean read	Clean base	Clean Rate (%)
Simulate01	Normal	664889900	83111056500	571476968	59657917639	71.78%
Simulate01	Tumor	633165724	79145540250	538312038	55353278301	69.94%
Simulate02	Normal	664889900	83111056500	571476968	59657917639	71.78%
Simulate02	Tumor	633165724	79145540250	538312038	55353278301	69.94%

Clean Rate (%) = Clean Data Size (bp)/Raw Data Size (bp)

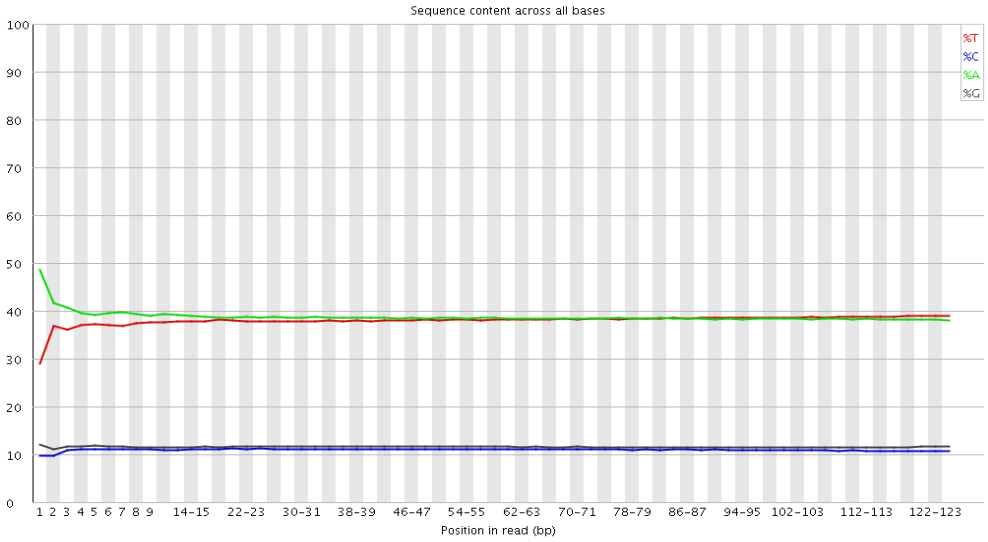

图 1.1: 样品 Simulate01 Clean reads 的碱基含量分布图 横坐标表示 reads 上碱基所在位置，纵坐标表示碱基比例。如果图中碱基分布不平衡则说明测序过程有异常情况发生。

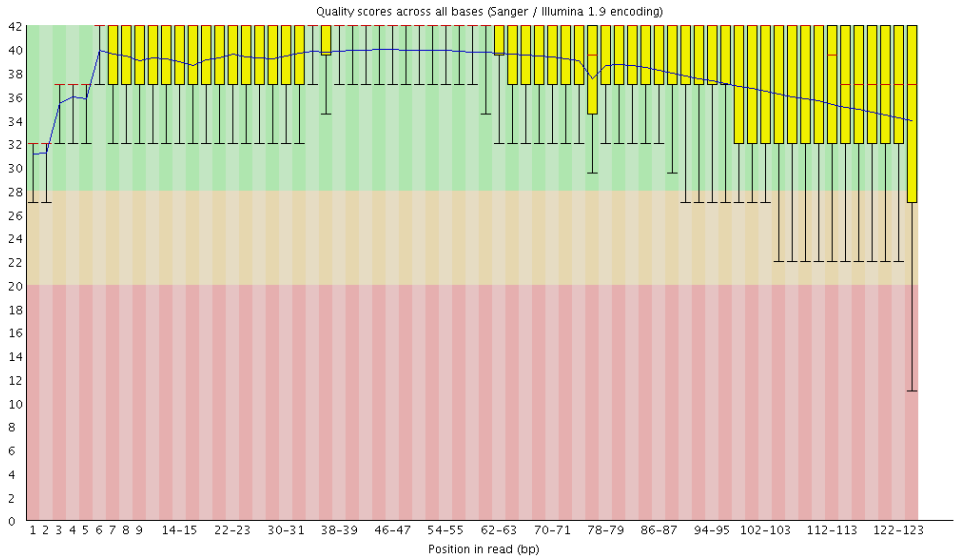

图 1.2: 样品 Simulate01 Clean reads 的碱基测序质量分布图 横坐标为 reads 上碱基位置；纵坐标为碱基测序质量。

在得到 clean data 之后，使用比对软件 BSMAP 将 reads 比对到参考基因组上，比对的统计结果如表 1-2 所示;然后根据需要对各个文库的 reads 进行去 duplication 处理。

表 1-2：比对结果统计

Sample Name	Sample Type	Total Reads	Pair Mapped	Pair Mapped Rate
Simulate01	Normal	420154240	394936188	94.00%
Simulate01	Tumor	398577754	365101772)	91.60%
Simulate02	Normal	420154240	394936188	94.00%
Simulate02	Tumor	398577754	365101772)	91.60%

表 1-3 是所有样品的测序深度和覆盖度情况。图 1.3 为样品 Simulate01 的测序深度分布图，理论上，其最高点对应的测序深度与全基因组平均覆盖深度一致或接近，这个分布图可以用于反映测序是否均匀。其他样品的测序深度分布图可在路径 /output/sample_name/sample_type/ 01.Data_Summary_and_QC 下查询。表 1-4 为所有样品在全基因组上的 C 位点覆盖度。

表 1-3：所有样品测序深度和覆盖度情况

Sample Name	Sample Type	Depth	Coverage
Simulate01	Normal	62.2%	45.42%
Simulate01	Tumor	55.86%	47.81%
Simulate02	Normal	62.2%	45.42%
Simulate02	Tumor	55.86%	47.81%

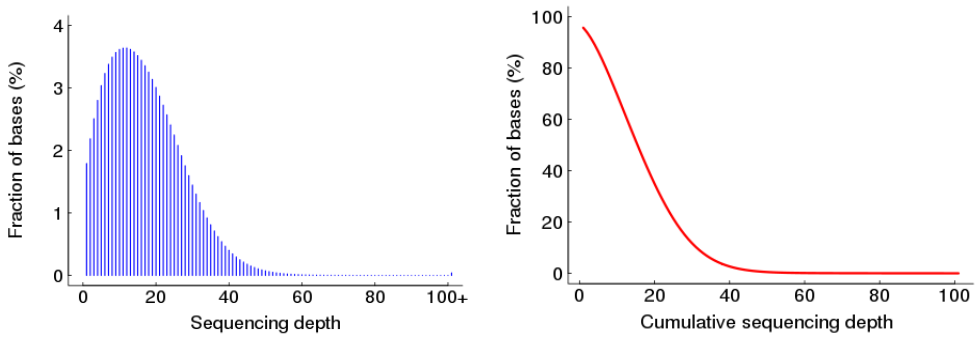

图 1.3：样品 Simulate01 测序深度覆盖度分布图

表 1-4：所有样品在全基因组上的 C 位点覆盖度

Sample Name	Sample Type	CG (%)	CHG (%)	CHH (%)
Simulate01	Normal	62.2%	45.42%	38.72%
Simulate01	Tumor	55.86%	47.81%	44.72%
Simulate02	Normal	62.2%	45.42%	38.72%
Simulate02	Tumor	55.86%	47.81%	44.72%

2 全基因组甲基化水平分析

用于分析的 DNA 样品为多细胞样品，因此 C 碱基的甲基化水平是一个 0%～100%范围内的数值，等于该 C 碱基上覆盖到的支持 mC 的序列数除以有效覆盖的序列总数，通常 CG 甲基化存在于基因和重复序列中，在基因表达调控过程中起到非常重要的作用。非 CG 类型的序列（CHG 和 CHH）在基因中十分少见，主要存在于基因间区和富含重复序列的区域，在沉默转座子过程中起关键作用。图 2-1 和表 2-1 为样品 Simulate01 在全基因组各染色体上的平均甲基化水平，表 2-2 为样品 Simulate01 在全基因组各类型调控元件范围内的甲基化水平。其余样品在全基因组各染色体上的平均甲基化水平与在全基因组各类型调控元件范围内的甲基化水平信息表可在路径 /output/sample_name/sample_type/02.Average_Methylation_Level_of_the_Whole_Genome 下查询。（加上点图 所有样品的三张合一）

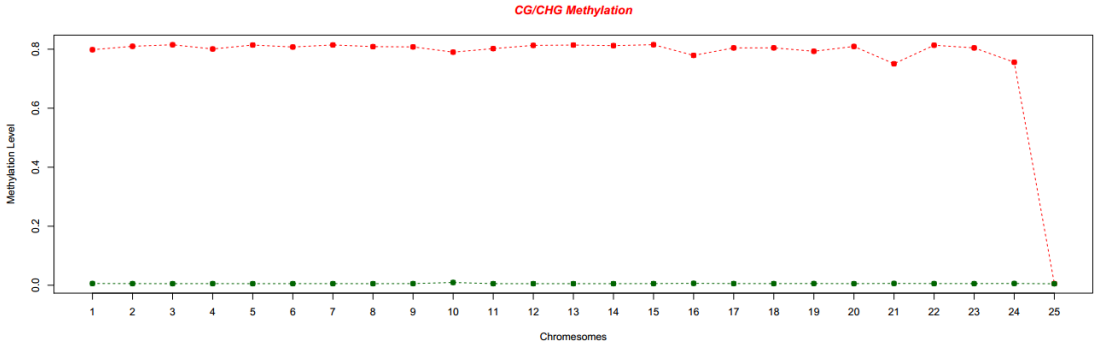

图 2.1：样品 Simulate01 在全基因组各染色体上的平均甲基化水平

表 2-1：样品 Simulate01 在全基因组各染色体上的平均甲基化水平

Chrosome	CG (%)	CHG (%)	CHH (%)
Chr1	80.93%	0.55%	0.55%
Chr2	81.78%	0.54%	0.55%
Chr3	81.63%	0.54%	0.55%
Chr4	80.94%	0.4%	0.54%
Chr5	81.23%	0.53%	0.54%
Chr6	80.78%	0.54%	0.55%
Chr7	81.73%	0.54%	0.54%
Chr8	81.42%	0.54%	0.54%
Chr9	81.16%	0.55%	0.54%
Chr10	81.59%	0.55%	0.54%
Chr11	80.31%	0.54%	0.54%
Chr12	81.24%	0.54%	0.54%
Chr13	81.46%	0.55%	0.56%
Chr14	81.14%	0.54%	0.54%
Chr15	81.66%	0.55%	0.54%
Chr16	81.36%	0.55%	0.54%
Chr17	81.47%	0.56%	0.54%
Chr18	81.62%	0.54%	0.55%
Chr19	79.41%	0.58%	0.54%
Chr20	80.97%	0.54%	0.54%
Chr21	80.48%	0.56%	0.56%
Chr22	82.18%	0.57%	0.54%
ChrX	81.64%	0.53%	0.53%
ChrY	76.52%	0.54%	0.55%
ChrM	0.58%	0.49%	0.48%
Total	81.22%	0.54%	0.55%

表 2-2：样品 Simulate01 在全基因组各类型调控元件范围内的甲基化水平

Different Genome Regions	CG (%)	CHG (%)	CHH (%)
3-UTR	80.98%	0.54%	0.54%
5-UTR	38.84%	0.53%	0.52%
CDS	80.61%	0.54%	0.51%
CpGIsland	20.01%	0.55%	0.56%
Downstream2k	78.75%	0.56%	0.53%
Genebody	82.38%	0.55%	0.54%
Intron	82.88%	0.55%	0.54%
Upstream2k	51.93%	0.55%	0.53%

3 甲基化 C 碱基中 CG, CHG 与 CHH 的分布比例

mCG，mCHG 和 mCHH 三种碱基类型的构成比例在不同物种中，甚至在同一物种不同样品中都存在很大差异。因此，不同时间、空间、生理条件下的样品会表现出不同的甲基化

图谱，各类型 mC(mCG、mCHG 和 mCHH)的数目，及其在全部 mC 的位点中所占的比例，在一定程度上反映了特定物种的全基因组甲基化图谱的特征。mCG、mCHG 和 mCHH 分别表示表示甲基化 CG、甲基化 CHG 和甲基化 CHH。三种碱基类型占比总和为 100%。

表 3-1 表示所有样品的 mCG、mCHG 和 mCHH 三种类型甲基化胞嘧啶的比例。图 3.1 表示样品 Simulate01 的不同序列类型甲基化 C 碱基的分布比例，其余样品的不同序列类型甲基化 C 碱基的分布比例图可在路径/output/sample_name/sample_type/03.Proportion_in_Total_Methyl_Cytosine_of_mCG_mCHG_and_mCHH 下查询。

表 3-1 ： mCG、mCHG 和 mCHH 三种类型甲基化胞嘧啶的比例

Sample Name	Sample Type	CG (%)	CHG (%)	CHH (%)
Simulate01	Normal	62.2%	45.42%	38.72%
Simulate01	Tumor	55.86%	47.81%	44.72%
Simulate02	Normal	62.2%	45.42%	38.72%
Simulate02	Tumor	55.86%	47.81%	44.72%

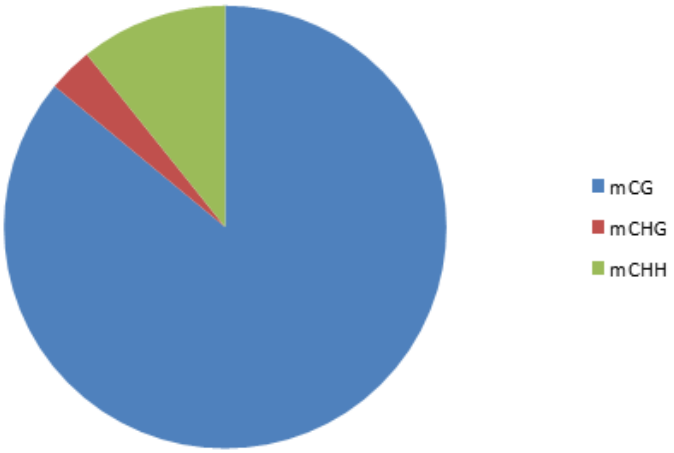

图 3.1 ： 样品 Simulate01 的不同序列类型甲基化 C 碱基的分布比例 蓝色部分为 mCG，红色为 mCHG，绿色为 mCHH。三者之和等于全基因组所有 mC（100%），即构成一个整圆。

4 甲基化的 CG，CHG，CHH 附近碱基的序列特征分析

在一些真核生物中，甲基化位点附近碱基的序列特征，对反映甲基化发生的序列偏向有指导意义[6]。为了研究序列特征与甲基化偏向性之间的联系，我们计算了甲基化位点上下游 9 个碱基（mC 位于第四个碱基）的甲基化百分比。图 4.1 表示样品 Simulate01 的 C 位点临近碱基的序列特征，其余样品的 C 位点临近碱基的序列特征图可在路径/output/sample_name/sample_type/04.Sequence_Preferences_for_Methylation_in_CG_CHG_and_CHH_Contexts 下查询。

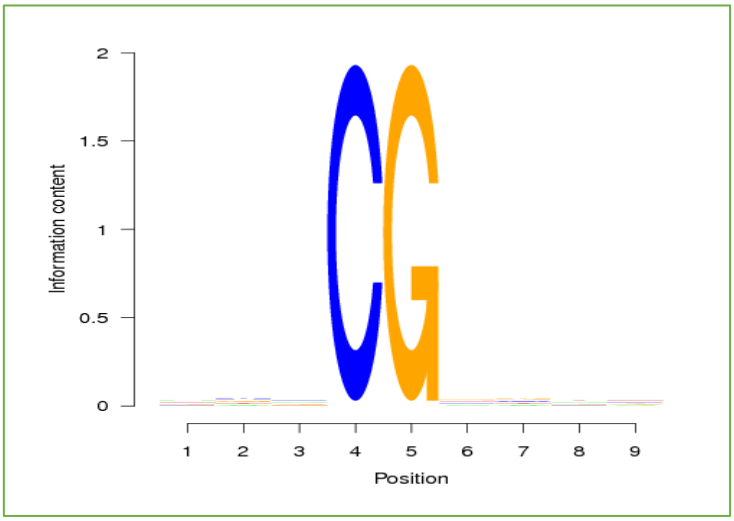

图 4.1 样品 Simulate01 的 C 位点临近碱基的序列特征。 图形横轴（x 轴）表示碱基位置，其中第四位上为用于分析的 C 碱基。纵轴（y 轴）为熵值（0 为最小值，表示四种碱基比例均匀，都为 25%，2 为最大值，表示四种碱基分布最不均匀，即只有一种特定碱基出现，如第四位的 C 与第五位的 G）。

5 染色体水平的甲基化 C 碱基密度分布

有研究证明非 CG 型的甲基化与 CG 型甲基化 C 的密度有很大的差异[3]。染色体亚端粒区域 DNA 甲基化水平通常较高，这一现象与端粒长度以及重组有十分重要的作用，此外还有基因表达和蛋白与 DNA 的相互作用都有紧密联系[3]。如图 5.1 为样品 Simulate01 染色体 chr1 上整体的甲基化水平图谱，图谱显示 DNA 甲基化的密度在整条染色体上变化很大。样品的其余染色体以及其余样品的各染色体的甲基化水平图谱可在路径 /output/sample_name/sample_type/05.Methylated_Cytosine_Density_of_Each_Chromosome 下查询。

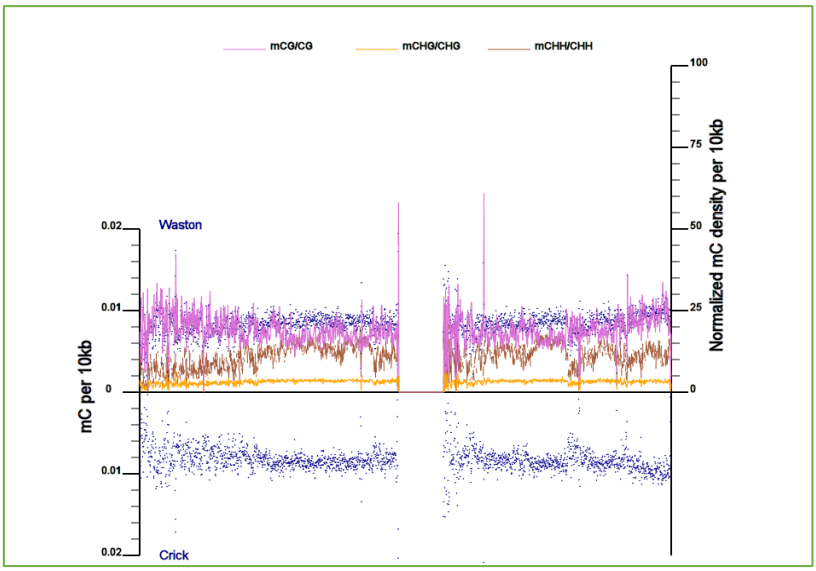

图 5.1 样品 Simulate01 chr1 染色体上甲基化 C 的密度分布 图形横轴表示 Chr1 染色体，从

左往右为染色体起点到终点。左边纵轴表示 10kb 为窗口计算得到的 mC 密度，以蓝点表示 mC 密度在染色体上的分布情况，右边纵轴表示标准化的 mC 比例，光滑曲线则表示不同类型甲基化 C 碱基(CG、CHG 和 CHH)的密度分布。

6 基因组不同转录元件中的 DNA 平均甲基化水平

为了深入地揭示 DNA 甲基化与基因表达的内在联系，所有编码基因序列被分成 7 种不同的转录元件区域，在此基础上对不同转录元件区域的平均甲基化水平进行统计。DNA 甲基化水平在不同功能区的分布特点有助于从全基因组水平去了解不同区域的 DNA 甲基化修饰的作用[7]. 图 6.1 为全基因组不同功能元件区域的甲基化平均水平分布图。其余样品的全基因组不同功能元件区域的甲基化平均水平分布图可在路径 /output/sample_name/sample_type/06.DNA_Methylation_Transcriptional_Units_at_Whole_Genome_Level 下查询。

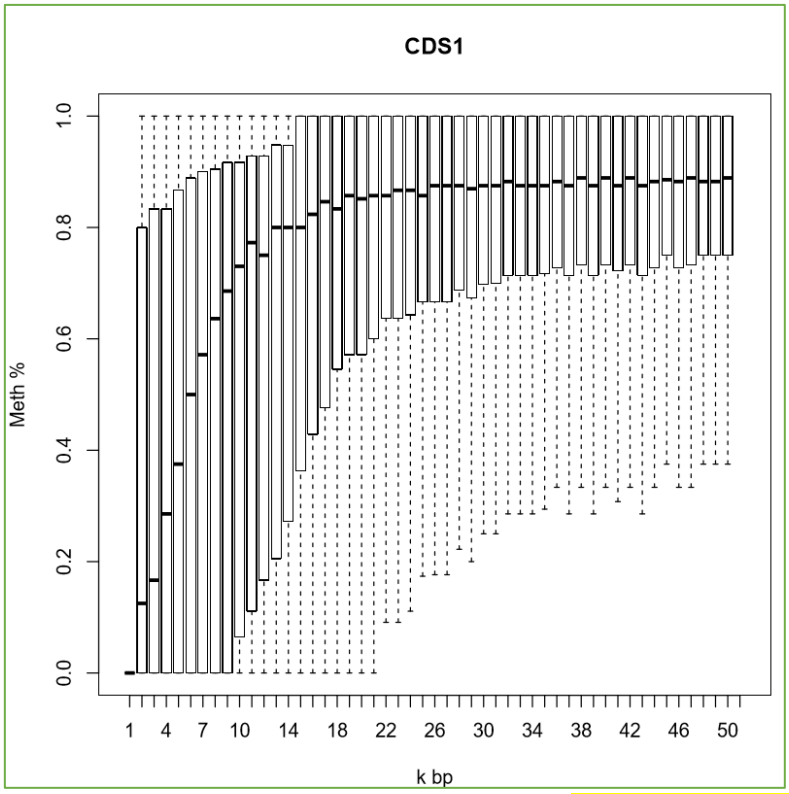

图 6.1 全基因组不同功能元件区域的甲基化平均水平分布 （将不同原件的图拼合成一个图）

7 ASM

8 DMR 的检测

差异甲基化区域（DMRs）是指不同样品中基因组表现出不同的甲基化模式的某些 DNA 片段。 DMR 与遗传印记相关，在个体中表现为与父本或母本的甲基化状态一致。甲基化的等位基因经常表现为沉默状态。亲本与子代甲基化模式的差异常常导致表观遗传缺陷 [7]，而人工繁殖技术可能会导致异常甲基化的比例升高，并导致疾病的发生。图 8.1 表示样品在全基因组不同功能元件区域的 DMR 分布示意图。其余样品在全基因组不同功能元件区域的

DMR 分布示意图可在路径 /output/sample_name/sample_type/08.Identification_and_Stats_of_DMRs 下查询。

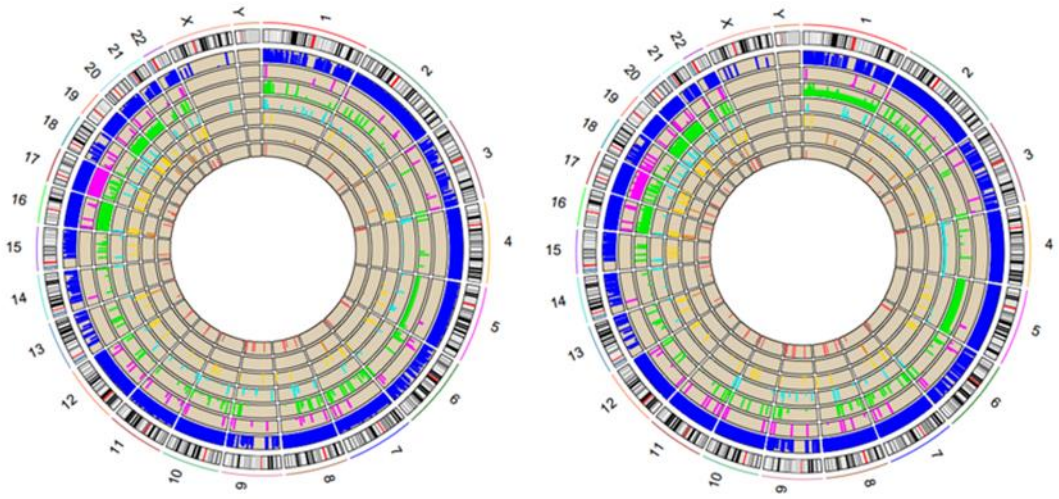

图 8.1 全基因组不同功能元件区域的 DMR 分布示意图 左侧为样品 N 的全基因组不同功能元件区域的 DMR 分布示意图，右侧为样品 T 的全基因组不同功能元件区域的 DMR 分布示意图。最外圈表示基因组染色体的位置，依次向内，第二圈表示 Intron 的甲基化率，第三圈表示 CDS 的甲基化率，第四圈表示 Upstream2K 的甲基化率，第五圈表示 CpGIsland 的甲基化率，第六圈表示 Downstream2k 的甲基化率，第七圈表示 5-UTR 的甲基化率，第八圈表示 3-UTR 的甲基化率。

9 DMR 相关基因的 GO 和 Pathway 分析

基因本体论（Gene ontology，GO）是所有物种中最主要的了解基因和基因产物属性的生物信息学分析手段，GO 分析能够用于鉴定基因产物的性能，它包含了三类基因功能信息：细胞组分（Cellular Component），分子功能（Molecular Function）和生物学过程（Biological Process）。为了探讨表观遗传变异在通路和生物学过程中起到的作用，我们对 DMR 相关的基因进行了 GO 和 Pathway 分析。

KEGG (Kyoto Encyclopedia of Genes and Genomes)是有关 Pathway 的主要公共数据库，该数据库整合了基因组、化学以及系统功能信息，特别是测序得到的基因集与细胞、生物体以及生态环境的系统性功能相关联。所有样品中的 DMR 相关基因均用 KEGG 数据库进行分析。

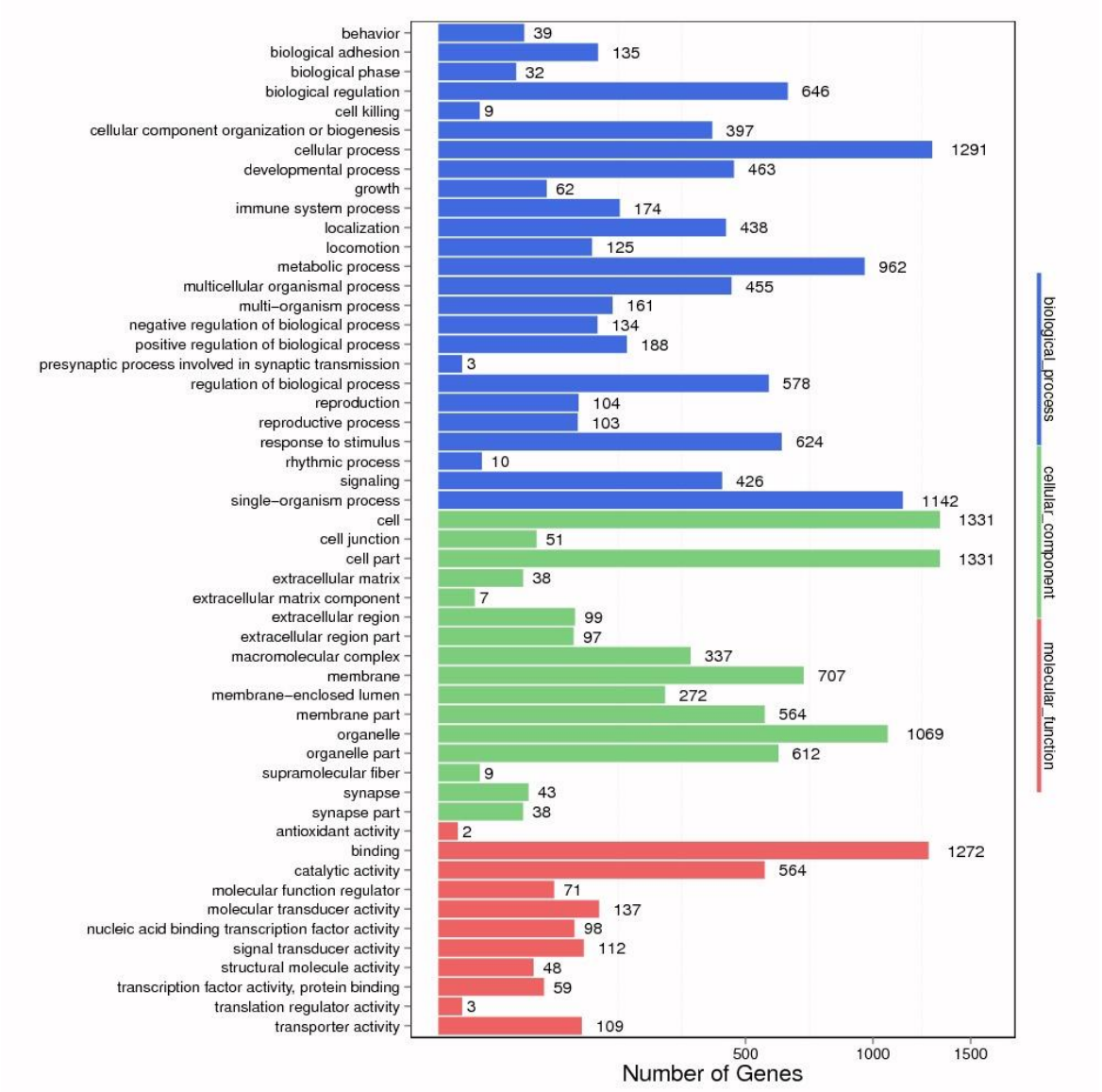

图 9.1：DMR 相关基因的 GO 聚类分析。图形横轴表示 DMR 相关基因的数量，纵轴表示各种 GO term，所有 GO term 分位三类，蓝色为生物学过程，绿色为细胞组分，红色为分子功能。

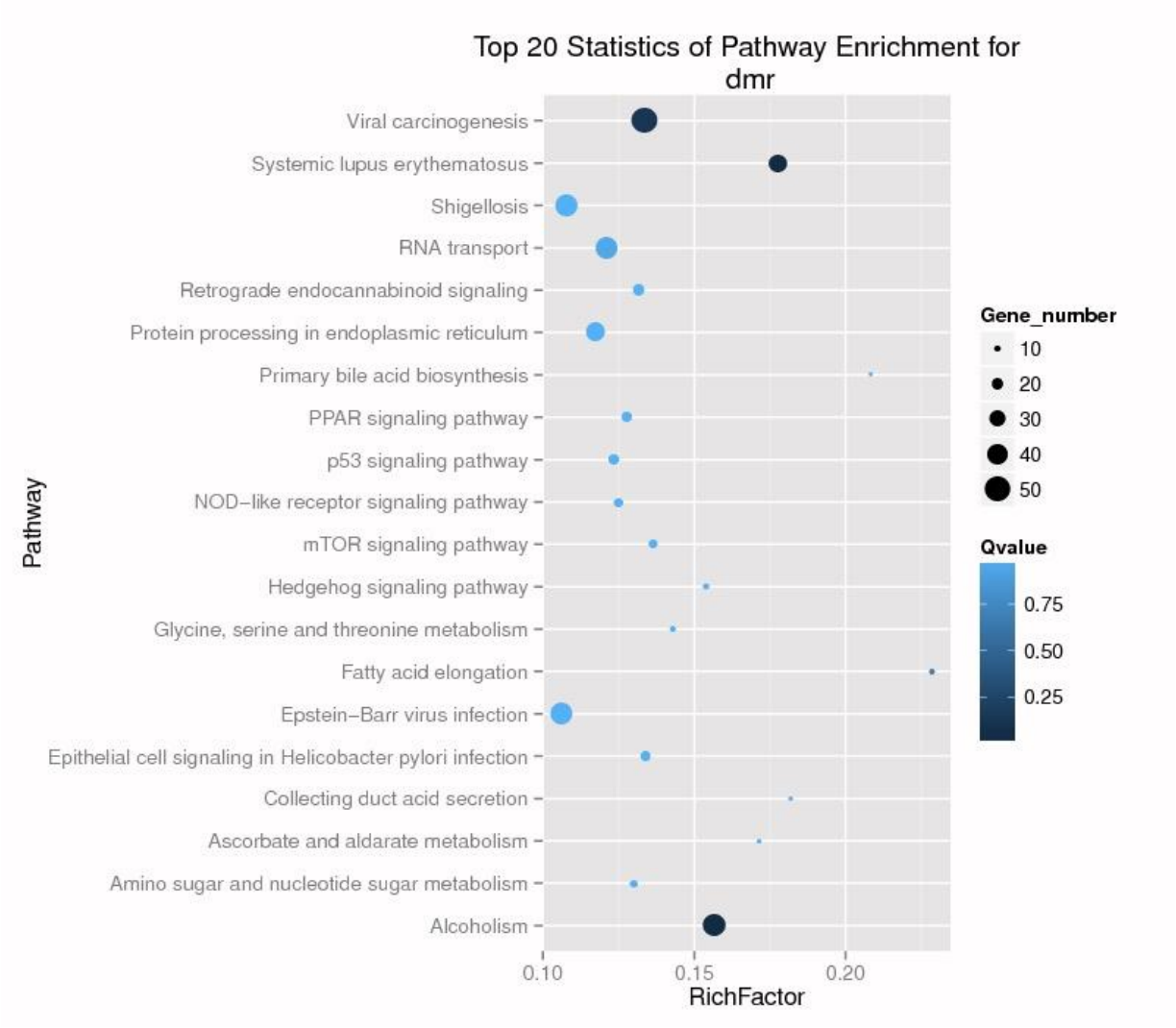

图 9.2：DMR 相关基因的 Pathway 功能显著性富集分析。 横轴（rich factor）是在通路中 DMR 相关基因占所有基因总数的比例，Rich-Factor 值越高则在该 pathway 中越富集。Q 为修正的 p value，其值为 0-1，Q 值越小，富集度越高。在此图中仅显示前 20 个富集通路。

分析方法

1 实验流程

2 信息分析流程

3 数据过滤

4 序列比对

5 甲基化水平

6 DMR 检测

7 甲基化水平程度差异

8 GO 注释

9 KEGG 通路富集

参考文献