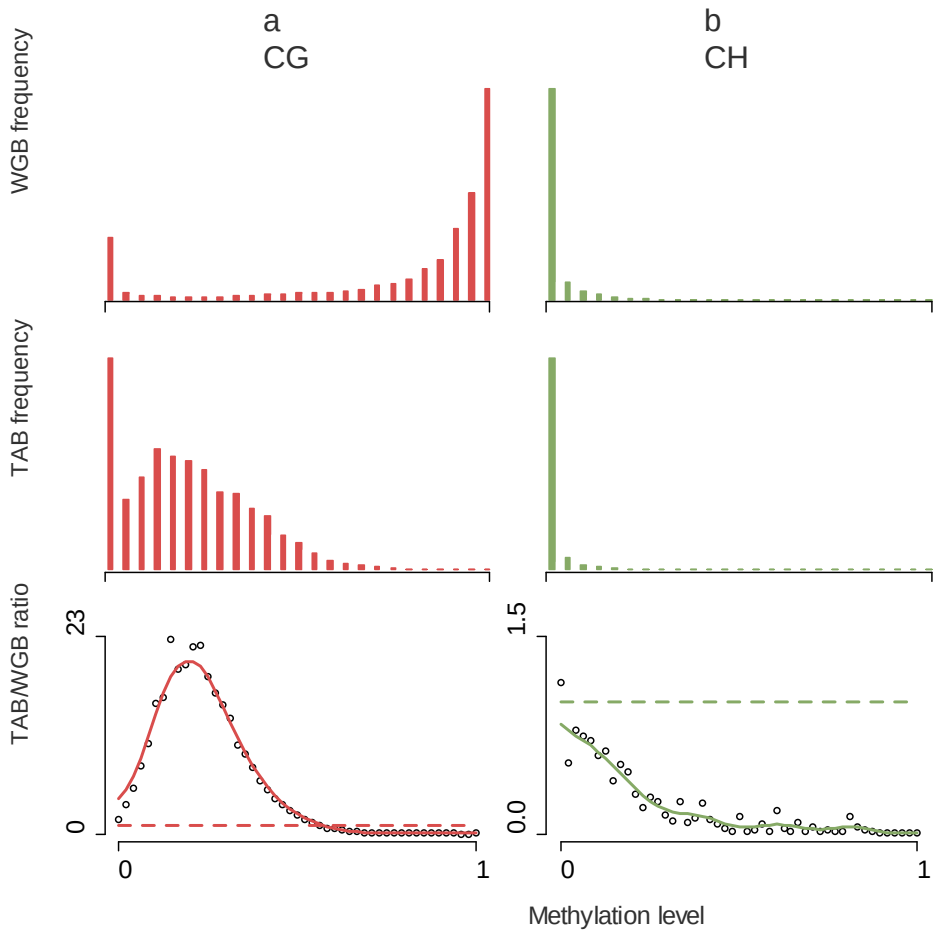
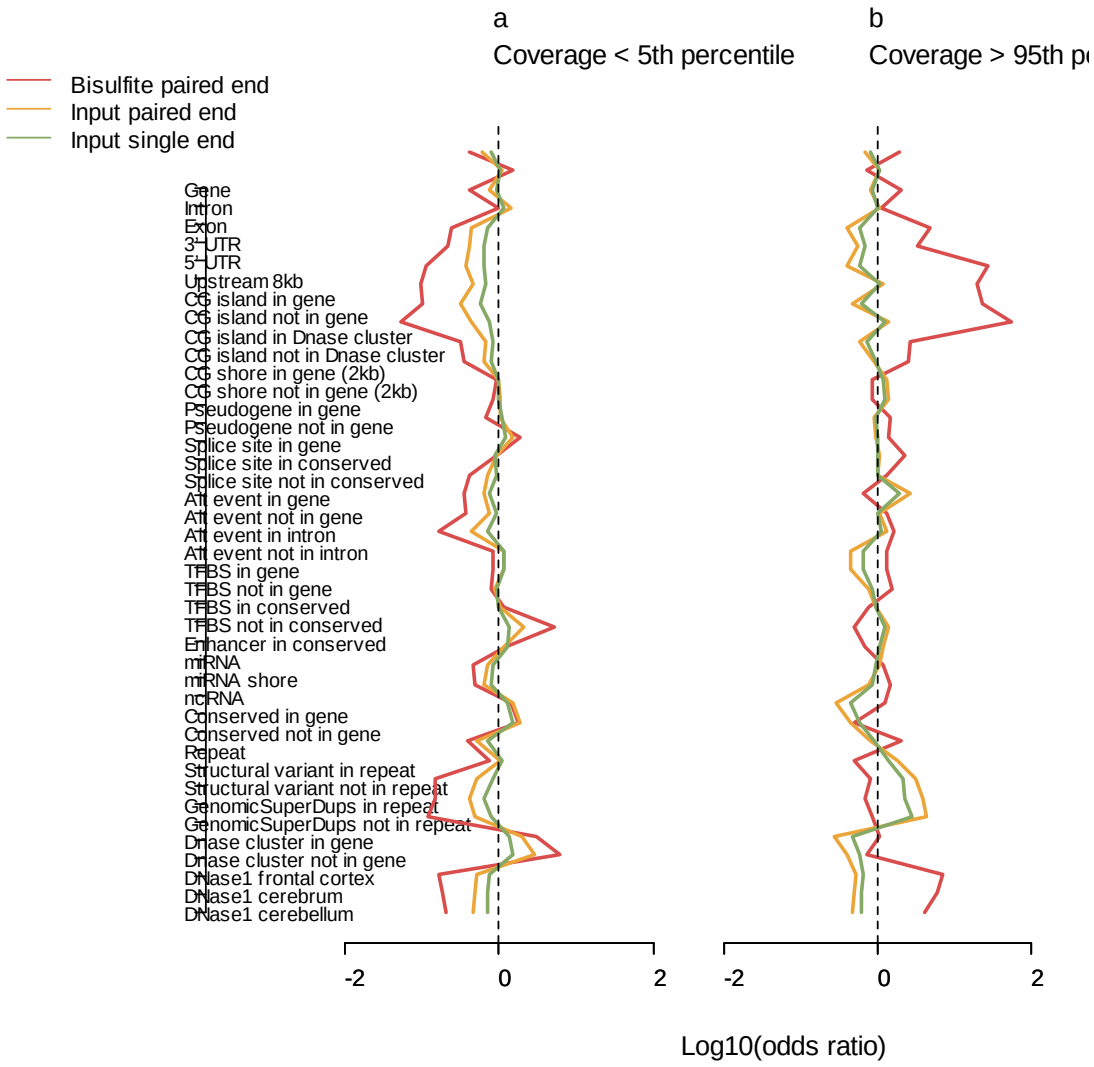
Enrichment methods provide a feasible approach to comprehensive and adequately powered investigations of the brain methylome

Supplemental figures and tables

Table of Contents

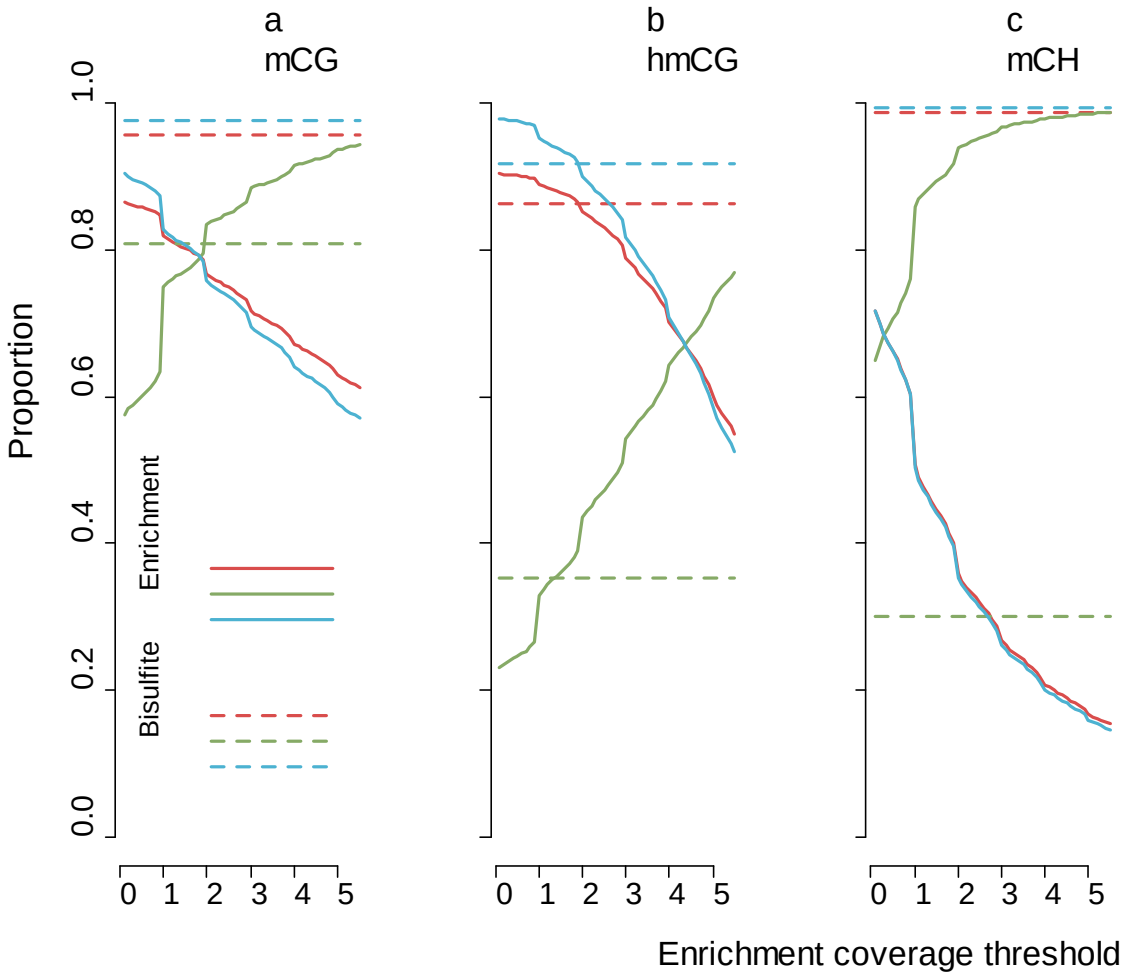


Supplemental Figure 1 WGB and TAB methylation estimates. Top two figures are histograms of WGB (mC + hmC) and TAB (hmC) methylation estimates. The bottom figures show for each break point the WGB/TAB ratio of the proportions displayed in the histograms. A kernel regression smoother was used to fit a trend line to these ratios. The dashed lines indicate ratios of one, implying equal proportions of sites with TAB and WGB methylation. (a) Substantial amounts of WGB and TAB methylation were observed in the CG context. Whereas WGB methylation levels were generally high, TAB showed more modest methylation levels. (b) The vast majority of the cytosines outside the CG context were not methylated. Part of the very low levels of WGB and TAB methylation we observed outside the CG context may reflect “noise” in the data. The bottom figure shows that if sites were methylated, this was predominantly because of WGB (i.e., mCH) and not TAB (hmCH).

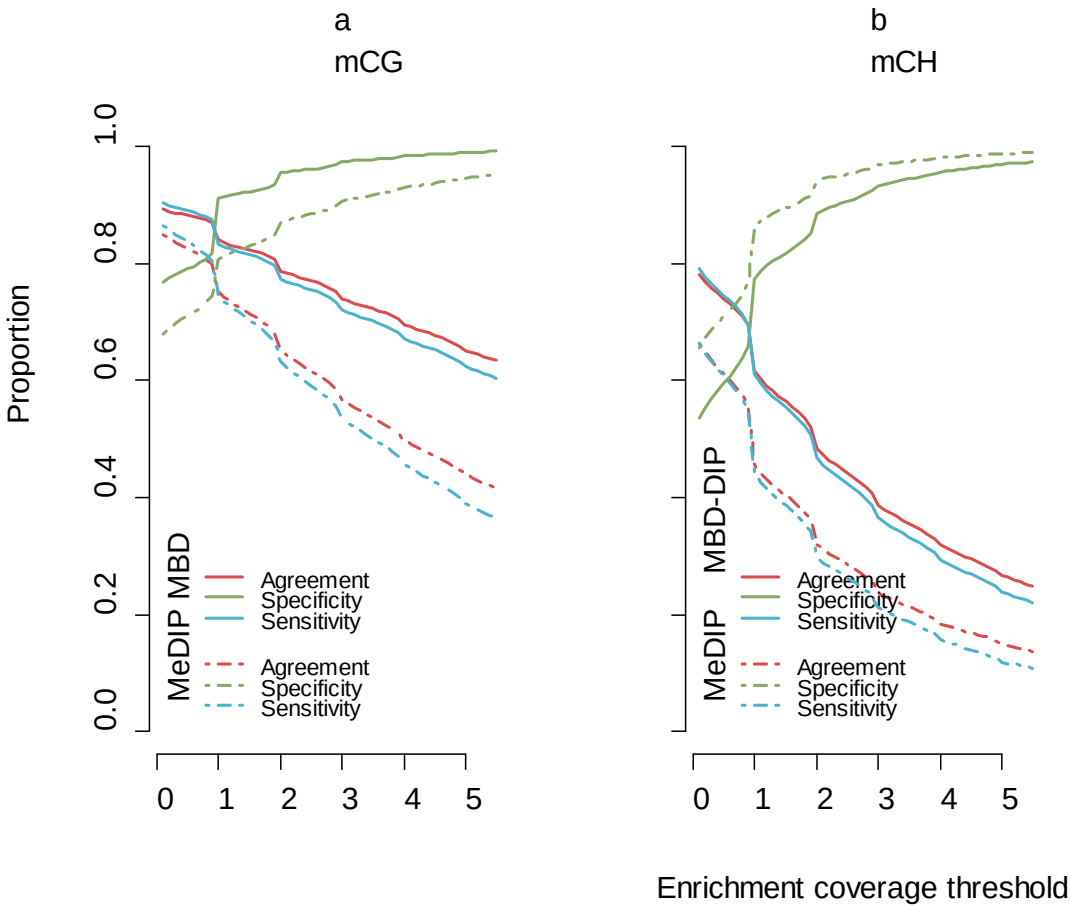


Supplementary Figure 2: Read coverage across genomic features. To study the possibility that mapping or library preparation may introduce differences between bisulfite and enrichment data, we plotted read coverage of genomic features for input and bisulfite samples. An input sample of unmodified DNA from the same subject was whole genome sequenced on the HiSeq X Ten System by Macrogen USA using 2x150 bp paired end libraries. To study the profiles of the single end libraries used for the enrichment methods, we also repeated the analysis using only data restricted to the first 75 bp of read one. Read coverage was calculated for all CGs and used to classify sites as either in the bottom (Coverage < 5th percentile) or top (Coverage > 95th percentile) 5%. Next, we calculated odds ratios to study whether sites located in the studied feature were more likely to have low or high coverage compared to sites not in this feature. The genomic regions used for comparisons were selected as described in section “Methylation profiles across genomic features”. An unbiased process should result in uniformly distributed read coverage across the genome. However, results show that this is not the case, where bisulfite methods are more severely biased. Moreover, profiles sometimes deviated for bisulfite and enrichment methods suggesting that some of the discrepancies observed between

the two methods may originate from this bias. Regions with too low read coverage are at risk for missing potentially relevant methylation sites. Panel a shows that this problem is more severe for bisulfite versus enrichment methods, and mainly involves regions that are CG dense or contain repetitive elements. Excess coverage means that reads are preferentially aligning to certain genomic regions thereby producing variation in the accuracy of methylation estimates (bisulfite methods) or erroneously suggesting the region is methylated (enrichment methods). Panel b shows that the bisulfite methods exhibited a bias for genic and CG dense regions, whereas the enrichment methods show substantially less genomic bias. The preference of bisulfite reads is likely results of the constitutively higher levels of methylation at these regions. Thus, these reads will more likely result in sequencer reads incorporating C base calls, which are easier to align due to higher homology to the native reference. As enrichment libraries contain inserts of native DNA, they are not affected by the loss of sequence complexity as observed in bisulfite treated DNA.



Supplementary Figure 3: Robustness of findings in neuronal cells. We used FACS to isolate the nuclei of neurons (NeuN+), and repeated all assays to study the robustness of our findings in bulk tissue. For this purpose, we calculated sensitivity (proportion of correctly identified methylated loci), specificity (proportion of correctly identified non-methylated loci), and overall agreement (the proportion of times the enrichment methods and bisulfite data arrived at the same conclusion regarding the methylation status). For all three forms of methylation (mCG, hmCG, and mCH in figures a, b, and c respectively), the overall pattern was very comparable to those obtained in bulk tissue (Figure 3a,b,c).



Supplementary Figure 4: Standard MeDIP versus MBD/MBD-DIP
To compare standard MeDIP study versus MBD/MBD-DIP we calculated their sensitivity, specificity, and overall agreement. Methylation status was determined using the bisulfite data. The calculation of the above indices was performed as described in **Fig. 3**. (**a**) For mCG, sensitivity/specific/agreement was consistently better for MBD versus standard MeDIP. (**b**) For mCH, sensitivity was substantially better for MBD-DIP, but at lower specificity than MeDIP. However, because the majority of sites contained methylated cystosines the overall agreement was consistently better for MBD-DIP.

Supplementary Table 1: Descriptive statistics for bisulfite assays

	WGB	TAB
Reads		
Total number reads (150 bp)	783,419,971	764,479,090
% read pairs aligned	67.3%	66.7%
Mappable mean depth (X)	27.7	26.8
CG sites		
Number of sites	26,216,350	26,288,978
% sites left after QC	92.7%	93.8%
CH sites		
Number of sites	1,021,843,871	1,023,407,003
% sites left after QC	84.2%	85.1%
Conversion controls		
Bisulfite conversion rate (CH to T)	99.2%	99.3%
Oxidation rate (mCG to T)	NA	96.9%
Normalized hmC protection rate	NA	91.4%

Note: All entries in the table involve the mean of the duplicates. Mappable mean read depth was calculated as {(total number reads) × (read length)}/reference size, with the reference size set to 2.86 billion and the read length set to 2x150. CG and CH sites were QC'ed at a coverage threshold equal to, or greater than, 5 reads. Conversion efficacy was estimated from non-mammalian methylated DNA controls with 100% methylated CG and hmC, respectively, that were spiked in with the human DNA sample. The first control consists of *SssI* treated Lambda DNA with 100% methylation at CG sites and 0% methylation at CH sites. In reality, we observed a 96.3% methylation level for CpGs in the mC control, and likely the result of incomplete methylation during the production of the control. For both WGB and TAB, the "bisulfite conversion rate" (the percentage of CH sites sequenced as T) provides an estimate of the success of the bisulfite conversion. For TAB, the "oxidation rate" (the percentage of CG sites sequenced as T) represents the success of the TET1 oxidation. The second control consists of pUC19 DNA with all C (both CG and CH) replaced by hmC through PCR amplification. For TAB, the "hmC protection rate" (the percentage of C sequenced as C) represents the rate of hmC protection from TET1 oxidation. Due to impurities in the commercial 5-hmC nucleotides, the pUC19 control typically does not receive 100% substitutions of C to hmC. However, the absolute level of hmC in the control can be obtained from WGB as hmC is not converted to uracil by bisulfite treatment. In our study the control contained 91.1% hmC. Therefore, the protection rate of hmC in TAB was normalized based on the WGB protection rate.

Supplementary Table 2: Descriptive statistics for enrichment assays

	MBD	hMe-Seal	MBD-DIP	MeDIP
Total reads (75 bp)	84,606,822	120,093,890	79,159,659	81,373,024
Reads aligned (% of total)	99.6%	99.5%	97.7%	98.5%
Reads after QC	69,825,915	115,577,408	74,871,049	70,742,080
Reads after QC (%)	82.5%	96.2%	94.6%	87.0%

Supplementary Table 3: Methylation patterns at fragment-sized loci

	Mean	SD	Median
mCG			
Number of sites	7.3	7.4	5.0
Mean	0.6	0.2	0.6
Variance	0.0	0.0	0.0
Sum	3.3	3.1	2.4
hmCG			
Number of sites	6.5	6.4	4.5
Mean	0.2	0.1	0.2
Variance	0.0	0.0	0.0
Sum	1.1	1.0	0.9
mCH			
Number of sites	73.3	27.8	73.8
Mean	0.0	0.0	0.0
Variance	0.0	0.0	0.0
Sum	2.1	1.2	1.9

Supplementary Table 4: Spearman correlations for bisulfite and enrichment duplicates

	Bisulfite × Bisulfite	Enrichment × Enrichment	Bisulfite × Enrichment	Bisulfite-Enrichment Ratio
Optimized panel				
mCG (MBD)	0.97	0.87	0.78	0.80
hmCG (hMe-Seal)	0.81	0.67	0.57	0.70
mCH (MBD-DIP)	0.63	0.71	0.48	0.77
Standard MeDIP				
mCG	0.97	0.82	0.50	0.51
mCH	0.63	0.58	0.42	0.67

Note: Correlations between bisulfite and enrichment assays were calculated as the mean of the four possible duplicate combinations (*e.g.* enrichment assay 1 or 2 versus bisulfite assay 1 or 2). The fourth column shows the ratio of the bisulfite x enrichment correlation and bisulfite duplicate correlation reported in column 1.

Supplementary Table 5: Tested genomic features and their backgrounds.

Label	Background
Gene	Rest of genome (all studied features excluded)
Intron	Rest gene
Exon	Rest gene
3’ UTR	Rest gene
5’ UTR	Rest gene
Upstream 8kb	Rest of genome (all studied features excluded)
CG island in gene	Rest gene
CG island not in gene	Rest gene
CG island in Dnase cluster	Rest DNase cluster
CG island not in Dnase cluster	Rest DNase cluster
CG shore in gene (2kb)	Rest gene
CG shore not in gene (2kb)	Rest gene
Pseudogene in gene	Rest gene
Pseudogene not in gene	Rest gene
Splice site in gene	Rest gene
Splice site in conserved	Rest conserved
Splice site not in conserved	Rest conserved
Alt event in gene	Rest gene
Alt event not in gene	Rest gene
Alt event in intron	Rest intron
Alt event not in intron	Rest intron
TFBS in gene	Rest gene
TFBS not in gene	Rest conserved
TFBS in conserved	Rest conserved
TFBS not in conserved	Rest conserved
Enhancer in conserved	Rest conserved
miRNA	miRNA shore
miRNA shore	Upstream 8kb
ncRNA	Rest of genome (all studied features excluded)
Conserved in gene	Rest gene
Conserved not in gene	Rest gene
Repeat	Rest of genome (all studied features excluded)
Structural variant in repeat	Rest repeat
Structural variant not in repeat	Rest repeat
GenomicSuperDups in repeat	Rest repeat
GenomicSuperDups not in repeat	Rest repeat
Dnase cluster in gene	Rest gene
Dnase cluster not in gene	Rest gene

DNase1 frontal cortex	Rest of genome (all studied features excluded)
DNase1 cerebrum	Rest of genome (all studied features excluded)
DNase1 cerebellum	Rest of genome (all studied features excluded)