Use of Agilent SureSelect Methyl-Seq to detect high value Differentially Methylated Regions (DMRs)

Application Note

Author

Richard S. Lee,
Fayaz Seifuddin,
Department of Psychiatry and
Behavioral Sciences,
Johns Hopkins University

Josh Zhiyong Wang,
Kyeong-Soo Jeong,
Agilent Technologies

Introduction

Differentially methylated regions (DMRs) often occur outside of promoters and CpG islands, where they are thought to be key contributors in epigenetic regulation of genes. Thus DMRs are extensively being studied with various methods including microarray, PCR, and more recently, next generation sequencing (NGS). SureSelect Methyl-Seq combines SureSelect target enrichment designed specifically for methylomic regions with subsequent bisulfite treatment to detect methylation events at the single base pair resolution.

This application note focuses on using open-source software, BSmooth, to detect DMRs in SureSelect Methyl-Seq data and independently validate these high value DMRs with bisulfite pyrosequencing, demonstrating the utility of SureSelect Methyl-Seq in effectively identifying DMRs.

Agilent Technologies

SureSelect Mouse Methyl-Seq used on Stress Model samples

Two SureSelect Methyl-Seq kits, available from Agilent Technologies, are 84 Mb Human Methyl-Seq and 109 Mb Mouse Methyl-Seq. The Mouse Methyl-Seq kit is specifically designed to target CpG islands, known tissue-specific DMRs, open regulatory annotations, and over 90 Mb Ensembl regulatory features such as CpG shores and shelves, DNase I hypersensitive sites, histone modification sites and transcription factor binding sites. (The Human Methyl-Seq kit design file is available at http://www.agilent.com/genomics/suredesign, while the Mouse Methyl-Seq design file is available upon request).

We have utilized an animal model of glucocorticoid exposure to evaluate the performance of the Mouse Methyl-Seq kit. The animals were administered 100 μg/mL corticosterone dissolved in 1% EtOH (CORT, N=8) or vehicle solution (1% EtOH, N=8) through their drinking water. Small volumes of tail blood (<25 μL) were collected weekly to confirm elevation of plasma CORT levels. Following four weeks of treatment, the animals were sacrificed and their blood and brain tissues were collected. Following red blood cell lysis, white blood cells were processed for isolation of genomic DNA.

Three to four μg of extracted gDNA from CORT (N=2) and Vehicle (N=2)-treated tissues were processed according to the Mouse Methyl-Seq target enrichment protocol and subsequently sequenced on an Illumina HiSeq2000. Target enrichment performance of the SureSelect Mouse Methyl-Seq kits for all four of the pilot samples processed (Table 1) indicates that a high, ~ 76%, on-target performance was achieved. For a design as large as the 109Mb Mouse methylome, at least 100 million sequence reads (10 Gb sequencing output) per sample are recommended. Although the 42 to 51 million sequence reads obtained per individual sample did not reach the suggested sequencing output of 100 million reads, the average 28x – 33x read depth obtained in this study did not prevent the detection of DMRs with the SureSelect Methyl-Seq design.

Sample	S1	S2	S3	S4
Total HQ uniquely mapped reads:	42M	51M	48M	47M
Percent duplicate reads:	0.2077	0.2264	0.2222	0.3359
Number of reads in targeted regions:	32M	39M	37M	36M
Percentage reads in targeted regions:	**75.90%**	**76.55%**	**76.52%**	**76.41%**
Percentage reads in regions ± 100bp:	82.91%	83.84%	83.40%	84.77%
Percentage reads in regions ± 200bp:	83.88%	84.86%	84.30%	85.89%
Average Read Depth:	**28.63**	**34.84**	**32.77**	**32.08**
Percentage of targeted bases covered by...				
...at least 1 read:	81.56%	82.12%	81.75%	81.33%
...at least 5 reads:	75.11%	76.35%	75.88%	75.40%
...at least 10 reads:	**68.39%**	**71.03%**	**70.07%**	**69.48%**
...at least 20 reads:	53.36%	58.96%	57.05%	56.49%
...at least 30 reads:	39.06%	46.63%	44.12%	43.71%

Table 1. SureSelect Mouse Methyl-Seq target enrichment performance (based on 42–51M sequence reads)

Detecting high value DMRs with BSmooth software on Methyl-Seq data

Mouse mm10 genome build was used as a reference during data analysis, as the Mouse Methyl-Seq sequencing FASTQ files were aligned using Bismark without adaptor trimming as previously described (Reference 1). Here Bismark v0.10.1 was used for alignment with the following command line and parameters:

```
bismark --bowtie2 -N <1> --output_dir <$output_dir> --bam -p <$threads> -L <22> --score_min <L,-0.6,-0.6> --chunkmbs <2048> <$Mouse_ucsc_mm10> -1 <$forward_reads> -2 <$reverse_reads>.
```

After alignment, PCR duplicates were discarded using the deduplicate_bismark perl script. Methylation information for individual cytosine was extracted using the Bismark_methylation_extractor with the following command line and parameters:

```
bismark_methylation_extractor --paired-end --no_overlap --report --bedGraph --counts --buffer_size 10G --no_header --cytosine_report --output <$outputdir> --genome_folder <$Mouse_ucsc_mm10_2/Sequence/WholeGenomeFasta/> $bamfile.
```

For each individual CpG dinucleotide, the coverage of methylated and unmethylated reads, as well as their genomic coordinates, are generated as an output file by the bismark_methylation_extractor, with the minimum coverage threshold set at 1. The same output file is generated for each individual sample, which can then be used as an input file in BSmooth. BSmooth is a pipeline for analyzing bisulfite sequencing data after initial analysis with Bismark and contains tools for data alignment, quality control and DMR identification (Reference 2 and 3).

The BSmooth pipeline is available through Bioconductor (http://www.bioconductor.org). Specifically, within the BSmooth pipeline, the BSSeq component is an R package used for smoothing methylation profiles and identifying DMRs. For the output files from Bismark_methylation_extractor, BSSeq incorporates a smoothing procedure to obtain reliable semi-local methylation estimates from low-coverage bisulfite sequencing data. Smoothing can be performed on a single sample, but typically data from biological replicates should be used whenever possible to allow for higher statistical power and accurate evaluation of DNA methylation. In this experiment, two mouse samples were used for each data point and BSmooth was used with the following command line and parameters:

```
> BS_blood_vehicle_cort_smoothed = BSmooth(BS_blood_vehicle_cort, ns = 20, h = 500, mc.cores = 4, verbose = TRUE, parallelBy = “sample”, maxGap = 10^8)
```

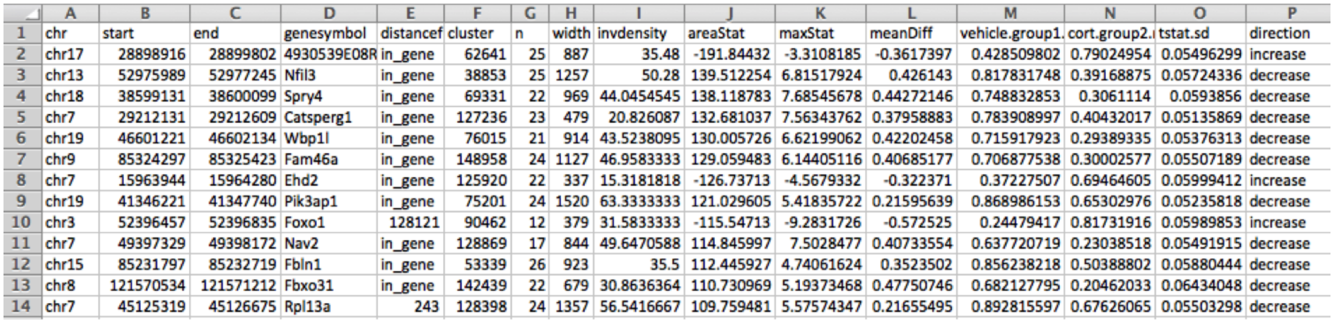
See abbreviations in Table 2 below.

Abbreviations	Description
BS_blood_vehicle_cort	An object of class BSSeq containing coverage of individual methylated and unmethylated CpG across all samples/replicates (union)
ns	The minimum number of methylation loci in a smoothing window
h	The minimum smoothing window, in bases
mc.cores	The number of cores used for running the BSmooth algorithm in a parallel cluster environment (either by chromosome or by sample)
maxGap	The maximum gap between two methylation loci, before the smoothing is broken across the gap. The default smoothes each chromosome separately.

Table 2: Abbreviations for smoothing parameters in BSmooth.

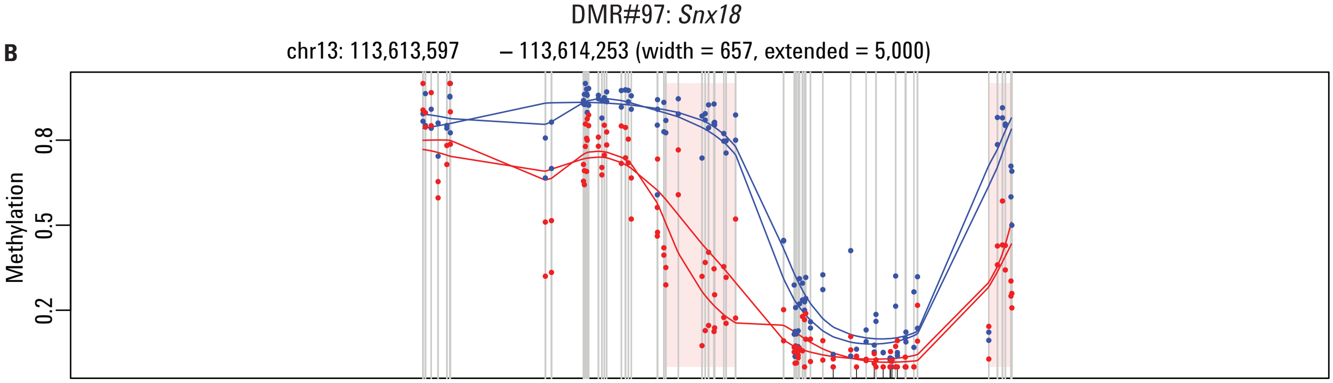
After smoothing, BSmooth uses biological replicates to estimate biological variation and identify DMRs. DMRs are ranked based on area statistics, which uses t-statistics performed at each CpG between the two comparison groups to calculate the overall smoothing differences. The BSmooth package enables users to define specific parameters, such as minimum number of CpGs, window size for smoothing and gap distance between DMRs. For this analysis, we used the default settings: sliding window size = 1,000 bp and minimum number of CpGs = 70. Using the DMR finder, we identified 5,048 unique DMRs that were separated from one another by at least 300 bp. Using the criteria that all of the analyzed CpGs contain at least 20 read sequences for each sample (the total average was ~40 reads per CpG) and each DMR had at least 3 CpGs with greater than 10% mean difference in methylation, we generated a candidate list of 2,634 non-redundant DMRs between VEHICLE- and CORT-treated groups. Both the text output (Fig. 1A) and plot of a typical DMR (Fig. 1B) are shown. The plot for the DMR associated with the *Snx18* gene shows percent methylation (y-axis) and genomic coordinate (x-axis) summaries for each CpG (gray vertical lines) in both groups (red dots=CORT, blue dots=VEHICLE). The blue and red lines represent a smoothed estimation of the overall DNA methylation for all CpGs within the DMR for these samples (pink shade).

A



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	chr	start	end	genesymbol	distancef	cluster	n	width	invdensity	areaStat	maxStat	meanDiff	vehicle.group1.	cort.group2.	tstat.sd	direction
2	chr17	28898916	28899802	4930539E08R	in_gene	62641	25	887	35.48	-191.84432	-3.3108185	-0.3617397	0.428509802	0.79024954	0.05496299	increase
3	chr13	52975989	52977245	Nfil3	in_gene	38853	25	1257	50.28	139.512254	6.81517924	0.426143	0.817831748	0.39168875	0.05724336	decrease
4	chr18	38599131	38600099	Spry4	in_gene	69331	22	969	44.0454545	138.118783	7.68545678	0.44272146	0.748832853	0.3061114	0.0593856	decrease
5	chr7	29212131	29212609	Catsperg1	in_gene	127236	23	479	20.826087	132.681037	7.56343762	0.37958883	0.783908997	0.40432017	0.05135869	decrease
6	chr19	46601221	46602134	Wbp1l	in_gene	76015	21	914	43.5238095	130.005726	6.62199062	0.42202458	0.715917923	0.29389335	0.05376313	decrease
7	chr9	85324297	85325423	Fam46a	in_gene	148958	24	1127	46.9583333	129.059483	6.14405116	0.40685177	0.706877538	0.30002577	0.05507189	decrease
8	chr7	15963944	15964280	Ehd2	in_gene	125920	22	337	15.3181818	-126.73713	-4.5679332	-0.322371	0.37227507	0.69464605	0.05999412	increase
9	chr19	41346221	41347740	Pik3ap1	in_gene	75201	24	1520	63.3333333	121.029605	5.41835722	0.21595639	0.868986153	0.65302976	0.05235818	decrease
10	chr3	52396457	52396835	Foxo1	128121	90462	12	379	31.5833333	-115.54713	-9.2831726	-0.572525	0.24479417	0.81731916	0.05989853	increase
11	chr7	49397329	49398172	Nav2	in_gene	128869	17	844	49.6470588	114.845997	7.5028477	0.40733554	0.637720719	0.23038518	0.05491915	decrease
12	chr15	85231797	85232719	Fbln1	in_gene	53339	26	923	35.5	112.445927	4.74061624	0.3523502	0.856238218	0.50388802	0.05880444	decrease
13	chr8	121570534	121571212	Fbxo31	in_gene	142439	22	679	30.8636364	110.730969	5.19373468	0.47750746	0.682127795	0.20462033	0.06434048	decrease
14	chr7	45125319	45126675	Rpl13a	243	128398	24	1357	56.5416667	109.759481	5.57574347	0.21655495	0.892815597	0.67626065	0.05503298	decrease

B



DMR#97: *Snx18*

chr13: 113,613,597 – 113,614,253 (width = 657, extended = 5,000)

Figure 1. Mouse Methyl-Seq DMR List Output and BSmooth Plot

Pyrosequencing to validate detected high value DMR

From this list, we chose two DMRs which ranked at #2 and #57 and were associated with *Nfil3* (Fig. 2A) and *Ltbr* (Fig. 2C) genes, respectively. Genomic coordinates from the DMR list were used to design outside and nested primers for bisulfite PCR. One of the nested primers was biotinylated to determine the percent methylation by pyrosequencing. Bisulfite conversion was performed on 250 ng of gDNA along with two rounds of PCR-amplification using outside and nested PCR primers with 40 cycles (94 °C for 1 min, 53 °C for 30 sec, 72 °C for 1 min). We successfully obtained ~250 bp amplicons for both DMRs, as verified by gel electrophoresis. Pyro Gold reagents were used to prepare the PCR products for pyrosequencing according to manufacturer's instructions (Qiagen). Several pyrosequencing primers were used to compare the percentage of methylation in five consecutive CpGs (yellow horizontal lines in Fig. 2A and 2C) per DMR amplicon with blood gDNA from 8 VEHICLE- and 8 CORT-treated animals. Pyrosequencing results for the five CpGs in *Nfil3* and *Ltbr* genes (Fig. 2B and 2D) confirm that those DMRs identified in SureSelect Methyl-Seq are of high value.

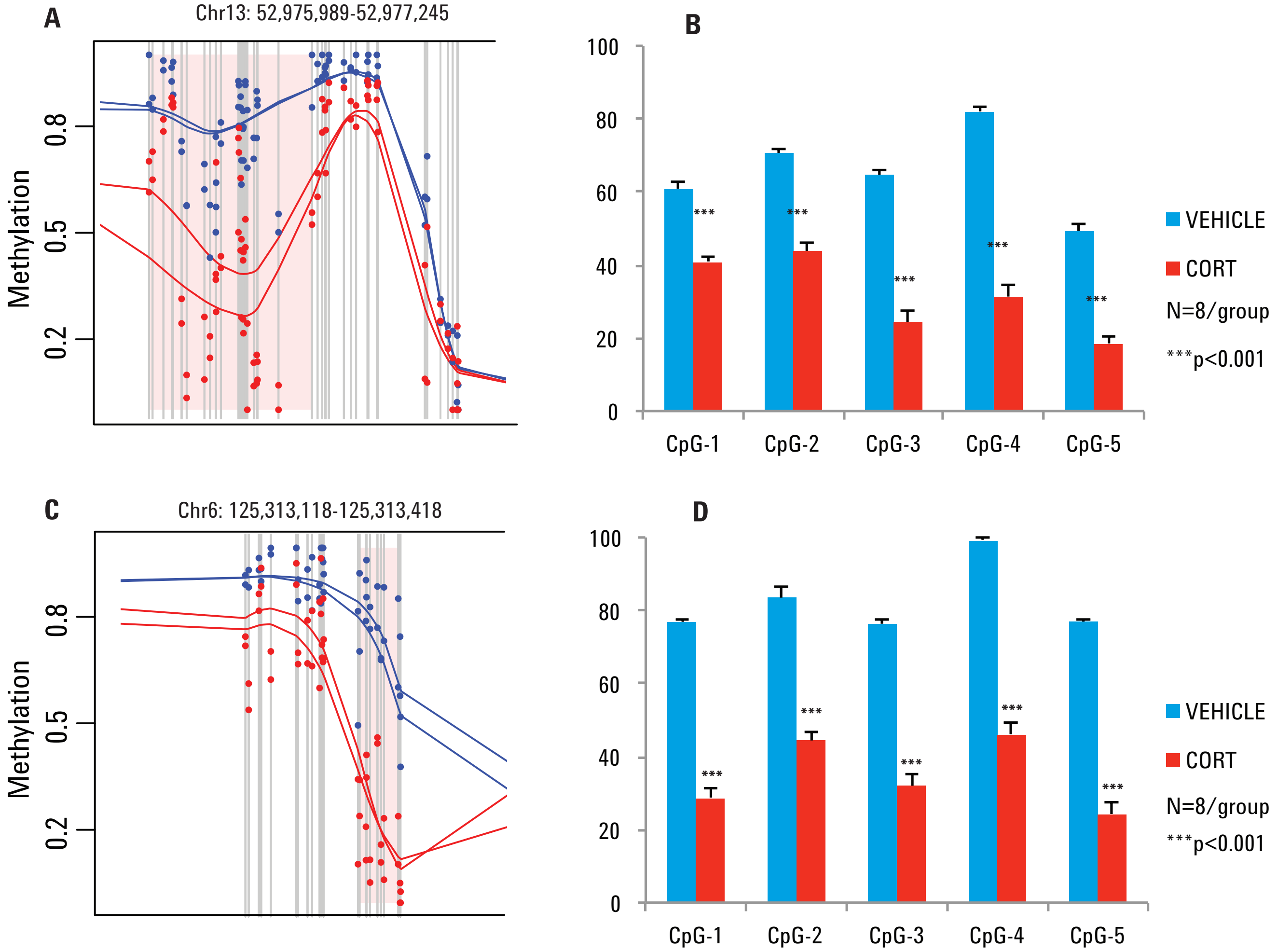


Figure 2. Mouse Methyl-Seq Results (A, C) and Pyrosequencing of DMRs in Blood (B, D)

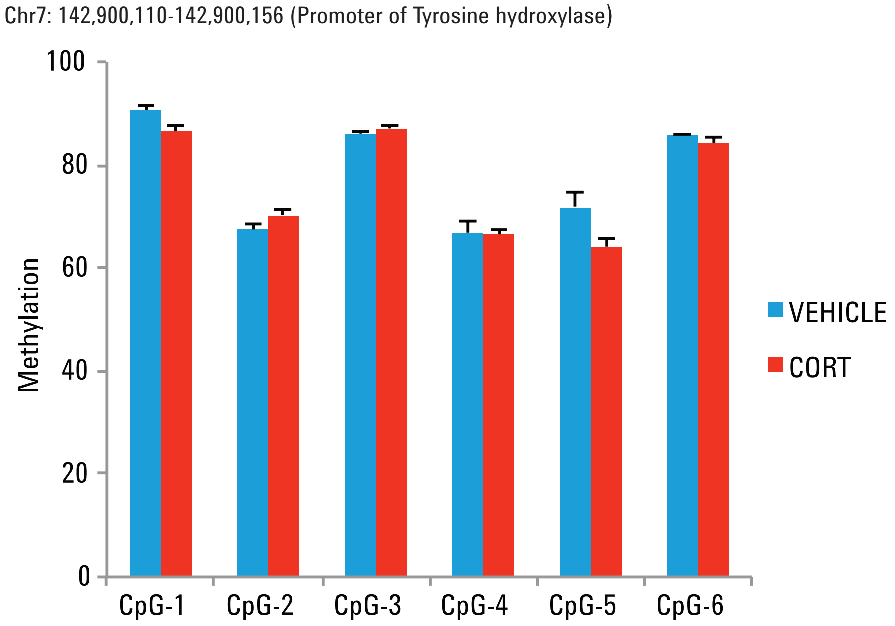


Figure 3. Pyrosequencing of a non-DMR in blood

Meanwhile, several pyrosequencing primers were also used to evaluate a 150 bp region upstream of the TSS of Mouse tyrosine hydroxylase, which did not show any methylation changes in SureSelect Methyl-Seq data. Pyrosequencing results showed that there were indeed no methylation changes in the CpG sites within this region (Fig. 3). Furthermore, this study also successfully validated a DMR located as far down the DMR candidate list as #456 (*Nr1d1*) that was deemed biologically interesting (data not shown).

Conclusions

Agilent’s SureSelect Methyl-Seq target enrichment platform offers a highly reliable and efficient method with single base pair resolution for studying methylated genomic regions. For the catalog Mouse or Human Methyl-Seq designs, a high on-target rate (usually over 75%) is routinely achieved. Flexibility to target any region of interest can also be achieved with custom Methyl-Seq designs. When performing methylation studies with both control and experimental groups, incorporating the BSmooth analysis pipeline allows for the reliable detection of DMRs from SureSelect Methyl-Seq data.

References

1. “Agilent SureSelect Human Methyl-Seq for the Quantitative Analysis of DNA Methylation with Single-Base Resolution” Agilent Technical Overview Publication Number 5991-0166EN, 2012

2. Hansen KD, Langmead B, Irizarry RA. BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol. 2012;13(10):R83. PMCID: 3491411.

3. Hansen KD, Timp W, Bravo HC, Sabunciyan S, Langmead B, McDonald OG, Wen B, Wu H, Liu Y, Diep D, Briem E, Zhang K, Irizarry RA, Feinberg AP. Increased methylation variation in epigenetic domains across cancer types. Nat Genet. 2011;43(8):768-75. PMCID: 3145050.

www.agilent.com/genomics/
Methyl-Seq

For research use only. Not for use in diagnostic procedures. This information is subject to change without notice.

PR7000-0467
Published in USA, May 3, 2016
Publication Number 5991-5298EN

Agilent Technologies