Humanizing Genomics
macrogen

Homo sapiens

Whole Genome Bisulfite Sequencing

Report

February 2016

Project Information

Client Name	Client Name
Company / Institution	Macrogen
Order Number	Order ID
Species	*Homo sapiens*
Reference	*UCSC hg19*
Type of Read	Paired-end
Read Length	151
Number of Samples	4
Library Kit	TruSeq DNA Methylation Kit
Library Protocol	TruSeq DNA Methylation Library Prep Guide, Part # 15066014 Rev. A
Reagent	HiSeq X Ten Reagent Kit v2.5
Sequencing Protocol	HiSeq X System Guide Part # 15050091 v01 RTA v2
Type of Sequencer	HiSeq X
Sequencing Control Software	HiSeq X Control Software v3.1
Comment	

Table of Contents

1. Experimental Methods and Workflow

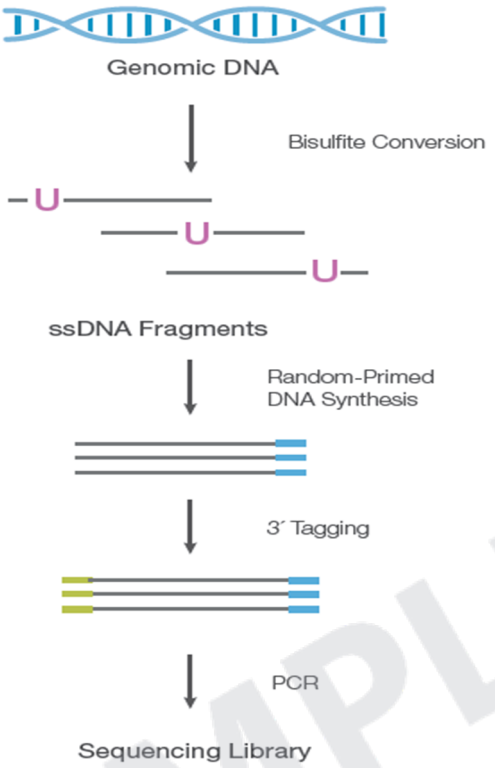


Figure 1. Workflow for TruSeq DNA Methylation Kit

Reference : http://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_truseq/truseq-dna-methylation/truseq-dna-methylation-library-prep-guide-15066014-a.pdf

1) Isolate the genomic DNA from sample of interest.

2) The sample is treated with bisulfite to convert unmethylated cytosines to uracils, while retaining those which are methylated before the TruSeq library prep procedure.

3) Bisulfite treated ssDNA fragments are randomly primed using a polymerase able to read uracil nucleotides to synthesize DNA strands containing a specific sequence tag.

4) The 3' ends of the newly synthesized DNA strands are then selectively tagged with a second specific sequence tag using a patented procedure. This process generates di-tagged DNA molecules with known sequence tags at their 5' and 3' ends.

5) The di-tagged DNA is enriched in PCR, resulting in double-stranded DNA(dsDNA) with the appropriate Illumina sequences for use on the HiSeq platform.

2. Analysis Methods and Workflow

Figure 2. Whole Genome Bisulfite Sequencing Analysis Workflow

1) After sequencing, the raw sequence reads are filtered based on quality. The adapter sequences are also trimmed off the raw sequence reads.

2) The trimmed reads are mapped to reference genome with BSMAP which is based on the SOAP(Short Oligo Alignment Program).

3) The only uniquely mapped reads are selected to sort and index using SAMtools(v 0.1.19). Afterwards, PCR duplicates are removed with Picard mark Duplicates(v1.118).

4) The methylation ratio of every single cytosine location is extracted from the mapping results using 'methylatio.py' script in BSMAP. The results of the coverage profiles were calculated as # of C /effective CT counts for each cytosine in CpG, CHH and CHG.

5) Each cytosine locus in CpG, CHH and CHG is annotated using table browser function of UCSC genome browser. Annotation includes functional location of each gene (promoter regions which are defined as -2kb upstream of the transcription start site, exons and introns etc.), transcripts ID, gene ID, strand and CpG island.

6) (Optional) As a further study, differentially methylated CpGs/CHHs/CHGs between groups with different conditions are filtered out through statistical hypothesis testing.

3. Summary of Data Production

3.1. Raw data Statistics

The total number of bases, reads, GC(%), Q20(%), and Q30(%) are calculated for the 4 samples. For example, in Control1, 1,318,427,422 reads are produced, and total read bases are 199 Gb. The GC content(%) is 29.77% and Q30 is 78.41%.

Table 1. Raw data Stats

Sample ID	Total read bases(bp)	Total reads	GC(%)	Q20(%)	Q30(%)
Control1	199,082,540,722	1,318,427,422	29.77	88.05	78.41
Control2	204,585,109,374	1,354,868,274	29.88	87.06	76.8
Test1	194,509,278,014	1,288,140,914	29.48	90.06	81.59
Test2	190,856,545,130	1,263,950,630	29.47	90.34	81.92

- Sample ID : Sample name.
- Total read bases : Total number of bases sequenced. (= Total reads x Read length)
- Total reads : Total number of reads.
- GC(%) : GC content.
- Q20(%) : Ratio of reads that have phred quality score over 20.
- Q30(%) : Ratio of reads that have phred quality score over 30.

3.2. Average Base Quality at Each Cycle

The quality of produced data is determined by the phred quality score at each cycle. Box plot containing the average quality at each cycle is created with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). The x-axis shows number of cycles and y-axis shows phred quality score. Phred quality score 20 means 99% accuracy and reads with score over 20 are accepted as good quality.

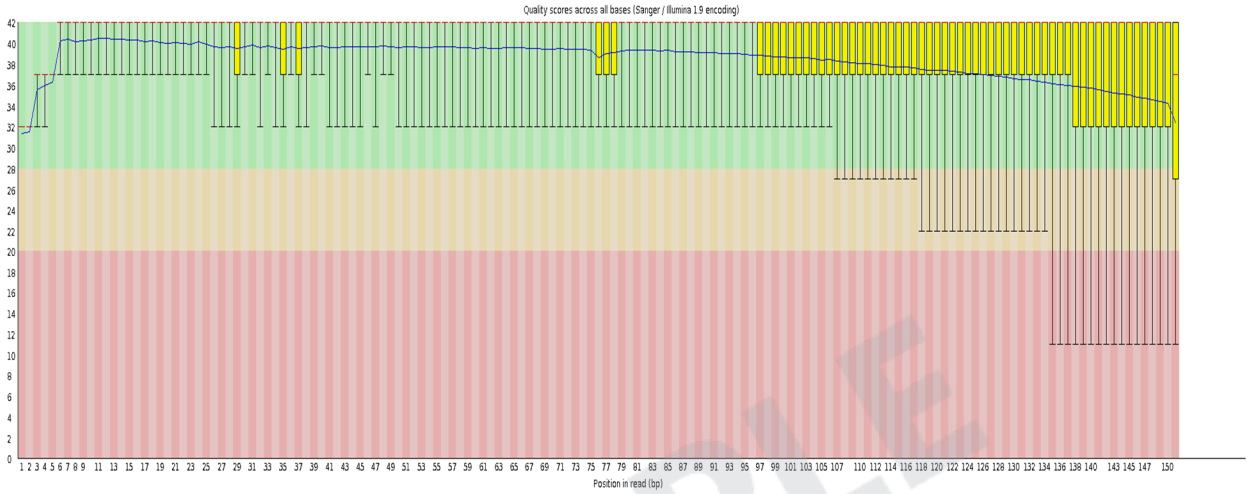


Figure 3. Read quality at each cycle of Control1 (read1)

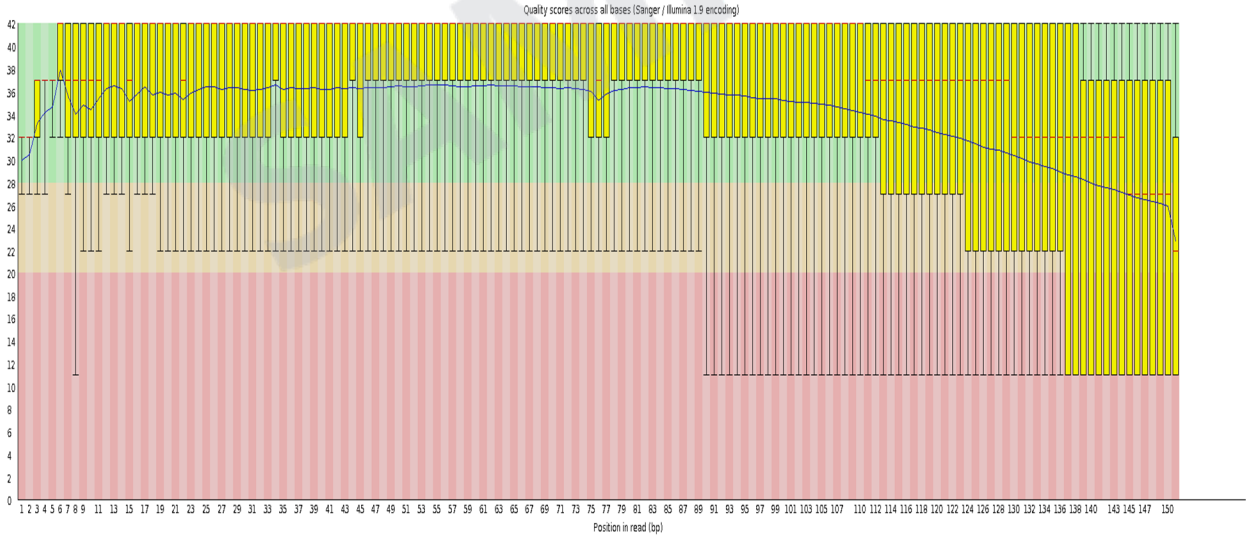


Figure 4. Read quality at each cycle of Control1 (read2)

- Yellow box : interquartile range (25-75%) of phred score at each cycle
- Red line : median phred score at each cycle
- Blue line : average phred score at each cycle
- Green background: Good quality.
- Orange background: Acceptable quality.
- Red background : Bad quality

3.3. Trimmed Data Statistics

Trimming process is done to eliminate adapter sequences and bases with low quality from each read using Trimmomatic program. The bases with low quality or N bases less than quality 3 from the ends of reads are trimmed. Also using sliding window method, bases of reads that does not qualify for a 4-base wide sliding window, and the average quality per window drops below 15 are trimmed. Afterwards, reads with length shorter than 36bp are dropped to produce cleaned data.

Table 2. Trimmed data Stats

Sample ID	Total read bases(bp)	Total reads	GC(%)	Q20(%)	Q30(%)
Control1	144,690,730,296	1,160,094,190	27.07	96.57	89.82
Control2	140,442,073,360	1,132,895,002	26.94	96.32	89.22
Test1	148,654,728,147	1,178,938,842	26.73	97.11	91.21
Test2	149,559,008,044	1,173,989,402	27.05	97.1	91.13

- Sample ID : Sample name.
- Total read bases : Total number of bases after trimming.
- Total reads : Total number of reads after trimming.
- GC(%) : GC content.
- Q20(%) : Ratio of reads that have phred quality score over 20.
- Q30(%) : Ratio of reads that have phred quality score over 30.

3.4 Average Base Quality at Each Cycle after Trimming

Figure5 shows base quality at each cycle after trimming.

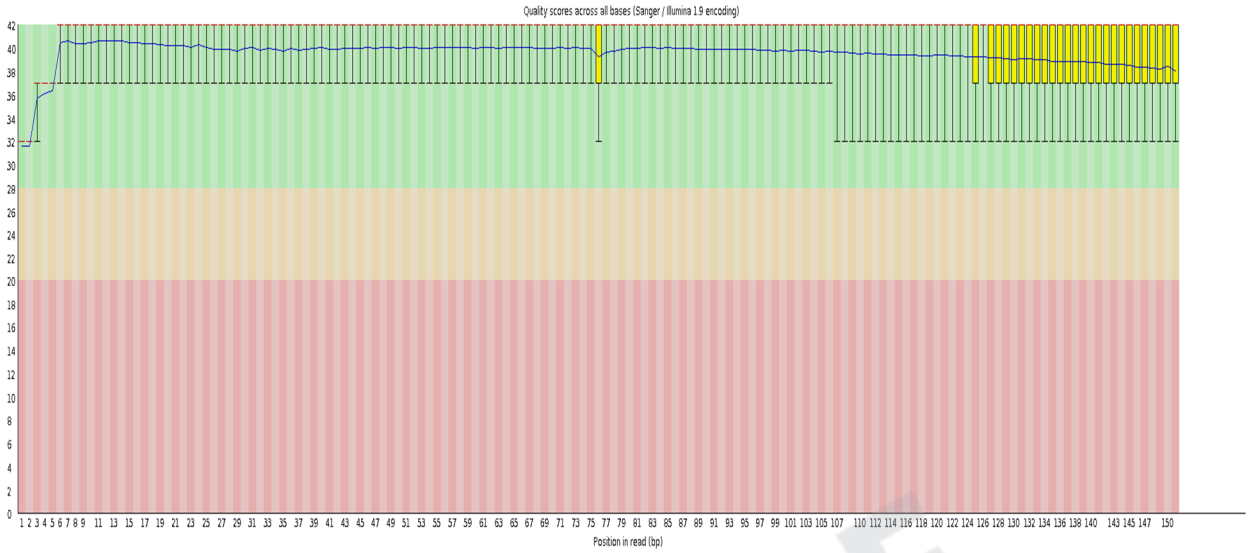


Figure 5. Read quality at each cycle of Control1 (read1) after trimming

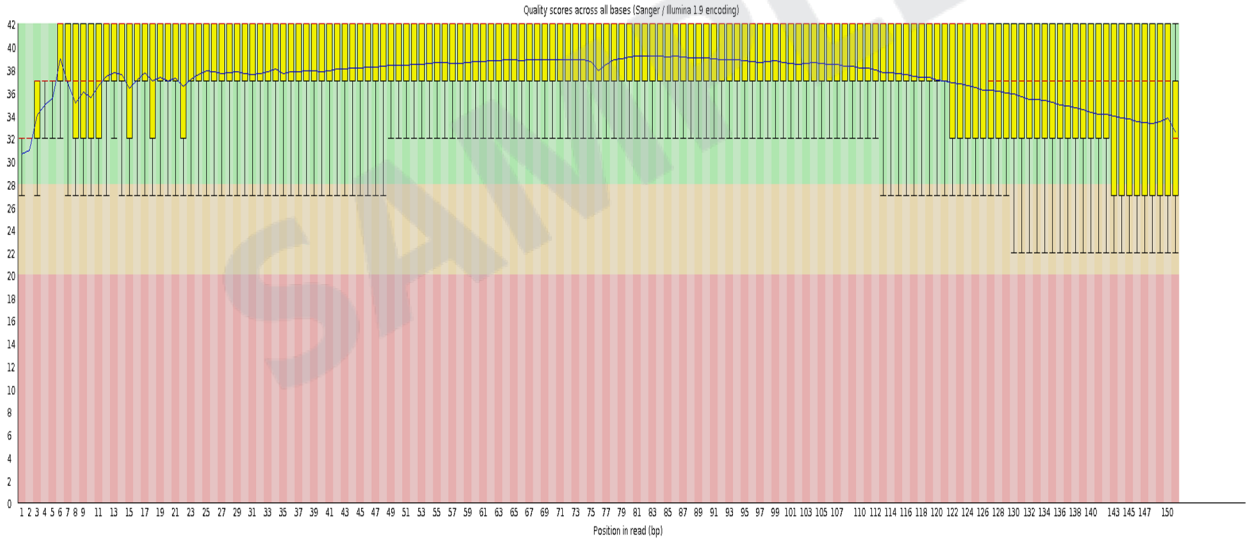


Figure 6. Read quality at each cycle of Control1 (read2) after trimming

- Yellow box : interquartile range (25-75%) of phred score at each cycle
- Red line : median phred score at each cycle
- Blue line : average phred score at each cycle
- Green background: Good quality.
- Orange background: Acceptable quality.
- Red background : Bad quality

3.5 Calculation of Bisulfite Conversion Rate

To estimate bisulfite conversion rate, unmethylated lambda phage DNA (Promega Cat# D1521) added to DNA prior to fragmentation. Typical range of spike-in level is from 0.1 - 0.5% (w/w)). Bisulfite conversion rate can be calculated as the following formula [1] at CG, CHG, CHH sites within the lambda genome.

[1] Bisulfite conversion rate(%) = # unmethylated read level measurements from lambda genome / (# methylated and # unmethylated read level measurements from lamda genome) x 100

Table 3. Estimated bisulfite conversion rate by sample

Sample ID	# methylated read level measurements from lambda genome	# unmethylated read level measurements from lambda genome	Estimated bisulfite conversion rate(%)
Control1	113400	8141255	98.63%
Control2	358917	19707244	98.21%
Test1	322638	7654438	95.96%
Test2	177765	12754076	98.63%

The bisulfite conversion rates were estimated to be 96%–99% across samples using lambda phage DNA as a spike-in control.

4. Quantification of Methylation Level

4.1 Mapping to Reference Genome

The cleaned reads were aligned to the *Homo sapiens hg19* using BSMAP based on the SOAP(Short Oligo Alignment Program). Table 3 shows the statistics obtained from SOAP algorithm. You can check the number of uniquely mapped reads, non-unique mapped read, deduplicated reads and analyzed reads in methylation calling.

* reference size(bp) : 3,095,693,983

Table 4. Mapping Data Stats

Sample ID	# of trimmed read bases(bp)	Average throughput depth of reference genome(X)	# of uniquely mapped reads (% out of trimmed reads)	# of suppressed non-unique mapped reads (% out of trimmed reads)	Deduplicated reads (deduplicated by Picard tools, % out of mapped reads)	Analyzed reads in BSMAP methylation calling
Control1	144,690,730,296	46.74	896,440,894 (77.27%)	40,295,942 (3.47%)	608,091,870 (67.83%)	608,091,849
Control2	140,442,073,360	45.37	870,940,188 (76.88%)	39,061,952 (3.45%)	604,775,900 (69.44%)	604,775,881
Test1	148,654,728,147	48.02	928,365,790 (78.75%)	42,358,802 (3.59%)	607,288,484 (65.41%)	607,288,462
Test2	149,559,008,044	48.31	923,908,150 (78.7%)	41,055,486 (3.5%)	589,495,930 (63.8%)	589,495,904

- Sample ID : Sample name.
- # of trimmed read bases(bp) : Total number of bases after trimming.
- Average throughput depth of reference genome
  : Calculated by # of trimmed read bases / reference genome size(ex. human reference size : 3,095,693,983).
- # of uniquely mapped reads (% out of trimmed reads)
  : Total uniquely mapped read count , percentage of uniquely mapped reads out of trimmed reads
- # of suppressed non-unique mapped reads
  : Reads suppressed by multiple mapping , percentage of non-unique mapped reads out of trimmed reads
- Deduplicated reads(%)
  : reads after removing PCR duplicates, percentage of deduplicated reads out of uniquely mapped reads
- Analyzed reads in BSMAP methylation calling : Used reads that used to extract methylation call for each locus.

4.2 Alignment QC

The evaluation of the quality of the alignment data(a BAM file) was performed with Qualimap2.2. The basic statistics of the alignment (ACGT Content, mean and standard overage by chromosome, insert distribution etc.) are summarized and several useful graphs are produced.

Below table 5 is the coverage of alignment for each sample.

Table 5. Mean Coverage

Sample ID	Mean Coverage(X)	Standard Deviation
Control1	26.07	33.81
Control2	25.86	33.39
Test1	25.96	32.02
Test2	25.52	35.38

The following figures are for Control1 sample, included as an example.

Figure 7 shows the coverage across reference. This figure consists of two plots. The upper plot provides the coverage distribution(red line) and coverage deviation across the reference sequence. The coverage is measured in X . The lower plot shows GC content across reference (black line) together with its average value (red dotted line).

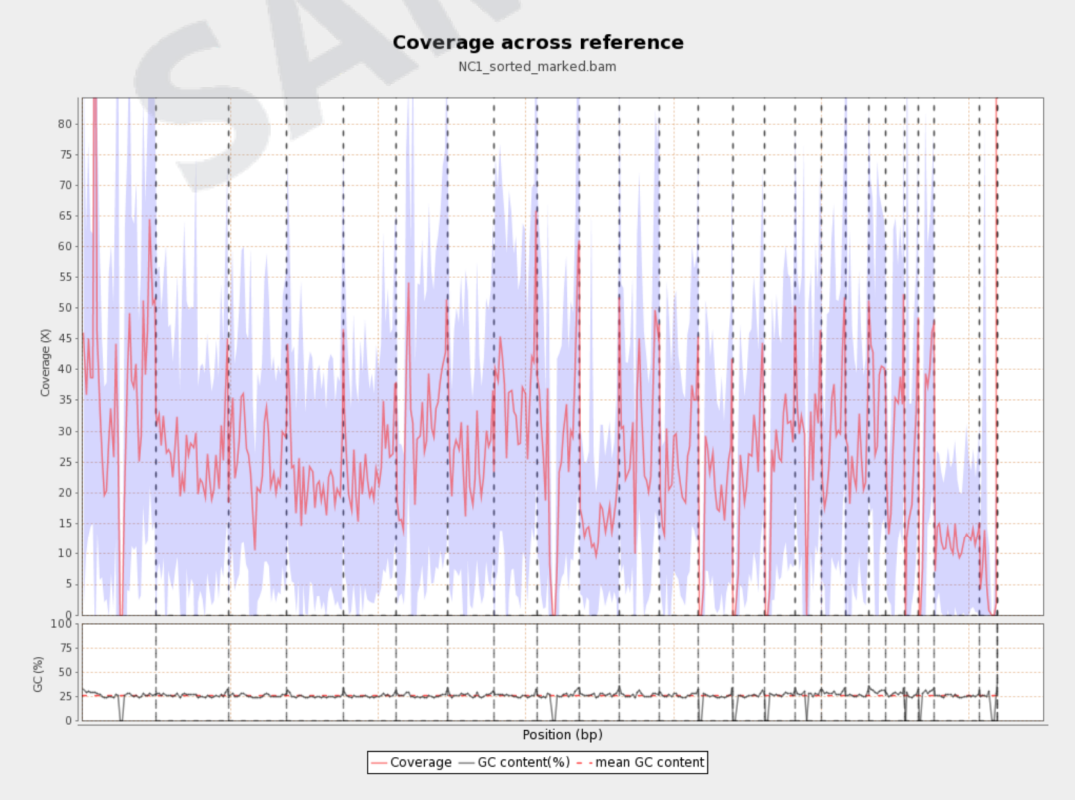


Figure 7. Coverage Across Reference for Control1

Figure 8 is a histogram of the number of genomic locations having a given coverage rate. The bins of the x-axis are conveniently scaled by aggregating some coverage values in order to produce a representative histogram also in presence of the usual NGS peaks of coverage.

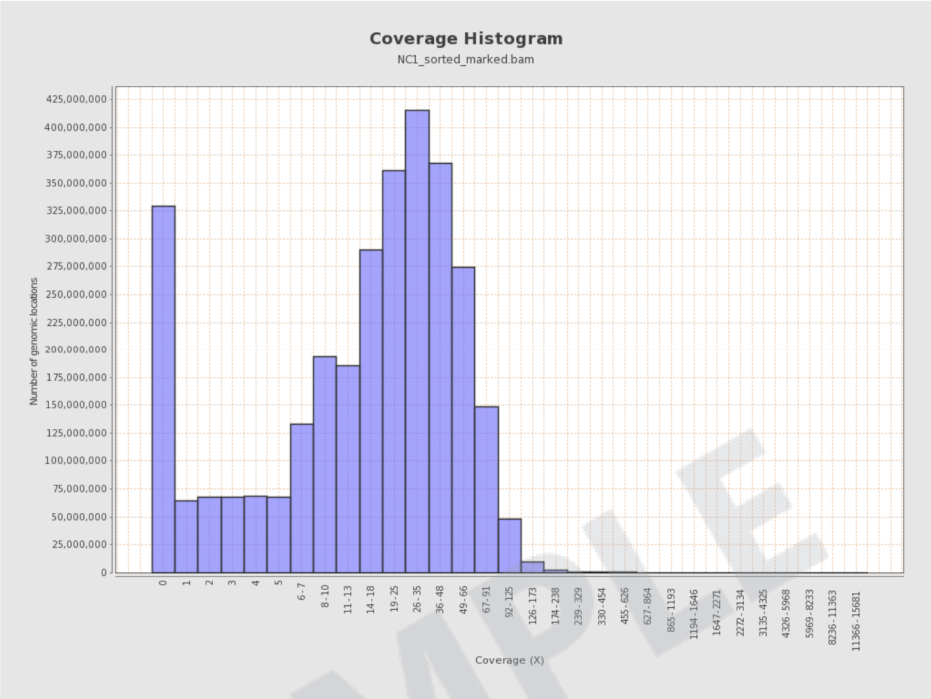


Figure 8. Coverage Histogram for Control1

Figure 9 shows the percentage of nucleotide content per position of the mapped reads.

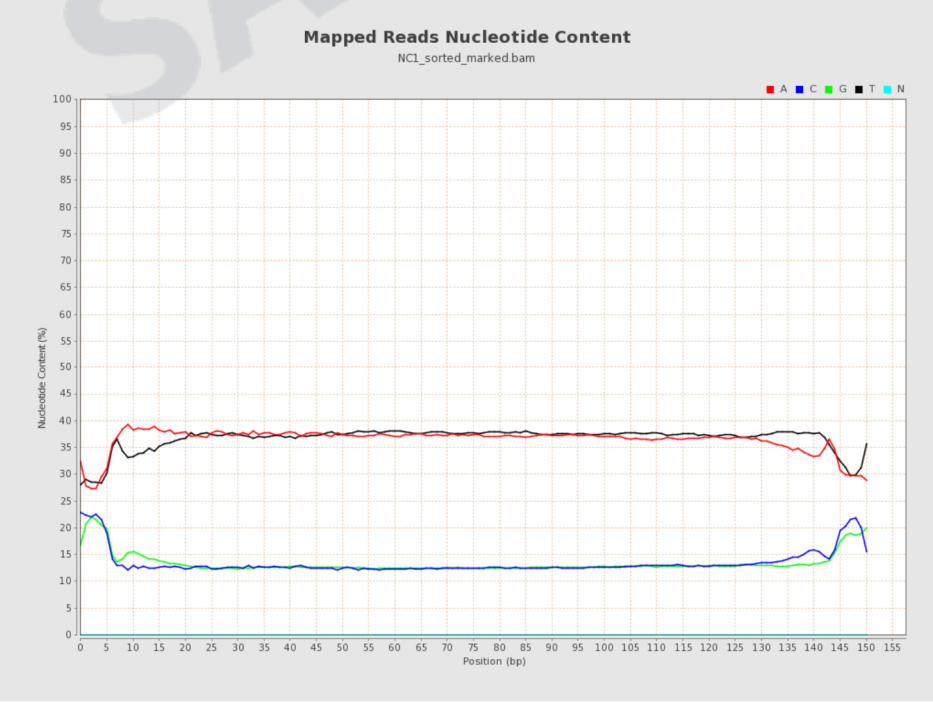


Figure 9. Mapped Reads Nucleotide Content for Control1

Figure 10 shows the distribution of GC content per mapped read.

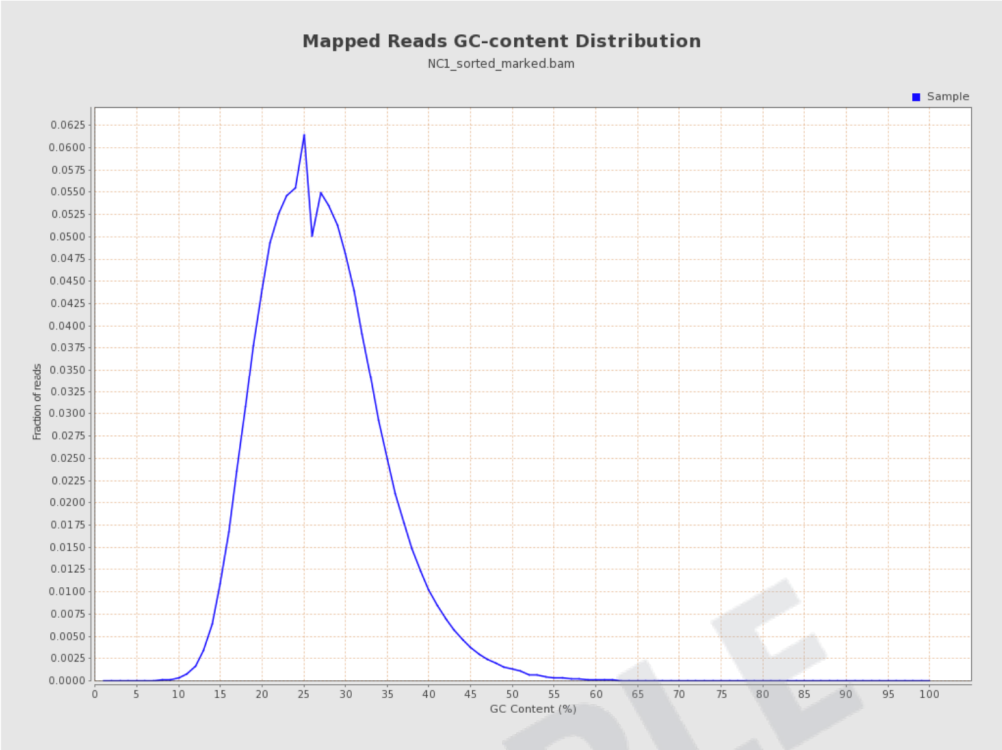


Figure 10. Mapped Reads GC Content Distribution for Control1

Figure 11 shows the histogram of insert size distribution.

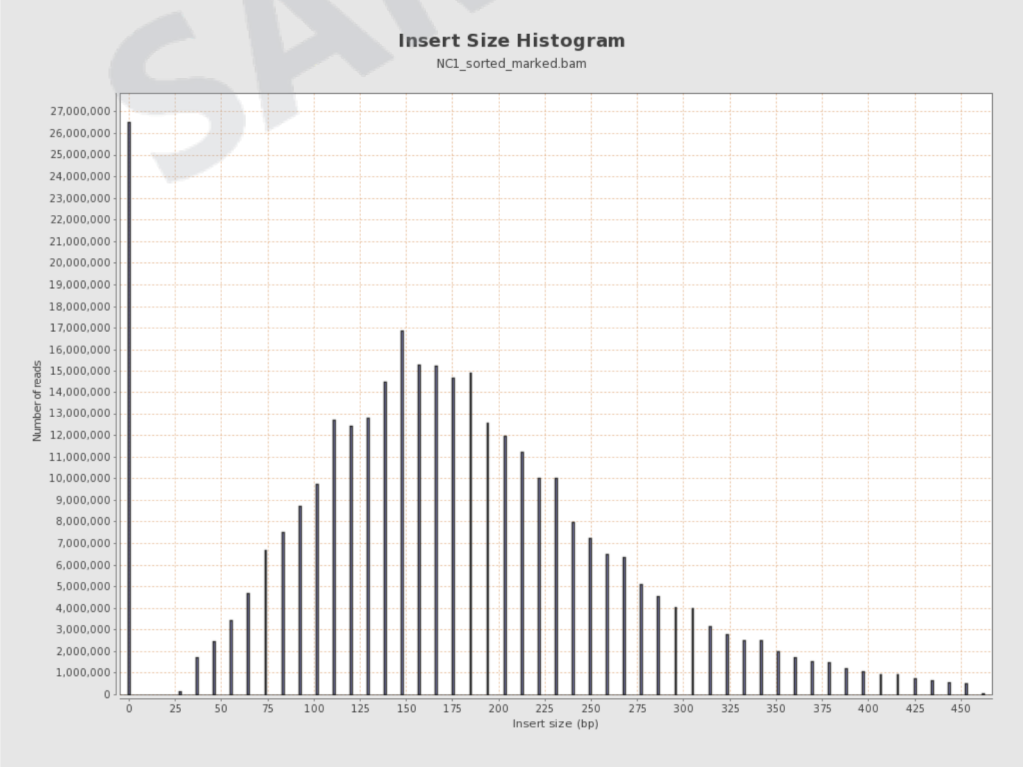


Figure 11. Histogram of insert size distribution for Control1

4.3 Methylation Level Calling

Methylation ratio of every single cytosine satisfying higher than 10 CT count are called using methyratio.py in BSMAP. For the regions covered by both ends of a read pair, only one read is used to call methylation. The results of coverage profiles are summarized as # of C /effective CT counts for each of the three sequence context(CG, CHG, and CHH).

Table 6. Methylated coverage in CpG

Sample ID	Total Coverage in CpG	Methylated coverage in CpG	Methyl%
Control1	766,820,583	539,498,282	70.36%
Control2	745,998,963	525,121,363	70.39%
Test1	731,143,274	516,127,424	70.59%
Test2	748,064,844	526,814,841	70.42%
Average	748,006,916	526,890,478	70.44%

Table 7. Methylated coverage in CHG

Sample ID	Total Coverage in CHG	Methylated coverage in CHG	Methyl%
Control1	2,905,895,255	74,934,396	2.58%
Control2	2,864,154,070	75,578,883	2.64%
Test1	2,823,477,435	69,878,692	2.47%
Test2	2,837,826,383	71,860,684	2.53%
Average	2,857,838,286	73,063,164	2.56%

Table 8. Methylated coverage in CHH

Sample ID	Total Coverage in CHH	Methylated coverage in CHH	Methyl%
Control1	8,072,937,786	133,677,394	1.66%
Control2	8,002,249,282	136,842,546	1.71%
Test1	7,899,907,267	123,464,094	1.56%
Test2	7,899,261,584	128,698,931	1.63%
Average	7,968,588,980	130,670,741	1.64%

Table 9 is an example of the methylation profile of every single cytosine for each sample. This table contains position information, methylation ratio, effective CT count, C count, CT count and functional annotation of genes.

Table 9. Methylation profile and annotations for each sample

Chr	chr1	chr1	chr1	chr1	chr1	chr1	chr1	chr1
Pos	10469	10470	10471	10472	10484	10485	10489	10490
Strand	+	-	+	-	+	-	+	-
Context	CG	CG	CG	CG	CG	CG	CG	CG
Methylation_ratio	0.667	1	1	1	1	1	1	0.75
eff_CT_count	3	2	5	2	5	3	5	8
C_count	2	2	5	2	5	3	5	6
CT_count	3	2	5	2	5	3	5	8
rev_G_count	2	5	2	5	2	5	5	5
rev_GA_count	2	5	2	5	2	5	5	5
CI_lower	0.208	0.342	0.566	0.342	0.566	0.438	0.566	0.409
CI_upper	0.939	1	1	1	1	1	1	0.929
Promoter_region	promoter	promoter	promoter	promoter	promoter	promoter	promoter	promoter
Promoter_transcript_id	NR_046018	NR_046018	NR_046018	NR_046018	NR_046018	NR_046018	NR_046018	NR_046018
Promoter_gene_id	DDX11L1	DDX11L1	DDX11L1	DDX11L1	DDX11L1	DDX11L1	DDX11L1	DDX11L1
strand	+	+	+	+	+	+	+	+
CGI_name	.	.	.	.	.	.	.	.
CGI_info (length;perCpg;perGc;obsExp)	.	.	.	.	.	.	.	.
Region	intergenic	intergenic	intergenic	intergenic	intergenic	intergenic	intergenic	intergenic
Gene	NONE(dist=I	NONE(dist=I	NONE(dist=I	NONE(dist=I	NONE(dist=I	NONE(dist=I	NONE(dist=I	NONE(dist=NONE),DDX11L1(dist=1384)

- Chromosome
- Coordinate (1-based)
- Strand
- Sequence context(CG | CHG | CHH)
- Methylation ratio, calculated as #C_counts / #eff_CT_counts
- Number of effective total C+T counts on this locus (#eff_CT_counts) #eff_CT_counts = #CT_counts * (#rev_G_counts / #rev_GA_counts)
- Number of total C counts on this locus (#C_counts)
- Number of total C+T counts on this locuso (#CT_counts)
- Number of total G counts on this locus of reverse strand (#rev_G_counts)
- Number of total G+A counts on this locus of reverse strand (#rev_GA_counts)
- Lower bound of 95% confidence interval of methylation ratio, calculated by Wilson score interval for binomial proportion.
- Upper bound of 95% confidence interval of methylation ratio, calculated by Wilson score interval for binomial proportion.
- Promoter_region ; the variant hits transcript TSS upstream 2k region.
- Promoter_transcript_id ; the transcript name (transcript_id from GTF).
- Promoter_gene_id ; the gene name (gene_id from GTF).
- Strand : + or - strand for transcript ID.
- CGI_name ; CpG Island, suffix number means Number of CpGs in island.
- CGI_info (Length;perCpg;perGc;obsExp); Island Length; Percentage of island that is CpG; Percentage of island that is C or G; Ratio of observed(cpgNum) to expected(numC*numG/length) CpG in island
- Region; the location where the variant hits(exons, intergenic regions, introns, or a non-coding RNA genes)
- Gene; the gene name.

5. Differentially Methylated CpGs Analysis Results

5.1 Data Analysis Quality Check and Preprocessing

There is a process to sort the differentially methylated CpGs among samples by methylation level of single cytosine. In preprocessing, the quality and similarity checks among samples perform in case of biological replicates exist.

5.1.1 Sample information and analysis design

Total 4 samples were used for analysis.

	Sample.ID	Sample.Group
1	Control1	NC
2	Control2	NC
3	Test1	P
4	Test2	P

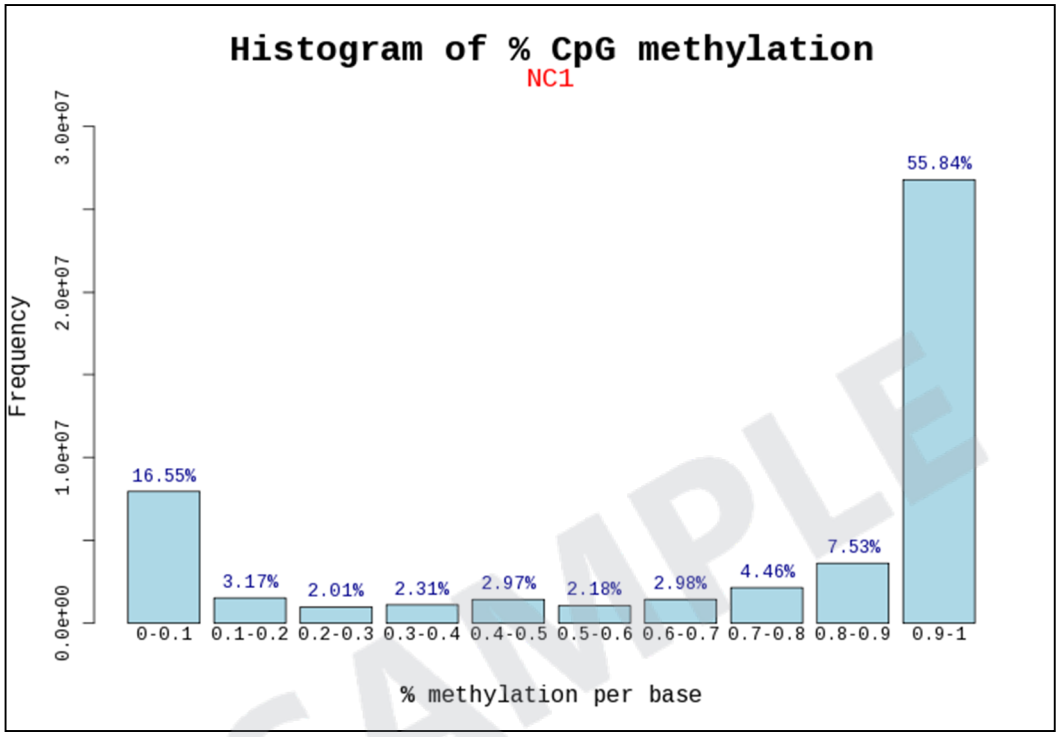
The statistical methods for each comparison pair are shown below.

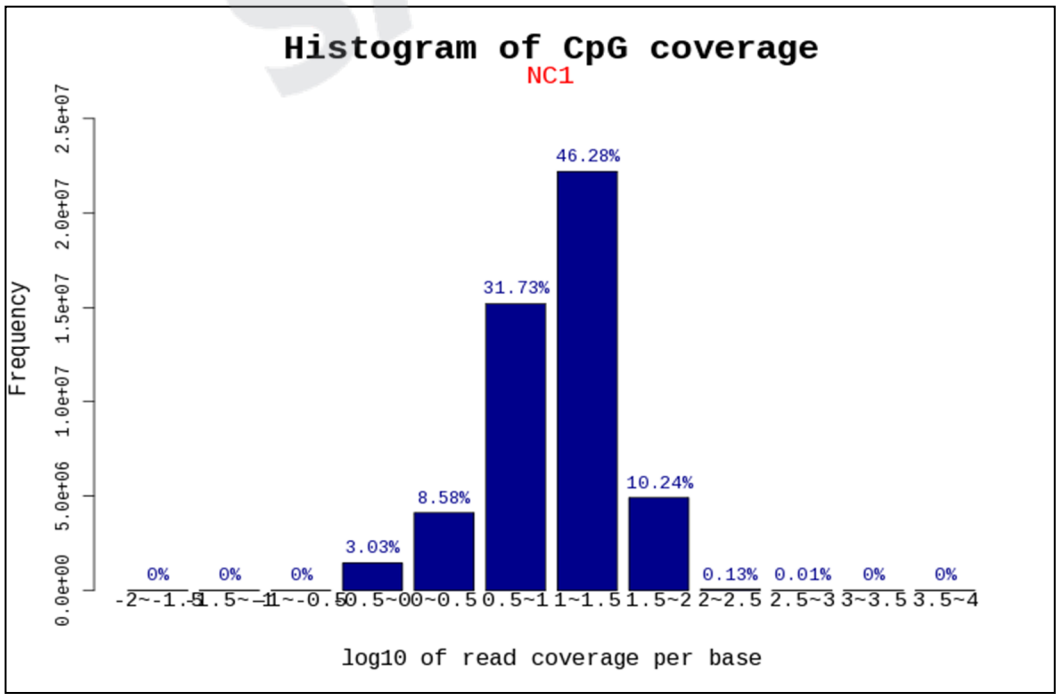
	Test vs. Control	Statistical Method
1	P_vs_NC	Difference of two groups(delta_mean), Independent T-test, Hierarchical Clustering

5.1.2 Data Quality Check

5.1.2.1 Individual % CpG methylation and coverage

The plots shown below are the examples of the histogram of percent methylation distribution and CpG coverage per sample.

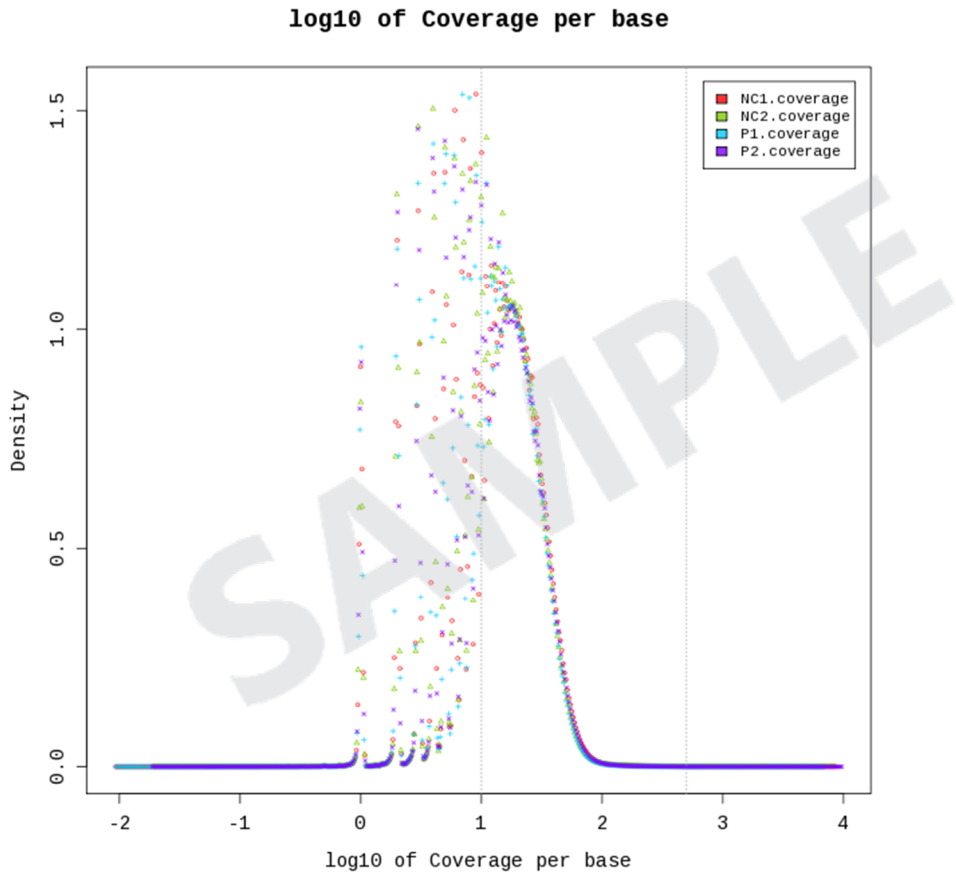




5.1.2.2 Filtering bases based on read coverage

For each CpG, 18,792,97 CpGs outside the lower read and higher read cut-offs in total CpG methylation sites of 47,983,977 are excluded leaving 46,1046,80 CpGs to be analyzed.

Lower read cut-off of 10 means that bases with coverage below 10X are discarded because of increasing the power of statistical tests.
Higher read cut-off or 500 means that bases with more than 500X in each sample are discarded.

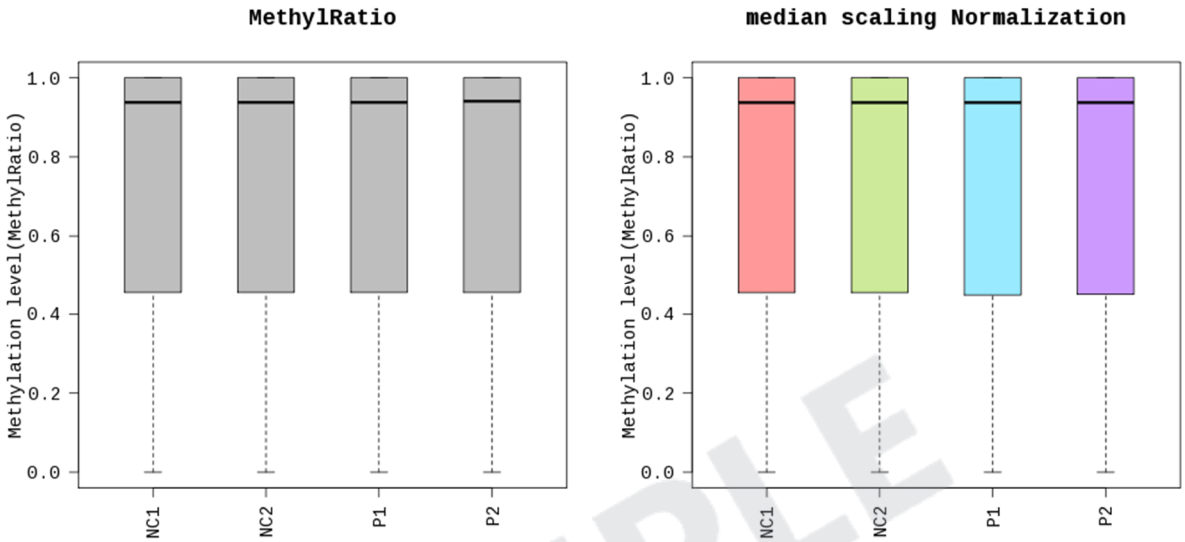


5.1.3 Data Normalization

Methylation levels over the whole genome are quantified by analyzing WGBS data using BSMAP with recommended parameters. Methylation ratio per single cytosine is calculated as # of C /effective CT counts for each of the three sequence context(CG, CHG, and CHH). To compare between clinical groups, MethylRatio values across samples are normalized using median scaling normalization method.

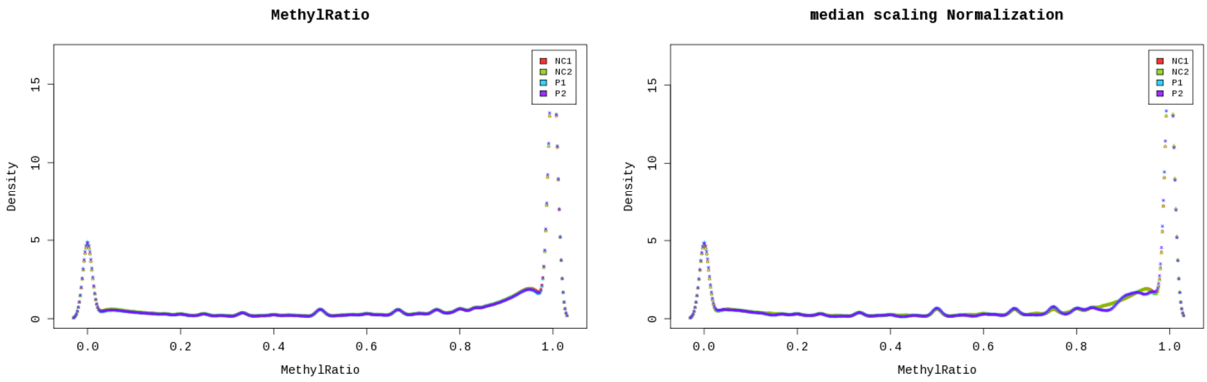
5.1.3.1 Boxplot of MethylRatio between samples

The boxplots shown below indicate the corresponding sample's methylation distribution based on percentile(median, 50 percentile, 75 percentile, maximum and minimum) of raw MethylRatio value and it's normalized value using median scaling normalization.



5.1.3.2 Density plot of MethylRatio per sample

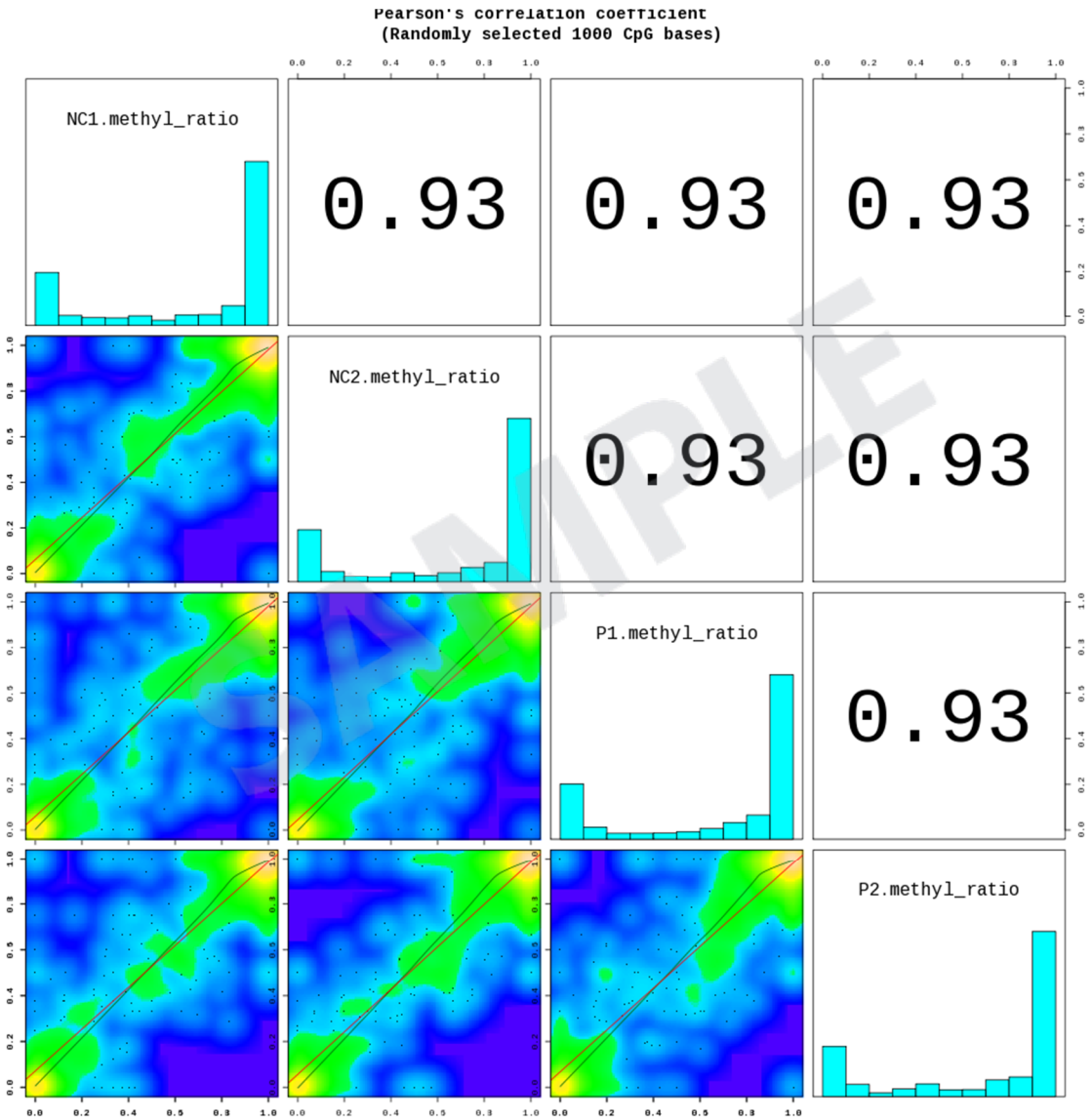
The following density plots show the corresponding samples methylation distribution before and after of median scaling normalization.



5.1.3.3 Correlation Analysis among samples

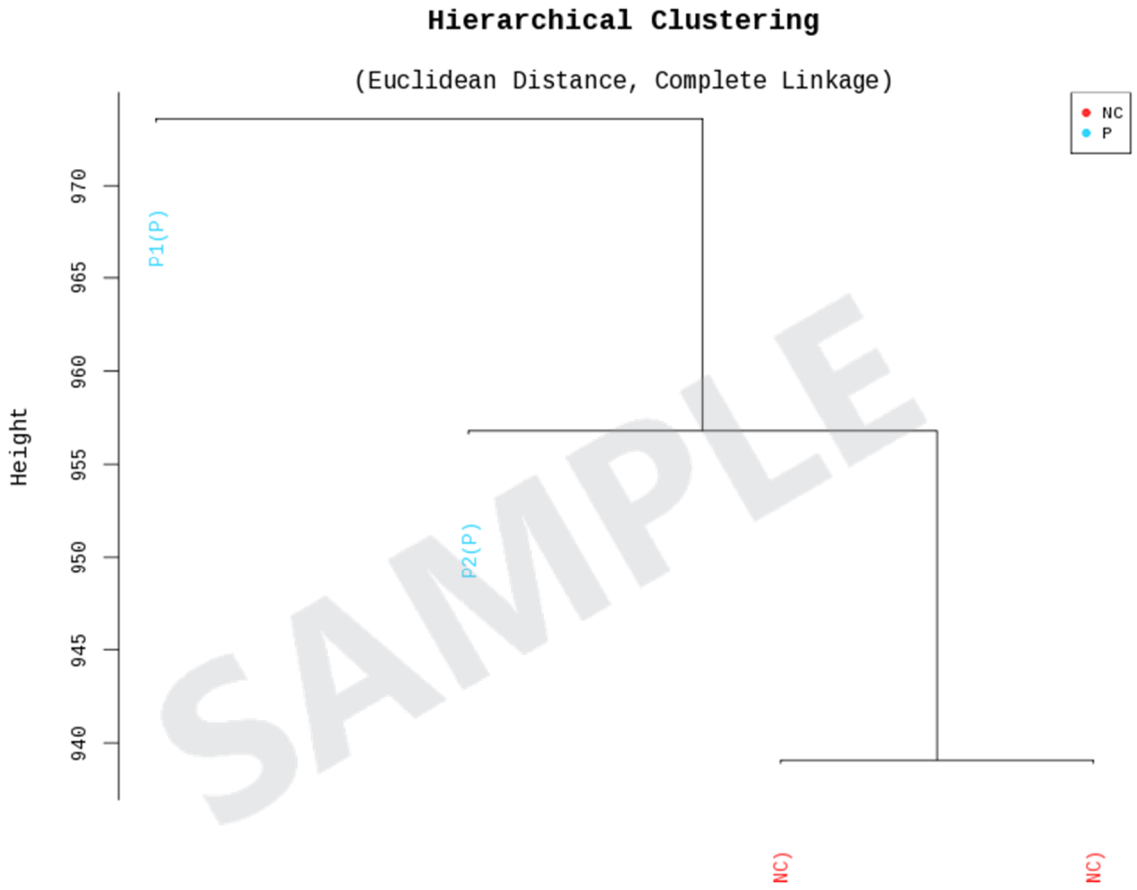
The similarity between samples is obtained through Pearson's coefficient of the MethylRatio value. Between the range from -1 to 1, if the value is close to 1, the samples are higher correlated each other.

Correlation matrix of all samples is as follows.



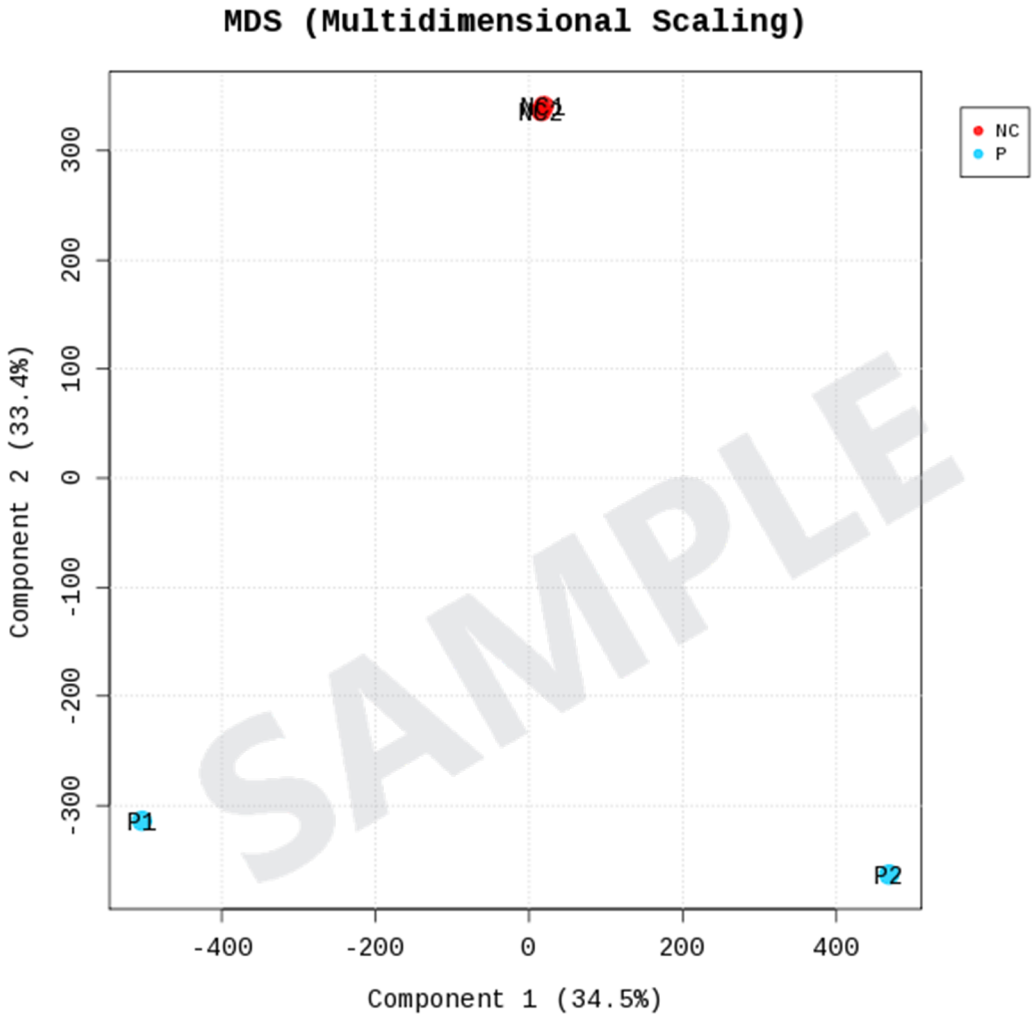
5.1.4 Hierarchical clustering Analysis

The samples are grouped together based on methylation similarities using their MethylRatio value (Distance metric=Euclidean distance, Linkage method= Complete linkage).



5.1.5 Multidimensional Scaling Analysis

Using each sample's MethylRatio value, the similarity between samples is graphically shown in a 2D plot to show the variability of the total data. This allows to identify any outlier samples, or similar expression patterns between sample groups.



5.2 Differentially Methylated CpGs Analysis Workflow

The steps of DM CpGs (Differentially Methylated CpGs) analysis are shown below.

1) MethylRatio values of single cytosine obtained from BSMAP program are used as the original raw data. The MethylRatio values of all samples are normalized using median scaling normalization to compare the methylation level between samples.

● Raw data
(Path: result_WGBS>[Sample Name]>* .CGmap)
: 47983977 CpGs, 4 samples

2) During data preprocessing, low coverage and high coverage bases are filtered. Afterwards, methylratio values are normalized across samples using median scaling normalization

● Processed data
(Path: DM_CpGs_result >[Comparison pair]>Result_[Comparison_pair].txt)

3) Statistical analysis is performed using the difference of two groups(delta_mean), independent t-test per comparison pair. The significant results are selected on conditions of |delta_mean|≥0.2 & raw p value <0.05.

● Significant data
(Path: DM_CpGs_result >[Comparison pair]>Sig.Result_[Comparison_pair].txt)

	# of total CpG	# of CpGs satisfying with |delta_mean|≥0.2 & raw.pval<0.05
Analysis1(P_vs_NC)	46104680	107226

4) For significant CpG list, hierarchical clustering analysis is performed to group the similar samples and CpGs. These results are graphically depicted using heatmap and dendogram.

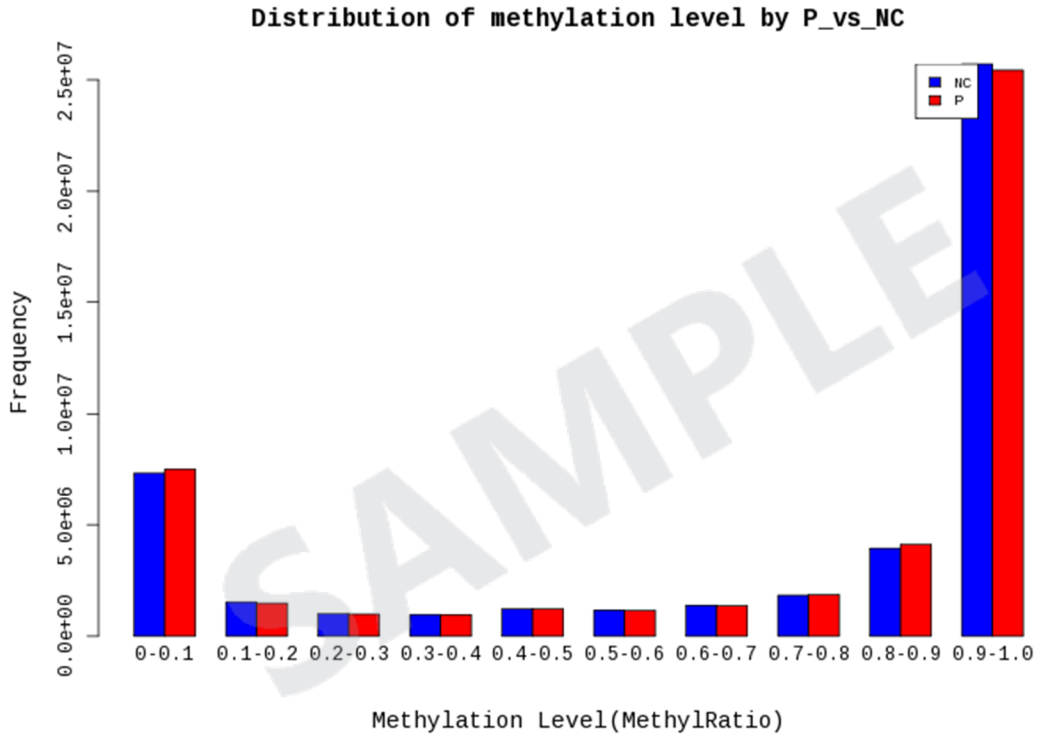
- Hierarchical Clustering (Euclidean Distance, Complete Linkage)
(Path: : DM_CpGs_result > Cluster image)

The following results are examples of P_vs_NC comparison pair.

5.3 Significant CpGs Results

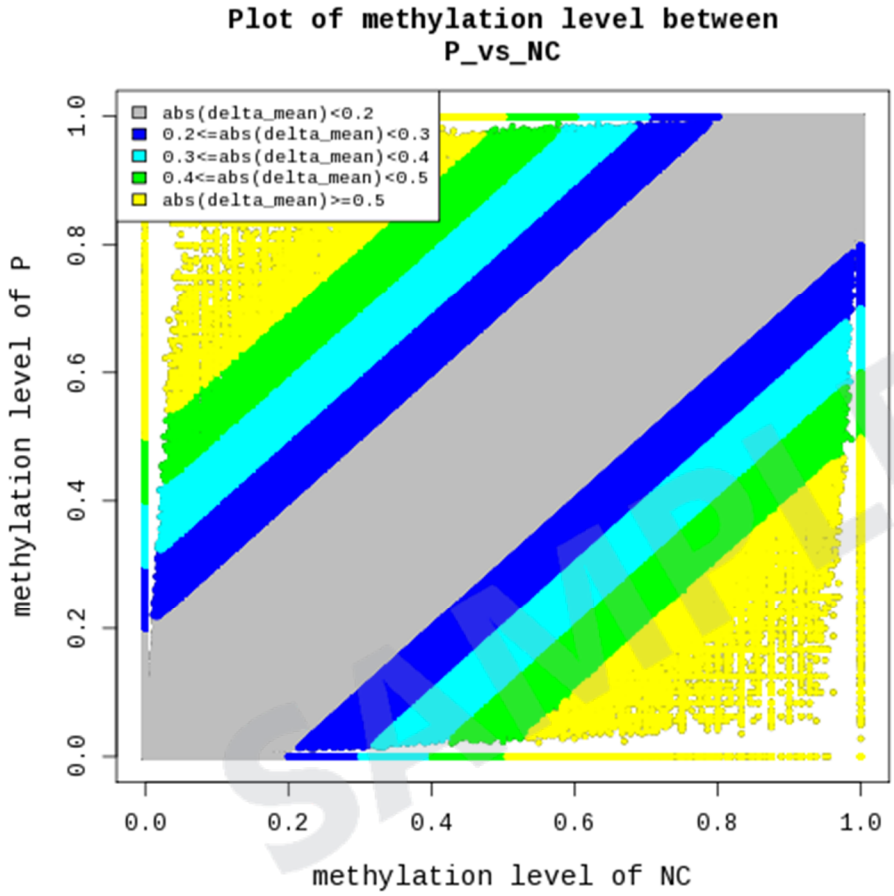
5.3.1 Distribution of methylation level between two groups

The plot below shows distribution of methylation level of each group for comparison pair.



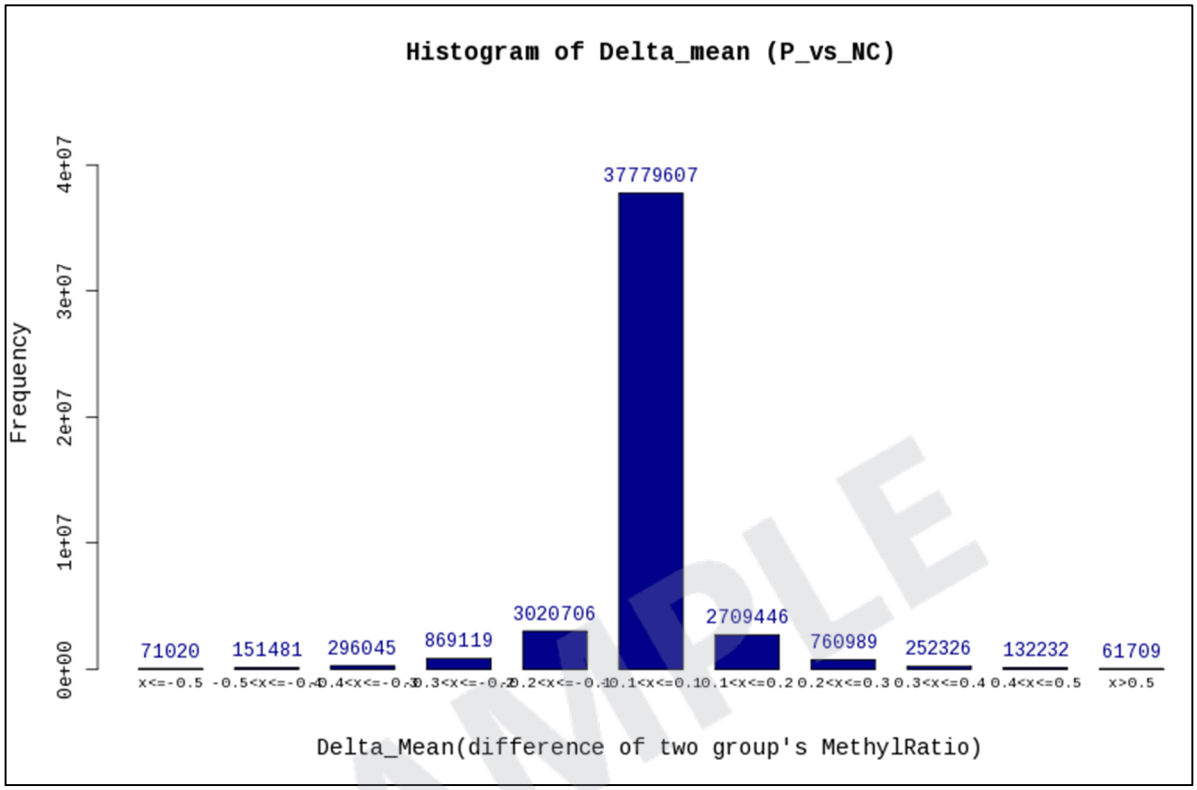
5.3.2 Scatter plot of expression level between two groups

The plot below shows methylation levels between comparison pair as a scatter plot. X-axis is control and Y-axis is average normalized value of the group.
Color coding is marked according to delta_mean value per each CpG site to distinguish the difference of the group.



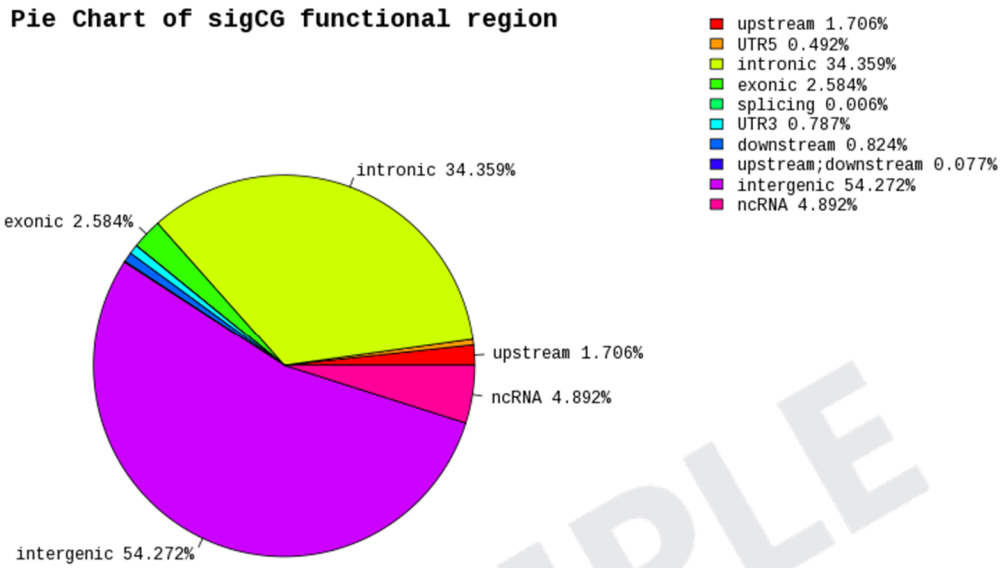
5.3.3 Histogram of delta mean

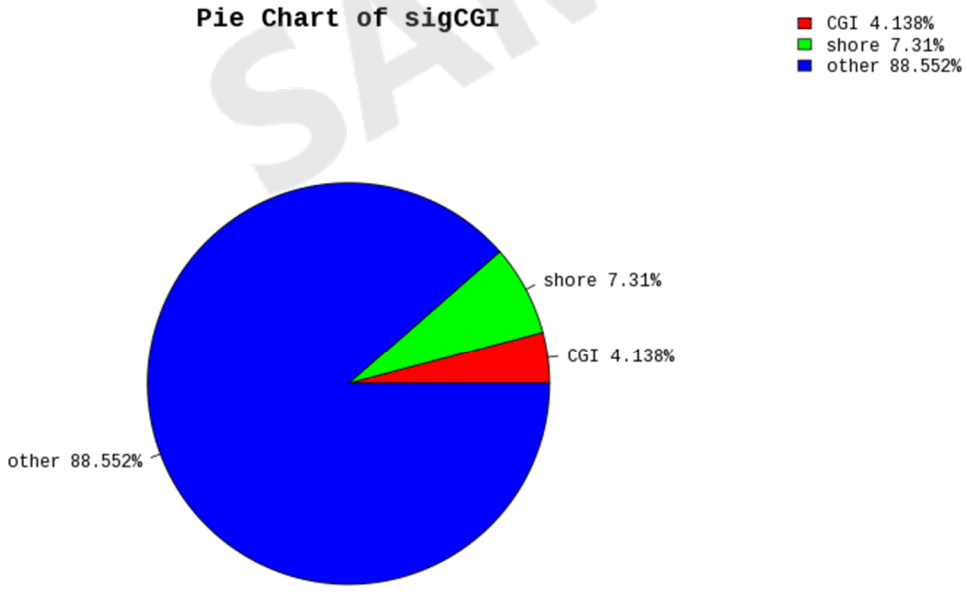
The plot below shows the difference between two group's means of methylation levels in the following histogram.



5.3.4 Pie chart of significant CpG's functional location

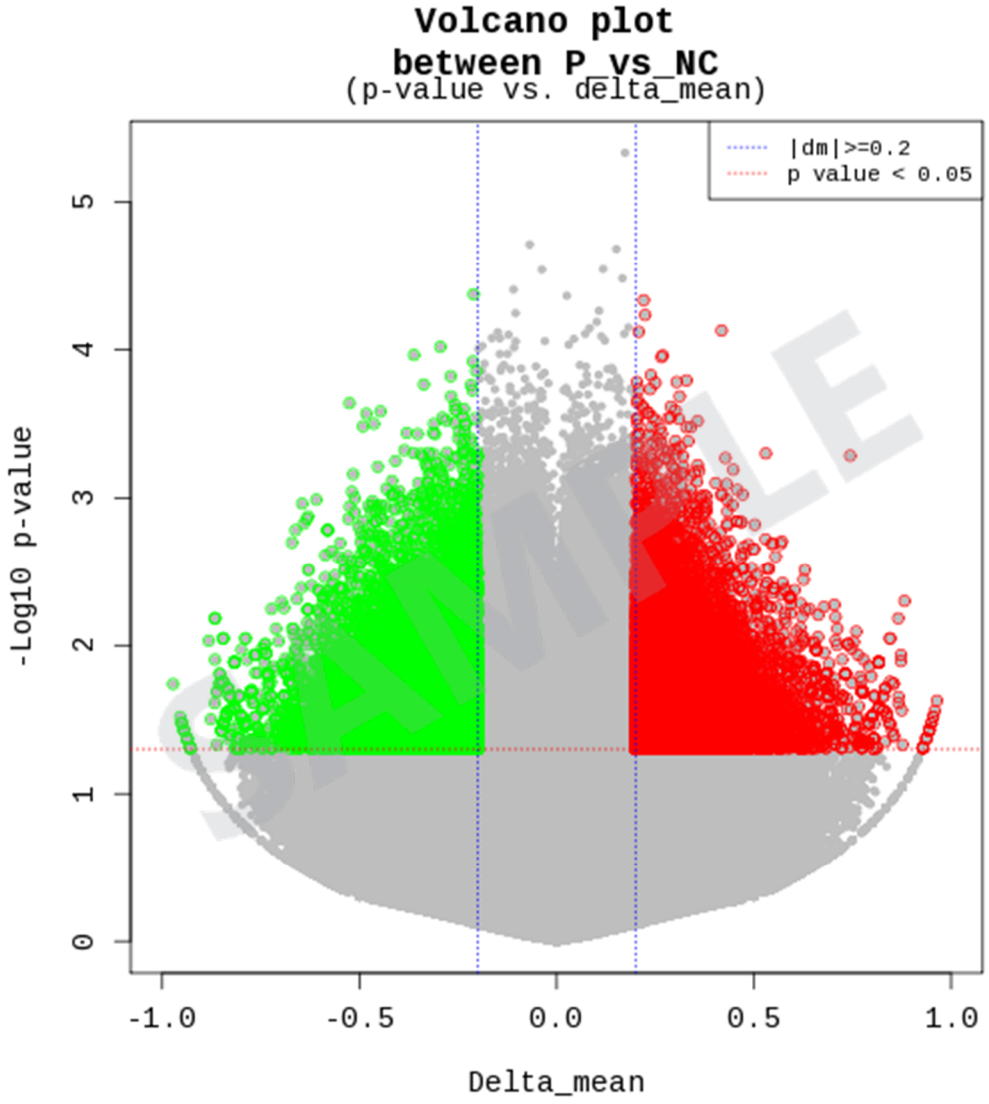
The plot below shows the percentage of differentially methylated CpGs overlapping with functional location(upstream, UTR5, intronic, exonic, splicing, UTR3, downstream etc.) or CpG island, shore and non-CpG island(other).





5.3.5 Volcano plot of methylation level of two groups. (in case of existing p-value)

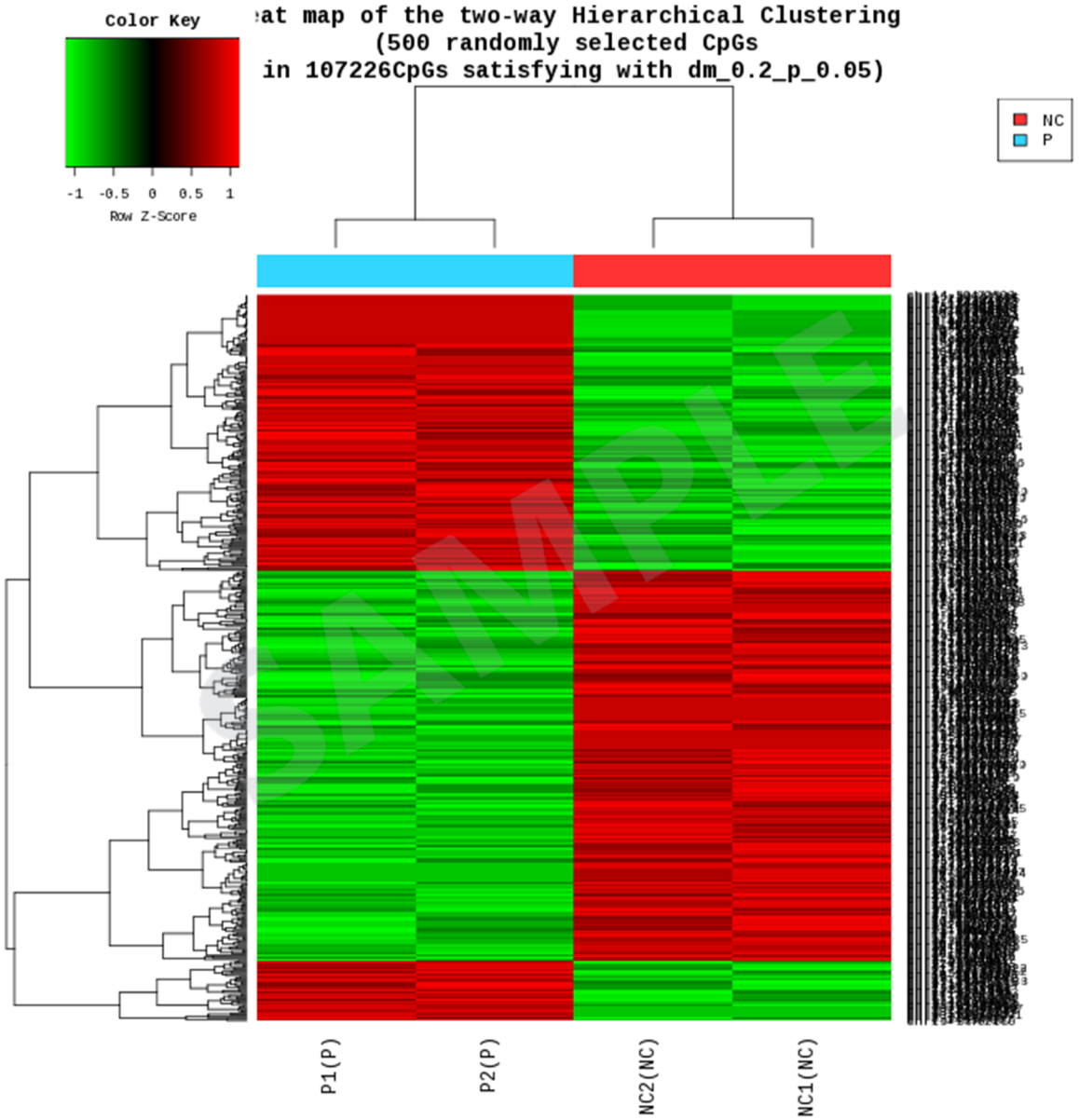
Delta_mean and p-value obtained from the comparison of the average for each group plotted as volcano plot. (X-axis: delta_mean, Y-axis: -log10 p-value)



5.3.6 Hierarchical Clustering Analysis

(Path: DM_CpGs_result > Cluster image)

This heatmap shows the result of hierarchical clustering analysis (Euclidean Method, Complete Linkage) which clusters the similar CpGs and samples by methylation level (normalized value) for the randomly selected 500 CpGs from significant DM CpG list.



6. Data Download Information

6.1. Raw Data

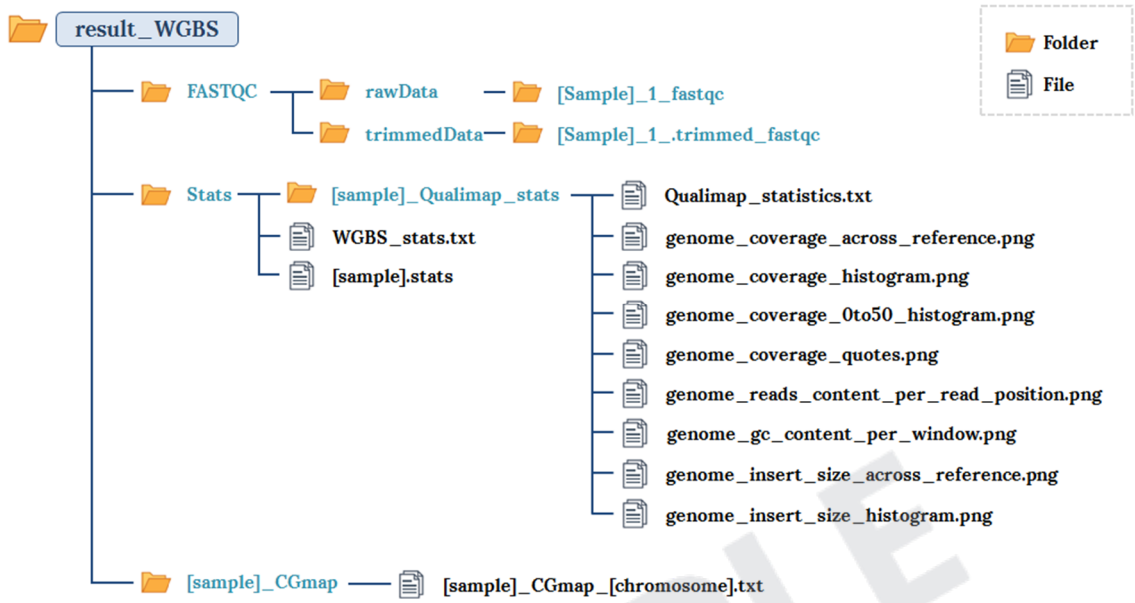
** Raw data & analysis results are sent by a hard drive.

FileName	File size	md5sum
Control1_1.fastq.gz	200G	2cc774c7efd446d98d212242dbbc04f8
Control1_2.fastq.gz	200G	e477c6dbb23c1c3127e451d62d5563f5
Control2_1.fastq.gz	200G	25a9419d8f107beabac8e5481693c4bb
Control2_2.fastq.gz	200G	7d770c082e2e3edb4ed2316746d1ce92
Test1_1.fastq.gz	200G	2cc774c7efd446d98d212242dbbc04f8
Test1_2.fastq.gz	200G	e477c6dbb23c1c3127e451d62d5563f5
Test2_1.fastq.gz	200G	25a9419d8f107beabac8e5481693c4bb
Test2_2.fastq.gz	200G	7d770c082e2e3edb4ed2316746d1ce92

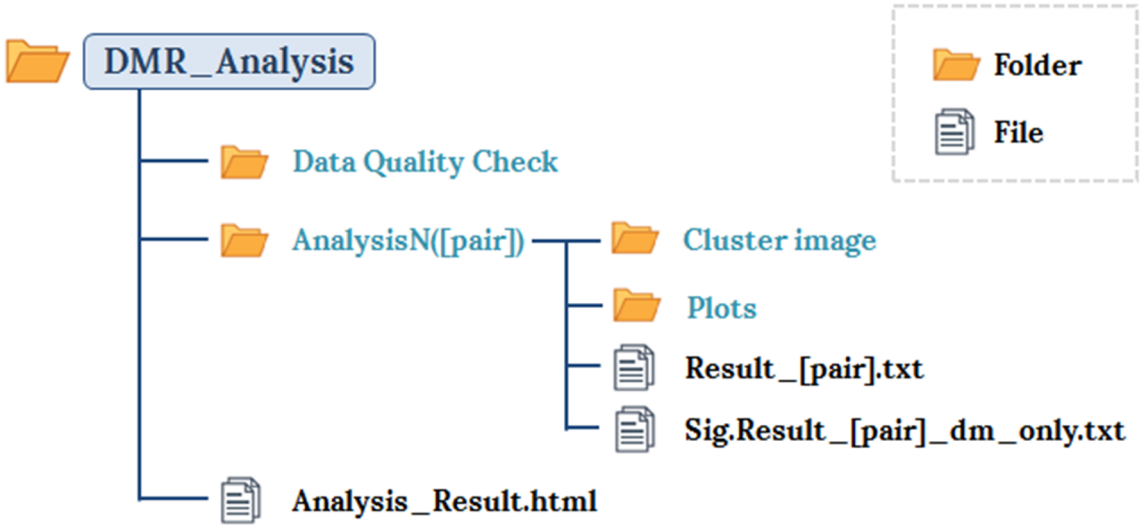
- fastq.gz : This is a zip file of raw data used in analysis
- md5sum : In order to verify the integrity of files, md5sum is used. If the values of md5sum are the same, there is no forgery, modification or omission.

6.2. Analysis Results

** Raw data & analysis results are sent by a hard drive.



File Name	File size	md5sum
result_WGBS.tar.gz	3.3G	2cc774c7efd446d98d212242dbbc04f8



Download link	File	md5sum
result_DM_CpGs.tar.gz	3.2G	e477c6dbb23c1c3127e451d62d5563f5

The data retention period is three months, please send an email (ngs@macrogen.com) or contact representative if you want longer retention period.

Appendix

1. FAQ

Q: I want to see the produced data. How can I open those files?

A: Large volume zip file that is provided by our company is not user-friendly in Windows environment, so it is recommended to use linux environment for smooth operation.

2. FASTQ File

Example of FASTQ

```
@HISEQ-MFG:501:HB0TFADXX:1:1101:1247:2183 1:N:0:
CTCAGCTAAATACTTTGACACCNGTANNANNNNNNNNNTNNNNNNNNNNN
+
@@@BDDDDHHHHFHIIIIIIII#3AC##########################
```

- FASTQ file is composed of four lines.
- Line 1 : ID line includes information such as flow cell lane information.
- Line 2 : Sequences line.
- Line 3 : Separator line (+ mark).
- Line 4 : Quality values line about sequences.

3. Phred Quality Score Chart

Phred quality score numerically expresses the accuracy of each nucleotide. Higher Q number signifies higher accuracy. For example, if Phred assigns a quality score of 30 to a base, the chances of having base call error are 1 in 1000. Phred Quality Score Q is calculated with -10log10P, where P is probability of erroneous base call.

Quality of phred score	Probability of incorrect base call	Base call accuracy	Characters
10	1 in 10	90%	!"#$%&'()*+
20	1 in 100	99%	,-./012345
30	1 in 1000	99.9%	6789:;h=i?
40	1 in 10000	99.99%	@ABCDEFGHIJ

- Encoding : Sanger Quality (ASCII Character Code=Phred Quality Value + 33)

4. Programs and databases used in Analysis

4. 1. FastQC v0.10.0

FastQC is a program that performs quality check on the raw sequences before analysis to make sure data integrity. The main function is importing BAM, SAM, FastQ files and providing quick overview on which section has problems. It provides such results as graphs and tables in html files.

4.2. Trimmomatic v0.32

Trimmomatic is a program that performs trimming depending on various parameters on illumina paired-end or single-end.

- ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.
- SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
- LEADING: Cut bases off the start of a read, if below a threshold quality.
- TRAILING: Cut bases off the end of a read, if below a threshold quality.
- CROP: Cut the read to a specified length.
- HEADCROP: Cut the specified number of bases from the start of the

read.

- MINLEN: Drop the read if it is below a specified length.

4.3. BSMAP v2.90

BSMAP is a short reads mapping software for bisulfite sequencing reads. Bisulfite treatment converts unmethylated Cytosines into Uracils (sequenced as Thymine) and leave methylated Cytosines unchanged, hence provides a way to study DNA cytosine methylation at single nucleotide resolution. BSMAP aligns the Ts in the reads to both Cs and Ts in the reference.
More information can be found here

(https://sites.google.com/a/brown.edu/bioinformatics-in-biomed/bsmap-for-methylation)

Reference

1. Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. *Reference Source*.

2. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, btu170.

3. Okonechnikov, K., Conesa, A., & García-Alcalde, F. (2016). Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data.Bioinformatics, 32(2), 292-294.

4. Xi, Y., & Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. BMC bioinformatics, 10(1), 1.

5. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841-842.

6. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics, 25(16), 2078-2079.

Humanizing Genomics
macrogen

Macrogen Korea

10F, 254 Beotkkot-ro,
Geumcheon-gu, Seoul
Rep. of Korea
Phone : +82-2-2113-7100

Contact

Web : www.macrogen.com
Lims : http://dna.macrogen.com