Supplementary methods

Preparation and high-throughput sequencing of libraries for other methods

With the same batch of genomic DNA extracted from T29 cell line and used to prepare the MB-seq library, we prepared MethylC-seq and MeDIP-seq libraries using standard procedures(Down et al., 2008; Lister et al., 2009). Final libraries were quantified by using Quant-iT PicoGreen dsDNA Kits and KAPA Illumina/Universal Library Quantification Kits (KAPA Biosystems, Inc., Woburn, MA) and sequenced on an Illumina HiSeq 2000.

Bisulfite read alignment and methylation site identification

The alignment of bisulfite-treated short reads to the reference genome hg18 was conducted as described by Maunakea *et al.* (Maunakea et al., 2010b). In brief, two read alignments were carried using the SOAP software to get the best hit for a given pair-end short read (Maunakea et al., 2010a). A straightforward seed-and-extension algorithm was then employed for the alignment, with two mismatches allowed in the seed (30 bp) and five mismatches in the whole read.

Methylcytosines were identified according to previously published criteria (Maunakea et al., 2010b). To ensure the reliability of 5mC identification, only bases with quality scores higher than 20 were considered for further analysis. Bisulfite conversion efficiency was calculated by using the C to T conversion rate for all cytosines in the CHH context (where H = A, T, or C). Even under the assumption that all 5mC in CHH nucleotides were products of conversion failure, the bisulfite conversion rate for each single-base resolution approach was >99%, which ensured that the false positive rate was <1%.

Genomic features

RepeatMasker annotations, CpG islands, and refGene coding loci features were downloaded from UCSC (Kent et al., 2002). GC content, CpG density, and CpG-oe values were calculated using custom Perl scripts. Promoters of genes were defined as

2kb regions, including 1kb upstream and 1kb downstream of TSS.

Identification of DMRs

We identified DMRs by comparing the methylomes of T29 and T29H using our previous developed software RRBS-Analyser. Candidate DMRs were determined as 200-bp non-overlapped sliding windows containing at least 5 CpG sites with a 1.25-fold change in the methylation level, t test p value <0.05, and FDR <0.05. In addition, for a valid DMR discovery, we required a difference value of a least 0.2 between the two samples. Two nearby candidate DMRs were either considered interdependent and joined into one continuous DMR, or considered independent. After iteratively merging interdependent DMRs, the final dataset of DMRs was made up of those that were independent of each other. Differentially methylated regions are listed in Table S9.

Transcriptome sequencing and analysis

Total RNA was extracted from T29 and T29H cell lines with the RNeasy Mini Kit (QIAGEN, Germany), respectively. The RNA quality was assessed using Bioanalyser 2100 (RNA nano kits, Agilent). mRNA-Seq libraries were generated from total RNA with polyA+ selection of mRNA using the TruSeq RNA Sample Prep Kit v2 (Illumina, San Diego, CA), and sequenced transcriptomes using the Illumina Hiseq 2000 in paired-end mode.

Following sequencing, we estimated gene expression levels and identified differentially expressed genes between the T29 and T29H samples. Sequencing adapters and low quality sequencing reads was excluded using the Trim Galore program. TopHat2 was used to align the reads to the genome, allowing up to two mismatches and a Phered-scaled mapping quality $\geqslant 4$. Detailed sequencing information is summarized in Table S10.

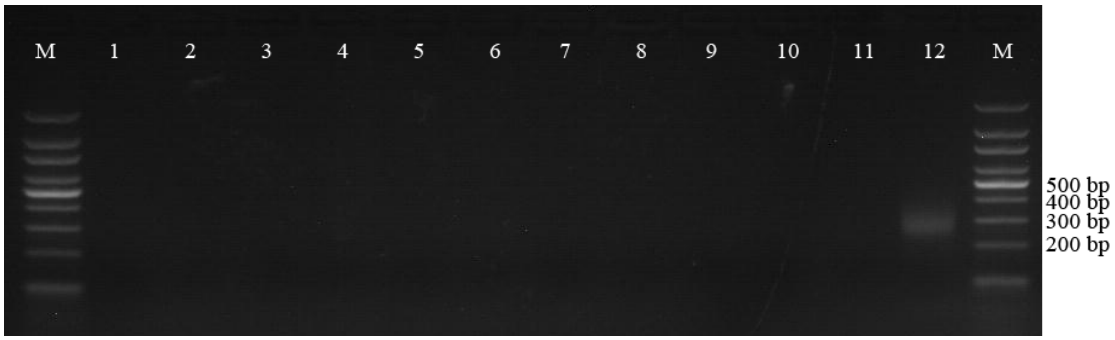
The bioconductor package edgeR was used to identify genes that were differentially expressed between T29 and T29H based on the number of mapped paired-end reads for each gene counted by the python package HTSeq

(http://www.htseq.org/). The edgeR package was suitable for data such as ours (with very few or even zero replicates) and so we ran edgeR for zero replicate testing, setting the argument BCV to 0.1 as suggested in the manual. Differentially expressed protein-coding genes and lncRNA genes were listed in Table S11 and S12, respectively.

Locus-specific bisulfite sequencing

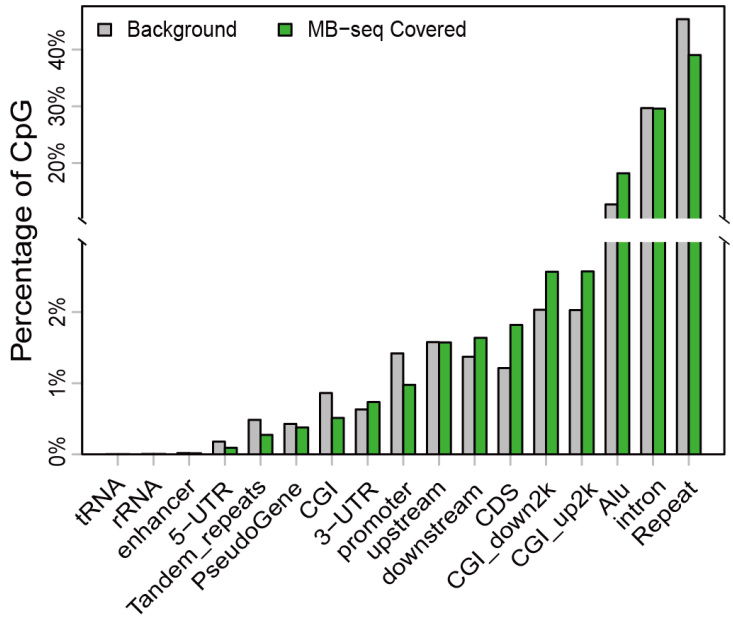
Genomic DNA from T29 and T92H was bisulfite treated with the EpiTect Bisulfite Kit (Qiagen, USA) using two rounds of the standard conversion. Approximately 50 ng of bisulfite-treated DNA was used to amplify mCpG DNA fragments, which were uniquely detected in MB-seq from T29 with nest PCR primers. The same approach was employed for amplification of the DMR DNA fragments identified in T29 and T29H. The full list of primers is available in Table S13. The amplified PCR products were gel purified, pooled in equimolar quantities, and libraries were constructed following the Illumina sequencing library preparation protocol. Libraries were sequenced on Miseq with a PE 250 bp sequencing kit and reads were aligned to the hg18 reference sequence.

Supplementary Figures

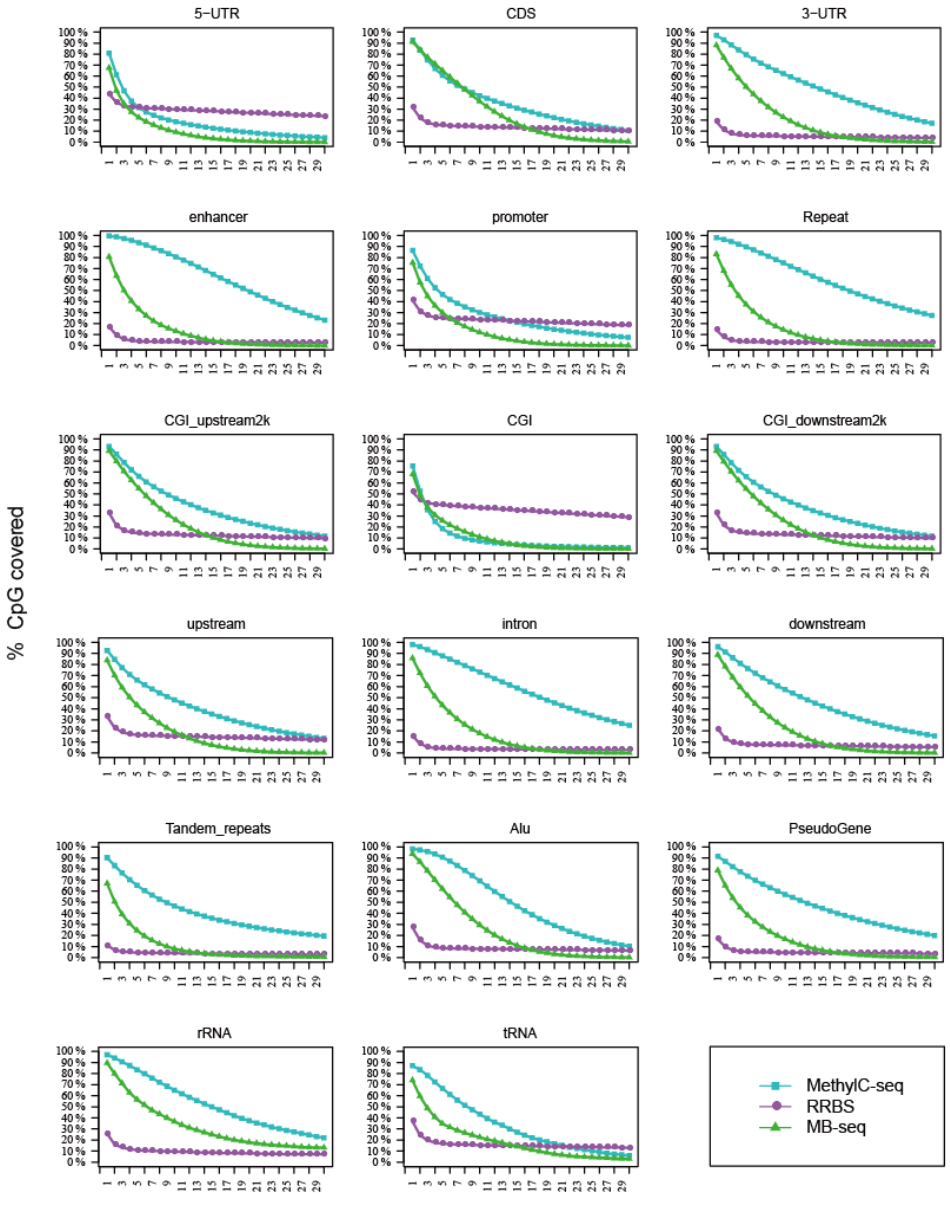


Supplementary Figure S1. Evaluation of PCR amplification efficiency for DNA containing CMS-modified cytosines with several commercial DNA polymerases. Fragmented DNA was end-repaired, A-tailed, and ligated to hydroxylmethylated Illumina adapters following the standard Illumina protocol. Two rounds of bisulfite treatment were used on hydroxylmethylated Illumina adapters ligated DNA with EpiTect Bisulfite Kit (Qiagen), converted the hydroxylmethylated cytosines to CMS. Ten nano gram CMS-modified DNA was amplified for 12 cycles with Illumina PCR primer PE1.0 and PE2.0, using 12 types of commercial DNA polymerase. Amplified products were purified with QIAquick PCR Purification Kit (QIAGEN), and loaded on a 2% agarose gel. Lane 12 (KAPA 2G Robust HotStart ReadyMix) shows efficient amplification in the scope of 200-500 bp, while lane 1-11 display failing amplification. Taq DNA polymerase used in Lane 1-10 were KAPA HiFi HotStart Uracil+ ReadyMix (KAPA, USA), KAPA HiFi Library Amplification Kit (KAPA, USA), Phusion High-Fidelity PCR master mix (Thermo, USA), NEBNext High-Fidelity 2X PCR Master Mix (NEB, USA), HotStar HiFidelity Polymerase Kit (Qiagen,German), GeneAmp Fast PCR Master Mix (Life Technologies, USA), JumpStart Taq DNA Polymerase (Sigma,USA), AptaTaq Fast PCR Master (Roche, Swiss), VeraSeqHigh Fidelity DNA Polymerase 2.0 (Enzymatics, USA), SapphireAmp Fast PCR Master Mix(Takara, Japan), PfuTurbo DNA Polymerase (Agilent, USA), respectively.

A

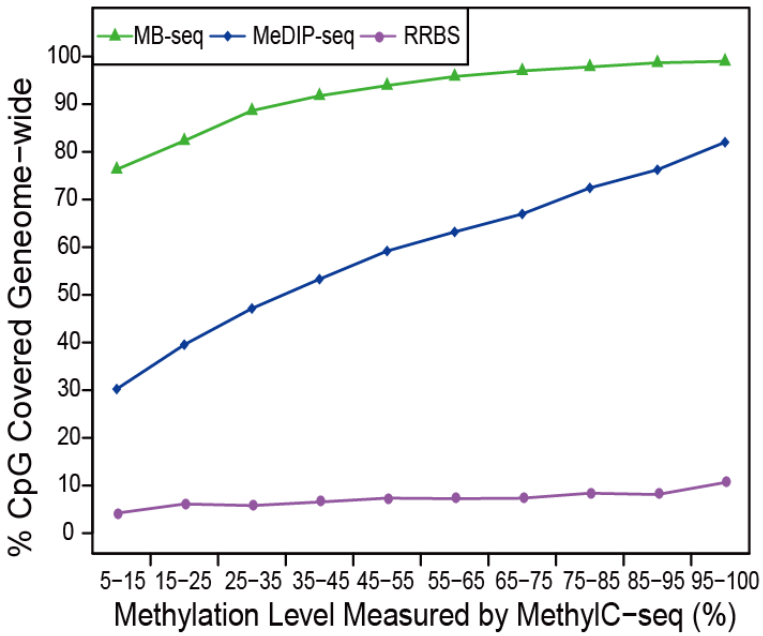

Background
MB−seq Covered
Percentage of CpG
0%
1%
2%
20%
30%
40%
tRNA
rRNA
enhancer
5−UTR
Tandem_repeats
PseudoGene
CGI
3−UTR
promoter
upstream
downstream
CDS
CGI_down2k
CGI_up2k
Alu
intron
Repeat

B

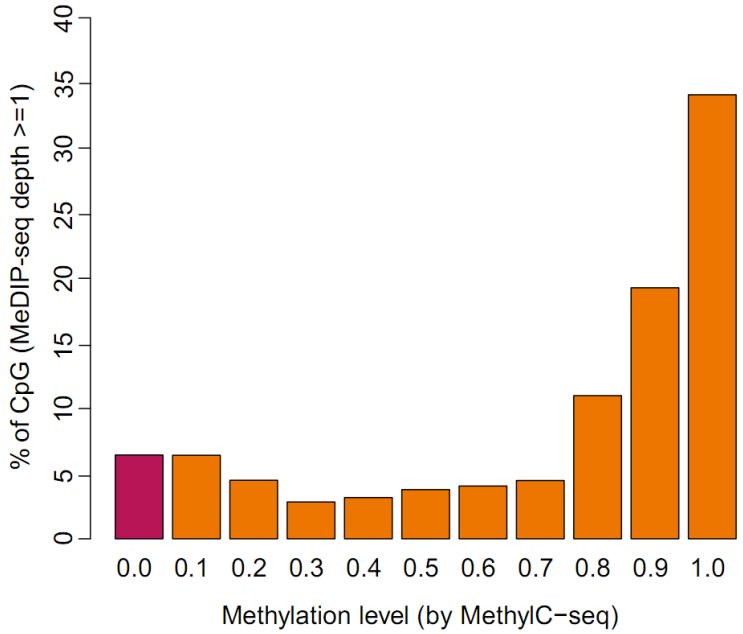

5−UTR
CDS
3−UTR
enhancer
promoter
Repeat
CGI_upstream2k
CGI
CGI_downstream2k
upstream
intron
downstream
Tandem_repeats
Alu
PseudoGene
rRNA
tRNA
% CpG covered
Read coverage threthold for CG
MethylC-seq
RRBS
MB-seq

Supplementary Figure S2. Coverage of genomic features in MB-seq. (A) Barplot for coverage of CpG of background and MB-seq covered in various genomic features. For each paired bars, left bar (grey) represents percentage of genomic background and right one (green) represents fraction of CpG covered by MB-seq. **(B)** CpG coverage as a function of read coverage threshold for MethylC-seq (cyan), RRBS (medium-orchid), and MB-seq (green). X-axis denotes sequencing depth and y-axis the fraction of CpGs that were at or above a given sequencing depth. The percentages of covered CpGs in 17 elements were plotted.

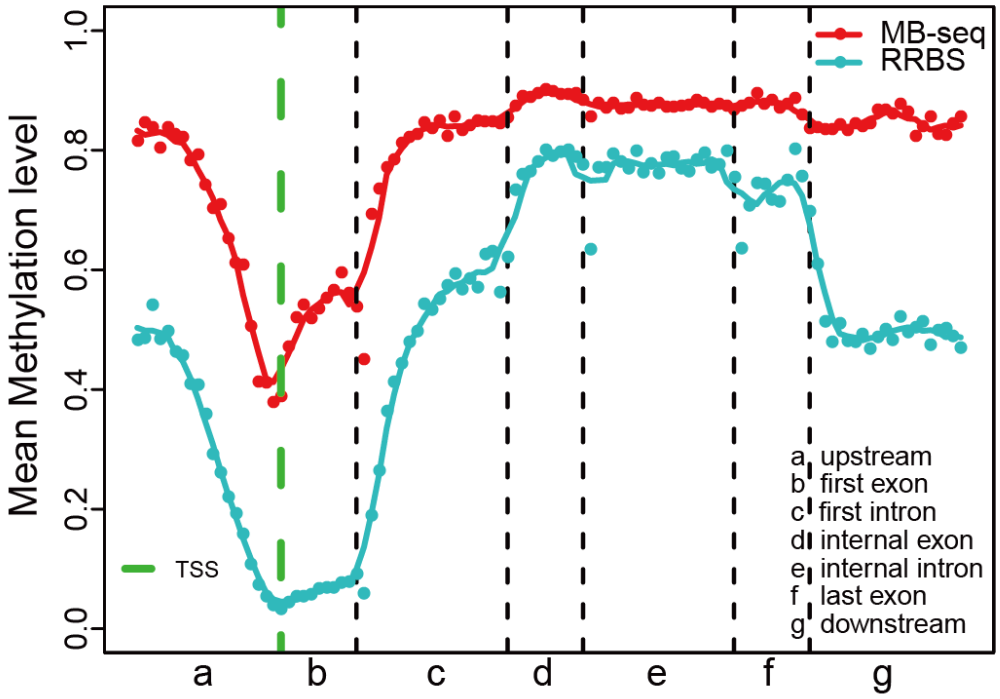
A



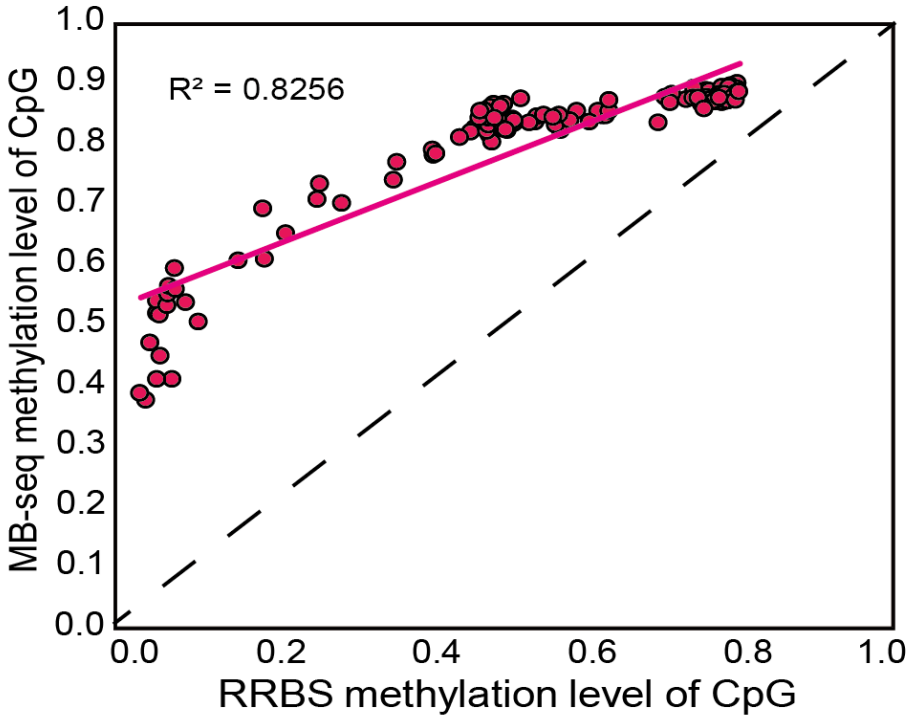
B



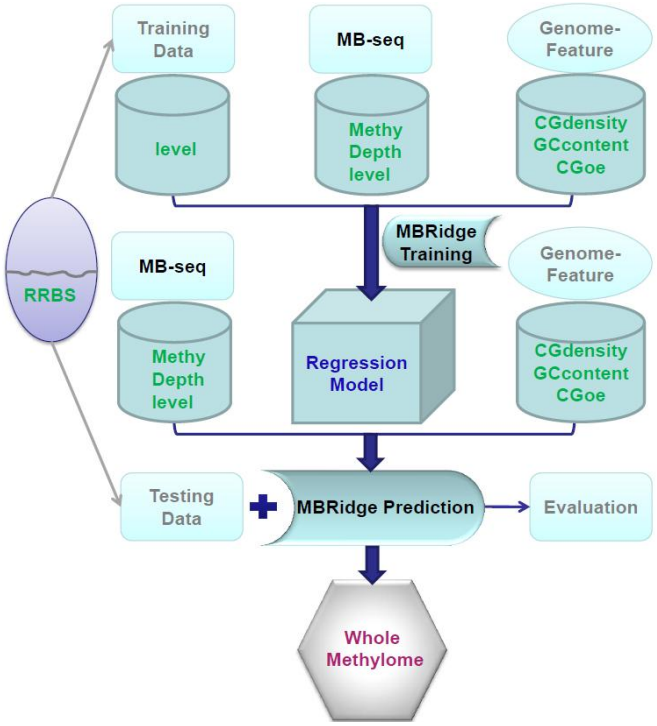
C


Mean Methylation level
0.0
0.2
0.4
0.6
0.8
1.0
MB-seq
RRBS
TSS
a upstream
b first exon
c first intron
d internal exon
e internal intron
f last exon
g downstream
a
b
c
d
e
f
g

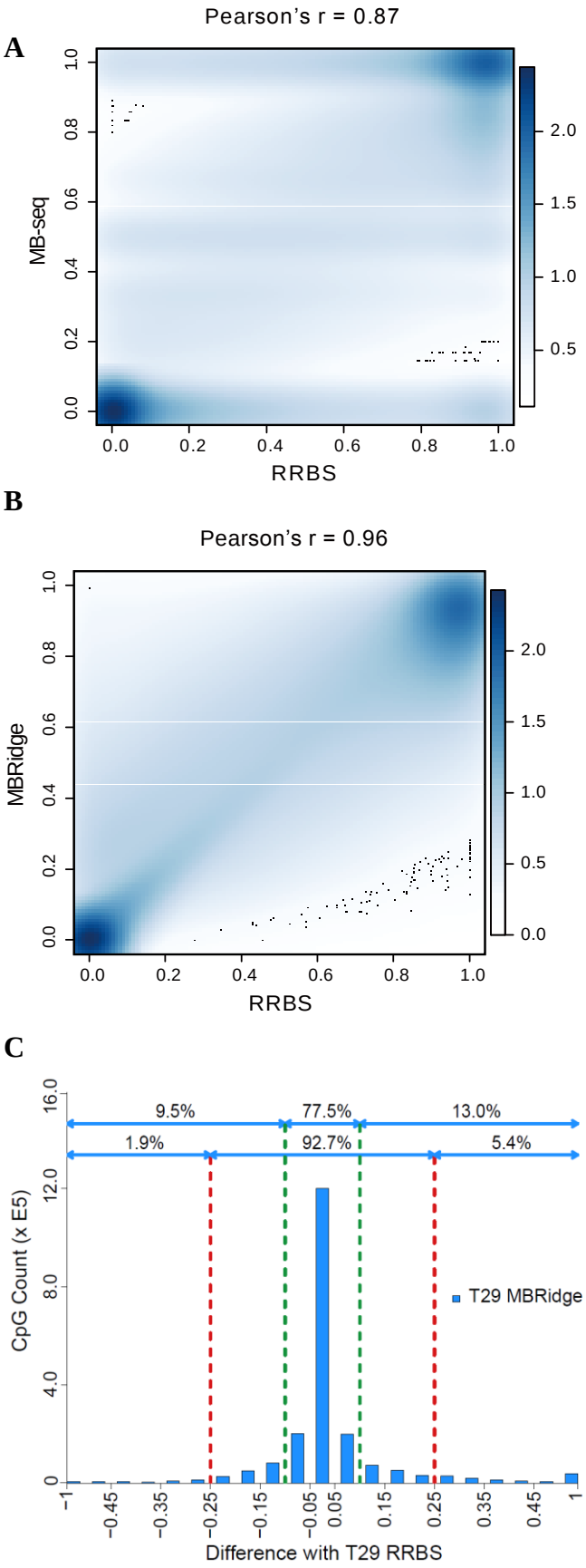
D


MB-seq methylation level of CpG
RRBS methylation level of CpG
R² = 0.8256
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

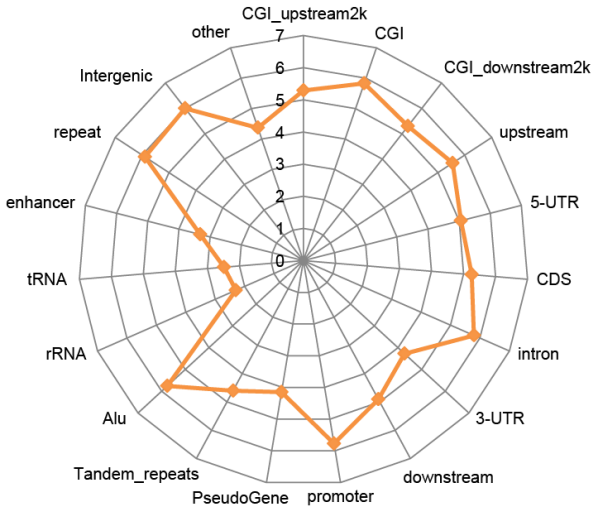
E



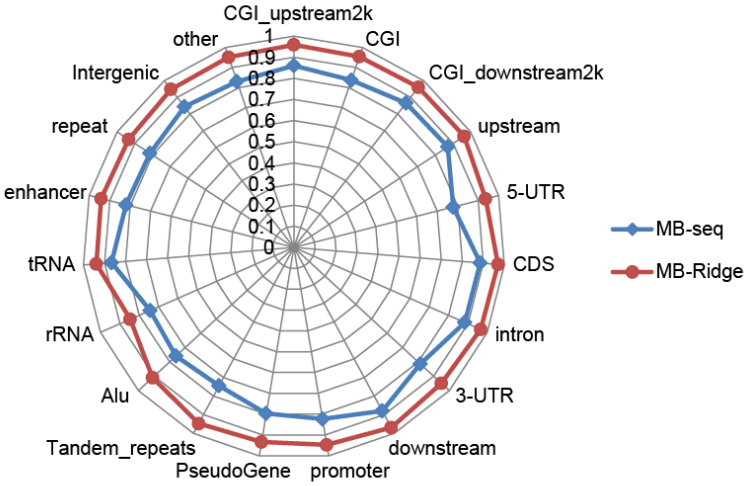
Supplementary Figure S3. Comparison of DNA methylation related information between RRBS and MB-seq. (A) Coverage of mCpG with different methylation level by MB-seq, MeDIP-seq and RRBS. Total mCpGs were split into 10 segments in x axis which based on the DNA methylation level measured by MethylC-seq (start with 5%, more 5% is defined methylated CpG). The coverage of mCpG of MB-seq, MeDIP-seq and RRBS at each segment of mCpGs with 10 different methylation levels were colored with green, blue and medium-orchid respectively. **(B)** Barplot representing the fraction of CpGs covered by MeDIP-seq for which the different methylation levels were measured by MethylC-seq. **(C)** Average methylation level of CpG in gene-associated regions using MB-seq and RRBS. Gene structure was divided into seven different functional regions and shown on x-axis. The y-axis was the average methylation level of CpGs in a 200 bp window. The green vertical line showed the mean location of the transcription start sites (TSS).The mean methylation level of MB-seq and RRBS were colored with red and cyan, respectively. Canonical gene was defined by seven different functional regions by black dash line. Green dash line represented the location of TSS. **(D)** Pearson correlation of the methylation level (average methylation level of CpGs in a 200 bp window from B) between MB-seq and RRBS. **(E)** Workflow of MBRidge.


A
Pearson's r = 0.87
MB-seq
RRBS
B
Pearson's r = 0.96
MBRidge
RRBS
C
9.5%
77.5%
13.0%
1.9%
92.7%
5.4%
CpG Count (x E5)
T29 MBRidge
Difference with T29 RRBS

D

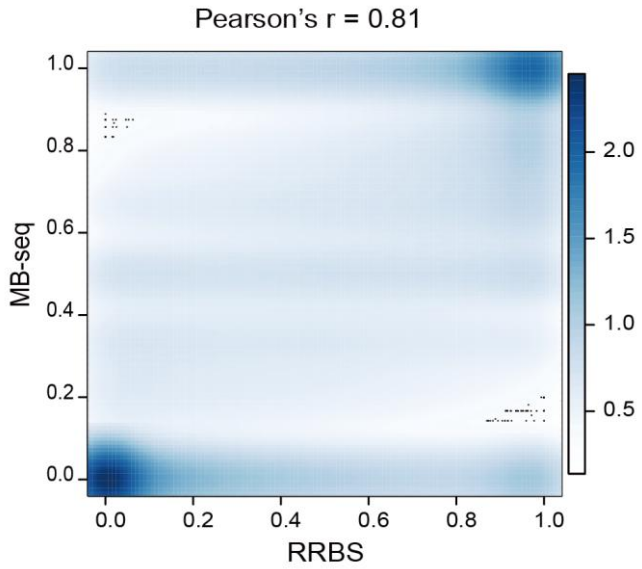


E

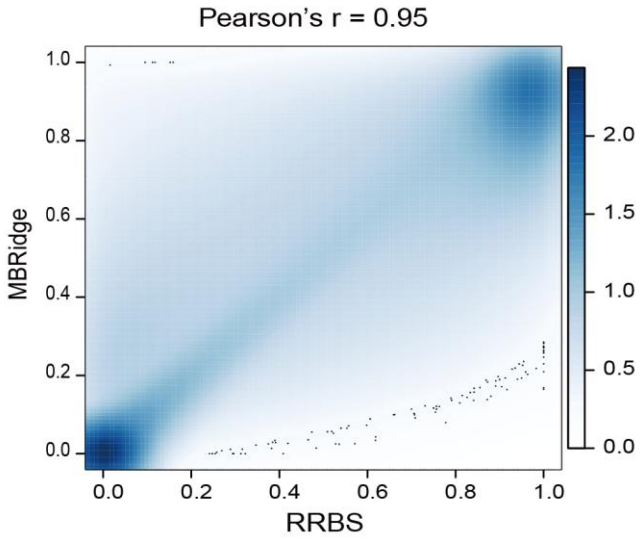


Supplementary Figure S4. Comprehensive evaluation for accuracy of before and after ridge regression for methylation levels from MB-seq in T29 cell line. (A-B) Scatter plots of PCC for comparison of methylation level through MB-seq and MBRidge with RRBS, respectively. The bar displayed a legend/color scale that describes the relative difference in numeric terms between different shades. The black dots exhibit the 100 most "sparse" points plotted over the smoothed density plot. **(C)** The number of CpGs as a function of the difference between MethylC-seq and MBRidge methylation levels—the two agree within 25% for 92.7% of the CpGs and within 10% for 77.5% of the CpGs. **(D)** The number of CpGs used for each comparison with RRBS on a log10 scale. **(E)** PCC for comparison of MBRdige, MB-seq with RRBS for annotated genomic features. Colour key presented methods: blue for MB-seq; red for MBRidge. The axes in each radar chart represented annotated genomic features.

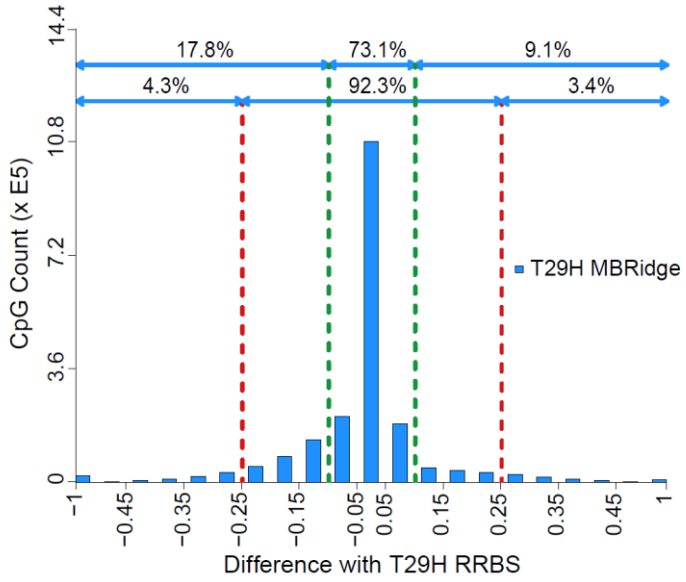
A


Pearson's r = 0.81
MB-seq
RRBS
1.0
0.8
0.6
0.4
0.2
0.0
2.0
1.5
1.0
0.5

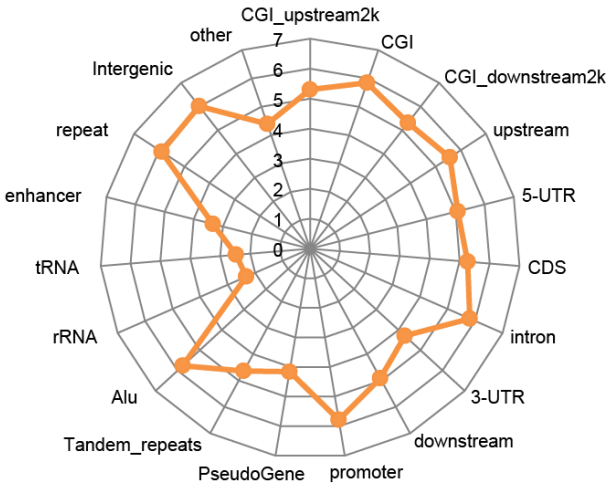
B


Pearson's r = 0.95
MBRidge
RRBS
1.0
0.8
0.6
0.4
0.2
0.0
2.0
1.5
1.0
0.5
0.0

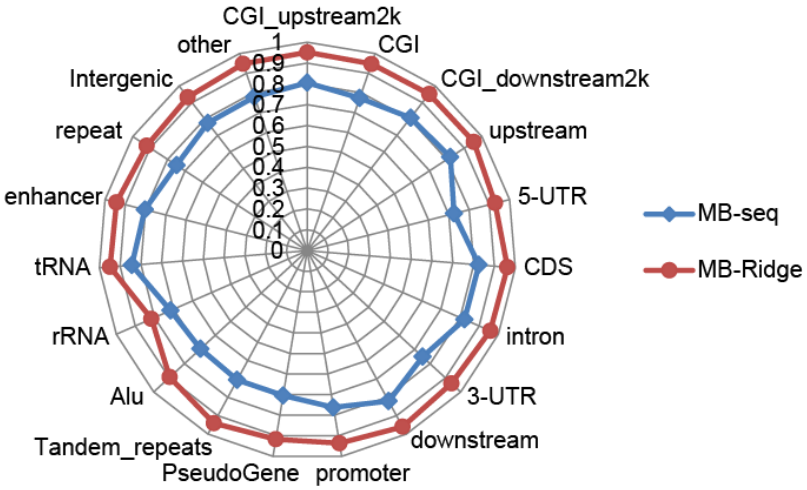
C


17.8%
73.1%
9.1%
4.3%
92.3%
3.4%
CpG Count (x E5)
14.4
10.8
7.2
3.6
0
T29H MBRidge
−1
−0.45
−0.35
−0.25
−0.15
−0.05
0.05
0.15
0.25
0.35
0.45
1
Difference with T29H RRBS

D

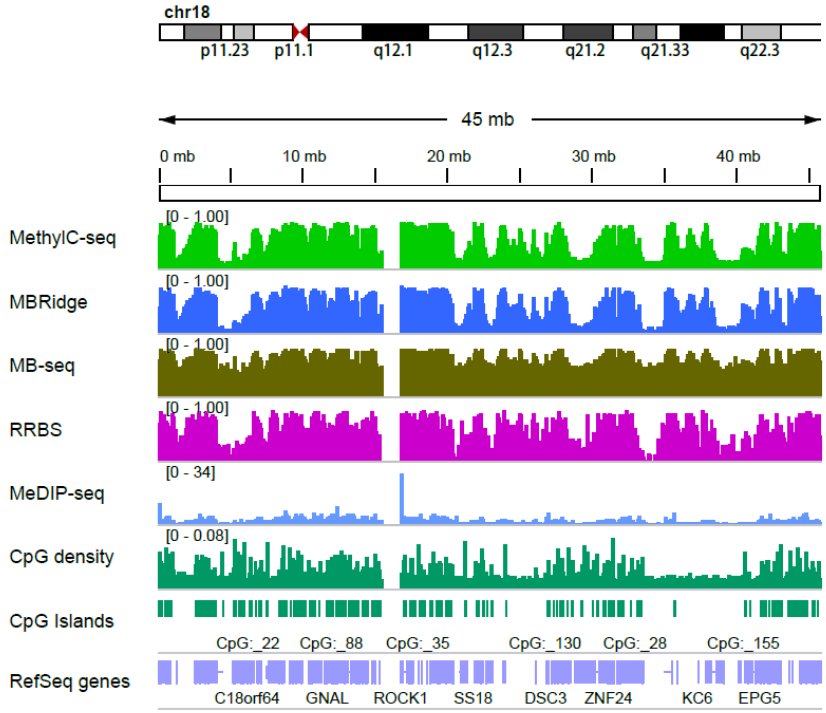


E

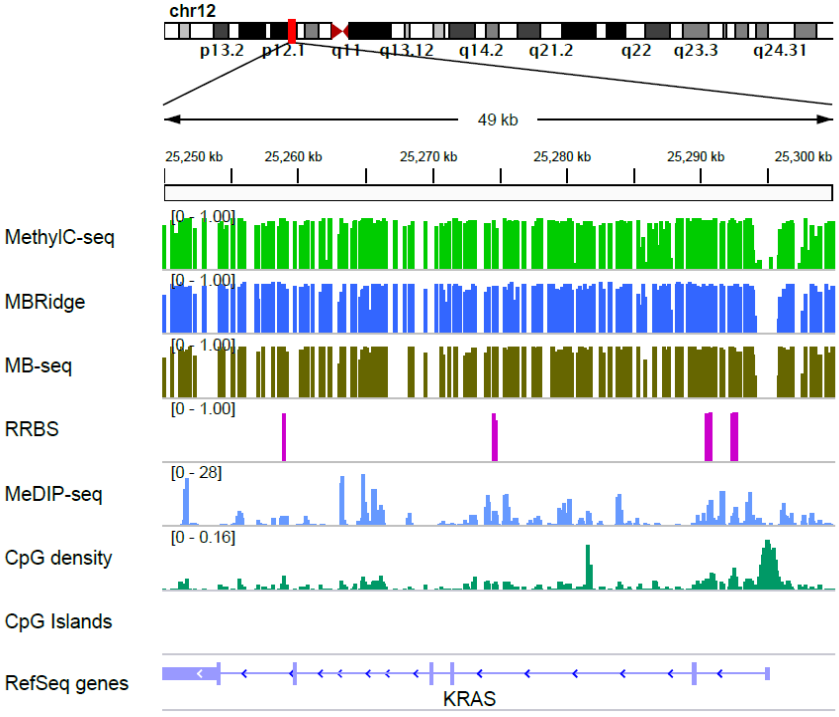


Supplementary Figure S5. Results of applying ridge regression to correct the methylation level from MB-seq in T29H cell line. (A-B) Scatter plots of PCC for comparison of methylation level through MB-seq and MBRidge with RRBS, respectively. The bar displayed a legend/color scale that describes the relative difference in numeric terms between different shades. The black dots exhibit the 100 most "sparse" points plotted over the smoothed density plot. **(C)** The number of CpGs as a function of the difference between RRBS and MBRidge methylation levels (T29H) —the two agree within 25% for 92.3% of the CpGs and within 10% for 73.1% of the CpGs. **(D)** The number of CpGs used for each comparison on a log10 scale. **(E)** PCC for comparison of MBRdige, MB-seq with RRBS for annotated genomic features. Color key presented methods: blue for MB-seq; red for MBRidge. The axes in each radar chart represented annotated genomic features.

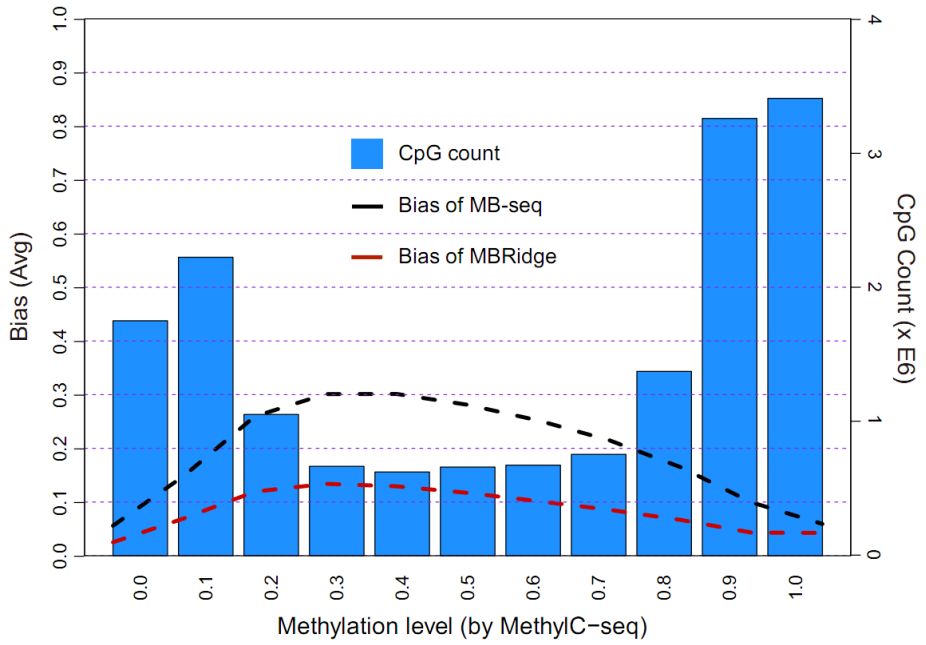
A



B

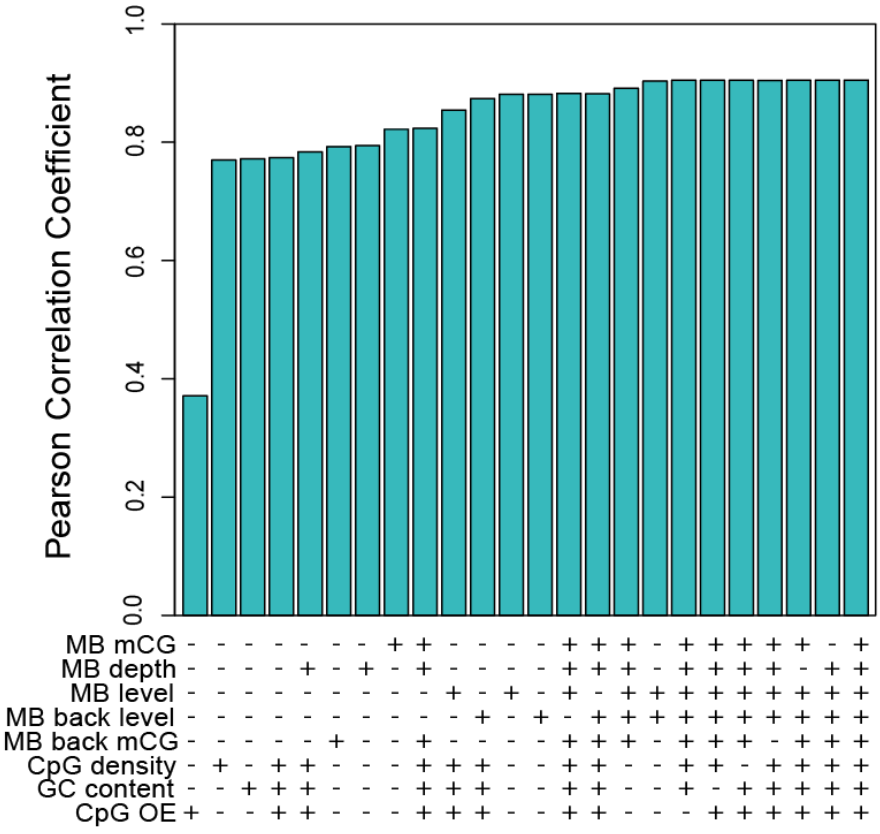


Supplementary Figure S6. Illustration of methylation level measured by different methods in T29 cell line. (A-B) The methylation level or coverage of MethylC-seq, MBRidge, MB-seq, RRBS and MeDIP-seq were displayed in different colored track across chromosome 18 (A) or gene KRAS (B).

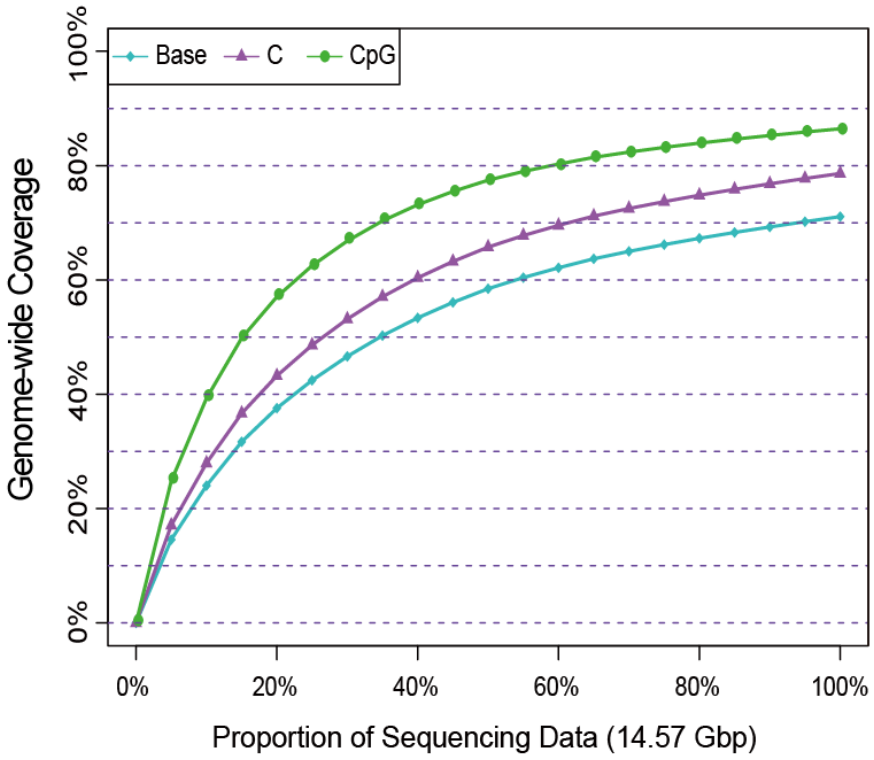


Supplementary Figure S7. Comprehensive evaluation of the methylation level derived from MB-seq and MBRidge in comparison with MethylC-seq in T29 cell line. Bias of MB-seq and MBRidge compared with MethylC-seq at different methylation level. Black line represented the mean difference value between MB-seq and MethylC-seq (left y-axis). Red line represented the mean difference value between MBRidge and MethylC-seq (left y-axis). Skyblue bars represented coverage of CpG at different methylation level (right y-axis).

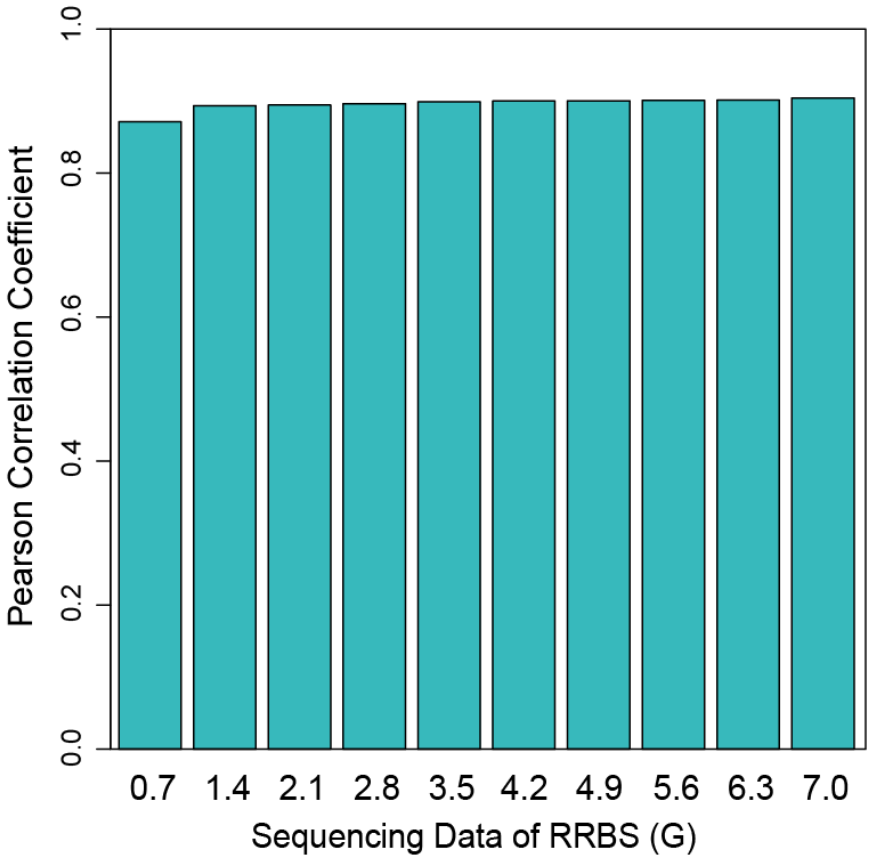
A



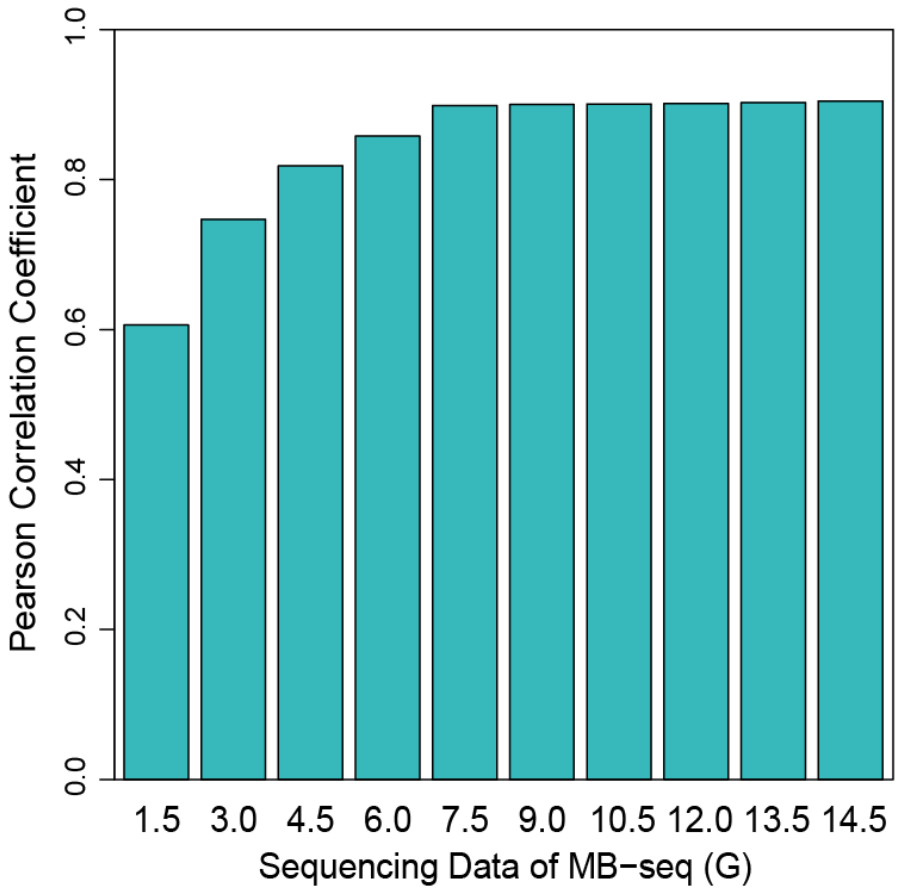
B


Base
C
CpG
Genome-wide Coverage
0%
20%
40%
60%
80%
100%
Proportion of Sequencing Data (14.57 Gbp)

C

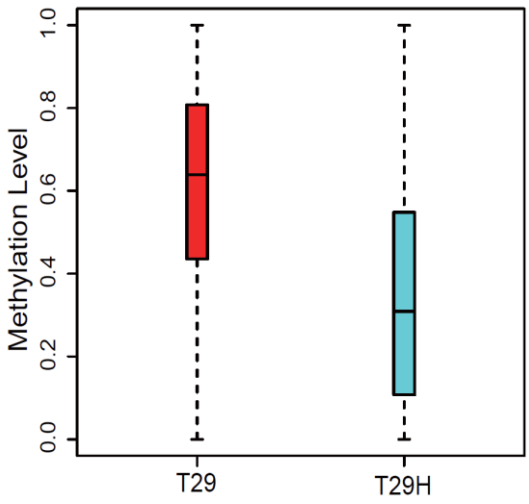

Pearson Correlation Coefficient
0.0
0.2
0.4
0.6
0.8
1.0
0.7
1.4
2.1
2.8
3.5
4.2
4.9
5.6
6.3
7.0
Sequencing Data of RRBS (G)

D

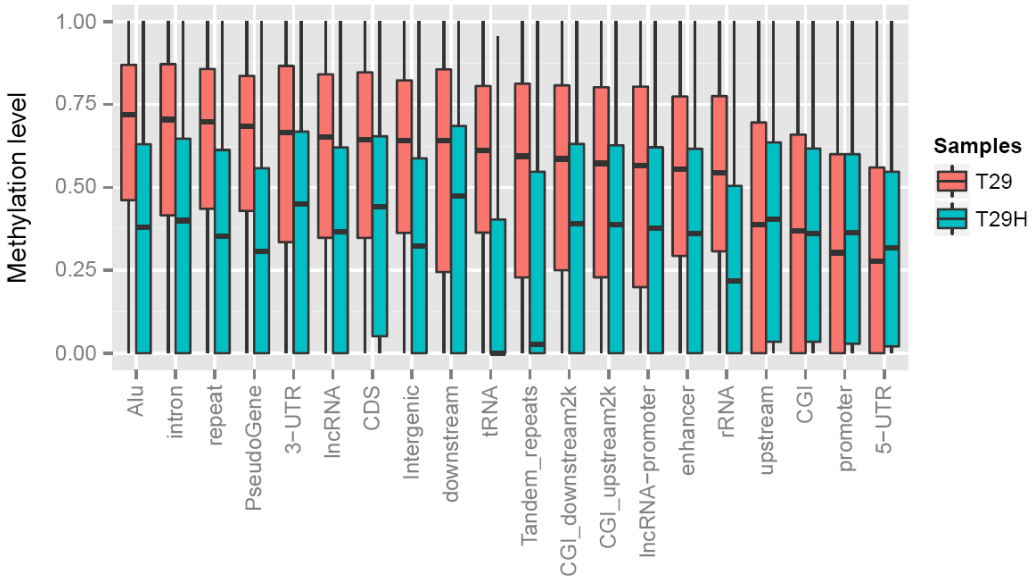


Supplementary Figure S8. Influence of predictors and data volume to MBRidge. **(A)** Bar plot representing the **c**ontribution of each predictor used in MBRidge by permutation test. "+" shows that the predictor was included while "-" shows that the predictor was excluded. PCC values in MBRidge compared with MethylC-seq were plotted on the y-axis **(B)** Genome-wide coverage of base, C, and CpG covered by at least 1 × reads, with increasing size of data set. Each curve represented base (cyan), C (medium orchid), and CpG (green). **(C–D)** Bar plot representing the accuracy of MBRidge for different randomly-selected MB-seq and RRBS data sets. PCC values for MBRidge vs. MethylC-seq were indicated on the y-axis.

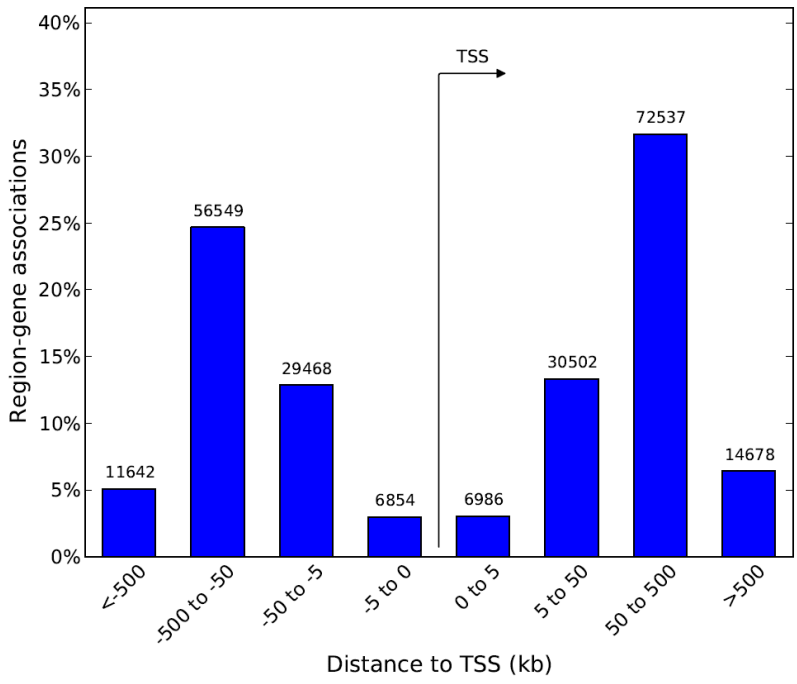
A


Methylation Level
0.0
0.2
0.4
0.6
0.8
1.0
T29
T29H

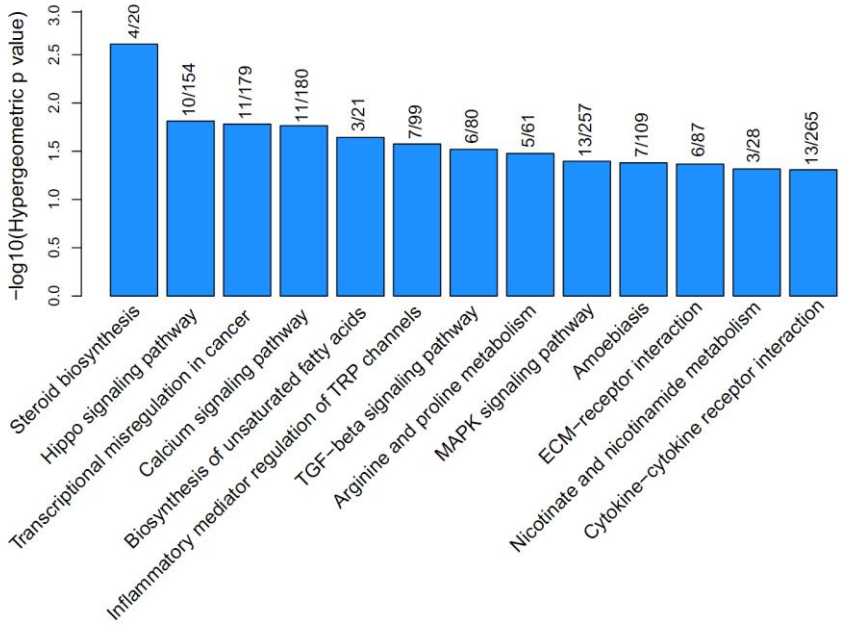
B


Methylation level
1.00
0.75
0.50
0.25
0.00
Alu
intron
repeat
PseudoGene
3−UTR
lncRNA
CDS
Intergenic
downstream
tRNA
Tandem_repeats
CGI_downstream2k
CGI_upstream2k
lncRNA−promoter
enhancer
rRNA
upstream
CGI
promoter
5−UTR
Samples
T29
T29H

C

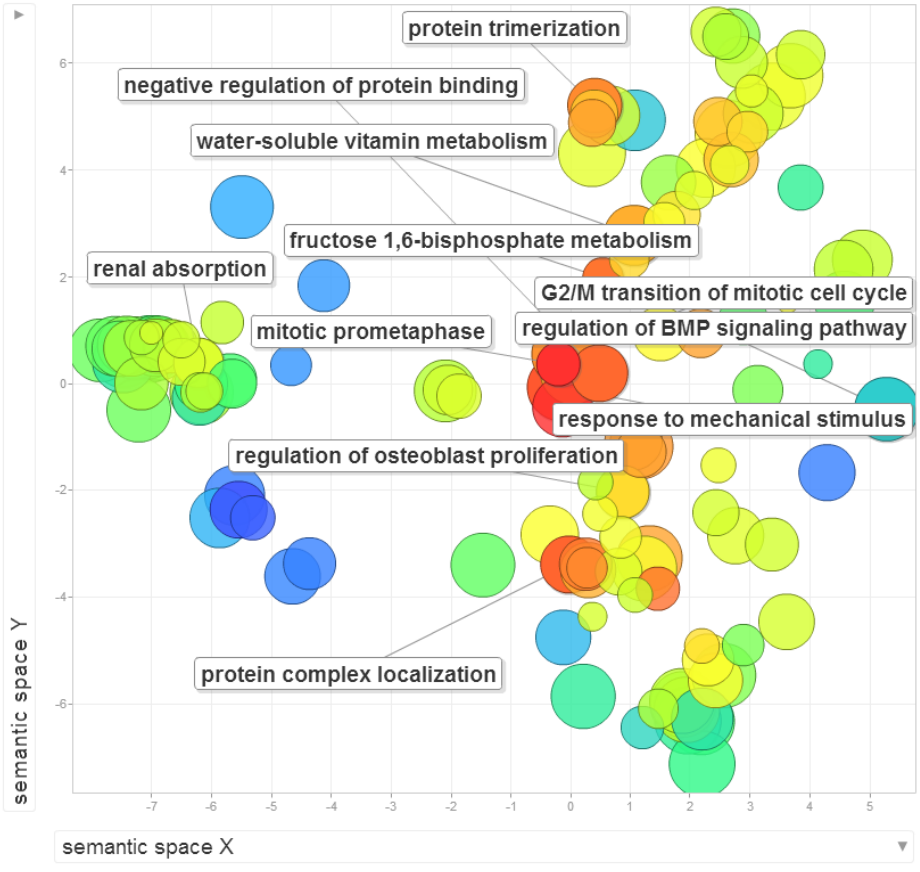

Region-gene associations
40%
35%
30%
25%
20%
15%
10%
5%
0%
TSS
11642
56549
29468
6854
6986
30502
72537
14678
<-500
-500 to -50
-50 to -5
-5 to 0
0 to 5
5 to 50
50 to 500
>500
Distance to TSS (kb)

D

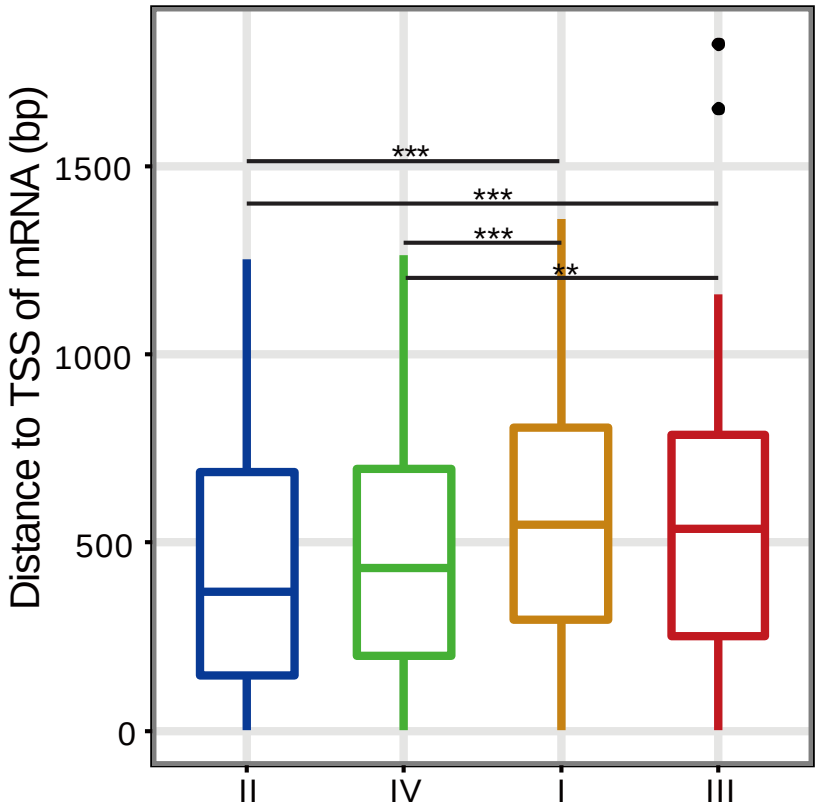


Pathway enrichments of methylation-regulated protein-coding genes

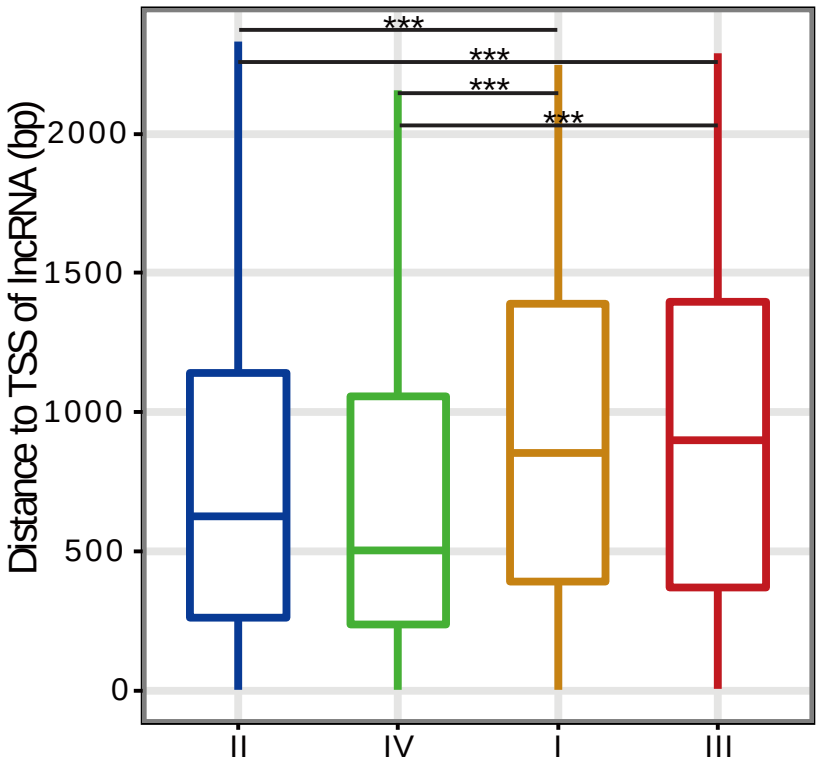
E



F

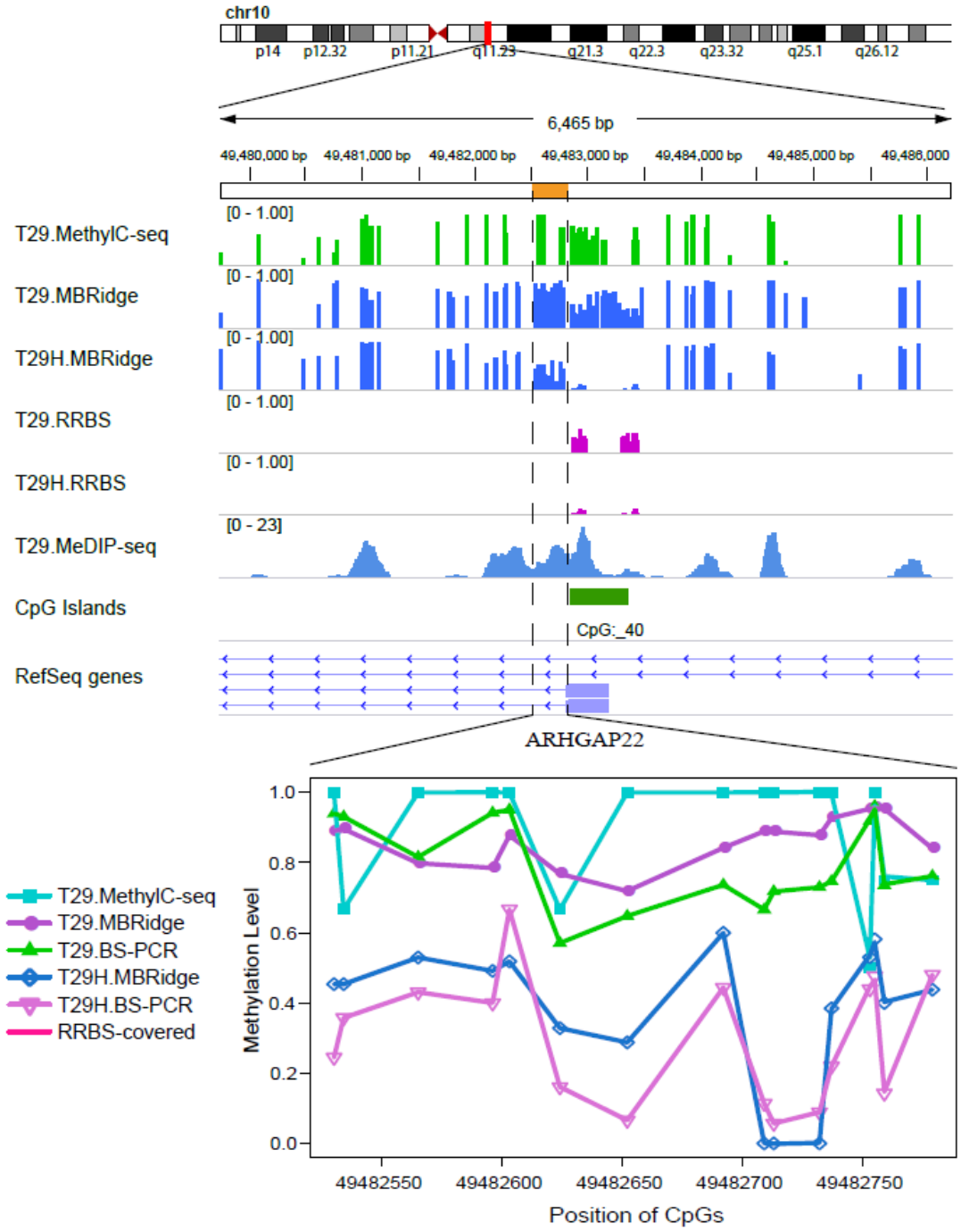

Distance to TSS of mRNA (bp)
1500
1000
500
0
***
***
***
**
II
IV
I
III

G

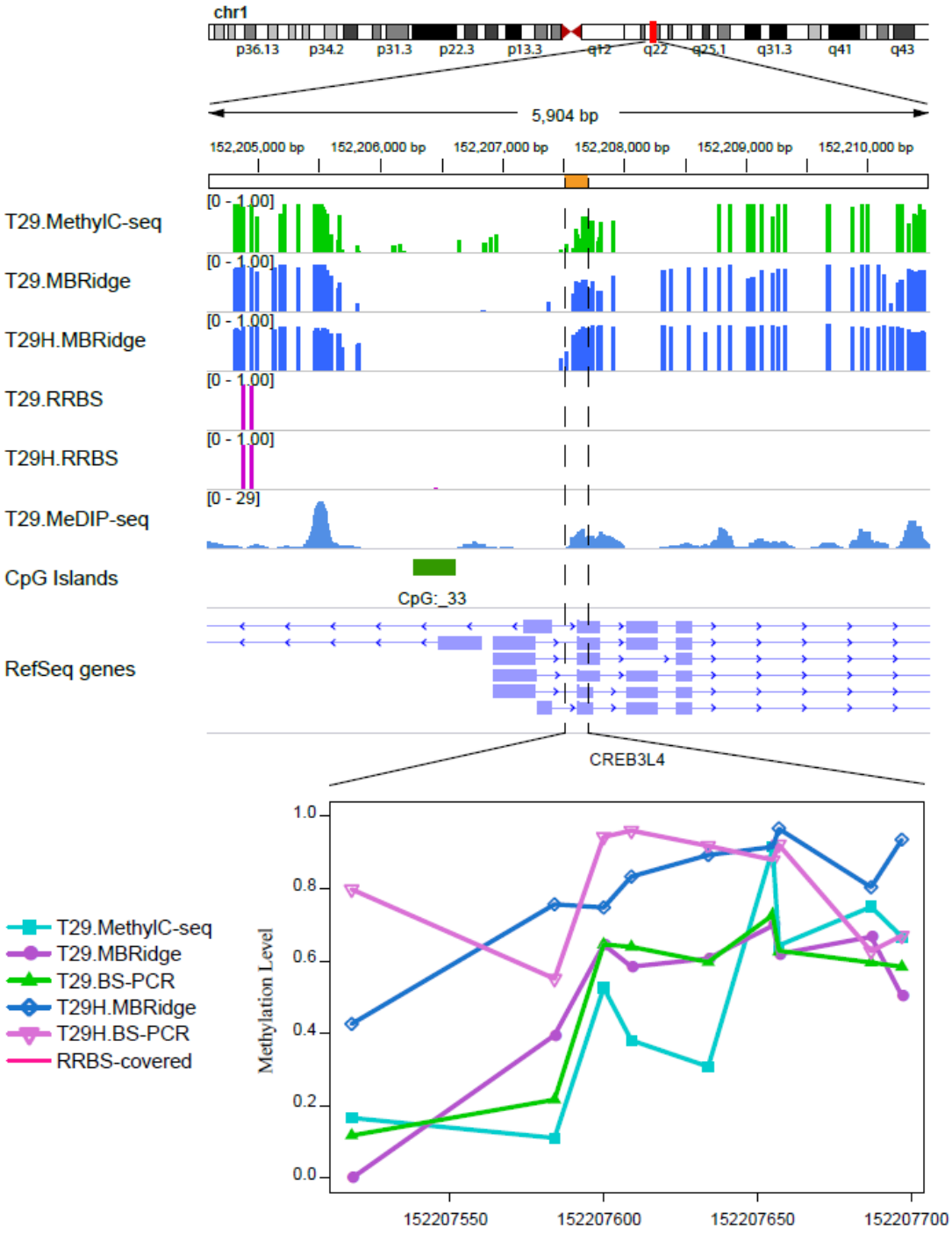

Distance to TSS of lncRNA (bp)
2000
1500
1000
500
0
***
***
***
***
II
IV
I
III

Supplementary Figure S9. Characteristics of DMRs derived from MBRidge data between T29 and T29H. (A-B) Boxplots of methylation levels of T29 (red) and T29H (violet) in DMR regions for genome-wide (A) and variant elements (B), respectively. The middle "box" represented the middle 50% of the dataset. Centre lines in the boxes represented medians and lines on both sides of the boxplot indicate dispersion for 99% of values. **(C)** Distribution of DMRs by orientation and distance to TSS. The distances of DMRs to TSS were showed in x-axis and the corresponding numbers of DMRs were showed in y-axis. The arrow represented the transcriptional orientation of genes. **(D)** Barplot showed the hierarchical order of the pathway enrichment of the inversely methylation-regulated genes based on their enrichment scores (-log [hypergeometric p value]). The fraction number upstream of each bar indicateed the percentage of genes overlapped with DMRs in each term of pathway. **(E)** Analysis of gene ontology for inversely methylation-regulated lncRNA genes by the tool REVIGO. **(F-G)** Boxplots of distances between center of DMRs and TSS for mRNA (F) or lncRNA (G) in four categories. The four categories were: (I) genes which significant increasing expression were positively correlated with significant increasing methylation; (II) genes which significant increasing expression were inversely correlated with significant decreasing methylation; (III) genes which significant decreasing expression were positively correlated with significant decreasing methylation; (IV) genes which significant decreasing expression were positively correlated with significant increasing methylation. The black dots denoted members outside of distribution range for 99% of values. Statistical analysis was performed using t-test for each comparison. Stars indicated level of significance by p-value (*<0.01, ** < .001, *** < .0001).

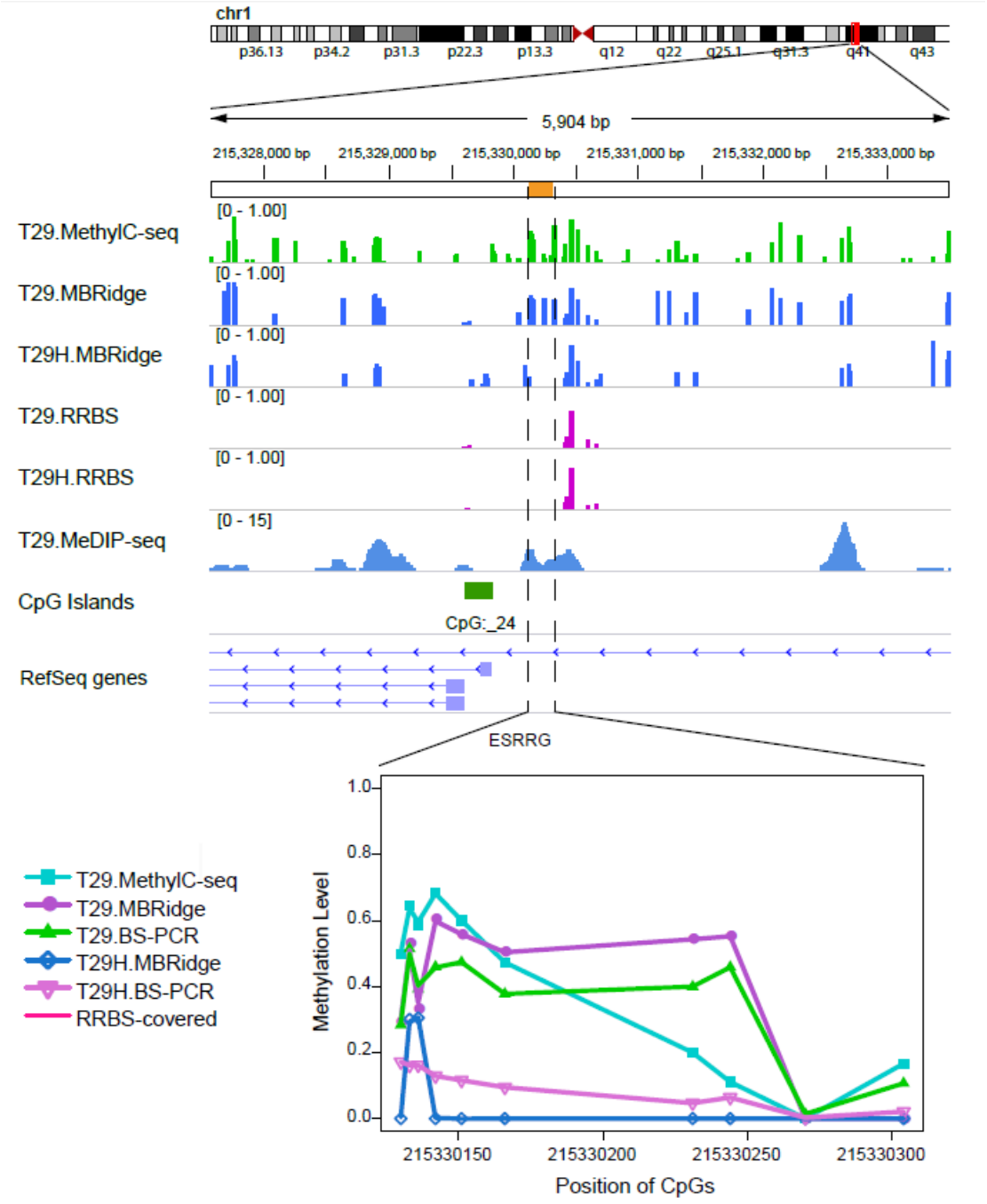
A. chr10:49482530-49482779 (ARHGAP22)



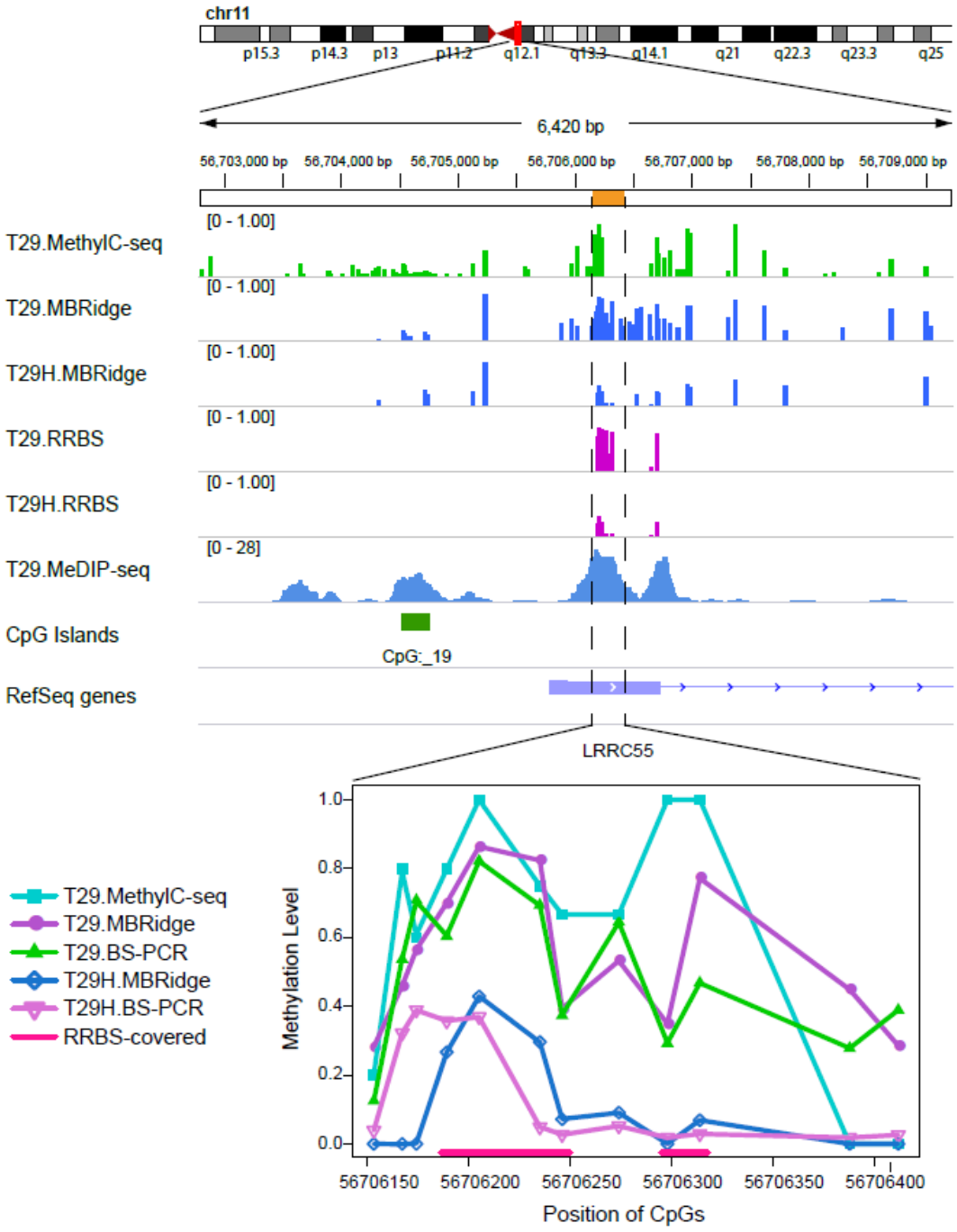
B. chr1: 152207518-152207697 (CREB3L4)



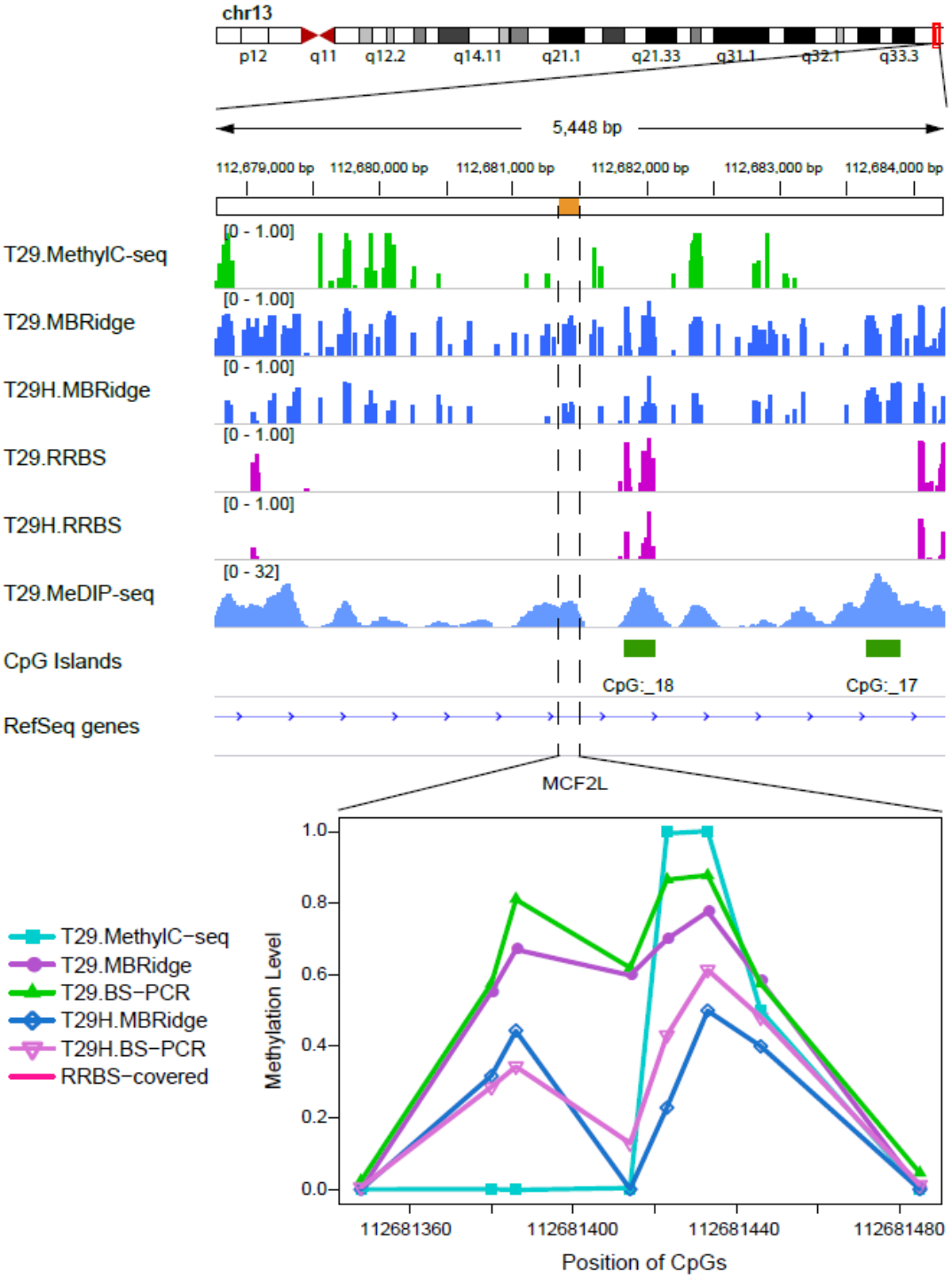
C. chr1: 215330130-215330304 (ESRRG)



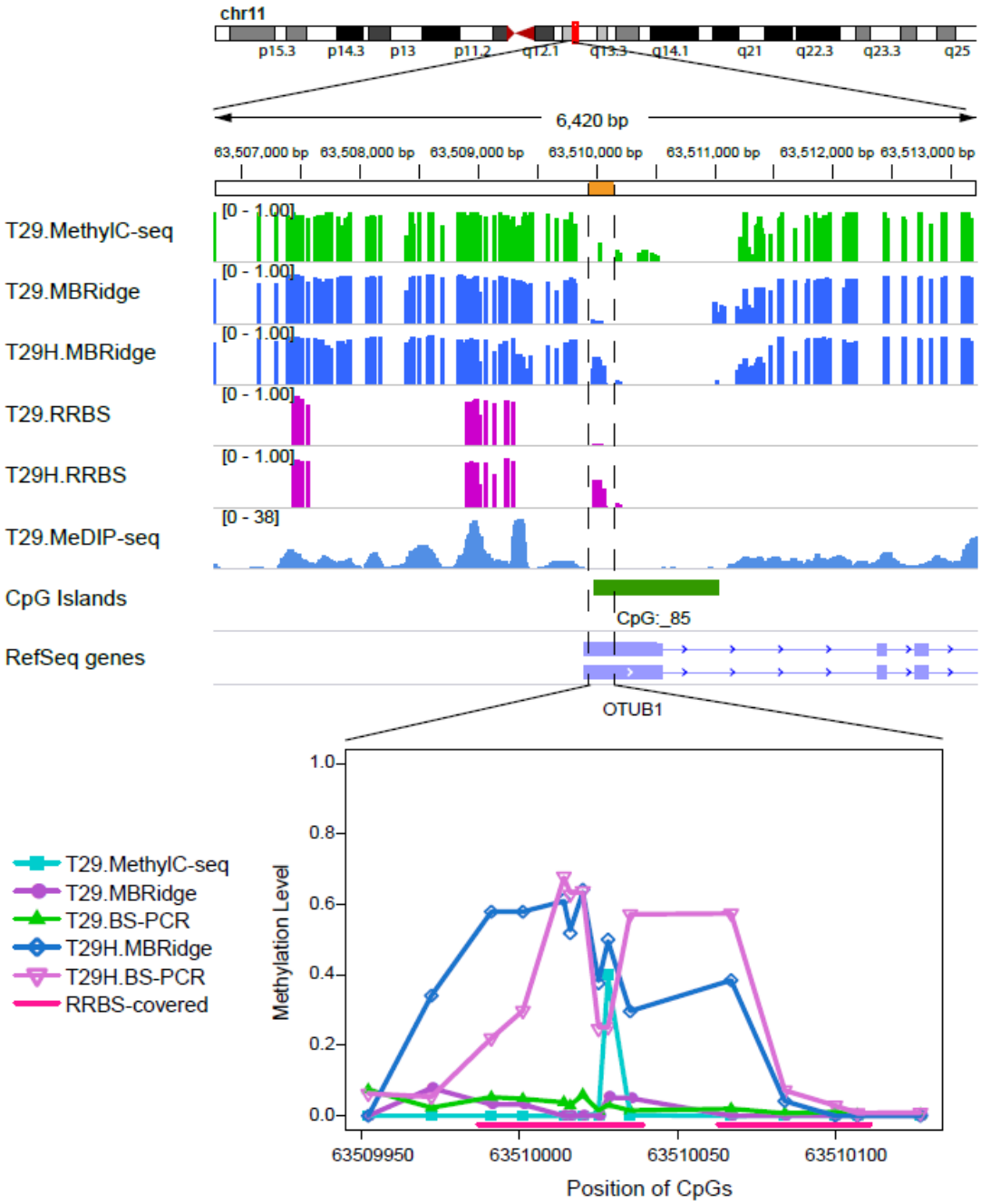
D. chr11: 56706153-56706412 (LRRC55)



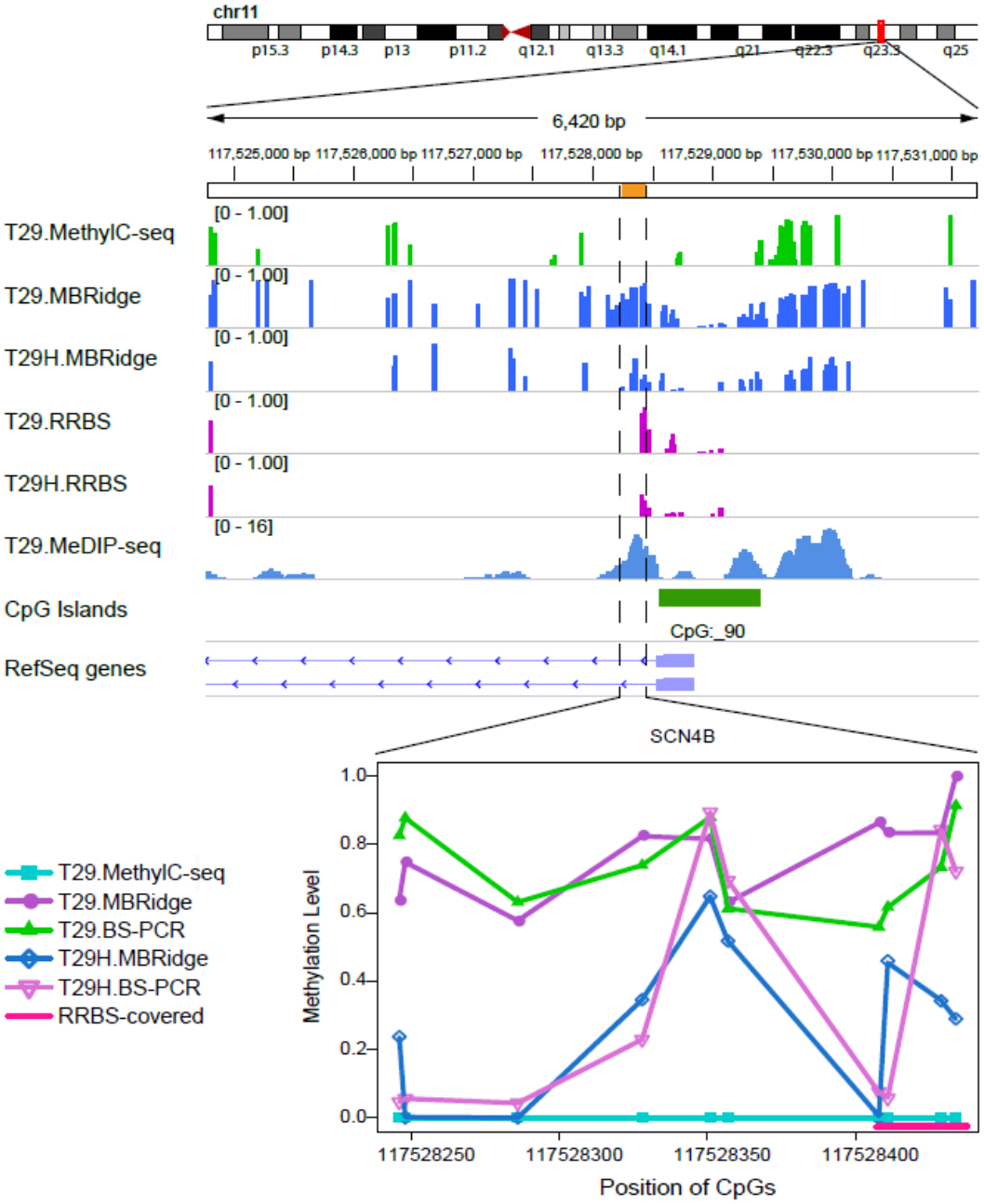
E. chr13: 112681348-112681485 (MCF2L)



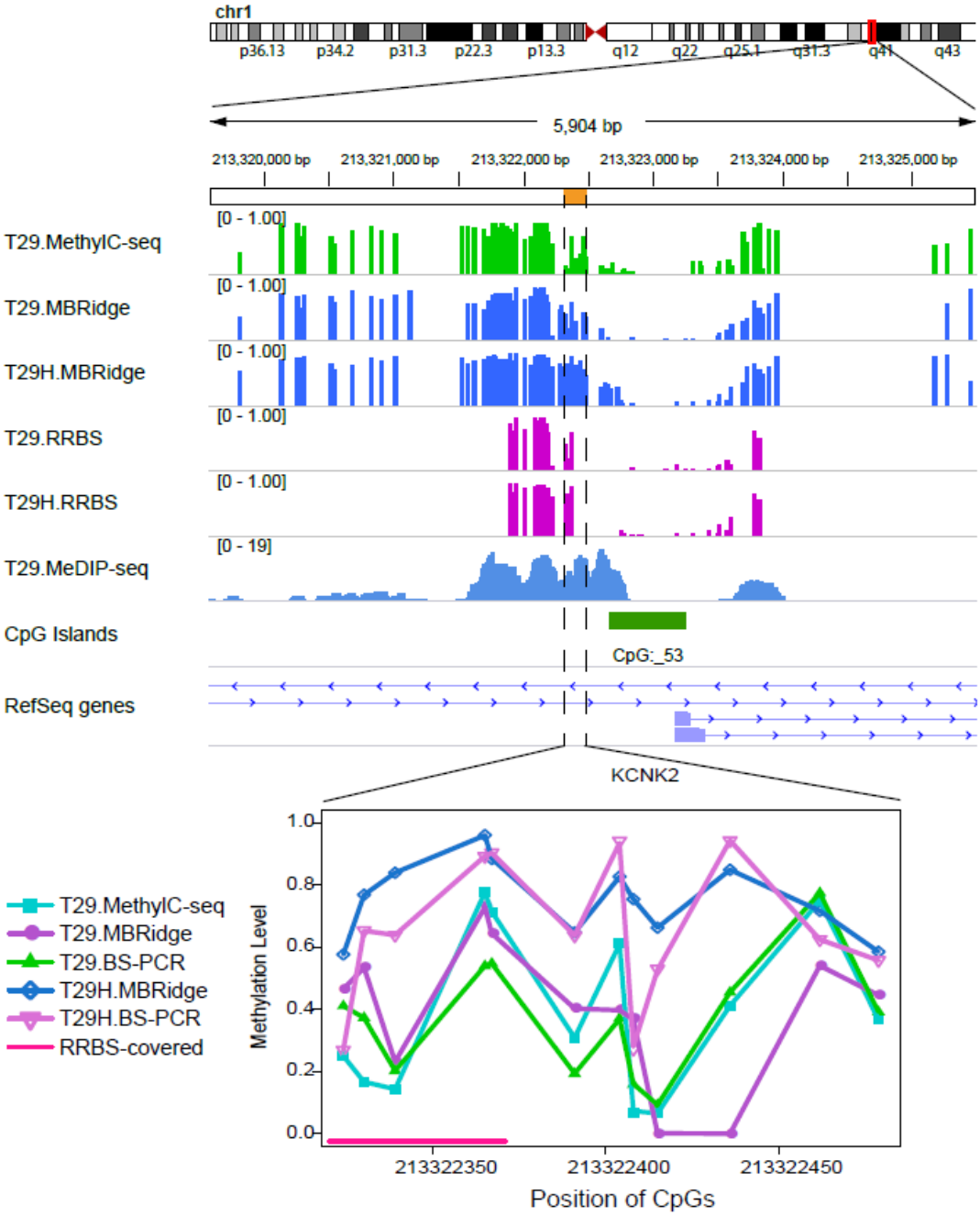
F. chr11: 63509952-63510127 (OTUB1)



G. chr11: 117528246-117528434 (SCN4B)



H. chr1: 213322324-213322479 (KCNK2)



Supplementary Figure S10. Experimental validation of DMRs. Genome browser views and line graphs of locations validated by location-specific bisulfite sequencing for DMRs between T29 and T29H (Fig.S10 A-H). Two additional locations were presented in the main text.

Supplementary References

Down, T.A., Rakyan, V.K., Turner, D.J., et al. (2008). A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nature Biotechnology *26*, 779-785.

Kent, W.J., Sugnet, C.W., Furey, T.S., et al. (2002). The human genome browser at UCSC. Genome research *12*, 996-1006.

Lister, R., Pelizzola, M., Dowen, R.H., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315-322.

Maunakea, A.K., Nagarajan, R.P., Bilenky, M., et al. (2010a). Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature *466*, 253-257.

Maunakea, A.K., Nagarajan, R.P., Bilenky, M., et al. (2010b). Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature *466*, 253-U131.