Genome-wide DNA Methylation Analysis Report

Demo Report

May 1, 2016

Contents

1 Introduction....1
2 Library Construction and Sequencing....1
2.1 Samples DNA Testing....2
2.2 Library Construction....2
2.3 Library Testing....3
2.4 Sequencing....3
3 Analysis Workflow....3
4 Analysis Result....4
4.1 Raw Data Quality Control....4
4.1.1 Raw Data....4
4.1.2 Distribution of Sequencing Error Rate....4
4.1.3 Distribution of Base Content....6
4.1.4 Raw Data Filtering....6
4.1.5 Summary of Sequencing Data....7
4.2 Alignment....8
4.2.1 Overview of Mapping Status....8
4.2.2 Read Coverage and Sequencing Depth....9
4.2.3 Visualization....13
4.3 Identification of Methylated Cytosines....13
4.3.1 Methylation of Cytosines....14
4.3.2 Summary of Genomic Methylation Profile....15
4.3.3 Methylation Level....17
4.3.4 Sequence Contexts of Methylated Cytosines....19
4.4 Comparative Analysis of Methylomes....20
4.4.1 DMS Analysis....20
4.4.2 DMR Analysis....20
4.4.3 DMP Analysis....28
4.4.4 PCA Analysis....34
4.4.5 Clustering....34
5 References....36
6 Appendix....37
6.1 software list....37

1 Introduction

As an important kind of epigenetic modifications, DNA methylation is widely studied, and it works with histone modifications to play significant roles in regulation of gene expression, chromatin conformation and so on. DNA methylation in vertebrates typically occurs at CpG sites (cytosine-phosphate-guanine sites, where a cytosine is directly followed by a guanine in the DNA sequence. (Goldberg AD, 2007), while there is a large proportion of non-CG methylation (CHH, CHG and H represents A, C and T, respectively) in plants (Jackson JP, 2002). This methylation results in the conversion of the cytosine to 5-methylcytosine (5mC).

In general, high DNA methylation could depress genes expression, and demethylation would recover genes expression. DNA methylation participates in many cell activities, including cell differentiation, tissue-specific expression, genomic imprinting, and X-chromosome inactivation (Bird A, 2002; Jones PA, 2001; Reik W, 2003). Abnormal DNA methylation can cause developmental abnormalities, tumors and other diseases. Thus DNA methylation plays a significant role in understanding gene expression and individual development as well as the mechanisms of diseases occurrence and development.

2 Library Construction and Sequencing

We performed cluster analysis among the samples based on the detected SNPs. The result can help us determine whether two paired samples were from the same patient.

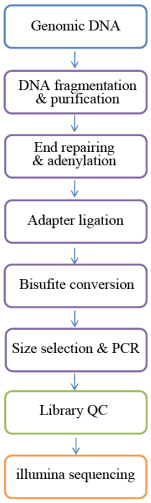

Figure 2.1 Cluster analysis among the samples

2.1 Samples DNA Testing

There are three methods of QC for DNA samples:

1) Agarose Gel Electrophoresis: testing DNA degradation and potential RNA contamination.

2) Nanodrop: testing DNA purity (OD_{260}/OD_{280}).

3) Qubit: determining DNA concentration.

2.2 Library Construction

After testing of the DNA samples, negative control is added into the DNA samples, and then they are fragmented into 200-300bp using Covaris S220. Subsequently, terminal repairing, A-ligation and methylation sequencing adapters ligation are performed to the DNA fragments. And the final DNA library is ready after bisulfite treatment (EZ DNA Methylation Gold Kit, Zymo Research. After BS treatment, unmethylated Cytosine will change into Uracil, while methylated Cytosine will stay unchanged). At last, size selection and PCR amplification are performed. The workflow is as follows:

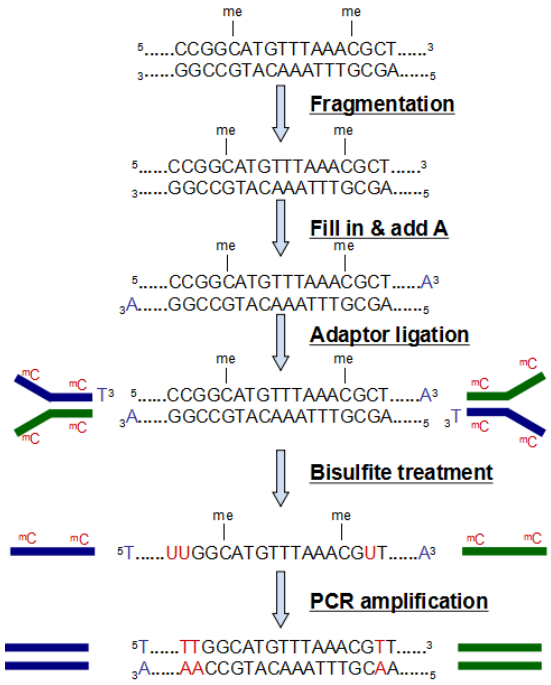

Figure 2.2 Library Construction for WGBS

2.3 Library Testing

Library concentration was first quantified using a Qubit2.0 fluorometer (Life Technologies), and then diluted to 1ng/μl before checking insert size on an Agilent 2100 and quantifying to greater accuracy by quantitative PCR (qPCR) (effective concentration of library > 2nM).

2.4 Sequencing

After passing library testing, different libraries would be pooled together and then fed into HiSeq devices according to effective concentration and expected data volume. The sequencing strategy is paired-end sequencing.

3 Analysis Workflow

The analysis workflow for sequencing data is as follows:

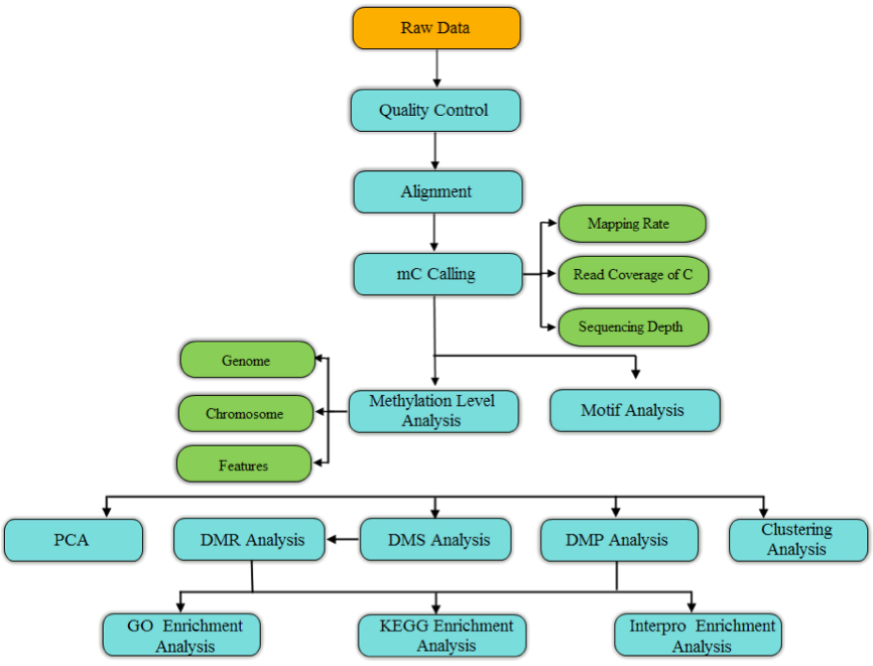

Figure 3.1 Analysis Workflow for WGBS

4 Analysis Result

4.1 Raw Data Quality Control

4.1.1 Raw Data

The original fluorescence images obtained from high throughput sequencing platforms are transformed to short reads by base calling. These short reads (Raw data) are recorded in FASTQ format, which contains sequence information (reads) and corresponding sequencing quality information.

Every read in FASTQ format is stored in four lines as follows:

```
@HWI-ST1328:93:H0VETADXX:1:1101:1390:2181
CATCAATGGCAATTGATTGCTTCCCGCTCTTGCTTGTTTCATCAGCTGTGCCTTTGC
CCTGCTTTTCAATAACAATCTTGTCATTCATCCCAAACTTGCA
+
CCCFFFFFGHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJHIIJJJJIJ
JJJJIIJIJJJJJIHHHHHHFFFFFFFEEEEEDDDDDDDDDD
```

Line 1 begins with a '@' character which is followed by a sequence identifier and an optional description. Line 2 shows the sequenced bases. Line 3 begins with a '+' character and is optionally followed by the same sequence identifier. Line 4 encodes the sequencing quality for each base in line 2, and contains the same number of characters as bases in line 2.

Illumina sequence identifier details:

EAS139	Unique instrument name
136	Run ID
FC706VJ	Flowcell ID
2	Flowcell lane
2104	Tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	Member of a pair, 1 or 2 (paired-end or mate-pair reads only)
Y	Y if the read fails filter (read is bad), N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	Index sequence

4.1.2 Distribution of Sequencing Error Rate

The ASCII value of every character at the fourth line minus 33 equals to the phred-scaled quality value of the corresponding sequenced base in the second line. The relationship between sequencing error rate (e) and base quality value (Q_{phred}) can be expressed by the following equation:

$$Q_{phred} = -10\log_{10}(e)$$

All relationships can be depicted by the following tables:

Table 4.1 Relationship between base quality and phred score (Illumina Casava v1.8)

Phred Score	Base Calling Error Rate	Base Calling Correct Rate	Q-sorce
10	1/10	90%	Q10
20	1/100	99%	Q20
30	1/1000	99.9%	Q30
40	1/10000	99.99%	Q40

Table 4.2 Examples of corresponding values among sequencing error rate (e), base quality value (Q_{phred}) and character.

Sequencing error rate	Sequencing quality value	Corresponding character
5%	13	.
1%	20	5
0.1%	30	?
0.01%	40	I

For WGBS, the distribution of sequencing error rate has two characters:

1) Sequencing error rate increase as the length of reads extended, which caused by the consumption of chemical reagents.

2) The first several bases in a read have a relatively high error rate, which maybe caused by incomplete combination of primers and DNA templates. Generally, the sequencing error rate of each base is less than 0.5%.

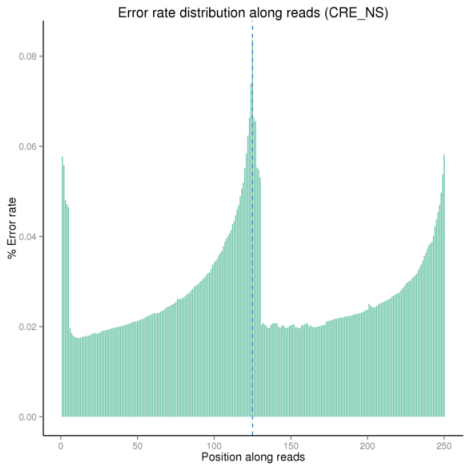

Figure 4.1 Distribution of Sequencing Error Rate

The x-axis represents the position of reads and the y-axis represents single base sequencing error rate. The first 125bp represents sequencing error rate distribution of read1; the last 125bp represents sequencing error rate distribution of read2.

4.1.3 Distribution of Base Content

In ordinary DNA library, theoretically, content of G bases should equal that of C bases, and content of A bases should equals that of T bases in each amplification cycle, which should be stable throughout the whole sequencing process. However, in the methylation library, C and G content will become low, and the strand-specific library strategy keeps information of strand orientation. Therefore, in the GC content distribution chart, C content of read1 is very low, while that of T is very high; G content of read2 is very low, while that of A is very high. In addition, distribution of complementary base pairs (A and T, G and C) in read1 and read2 shows an X-type feature.

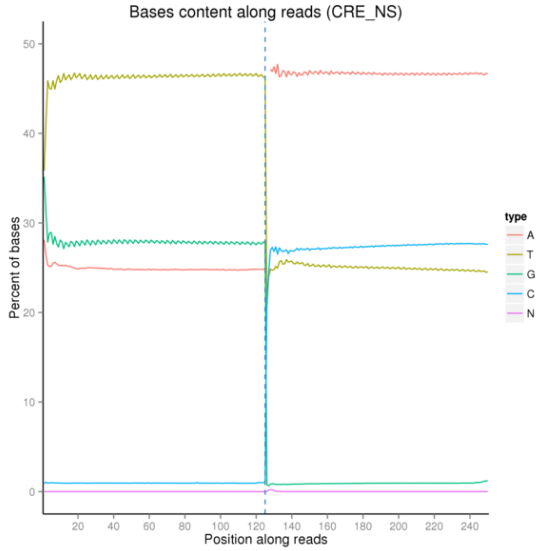

Figure 4.2 Distribution of GC Content

The x-axis represents the position of reads and the y-axis represents percentage of each base. Different colors stand for different base types.

4.1.4 Raw Data Filtering

Raw reads are filtered to remove reads containing adapters or reads of low quality,so that downstream analyses are based on clean reads. The filtering process is as follows:

1) Discard reads with adapter contamination.

2) Discard reads when uncertain nucleotides comprise more than 10 percent of either read (N > 10%).

3) Discard reads when low quality nucleotides (base quality score is less than 20) comprise more than 50 percent of the read..

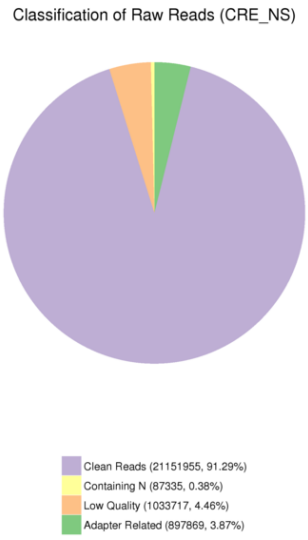

Figure 4.3 Raw Data Filtering

Results are shown as percentage of total raw reads. Clean Reads, reads that have passed quality control. Containing N, reads in which uncertain nucleotides comprise more than 10 percent of the read.

Low Quality, reads in which low quality nucleotides comprise more than 50 percent of the read. Adapter Related, reads that have adapter contamination.

4.1.5 Summary of Sequencing Data

Summary of Sequencing Data is described in table 1.1 in detail.

Table 4.3 Summary of Sequencing Reads

Sample name	Raw Reads	Raw Bases	Clean Reads	Clean Bases	Error Rate	Q20	Q30	GC Content	BS Conversion Rate
CRE_NS	46341752	5.79G	42303910	5.29G	0.03%	94.90%	89.96%	28.54%	99.66%
CRE_NSR	41498450	5.19G	38065638	4.76G	0.03%	94.70%	89.69%	26.68%	99.62%
CRE_CT	49138594	6.14G	40134650	5.02G	0.04%	93.30%	88.09%	23.43%	99.68%

Detail statistics of sequencing data:

1) Sample name: the name of sample
2) Raw Reads: the original sequencing reads count
3) Raw Bases: raw reads multiply reads length to get the number of raw bases, and the unit is G
4) Clean Reads: number of reads after filtering
5) Clean Bases: clean reads number multiply read length, saved in G unit
6) Error Rate: average sequencing error rate, which is calculated by $Q_{phred}=-10\log_{10}(e)$
7) Q20: percentage of bases whose correct base recognition rates are greater than 99% in total bases
8) Q30: percentage of bases whose correct base recognition rates are greater than 99.9% in total bases
9) GC content: percentage of G and C in total bases
10) BS conversion Rate: percentage of T changed from C by bisulfite

4.2 Alignment

Bismark software (Krueger, 2011) is used to perform alignments of bisulfite-treated reads to a reference genome. Cs are transformed to Ts and Gs are transformed to As (reverse complementary) in sequencing reads and reference genome by Bismark, and then these sequencing reads are aligned to similarly converted versions of the genome separately.

The processes of Bismark performing alignment is as follows:

1) Perform C-to-T and G-to-A conversions for both sequencing reads and reference genome.

2) Align converted reads to the converted genome.

3) Choose the best alignment from the parallel four alignment processes and consider it as the final result.

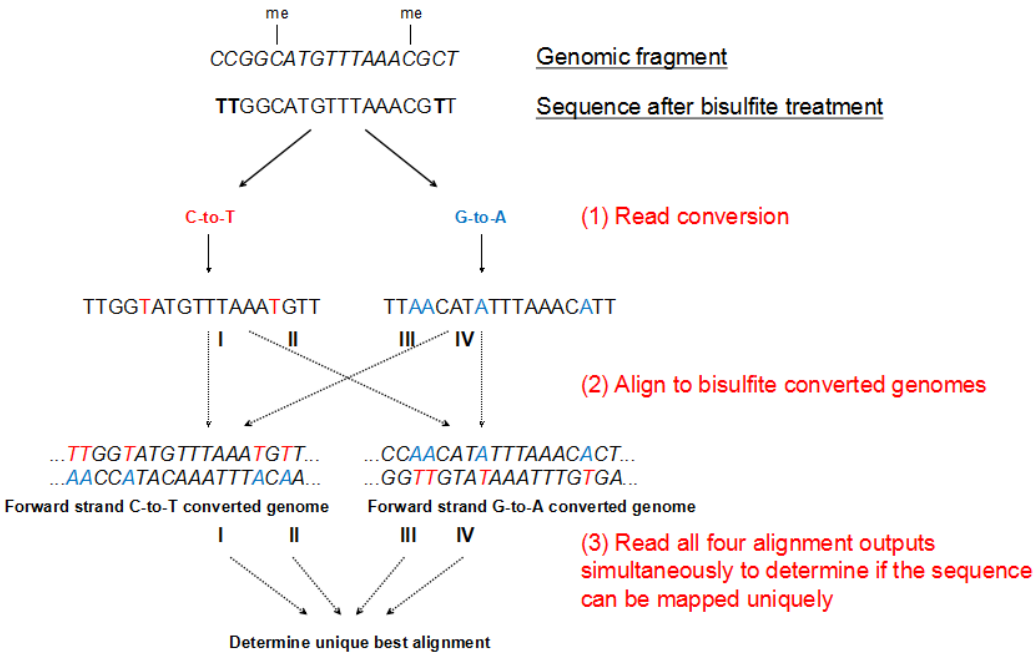

Figure 4.4 Bismark Alignment Strategy

Genomic fragment: shown in italics; Sequence after bisulfite treatment: shown in Roman letters.

4.2.1 Overview of Mapping Status

Mapping status is described as follows:

Table 4.4 Overview of Mapping Status

Samples	Total reads	Mapped reads	Mapping rate(%)	Duplication rate(%)	1x C coverage(%)	5x C coverage(%)
CRE_NS	21151955	17250119	81.55	6.85	81.49	77.54
CRE_NSR	19032819	15594856	81.94	6.45	81.41	76.68
CRE_CT	20067325	14492045	72.22	5.47	82.39	77.22

Details:

1) Samples: sample names

2) Total reads: number of total filtered clean reads

3) Mapped reads: number of uniquely mapped reads

4) Mapping rate(%): unique mapping rate of reads

5) Duplication rate(%): percentages of duplicated reads in total clean reads

6) 1x C Coverage(%): percentages of bases covered by at least 1 reads in the genome

7) 5x C Coverage(%): percentages of bases covered by at least 5 reads in the genome

Table 4.5 Methylation Status of Cytosine Sites

C(Mb)	mC(Mb)	mC percent(%)	CG(Mb)	mCG(Mb)	mCG percent(%)	CHG(Mb)	mCHG(Mb)	mCHG percent(%)	CHH(Mb)	mCHH(Mb)	mCHH percent(%)
752.6	21.0	2.80	205.3	9.1	4.45	180.0	4.3	2.42	367.3	7.6	2.06
689.3	4.7	0.69	180.0	2.7	1.47	159.9	0.6	0.40	349.4	1.5	0.42
676.4	4.1	0.61	157.5	2.1	1.30	142.7	0.6	0.39	376.2	1.5	0.40

Details:

1) C(Mb): the total number of C bases mapped to the genome

2) mCX(Mb): the number of methylated C bases mapped to the genome

3) mC percent(%): the percentage of methylated C bases mapped to the genome

4) CG(Mb): the total number of C bases mapped to the CG regions

5) mCG bases: the number of methylated C bases mapped to the CG regions

6) mCG percent(%): the percentage of methylated C bases mapped to the CG regions

7) CHG(Mb): the total number of C bases mapped to the CHG regions

8) mCHG(Mb): the number of methylated C bases mapped to the CHG regions

9) mCHG percent(%): the percentage of methylated C bases mapped to the CHG regions

10) CHH(Mb): the total number of C bases mapped to the CHH regions

11) mCHH(Mb): the number of methylated C bases mapped to the CHH regions

12) mCHH percent(%): the percentage of methylated C bases mapped to the CHH regions

4.2.2 Read Coverage and Sequencing Depth

4.2.2.1 Distribution of Reads Coverage in Genome

Calculate reads coverage of each base in genome. Then get distribution of reads coverage and cumulative distribution of reads coverage separately.

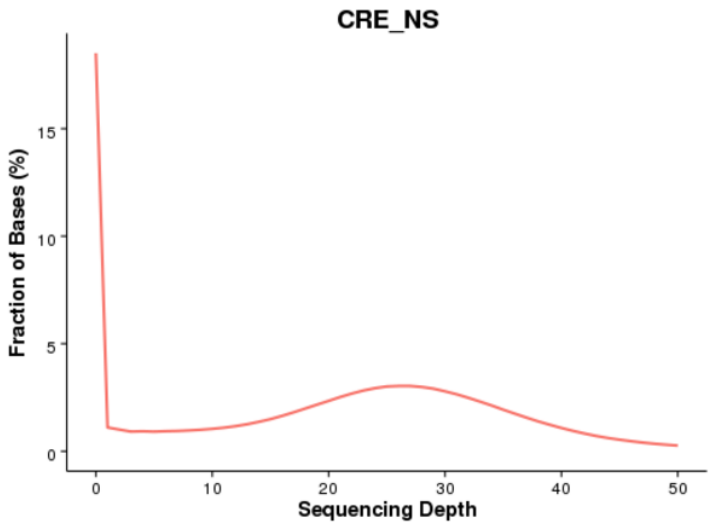

Figure 4.5 Distribution of Reads Coverage in Genome

The x-axis represents sequencing depth and y-axis represents the faction of bases with corresponding depth.

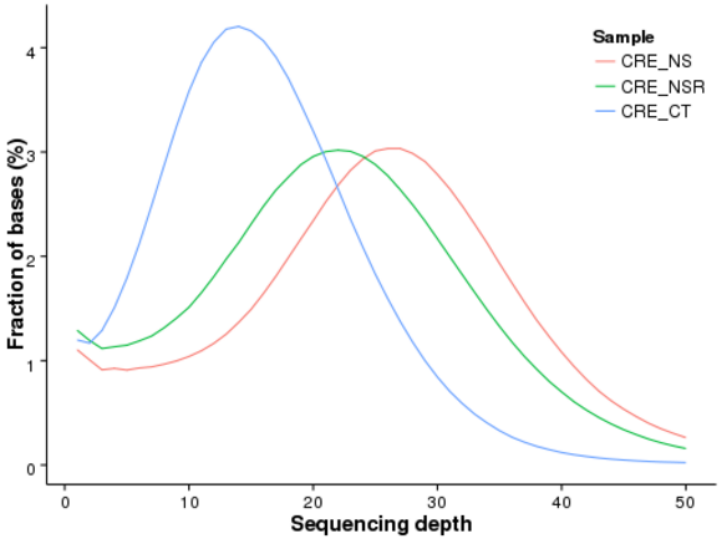

Figure 4.6 Distribution of All Sample Reads in Chromosomes

The x-axis represents sequencing depth and y-axis represents the faction of bases with corresponding depth. Different colors represent different samples.

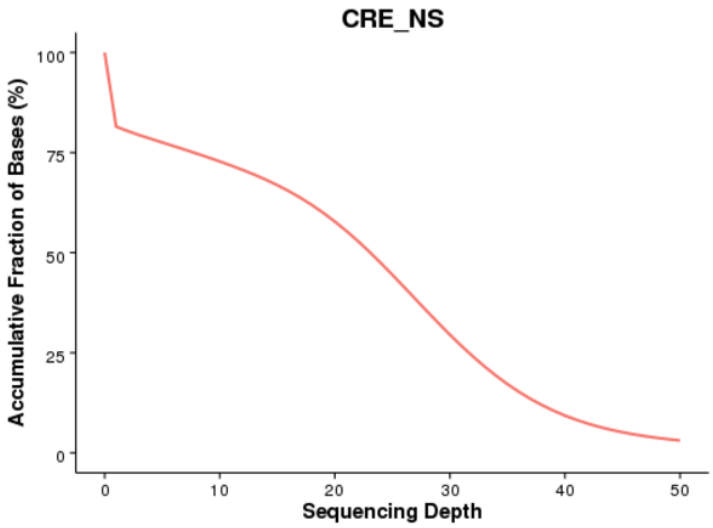

Figure 4.7 The Accumulative Distribution of Reads Coverage in Genome

The x-axis represents sequencing depth and y-axis represents the accumulative faction of bases above corresponding depth.

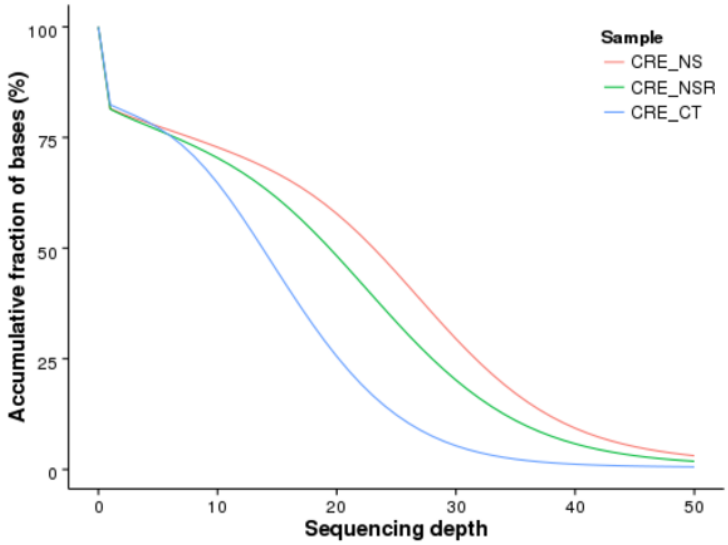

Figure 4.8 The Accumulative Distribution of All Sample Reads Coverage in Genome

The x-axis represents sequencing depth and y-axis represents the faction of bases above corresponding depth. Different colors represent different samples.

4.2.2.2 Distribution of Mapped Reads in Chromosomes

Calculate number of reads mapped to each chromosome, and get distribution of average read depth and coverage:

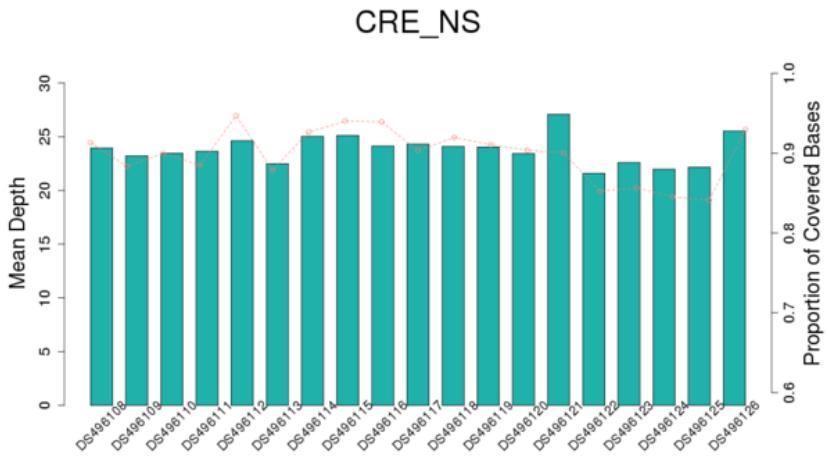

Figure 4.9 Distribution of Reads in Chromosomes

The x-axis represents chromosome ID, and y-axis represents the average depth of bases on each chromosome shown in the left histogram plot, the average coverage of bases on each chromosome is shown in the scatter diagram.

4.2.2.3 Cumulative Distribution of Cytosines

Calculate read coverage of cytosines in different contexts (CpG, CHH, CHG) and draw accumulative coverage distribution diagrams of bases in different contexts.

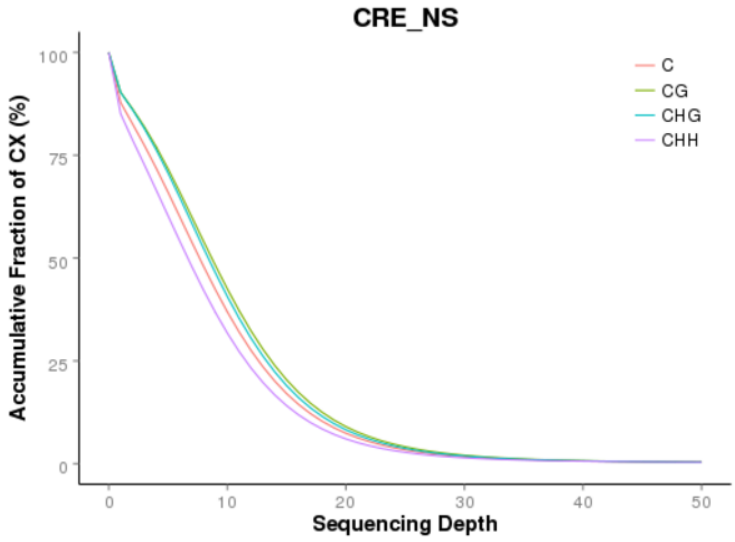

Figure 4.10 Cumulative Distribution of Cytosines

The x-axis represents sequencing depth and the y-axis represents the accumulative faction of bases above corresponding depth. Different colours stand for different contexts.

4.2.3 Visualization

Files are provided in BW format, a binary format of BEDGRAPH files that contains mapping results information, and the corresponding reference genome and gene annotation file for some species. The Integrative Genomics Viewer (IGV) is recommended for visualizing methylation. The IGV has several features:

1) it displays the positions of single or multiple reads in the reference genome, as well as reads distribution in annotated exons, introns or intergenic regions in adjustable scale;

2) it displays reads abundance in different regions to demonstrate their methylation level of different regions in adjustable scale;

3) it provides annotation information for both genes and splicing isoforms;

4) it displays annotations downloaded from remote servers and/or imported from local machines.

Figure 4.11 IGV Visualization

4.3 Identification of Methylated Cytosines

After alignment, Bismark (Krueger, 2011) is used to identify methyalted sites. Firstly, PCR duplicates must be removed and then methylation states of cytosines can be judged by comparison of reads base and the reference genome base at the same position. If the read base is C, it suggests that this cytosine is methylated; otherwise, it is unmethylated. Algorithm of Bismark identifying methylation sites is as follows.

In next generation sequencing, the sequencing depth of each site is different, so the ratio of methylated and unmethylated sites is different. However, the probability of each methylated site should be subjected to the binomial distribution. Binomial distribution, the repeat n times Bernoulli Experiment, is a widely used probability distribution of discrete random variables. In analysis process, in order to identify the true methylated sites, methylated and unmethylated counts at each site from Bismark output is tested by binomial distribution. Assuming the number of methylated cytosine is x in a certain site, where the read coverage is n and BS conversion rate is p, so the reliability of x methylated cytosine needs to be tested in above conditions. In order to find accurate methylated sites, a set of thresholds are set in analysis process (Ehsan Habibi et al., 2013, Casey A.Gifford et al., 2013) :

1) the sequencing depth is equal to or greater than five;

2) q-value equals to or is less than 0.05. Algorithm of Bismark Identifying Methylation Sites is as follows:

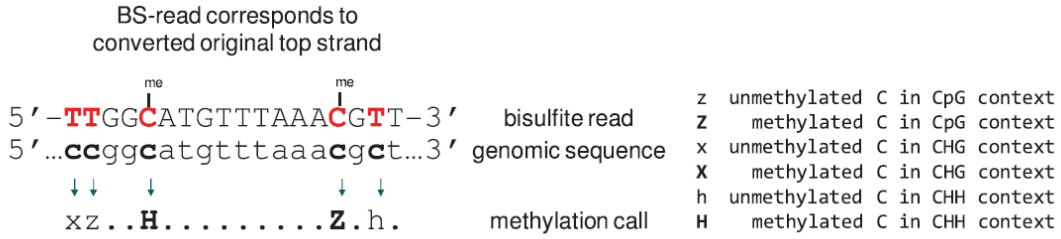

Figure 4.12 Bismark algorithm

4.3.1 Methylation of Cytosines

For the methylated sites, the methylation level is calculated using the following formula: ML=mC/(mC+umC) . ML represents the methylation level; mC and umC represent the number of mythylaed and unmethylated cytosines, respectively. Due to incomplete bisulfite conversion rate, the methylation level needs to be corrected using the following formula: ML_corrected=(ML-r)/(1-r). ML_corrected represents the corrected mythylation level, r represents Bisulfite non-conversion rate (Lister et al. 2013). The following methylation level is all corrected. Information of methylated cytosine are shown in the following table:

Table 4.6 Methylation Level of Cytosines

Chromosome	Position	Strand	Methylation Level	mC count	umC count	Pvalue	Corrected pvalue	Context
BK000554	9	+	0.204	43	168	4.2235e-62	3.8741e-59	CG
BK000554	10	-	0.161	5	26	6.8890e-08	6.2643e-06	CG
BK000554	41	+	0.451	317	386	0.0000e+00	0.0000e+00	CG
BK000554	42	-	0.056	14	237	3.6462e-13	5.0797e-11	CG
BK000554	48	+	0.289	246	605	0.0000e+00	0.0000e+00	CG
BK000554	49	-	0.275	82	216	6.8038e-129	1.0319e-125	CG

1) Chromosome: chromosome ID

2) Position: cytosine coordinate at genome

3) Strand: strand information of the cytosine

4) Methylation level: corrected methylation level at the site

5) mC count: the number of methylated cytosines

6) umC count: the number of unmethylated cytosines

7) Pvalue: p value of binomial distribution test

8) Corrected pvalue: adjusted p value

9) Context: contexts in which the cytosine is located

4.3.2 Summary of Genomic Methylation Profile

In different sequence context (CpG, CHH, CHG, H represents A, C, T), percentages of methylated cytosines in their contexts are shown in following table:

Table 4.7 Summary of Genomic Methylation Profile

Samples	mC percent(%)	mCpG percent(%)	mCHG percent(%)	mCHH percent(%)
CRE_NS	0.42%	1.26%	0.06%	0.11%
CRE_NSR	0.29%	1.03%	0.00%	0.01%
CRE_CT	0.25%	0.90%	0.00%	0.01%

Details:

1) Samples: sample name

2) mC percent(%): percent of methylated cytosines in all of the genomic cytosines

3) mCpG percent(%): percent of methylated cytosines in CG regions

4) mCHG percent(%): percent of methylated cytosines in CHG regions

5) mCHH percent(%): percent of methylated cytosines in CHH regions

In different contexts, number of methylated cytosines and the percentages of methylated cytosines in total methylated cytosines are shown in Table 4.8 and Figure 4.13:

Table 4.8 Number and Percentage of Methylated Cytosines

Samples	mC	mCG	mCHG	mCHH
CRE_NS	280906 (100%)	235339 (83.78%)	10737 (3.82%)	34830 (12.40%)
CRE_NSR	194169 (100%)	191843 (98.80%)	669 (0.34%)	1657 (0.85%)
CRE_CT	171262 (100%)	167674 (97.90%)	702 (0.41%)	2886 (1.69%)

Details:

1) Samples: sample names

2) mC: number and percent of methylated cytosines in genome

3) mCG: number of methylated cytosines in CG context and its percentage in total methylated cytosines

4) mCHG: number of methylated cytosines in CHG context and its percentage in total methylated cytosines

5) mCHH: number of methylated cytosines in CHH context and its percentage in total methylated cytosines

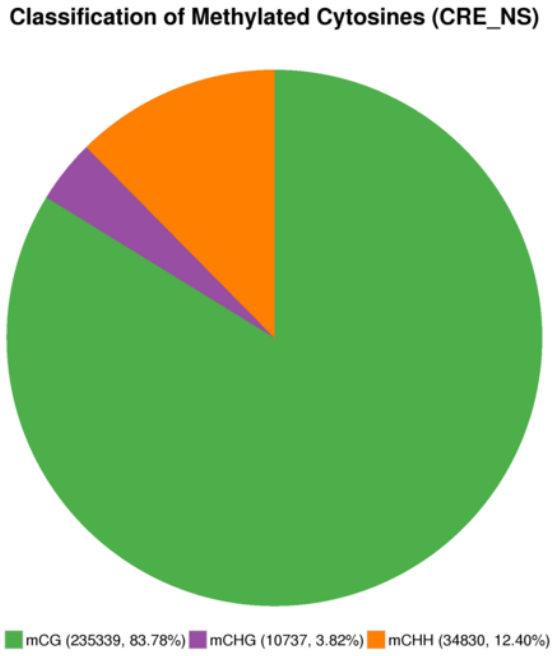

Figure 4.13 The Precentage Distribution of classified Methylated Cytosine

Different colours represent methylated cytosines in different contexts, the area represents the percentage of methylated cytosines in the corresponding context.

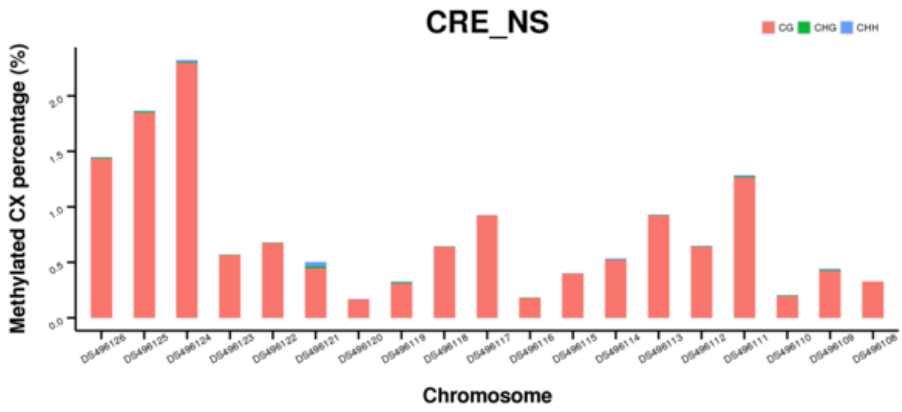

Figure 4.14 Distribution of Methylated Cytosine Proportion in Chromosomes in Different Contexts

The x-axis represents chromosome ID, and the y-axis represents methylated cytosines proportion in corresponding chromosomes. Different colours represent different contexts.

4.3.3 Methylation Level

4.3.3.1 Methylation Level of the Whole Genome

Average methylation level of the whole genome in each sample is calculated, and the result is shown in the following figure:

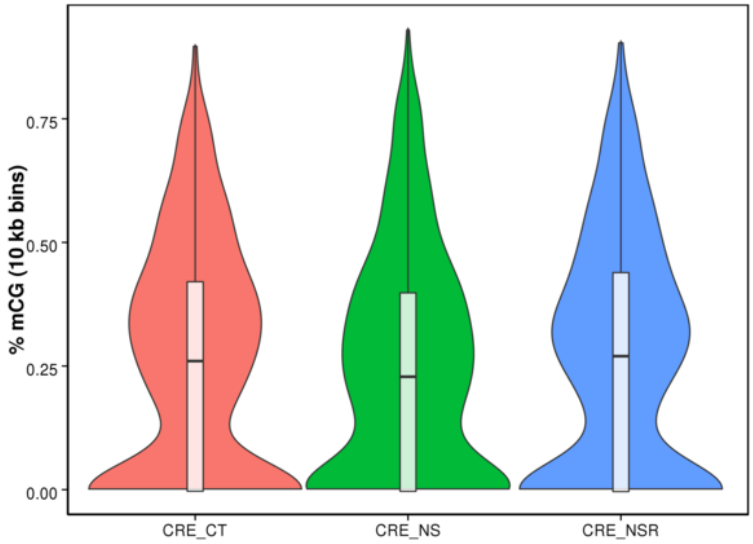

Figure 4.15 Distribution of the Whole Genome Methylation Level in Different Samples

The x-axis represents the different sample names, and the y-axis represents the methylation level with 10kb as a bin. The width of each violin represents the number of methylated sites in corresponding methylation level.

4.3.3.2 Methylation Level of Chromosomes

Average methylation level on each chromosome is calculated, and the result is shown in the following figure:

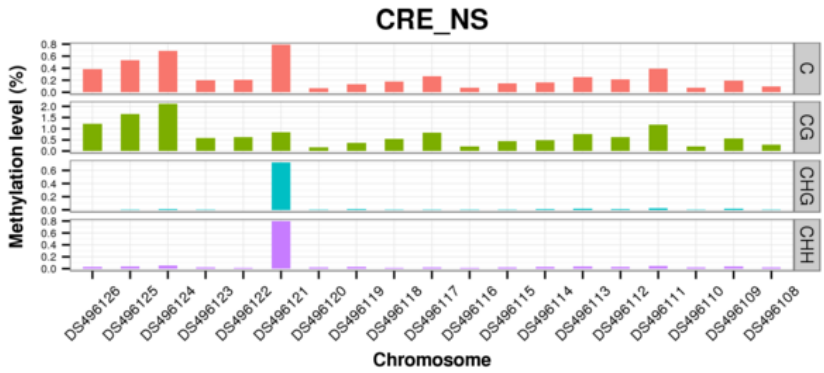

Figure 4.16 Distribution of Average Methylation Level in Chromosomes

The x-axis represents the chromosomes ID, and the y-axis represents the average methylation level in corresponding chromosomes. Different colours stand for different contexts.

4.3.3.3 Methylation Level of Genomic Features

The proportions of methylated sites in different cytosine contexts are calculated in different functional genomic regions (Promoter region is the 2kb region above TTS; 5' UTR, exon and intron are obtained from the Ensemble gene structure annotation files), the average methylation level distribution of all samples in genomic functional regions are shown in the following figure:

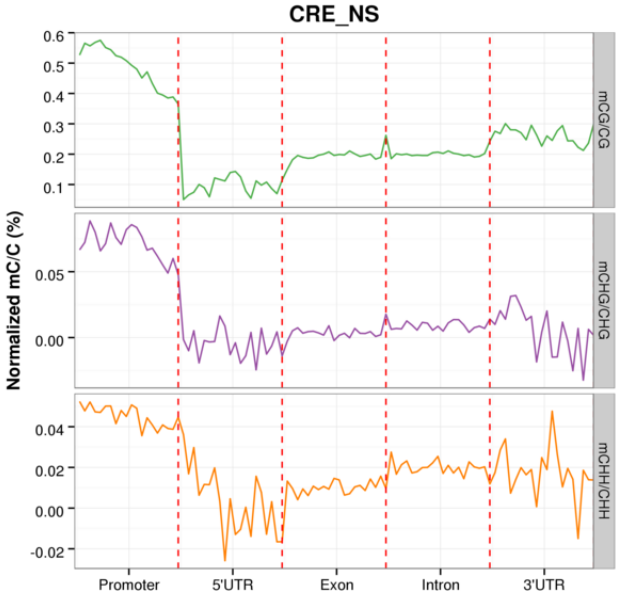

Figure 4.17 Methylation Level Distribution of All samples in Genomic Features

The x-axis represents different genomic elements, and the y-axis represents methylation level. The various functional areas of each gene would be divided into 20 bins, and then average methylation level within each bin region was calculated. Different colours stand for different contexts (CpG, CHG, CHH).

4.3.4 Sequence Contexts of Methylated Cytosines

In CG contexts, high methylated sites are defined as the methylation level is higher than 75%, while low methylated sites are defined as the methylation level is lower than 75%.

In non-CG contexts, high methylated sites are defined as the methylation level is higher than 25%, while low methylated sites are defined as the methylation level is lower than 25%. In order to study the characters of up-stream and down-stream sequences of methylated cytosines in different contexts, logo plots are drawn within 9 bp sequence around the positions of methylated cytosines, and the results are shown in the following figure:

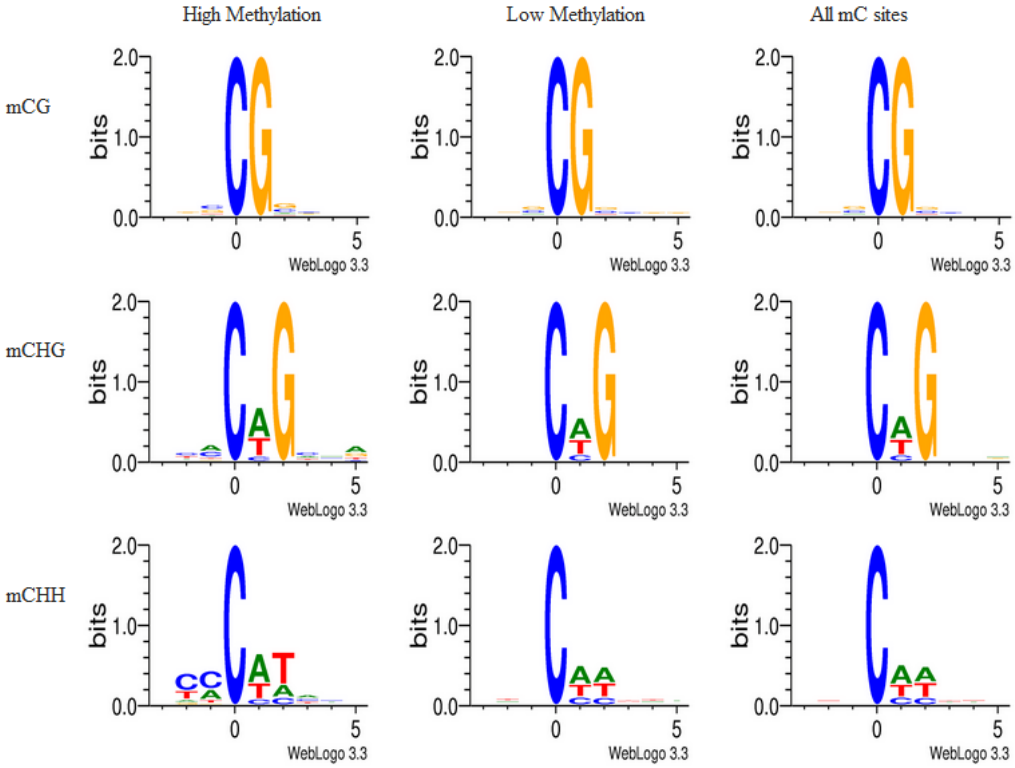

Figure 4.18 Logo Plots of Bases around Methylated Cytosines in Different Sequence Contexts.

The x-axis represents base positions, C position is defined as zero and the y-axis represents enrichment degree of bases; different colours represent different base types. According to the order from above to down, and lift to right, they are Weblog plots of CG high methylated sites, CG low methylated sites, CG methylated sites, CHG high methylated sites, CHG low methylated sites, CHG methylated sites, CHH high methylated sites, CHH low methylated sites and CHH methylated sites.

4.4 Comparative Analysis of Methylomes

Methylation profiles are compared in multiple samples, including Differentially Methylated Site (DMS) analysis, Differentially Methylated Region (DMR) analysis as well as Differentially Methylated Promoter (DMP) analysis.

4.4.1 DMS Analysis

Table 4.9 DMS analysis results

chr	start	end	strand	A_methyl	A_non_methyl	A_depth	A_methyl_ratio	B_methyl	B_non_methyl	B_depth	B_methyl_ratio	ratio_diff	pvalue	FDR
Chr01	1483	1484	+	5	17	22	0.23	9	6	15	0.6	0.37	0.037932496	0.044446308
Chr01	2120	2121	+	2	4	6	0.33	17	2	19	0.89	0.56	0.015132693	0.028004284
Chr01	3063	3064	+	2	12	14	0.14	13	6	19	0.68	0.54	0.003988915	0.012436288
Chr01	3647	3648	+	8	9	17	0.47	16	3	19	0.84	0.37	0.032773629	0.041562059
Chr01	4698	4699	+	2	18	20	0.1	6	4	10	0.6	0.5	0.007234844	0.018174015
Chr01	7887	7888	+	1	19	20	0.05	5	7	12	0.42	0.37	0.018499391	0.031155796
Chr01	11838	11839	+	0	9	9	0	14	8	22	0.64	0.64	0.001305154	0.005912956
Chr01	23421	23422	+	9	7	16	0.56	13	1	14	0.93	0.37	0.039458049	0.045520374
Chr01	28521	28522	+	0	32	32	0	3	11	14	0.21	0.21	0.02397892	0.035367747

1) chr: chromosome name

2) start: DMS position in the chromosome (0-based)

3) end: DMS position in the chromosome (1-based)

4) strand: DNA strand which DMS resides (in "+" or "-", though because CG is symmetric, "+" is always used)

5) A_methyl: number of reads that support methylation state of this site in Sample A

6) A_non_methyl: number of reads that support non-methylation state of this site in Sample A

7) A_depth: num of A_methyl and A_non_methyl, which is total number of reads that cover this site in Sample A

8) A_methyl_ratio: the ratio of A_methyl to A_depth, represents the methylation level of this site in Sample A

9) B_methyl: counterpart of A_methyl in Sample B

10) B_non_methyl: counterpart of A_non_methyl in Sample B

11) B_depth: counterpart of A_depth in Sample B

12) B_methyl_ratio: counterpart of A_methyl_ratio in Sample B

13) ratio_diff: the absolute value of the difference between A_methyl_ratio and B_methyl_ratio

14) pvalue: statistically calculated p-value

15) FDR: corrected p-value using FDR method for multiple testing

4.4.2 DMR Analysis

There exists DMRs (Differentially Methylated Regions) among different biological samples under different conditions, which means DNA methylation profiles among samples show differences within these regions. As an important epigenetic change sign, DMR may be involved in regulation of differentially expressed genes to affect biological processes. swDMR (http://122.228.158.106/swDMR/) is used to detect DMRs, which is suitable for BS-seq of two or more samples, including DMR identification, annotation and visualization. Based on methylation information of each site, the software use sliding-window approach to scan the whole genome. Its workflow mainly includes:

1) Set sliding window size (1000bp) and step length (100bp), choose regions with average methylation level difference or the fold change being greater than the cut-off value, and with number of cytosines containing within the region being greater than the cut-off value.

2) The software integrates several useful statistics methods to perform significant difference test, and the regions with significant differences are considered as potentials DMRs.

3) Perform hypothesis testing for methylation information of next window with the given step length.

4) Perform the above steps again and again to obtain potential DMRs in the whole genome.

5) Use FDR value to adjust all of the p-values.

6) Merge potential DMRs overlapping with each other into one region, and perform hypothesis testing once again. The merged DMR will be considered as the final result.

4.4.2.1 Identification and Statistics of DMRs

DMR analysis results are shown in Table 4.10. Figure 4.19 and Figure 4.20 show DMRs length distribution, while Figure 4.21 show distributions of average DNA methylation level of DMRs, respectively.

Table 4.10 DMR Results

DMR_id	chr	start	end	length	strand	DMS_num	C_num	DMS_ratio	methyl_level_in_A	methyl_level_in_B	methyl_level_diff	pvalue	FDR
DMR_Chr01_3648201_+	Chr01	3648201	3649300	1100	+	6	11	0.545454545	0.042941035	0.365422739	-0.322481704	0.001054369	0.001207732
DMR_Chr01_4255401_+	Chr01	4255401	4256500	1100	+	8	13	0.615384615	0.889575393	0.417968654	0.471606739	3.65E-06	5.76E-06
DMR_Chr01_48984701_+	Chr01	48984701	48985800	1100	+	6	11	0.545454545	0.256772016	0.744327731	-0.487555715	7.21E-06	1.08E-05
DMR_Chr03_19789701_+	Chr03	19789701	19791100	1400	+	19	21	0.904761905	0.669007243	0	0.669007243	6.61E-11	1.74E-10

Details:

1) DMR_id: Unique identification for DMR (including chromosome name, start position as well as strand information)

2) chr: chromosome name

3) start: start position of DMR in the chromosome

4) end: end position of DMR in the chromosome

5) length: length of the DMR

6) strand: DNA strand which DMR resides (in “+” or “-”, though because CG is symmetric, “+” is always used)

7) DMS_num: number of DMSs identified from comparing Sample A and Sample B within the DMR

8) C_num: total number of cytosine in both Sample A and Sample B within the DMR

9) DMS_ratio: ratio of DMS_num to C_num

10) methyl_level_in_A: methylation level in the DMR for Sample A

11) methyl_level_in_B: methylation level in the DMR for Sample B

12) methyl_level_diff: difference of methylation level between Sample A and Sample B in the DMR

13) pvalue: statistically calculated p-value

14) FDR: corrected p-value using FDR method for multiple testing

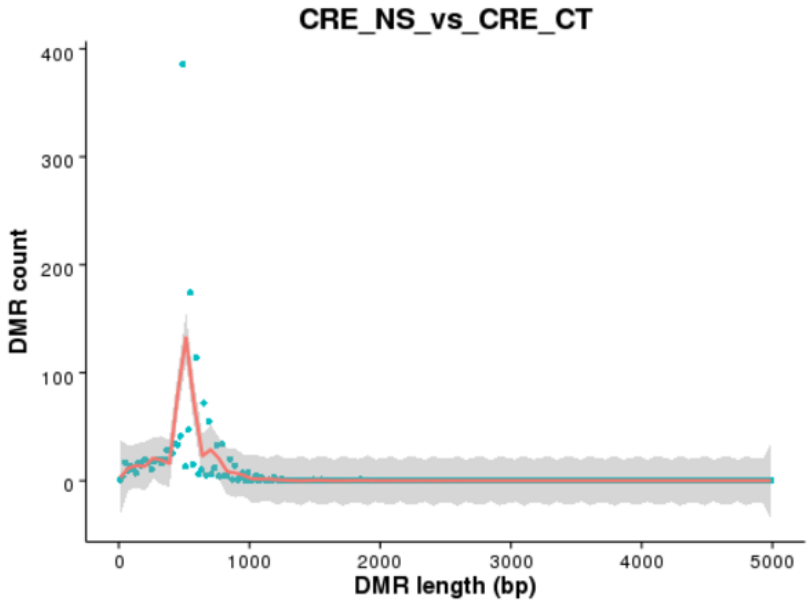

Figure 4.19 Distribution of DMR Length

The x-axis represents DMR Length, and y-axis represents DMR counts with corresponding length

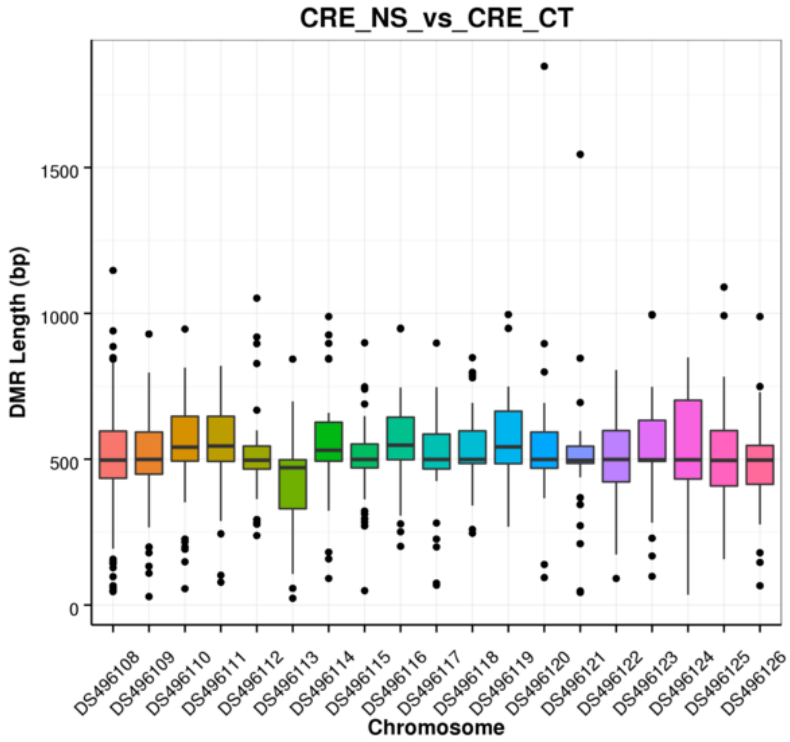

Figure 4.20 Distribution of DMR Length in Chromosomes

The x-axis represents chromosome ID, and y-axis represents distribution of DMR length in the corresponding chromosome.

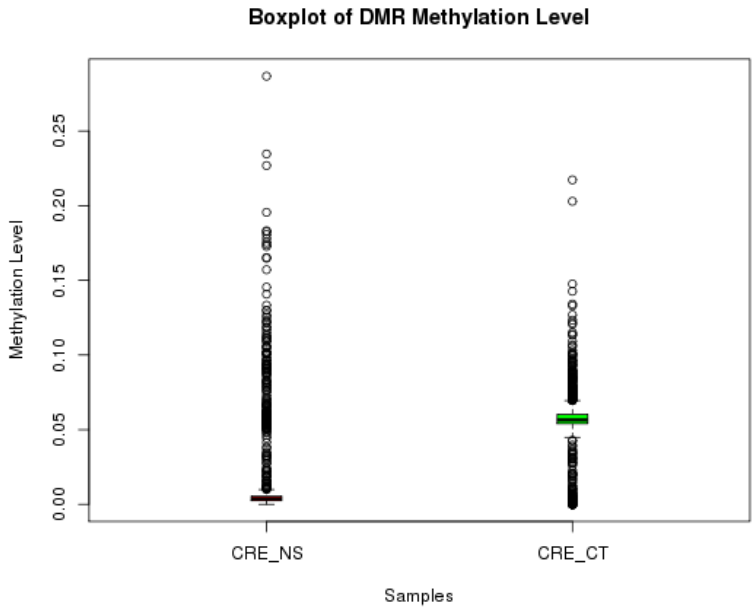

Figure 4.21 Distribution of DMR Methylation Level in different samples

The x-axis represents samples, and y-axis represents DNA methylation level.

4.4.2.2 Annotation of DMRs

DMRs are the regions show methylation level difference in different samples, which to some extent reflects transcriptional regulation differences between samples. Based on the important biology significance of DMRs, their structural annotation works have been done. When the location of DMRs and certain genomic functional elements are overlapped with each other, the related genes would be selected and defined as DMR related genes. The results are as follows:

Table 4.11 Annotation of DMR Results

DMR id	Chromosome	DMR start	DMR end	Gene id	Genomic feature	Gene name
DMR_DS496108_1408480	DS496108	1408480	1408758	CHLREDRAFT_101596	exon,intron	TRM2C
DMR_DS496119_539861	DS496119	539861	540460	CHLREDRAFT_170025	promoter	FAP141
DMR_DS496108_5289824	DS496108	5289824	5290334	CHLREDRAFT_206044	exon,intron	ASP2
DMR_DS496133_1264863	DS496133	1264863	1265508	CHLREDRAFT_49544	exon,intron	PRFA2
DMR_DS496108_2187686	DS496108	2187686	2188087	CHLREDRAFT_171688	exon,intron	SSA14
DMR_DS496129_1001493	DS496129	1001493	1002137	CHLREDRAFT_130038	exon,intron	MMP3

Details:

1) DMR id: DMR ID

1) Chromosome: chromosome ID

2) DMR start: DMR start position

3) DMR end: DMR end position

4) Gene id: DMR related gene ID

5) Genomic feature: the type of gene structure in a DMR

4.4.2.3 GO Enrichment Analysis of DMRs Related Genes

Gene Ontology (GO, http://www.geneontology.org/) is an international standard classification system for gene function. Implementing Go enrichment analysis to DMRs related genes contributes to dig out the biological processes of biological problems. GO enrichment analysis is performed by GOseq (Young et al, 2010), which is based on Wallenius non-central hyper-geometric distribution. Its characteristics are: the probability of drawing an individual from a certain category is different from that of drawing it from outsides of the category, and this difference is obtained from estimating of the preference of gene length. Thus, this method could estimate GO Term enrichment probabilities of DMRs related genes. GO enrichment analysis results of DMR related genes are as follows:

Table 4.12 GO Enrichment Analysis of DMRs Related Genes

GO accession	Description	Term type	Over represented p-value	Corrected p-value	DMR genes item	DMR genes list
GO:0009539	photosystem II reaction center	cellular_component	2.1735e-09	9.8658e-06	9	651
GO:0019843	rRNA binding	molecular_function	5.9922e-09	9.8658e-06	18	651
GO:0009536	plastid	cellular_component	6.6234e-09	9.8658e-06	109	651
GO:0009507	chloroplast	cellular_component	7.8832e-09	9.8658e-06	107	651
GO:0009521	photosystem	cellular_component	9.1333e-09	9.8658e-06	21	651
GO:0009523	photosystem II	cellular_component	4.4228e-08	3.9812e-05	17	651

Details:

1) GO_accession: the unique entry number of Gene Ontology

2) Description: detail description of Gene Ontology

3) Term type: GO types, including cellular-component, biological-process, and molecular-function

4) Over represented p-value: P-value in hypergenometric test

5) Corrected p-value: corrected P-value, if it is less than 0.05, then it is significant enrichment

6) DMR gene item: The number of genes related to this GO term

7) DMR gene list: The number of DMRs related genes annotated by GO

GO enrichment bar chart of DMRs related genes can reflect the distribution of DMRs related genes counts in enriched GO terms of biological process, cellular component and molecular function.

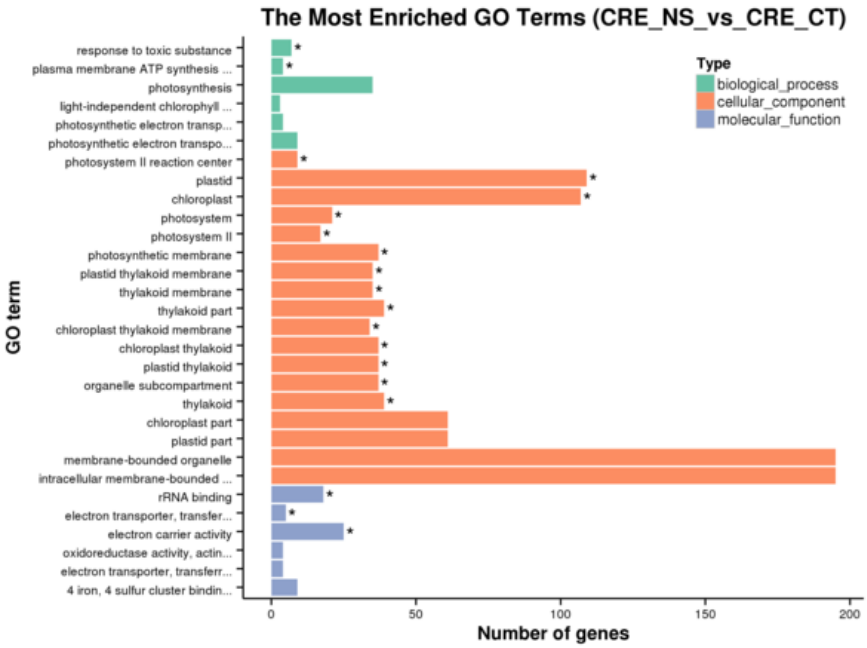

Figure 4.22 GO Enrichment Bar Plot

The y axis represents GO terms enriched, and the x axis represents the number of genes related to DMRs in terms. Different colours are used to distinguish biological process, cellular component and molecular function.

Directed Acyclic Graph (DAG) is a way to show the results of GO enrichment of DMRs related genes. The branches represent the containment relationships, and the range of functions gets smaller and smaller from top to bottom. Generally, the top ten GO enrichment results are selected as the master nodes in directed acyclic graph, showing the associated GO terms together via the containment relationship, and the degree of colours represent the extent of enrichment. DAG figures of biological process, molecular function and cellular component are shown below, respectively.

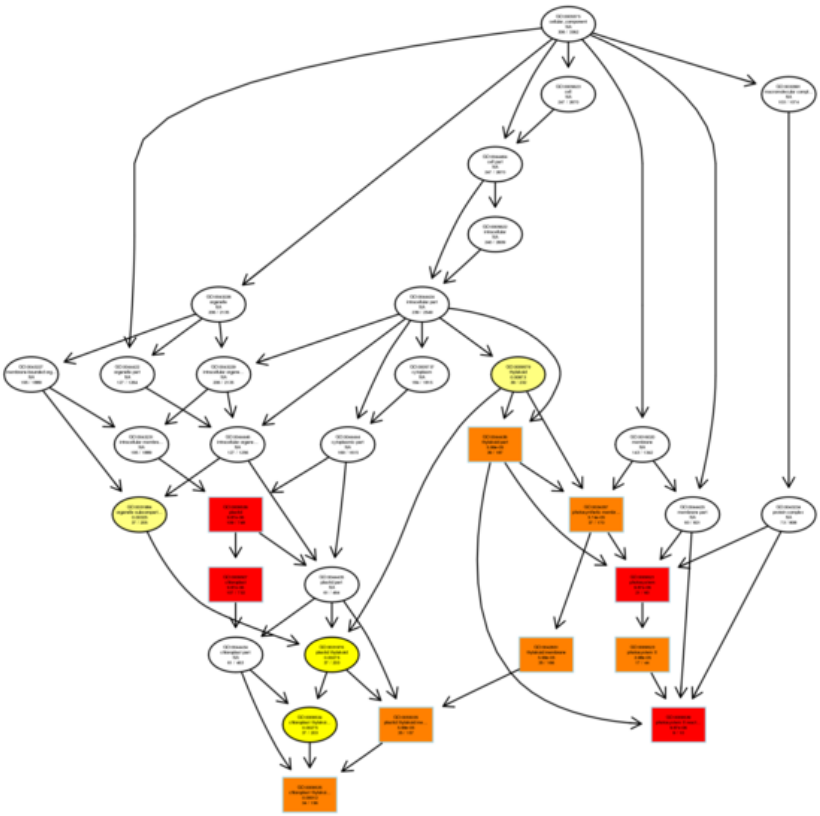

Figure 4.23 DAG Figures of GO Enrichment of DMRs Related Genes

Each node represents a GO term, boxes represent the top10 enrichment GO terms, and the degree of colours represent extent of enrichment. The darker the colour is, the higher is the enrichment level of the term. The term name and p-value of each term are presented on the node.

4.4.2.4 KEGG Enrichment Analysis of DMRs Related Genes

The interactions of multiple genes may be involved in certain biological functions. Pathway significant enrichment could determine the most important biological metabolic pathways and signal transduction pathways that DMR related genes are involved in. KEGG (Kyoto Encyclopedia of Genes and Genomes) is the main public database of the related pathways (Kanehisa, 2008). Pathway enrichment analysis identifies significantly enriched metabolic pathways using super geometry inspection compared with the whole genome background. We performed KEGG enrichment analysis of genes related to DMR. The results of KEGG metabolic pathway enrichment are as follows:

Table 4.13 KEGG Enrichment Analysis of DMRs Related Genes

Term	ID	DMR genes number	Background number	P-value	Corrected p-value
Photosynthesis	KEGG PATHWAY	cre00195	28	66	7.0764e-11
Ribosome biogenesis in eukaryotes	KEGG PATHWAY	cre03008	10	55	2.1398e-02

Details:

1) Term: the description of KEGG pathways

2) ID: KEGG ID

3) DMR genes number: Number of genes related to DMR with pathway annotation

4) Background number: Number of all reference genes with pathway annotation

5) P-value: P-value in hypergenometric test

6) Corrected P-value: Corrected P-value, pathway with Corrected P-value <0.05 are significantly enriched

Scatter diagram is a graphical display way of KEGG enrichment analysis results. In this plot, enrichment degree of KEGG can be measured through Rich factor, Qvalue and genes counts enriched to this pathway. Rich factor is the ratio of DMRs related genes counts to this pathway in the annotated genes counts. The more the Rich factor is, the higher the degree of enrichment is. Qvalue is the adjusted p-value after multiple hypothesis testing, and its range is from 0 to 1. The more the qvalue is close to zero, the higher the significant level of the enrichment is. The top 20 most significant enriched pathways are chosen in KEGG scatter plot, and if the enriched pathways counts is less than 20, then put all of them into the plot. KEGG enrichment scatter diagram is shown below:

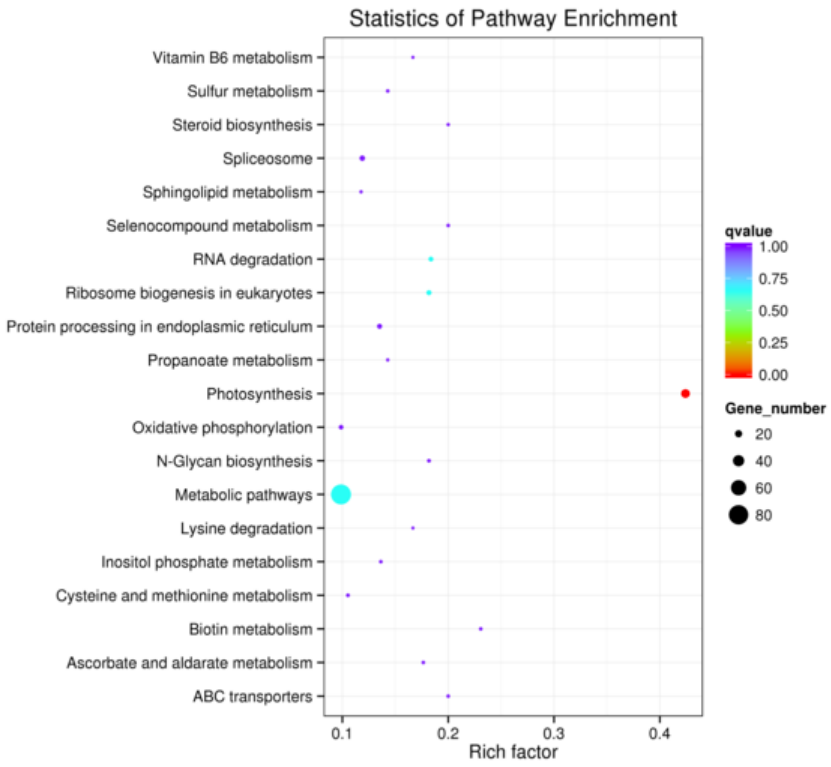

Figure 4.24 KEGG Enrichment Scattered Plot of DMRs Related Genes

The x-axis represents Rich factor, and the y-axis represents pathway name. The size of points stand for DMRs related genes counts and the colours stand for different Qvalues range.

4.4.2.5 InterPro Protein Domain Enrichment Analysis of DMRs Related Genes

InterPro (http://www.ebi.ac.uk/interpro/) is a resource that provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites. It integrated data from multiple databases, such as Pfam and PANTHER, aiming to provide the users with information as complete as possible. InterPro protein domain enrichment analysis may reveal the biological functional patterns of DMRs related genes, being helpful to understand how DNA methylation influence biological pathways. In the enrichment analyais, Fisher's exact test as well as FDR multiple-testing-correction method are used to identify the statistically significantly enriched InterPro domain signatures. After FDR correction, p-values less than 0.05 are considered as significant.

Table 4.14 InterPro protein domain enrichment analysis results of DMR related genes

term	these_with_this_term	these_without_this_term	those_with_this_term	those_without_this_term	odds_ratio	pvalue	FDR
U box domain	30	14195	35	8366	0.51	0.006636409	0.016460386
Riboflavin synthase-like beta-barrel	13	14212	17	8384	0.45	0.036014287	0.04050865
Zinc finger, NHR/GATA-type	13	14212	17	8384	0.45	0.036014287	0.04050865
Ribosomal protein S4/S9	1	14224	5	8396	0.12	0.029223476	0.034579459

Details:

1) term: InterPro domain signature

2) these with this term: number of genes annotated with this term in the selected gene set

3) these without this term: number of genes not annotated with this term in the selected gene set

4) those with this term: number of genes annotated with this term in the background (unselected) gene set

5) those without this term: number of genes not annotated with this term in the background (unselected) gene set

6) odds ratio: (these with this term / these without this term) / (those with this term / those without this term)

7) pvalue: p-values produced by Fisher's exact test

8) FDR: corrected p-values using FDR method

4.4.3 DMP Analysis

A Differentially Methylated Promoter (DMP) is a gene promoter region showing significantly different methylation levels between two samples. Methlylation of cytosines in a promoter region may suppress the expression of corresponding gene, which is one important way to regulate gene expression. DMP analysis is performed to all classes of cytosine methylation, including CG, CHG, CHH and C. Fisher's exact test and FDR correction method are used to conduct the analysis. Promoter regions showing methylation level difference no less than 0.2 with corrected p-values less than 0.05 are considered as significantly DMPs.

The main procedure of identifying DMPs is as follows:

1) A 2kb upstream region of TSS is defined as a promter region, in which all cytosines are identified

2) Cytosine methylation levels are calculated for each sample in the same promoter region.

3) Using statistical test as well as multple testing correction to identify DMPs as previous explained.

4.4.3.1 Identification and Statistics of DMPs

DMP results are listed as follows:

Table 4.15 DMP results

DMP_id	Gene_name	A_methyl	A_non_methyl	A_depth	A_methyl_ratio	B_methyl	B_non_methyl	B_depth	B_methyl_ratio	methyl_ratio_diff	pvalue	FDR
ATMG00010.1|ATMG00010|promoter	ORF153A	51	710	761	0.07	664	1476	2140	0.31	0.24	2.36E-48	1.65E-47
ATMG00110.1|ATMG00110|promoter	CCB206	66	305	371	0.18	271	388	659	0.41	0.23	5.26E-15	1.00E-14
ATMG00120.1|ATMG00120|promoter	ORF143	145	522	667	0.22	3	491	494	0.01	0.21	6.08E-34	3.19E-33
ATMG00130.1|ATMG00130|promoter	ORF121A	40	417	457	0.09	196	481	677	0.29	0.2	1.12E-17	2.35E-17
ATMG00140.1|ATMG00140|promoter	ORF167	28	58	86	0.33	3	107	110	0.03	0.3	8.67E-09	1.30E-08

Details:

1) DMP_id: Unique identification for DMP (including transcript_id and gene_id)

2) Gene name: corresponding gene name

3) A_methyl: number of reads that support methylation states of all cytosine sites in Sample A

4) A_non_methyl: number of reads that support non-methylation states of all cytosine sites in Sample A

5) A_depth: num of A_methyl and A_non_methyl, which is total number of reads that cover all cytosine sites in Sample A

6) A_methyl_ratio: the ratio of A_methyl to A_depth, represents the methylation level of the promoter region in Sample A

7) B_methyl: counterpart of A_methyl in Sample B

8) B_non_methyl: counterpart of A_non_methyl in Sample B

9) B_depth: counterpart of A_depth in Sample B

10) B_methyl_ratio: counterpart of A_methyl_ratio in Sample B

11) ratio_diff: the absolute value of the difference between A_methyl_ratio and B_methyl_ratio

12) pvalue: statistically calculated p-value

13) FDR: corrected p-value using FDR method for multiple testing

4.4.3.2 GO Enrichment Analysis of DMP Genes

Gene Ontology (GO, http://www.geneontology.org/) is an international standard classification system for gene function. Implementing Go enrichment analysis to DMP genes contributes to dig out the biological processes of biological problems. GO enrichment analysis is performed by GOseq (Young et al, 2010), which is based on Wallenius non-central hyper-geometric distribution. Its characteristics are: the probability of drawing an individual from a certain category is different from that of

drawing it from outsides of the category, and this difference is obtained from estimating of the preference of gene length. Thus, this method could estimate GO Term enrichment probabilities of DMP genes. GO enrichment analysis results of DMP genes are as follows:

Table 4.16 GO Enrichment Analysis of DMP Genes

GO accession	Description	Term type	Over represented p-Value	Corrected p-Value	DMP genes item	DMP genes list
GO:0051536	iron-sulfur cluster binding	molecular_function	2.521e-09	5.5475e-06	61	4088
GO:0051540	metal cluster binding	molecular_function	2.521e-09	5.5475e-06	61	4088
GO:0006184	obsolete GTP catabolic process	biological_process	4.7246e-09	6.931e-06	33	4088
GO:0051537	2 iron, 2 sulfur cluster binding	molecular_function	6.6976e-09	7.369e-06	26	4088

Details:

1) GO_accession: the unique entry number of Gene Ontology

2) Description: detail description of Gene Ontology

3) Term type: GO types, including cellular-component, biological-process, and molecular-function

4) Over represented p-value: P-value in hypergenometric test

5) Corrected p-value: corrected P-value, if it is less than 0.05, then it is significant enrichment

6) DMP gene item: The number of DMP genes annotated to this GO term

7) DMP gene list: The number of DMP genes annotated by GO

GO enrichment bar chart of DMP genes can reflect the distribution of DMP genes counts in enriched GO terms of biological process, cellular component and molecular function.

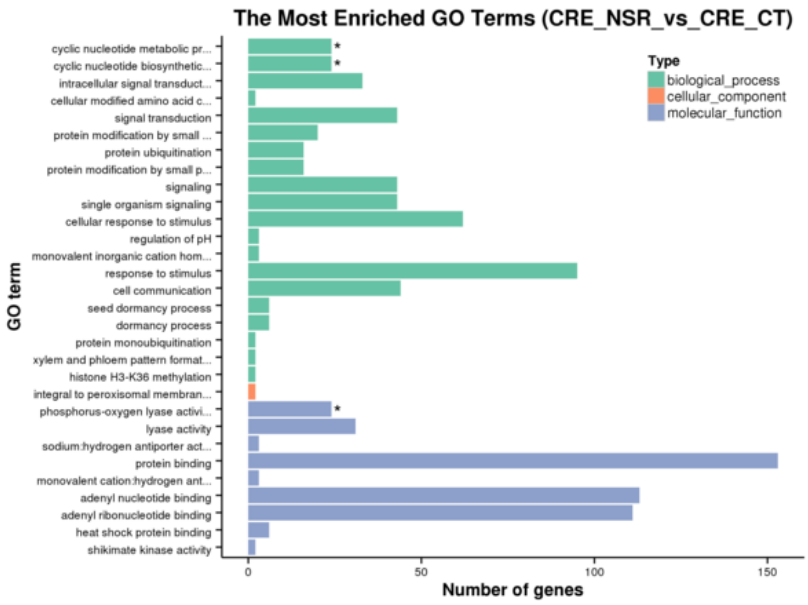

Figure 4.25 GO Enrichment Bar Plot

The y axis represents GO terms enriched, and the x axis represents the number of DMP genes in terms. Different colours are used to distinguish biological process, cellular component and molecular function.

Directed Acyclic Graph (DAG) is a way to show the results of GO enrichment of DMP genes. The branches represent the containment relationships, and the range of functions gets smaller and smaller from top to bottom. Generally, the top ten GO enrichment results are selected as the master nodes in directed acyclic graph, showing the associated GO terms together via the containment relationship, and the degree of colours represent the extent of enrichment. DAG figures of biological process, molecular function and cellular component are shown below, respectively.

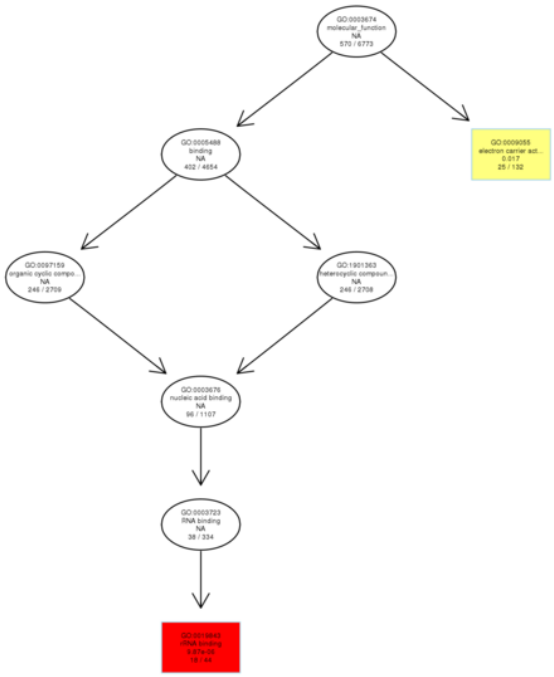

Figure 4.26 DAG Figures of GO Enrichment of DMP Genes

Each node represents a GO term, boxes represent the top10 enrichment GO terms, and the degree of colours represent extent of enrichment. The darker the colour is, the higher is the enrichment level of the term. The term name and p-value of each term are presented on the node.

4.4.3.3 KEGG Enrichment Analysis of DMP Genes

The interactions of multiple genes may be involved in certain biological functions. Pathway significant enrichment could determine the most important biological metabolic pathways and signal transduction pathways that DMP genes are involved in. KEGG (Kyoto Encyclopedia of Genes and Genomes) is the main public database of the related pathways (Kanehisa, 2008). Pathway enrichment analysis identifies significantly enriched metabolic pathways using super geometry inspection compared

with the whole genome background. We performed KEGG enrichment analysis of DMP genes. The results of KEGG metabolic pathway enrichment are as follows:

Table 4.17 KEGG Enrichment Analysis of DMP Genes

#Term	Database	ID	Sample number	Background number	P-Value	Corrected P-Value
Photosynthesis	KEGG PATHWAY	tcc00195	60	70	8.56000183557e-10	9.92960212927e-08
Carbon metabolism	KEGG PATHWAY	tcc01200	107	216	2.39429119521e-06	0.000138868889322
Glyoxylate and dicarboxylate metabolism	KEGG PATHWAY	tcc00630	36	56	0.000157042843536	0.00607232328339
Carbon fixation in photosynthetic organisms	KEGG PATHWAY	tcc00710	38	64	0.000340979549268	0.0091725280072

Details:

1) Term: the description of KEGG pathways

2) ID: KEGG ID

3) DMR genes number: Number of DMP genes with pathway annotation

4) Background number: Number of all reference genes with pathway annotation

5) P-value: P-value in hypergenometric test

6) Corrected P-value: Corrected P-value, pathway with Corrected P-value <0.05 are significantly enriched

Scatter diagram is a graphical display way of KEGG enrichment analysis results. In this plot, enrichment degree of KEGG can be measured through Rich factor, Qvalue and genes counts enriched to this pathway. Rich factor is the ratio of DMP genes counts to this pathway in the annotated genes counts. The more the Rich factor is, the higher the degree of enrichment is. Qvalue is the adjusted p-value after multiple hypothesis testing, and its range is from 0 to 1. The more the qvalue is close to zero, the higher the significant level of the enrichment is. The top 20 most significant enriched pathways are chosen in KEGG scatter plot, and if the enriched pathways counts is less than 20, then put all of them into the plot. KEGG enrichment scatter diagram is shown below:

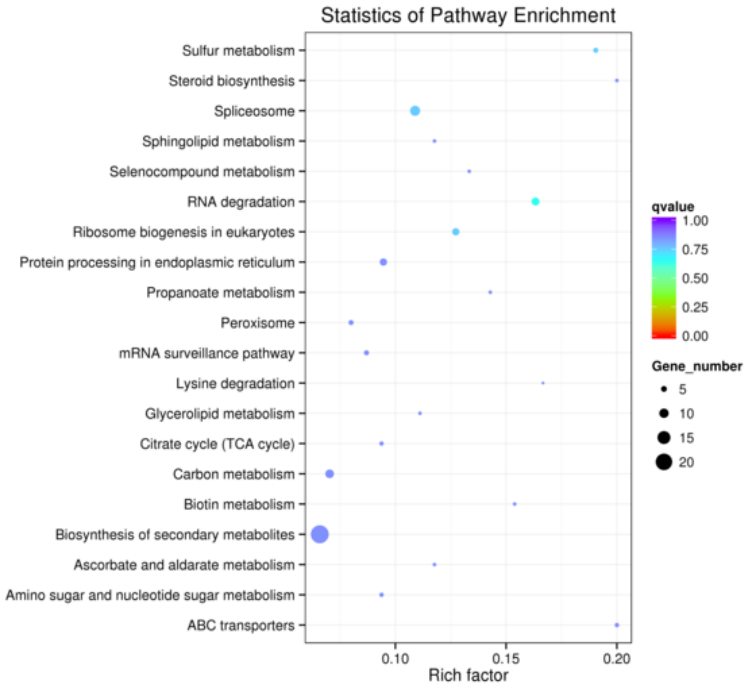

Figure 4.27 KEGG Enrichment Scattered Plot of DMP Genes

The x-axis represents Rich factor, and the y-axis represents pathway name. The size of points stand for DMP genes counts and the colours stand for different Qvalues range.

4.4.3.4 InterPro Protein Domain Enrichment Analysis of DMRs Related Genes

InterPro (http://www.ebi.ac.uk/interpro/) is a resource that provides functional analysis of protein sequences by classifying them into families and predicting the presence of domains and important sites. It integrated data from multiple databases, such as Pfam and PANTHER, aiming to provide the users with information as complete as possible. InterPro protein domain enrichment analysis may reveal the biological functional patterns of DMP genes, being helpful to understand how DNA methylation influence biological pathways. In the enrichment analyais, Fisher's exact test as well as FDR multiple-testing-correction method are used to identify the statistically significantly enriched InterPro domain signatures. After FDR correction, p-values less than 0.05 are considered as significant.

Table 4.18 InterPro protein domain enrichment analysis results of DMP genes

term	these_with_this_term	these_without_this_term	those_with_this_term	those_without_this_term	odds_ratio	pvalue	FDR
U box domain	30	14195	35	8366	0.51	0.006636409	0.016460386
Riboflavin synthase-like beta-barrel	13	14212	17	8384	0.45	0.036014287	0.04050865
Zinc finger, NHR/GATA-type	13	14212	17	8384	0.45	0.036014287	0.04050865
Ribosomal protein S4/S9	1	14224	5	8396	0.12	0.029223476	0.034579459

Details:

1) term: InterPro domain signature

2) these with this term: number of genes annotated with this term in the selected gene set

3) these without this term: number of genes not annotated with this term in the selected gene set

4) those with this term: number of genes annotated with this term in the background (unselected) gene set

5) those without this term: number of genes not annotated with this term in the background (unselected) gene set

6) odds ratio: (these with this term / these without this term) / (those with this term / those without this term)

7) pvalue: p-values produced by Fisher's exact test

8) FDR: corrected p-values using FDR method

4.4.4 PCA Analysis

With more than 9 samples, Principal Component Analysis (PCA) using methylation levels of functional gene regions from different samples can be performed.

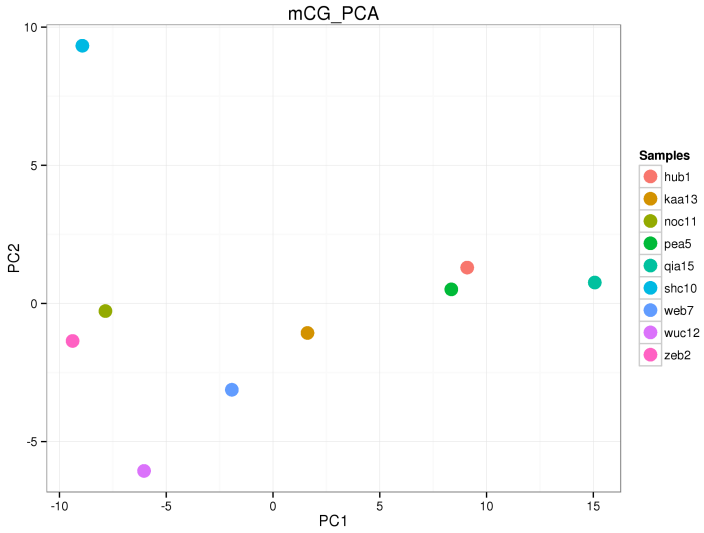

Figure 4.28 PCA analysis, samples are plotted to the two dimentional space defined by the first two componets

4.4.5 Clustering

Clustering analysis contributes to analyze the relationship among all the samples. Moreover, it is helpful in identifying differential trends among samples and further to study the biological significance of these differences. In the process of methylation level analysis, the largest regional differences in methylation level between two samples are selected for clustering analysis (Smallwood, 2014). We adopt hierarchical clustering methods to analyze different samples, and the results are shown in Figure 4.29. Clustering analyses are preformed based on the methylation level of different genomic functional regions, and the results are shown in the Figure 4.30:

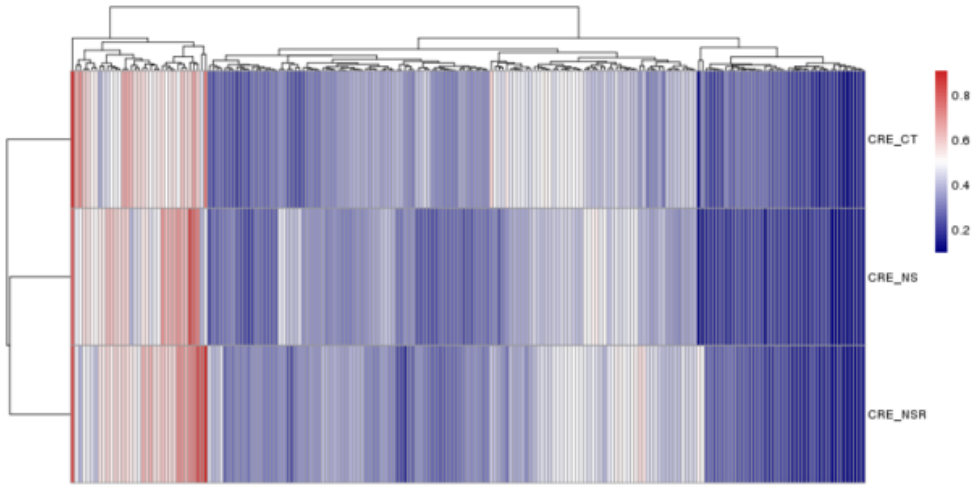

Figure 4.29 Clustering results of different samples

Each line stands for the different regions, each column stands for different samples. Different colours stand for different methylation level values

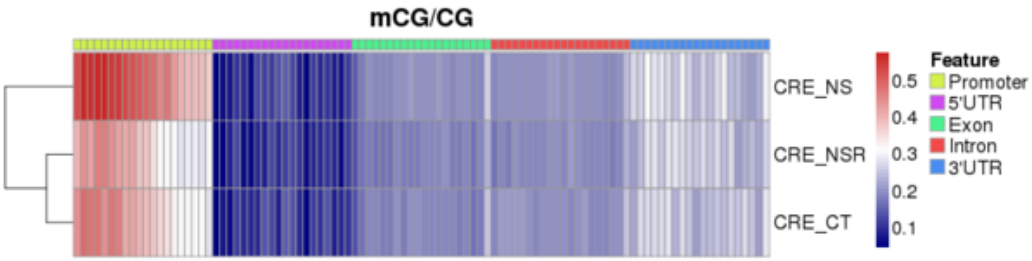

Figure 4.30 Clustering results of different contexts of various samples

5 References

Bird A. (2002) DNA methylation patterns and epigenetic memory. Genes Dev, 16: 6-21.

Langmead B, Salzberg SL. (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods, 9(4): 357-9. (Bowtie2)

Casey AG, Michael JZ, Hongcang Gu, et al.(2013) Transcriptional and Epigenetic Dynamics during Specification of Human Embryonic Stem Cells. Cell, 153:1149-1163.

Ehsan Habibi, Arie BB, Julia Arand, et al. (2013) Whole-Genome Bisulfite Sequencing of Two Distinct Interconvertible DNA Methylomes of Mouse Embryonic Stem Cells. Cell Stem Cell, 13:360-369.

Lister R, Pelizzola M, Dowen RH, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature, 462(7271): 315-22

Lister R, Mukamel EA, Nery JR, Urich M, Puddifoot CA, Johnson ND, Lucero J, Huang Y, Dwork AJ, Schultz MD, et al. (2013) Global epigenomic reconfiguration during mammalian brain development. Science 341:1237905.

Mao X, Cai T, Olyarchuk JG, et al. (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics, 21(19): 3787-93.(KOBAS)

Goldberg AD, Allis CD, Bernstein E, et al. (2007) Epigenetics: a landscape takes shape. Cell, 128(4): 635-8.

Jackson JP, Lindroth AM, Cao X, et al. (2002) Control of CpNpG DNA methylation by the KPYPTONITE histone H3 mehyltransferase. Nature, 416:556-60.

Jones PA, Takai D. (2001) The role of DNA methylation in mammalian epigenetics. Science, 293(5532): 1068-70. Kanehisa M, Araki M, Goto S, et al. (2008). KEGG for linking genomes to life and the environment. Nucleic Acids Res, 36(Database issue): D480-4.(KEGG)

Krueger F, Andrews SR. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics, 27(11): 1571-2. (Bismark)

Reik W, Santos F, Dean W. (2003) Mammalian epigenomics: reprogramming the genome for development and therapy. Theriogenology, 59: 21-32.

Vertino PM, Gao J, Baylin SB, et al. (1996) De novo methylation of CpG island sequences in human fibroblasts overexpressing DNA (cytosine-5-)-methyltransferase. Molecular and Cellular Biology, 16(8): 4555-65.

Young MD, Wakefield MJ, Smyth GK, et al. (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol, 11(2): R14.(GOseq)

Smallwood SA, Lee HJ, Angermueller C, et al. (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods, 11(8): 817-20.

swDMR: a sliding window approach to identify differentially methylated regions based on whole-genome bisulfite sequencing

6 Appendix

6.1 software list

Softwares

Analysis	Software (version)	Parameters	Remarks
mapping	Bismark (0.12.5)	--score_min L, 0, -0.2	bowtie2 (2.2.5) as aligner engine
	BSMAP (2.88)	-v 0.04	
DMR analysis	swDMR (1.07)	pvalue 0.05, fdr 0.05, coverage 5, fold 2.0, diff 0.1	Fisher's exact test with FDR multiple test correction
	Bioconductor (2.13) (BSSeq Package)	lowQ 0.025, highQ 0.975, mdiff = 0.1	Only to samples with replicates
DMS analysis	in house script	corrected p-value<0.05, diff_level≥0.2	Fisher's exact test with FDR multiple test correction
DMR analysis	in house script	corrected p-value<0.05, diff_level≥0.2	Student's t test with FDR multiple test correction
DMP analysis	in house script	corrected p-value<0.0, diff_level≥0.2	Fisher's exact test with FDR multiple test correction
GO enrichment	GOseq, topGO, Bioconductor (2.13)	Corrected P-value < 0.05	
KEGG enrichment	KOBAS (2.0)	Corrected P-value < 0.05	
Interpro enrichment	in house script	Fisher's exact test with FDR multiple test correction	