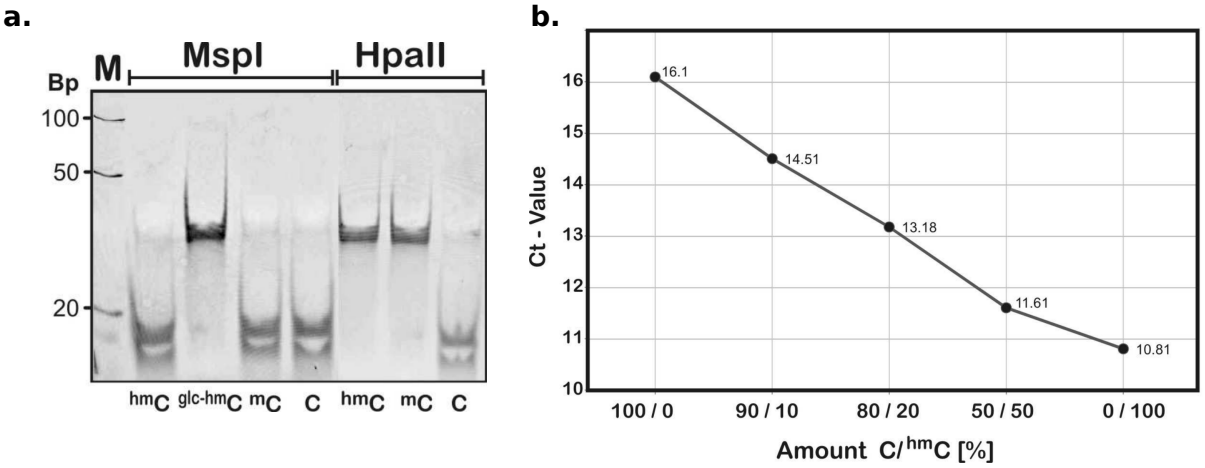
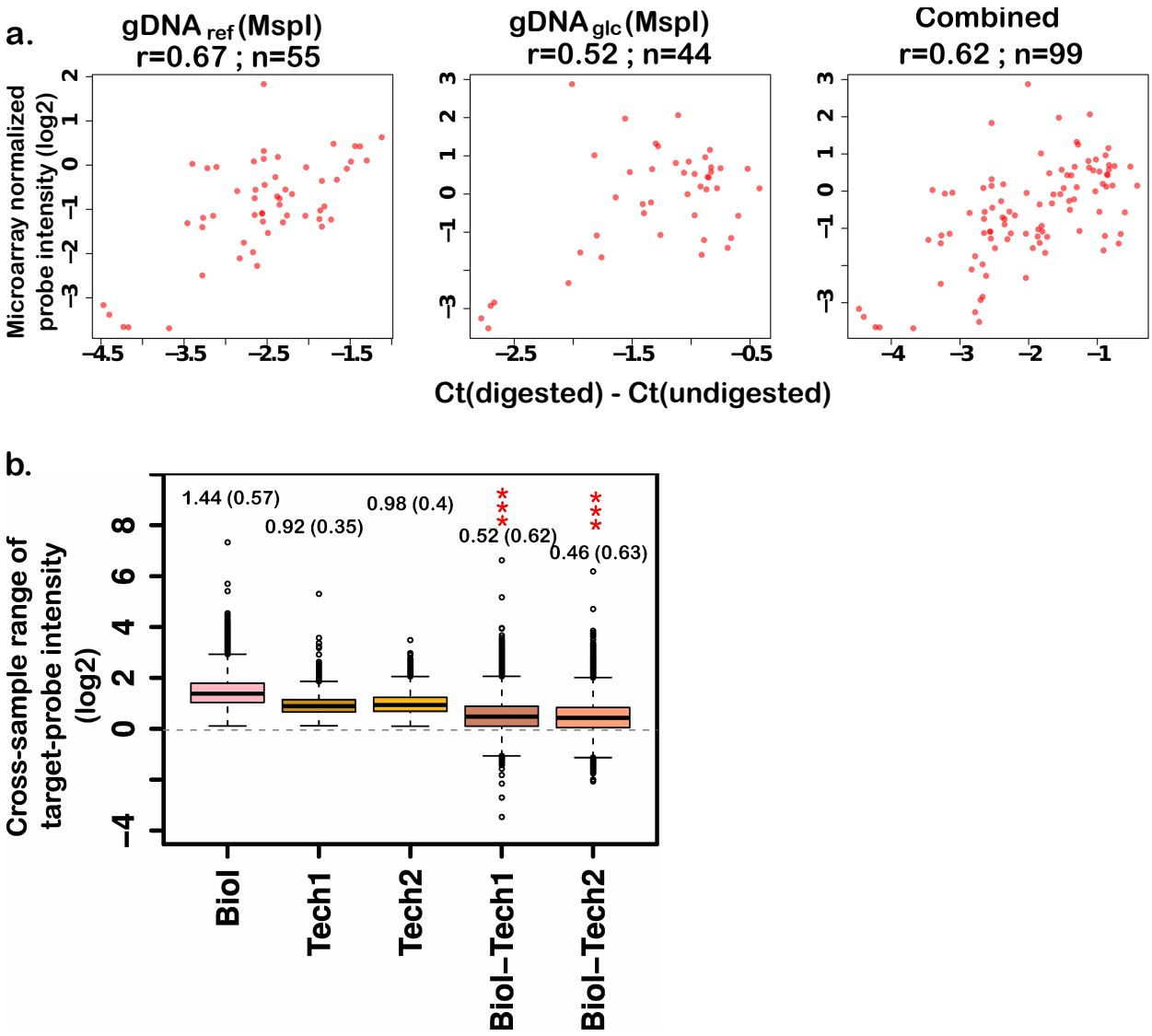
Supplementary Figures

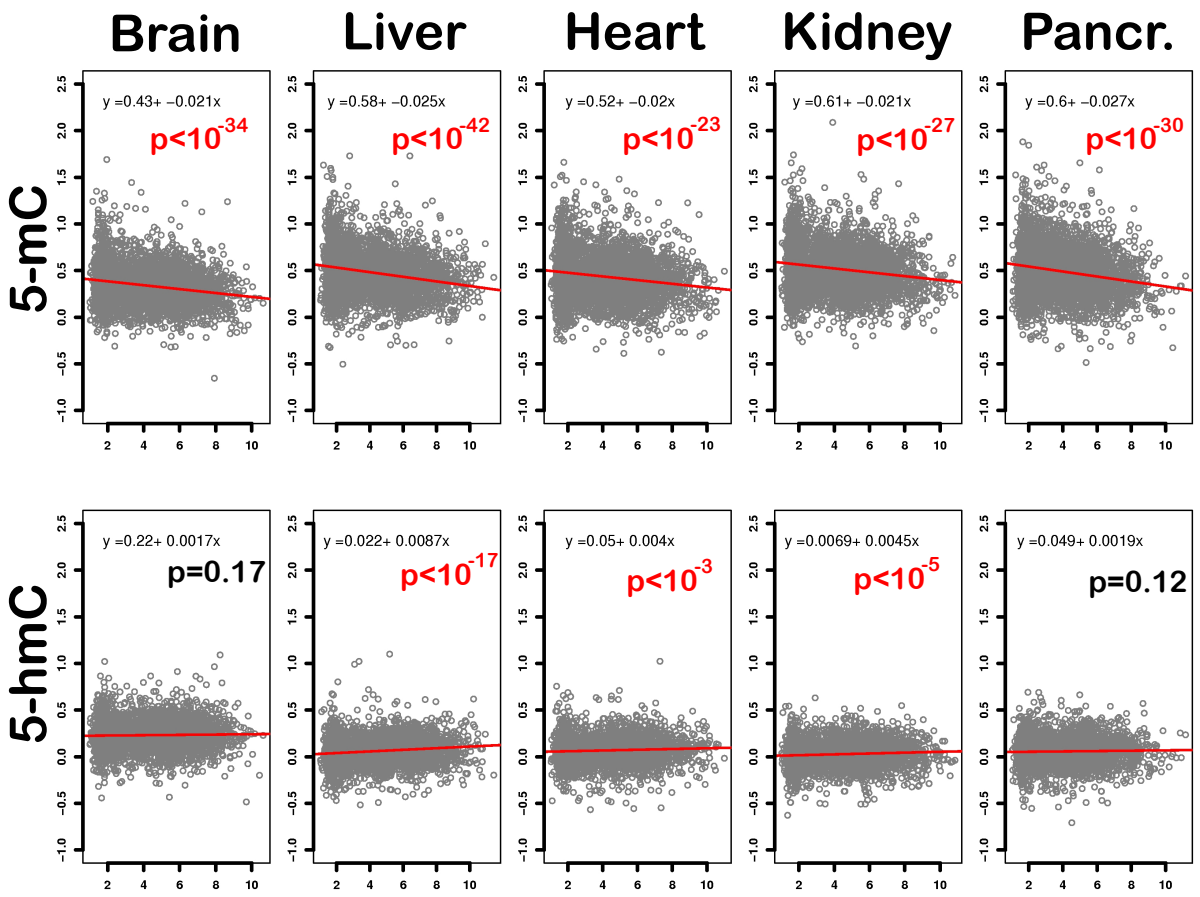


Supplementary Figure 1: Validation for BGT-glucosylation as an assay for 5-hmC measurement. (a) Effect of glucosylation treatment on a 31-mer DNA duplex containing 5-hmC, 5-mC or C on one strand of a CCGG target site (modified cytosine is underlined); details in Supplementary Note 1. Cleavage by MspI is only blocked by glucosylation of the 5-hmC residue (from left: lanes 1–4); HpaII digestion is inhibited when either modification, 5-mC or 5-hmC, is present at the CpG site (lanes 5–7). M – FastRuler™Ultra Low Range DNA Ladder (Fermentas). (b) Standard curve for 5-hmC estimates from real-time PCR. A 200 bp DNA fragment containing one 5-hmC-modified MspI site ($C^{hm}CGG$) was spiked in different amounts into a quantity of unmodified DNA of the same sequence (x-axis; 10ng total amount). The total DNA was subjected to BGT-glucosylation (Online Methods) and subsequent treatment with 10 U of MspI at 37^0C for 16 hrs, followed by quantitative real-time PCR. Quantitative PCRs were performed on a Rotor-Gene 6000 (Corbett Research) instrument using MaximaTM SYBR Green qPCR Master Mix (Fermentas) (Supplementary Note 1). The threshold cycle (Ct) values of the corresponding DNA mixtures are shown as inset.

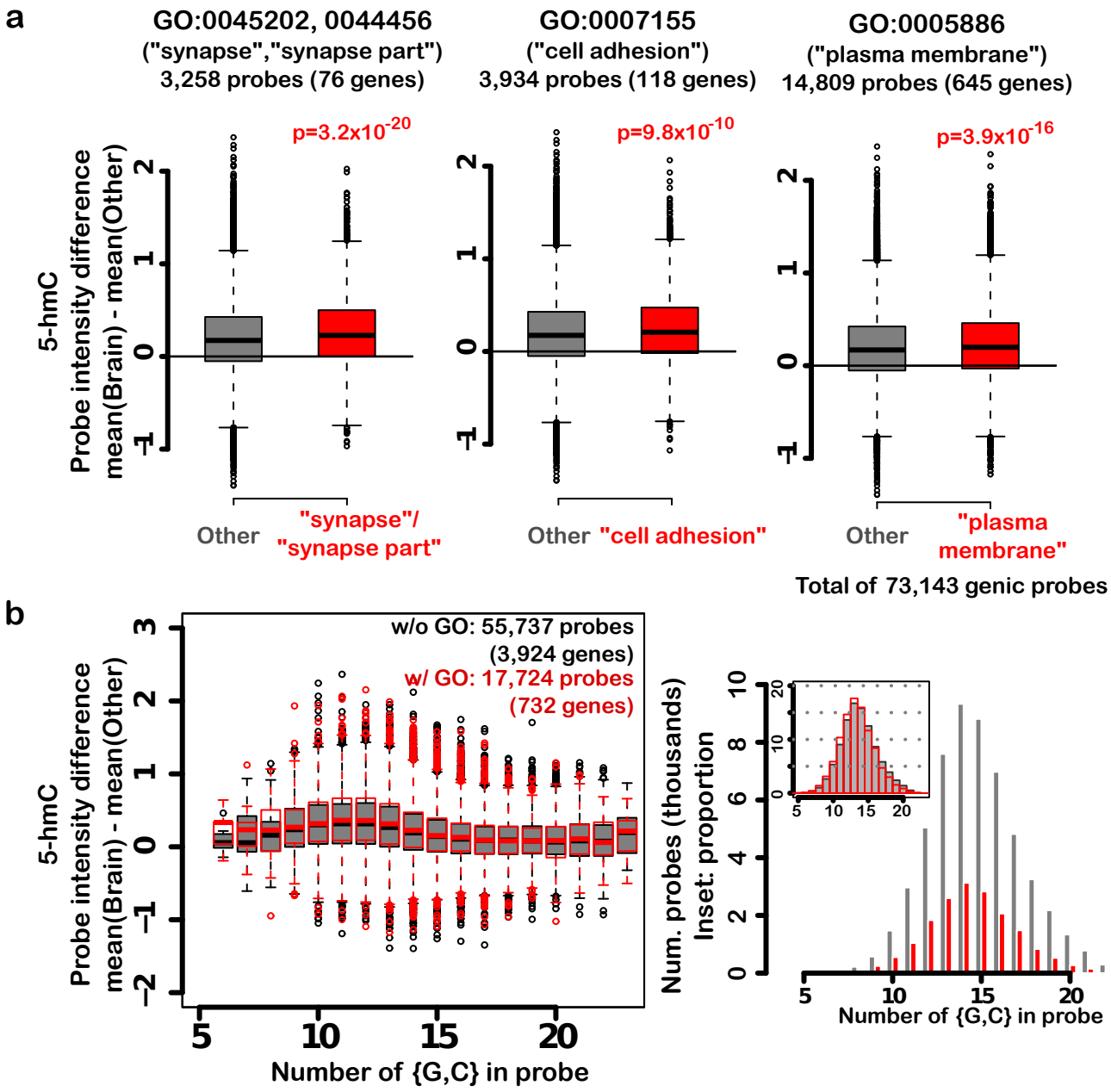


Supplementary Figure 2: Microarray-based validation of 5-hmC assay. (a) Correlation of digestion efficacy as measured by quantitative PCR (qPCR) and by microarray. Microarray single probe intensity (y-axis) is plotted with qPCR measures (Ct value; x-axis) at 11 arbitrarily-selected loci (Supplementary Table 3). Loci for qPCR have the property that the target site (CCGG) lay directly on a microarray probe; for each locus, DNA from 4-5 individuals was separately qPCR-amplified. Shown are correlations for (left) changes in unglucosylated genomic DNA following MspI digestion (gDNA (MspI)), (middle) changes in glucosylated genomic DNA following MspI digestion (glc-gDNA (MspI)), and (right) data from both conditions combined. The Ct has an inverse relationship with the amount of DNA fragment at the start of qPCR; i.e. a greater Ct value reflects a lower starting template to be PCR-amplified. Each dot shows individual-level (not sample-averaged) data; *n* denotes number of data points and *r* is the correlation coefficient. (b) Microarray analysis results in biological variability that exceeds technical variability. Each boxplot shows the distribution of the range of target-probe intensities. Data is shown for MspI-treated unglucosylated genomic DNA. Six biological replicates ("Biol", pink) were compared to each of two technical replicates ("Tech1" and "Tech2"; brown, orange). Each dot measures the cross-sample range (max – min) intensity for target probes on human chromosome 5 (27,546 probes). Mean (sd) shown above each boxplot. The range of probe intensities is greater for biological replicates ("Biol – Tech1" (dark salmon); "Biol – Tech2"

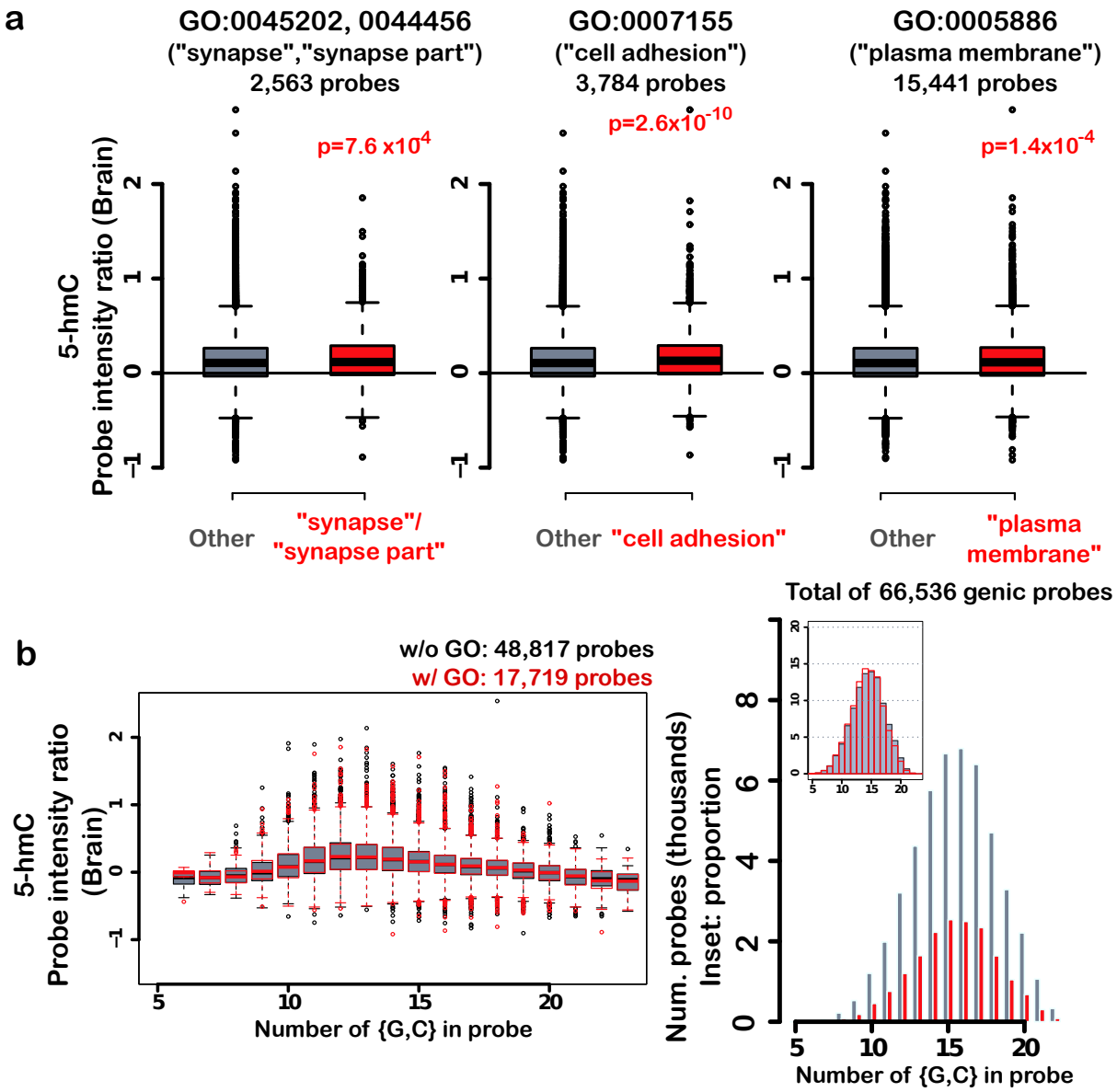
(light salmon)); one-sample t-test in both cases results in p-values $< 10^{-16}$.



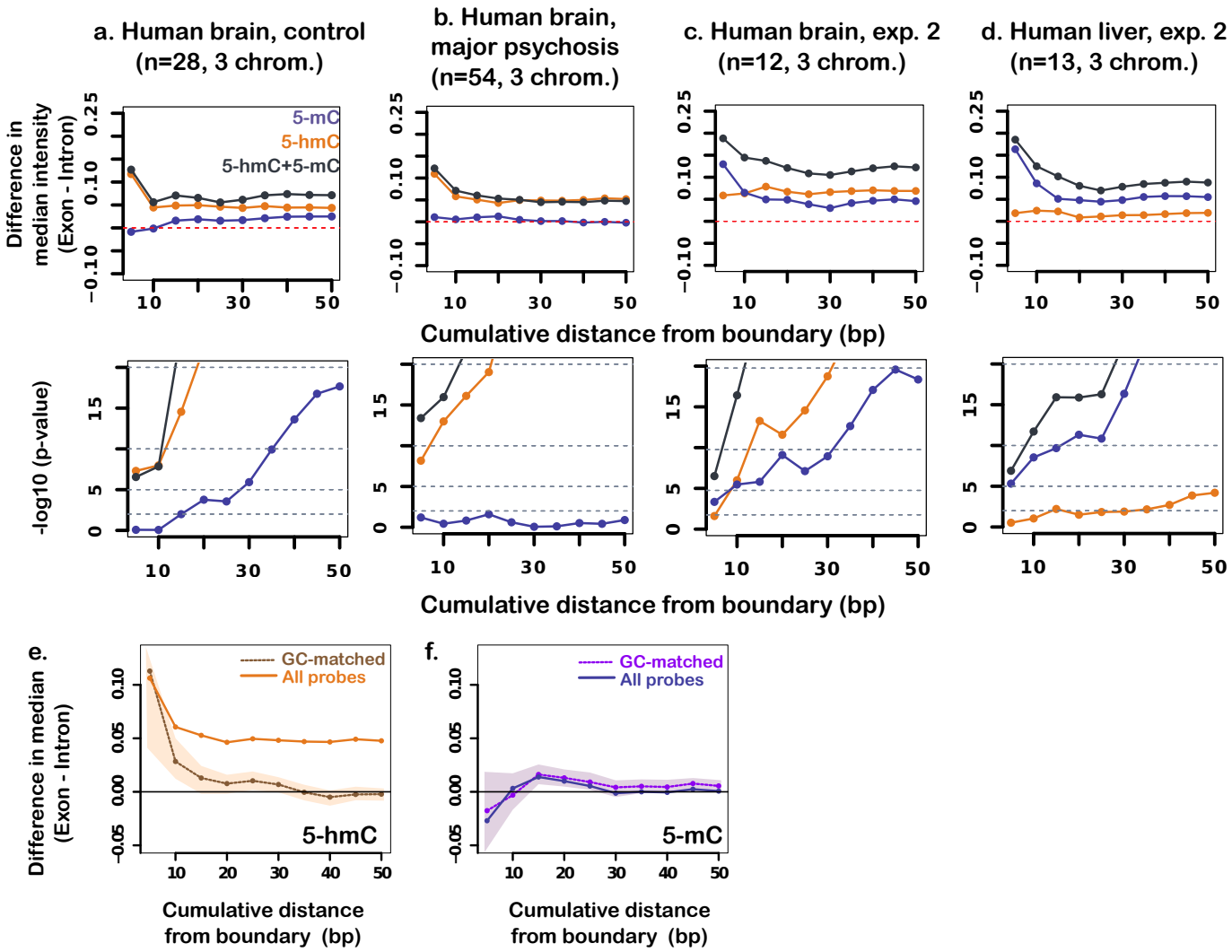
Supplementary Figure 3: Linear regression of steady-state mRNA levels with mean genic intensity of DNA modifications. Plots show gene-averaged (mean) mRNA levels (x-axis) against averaged probe intensities for corresponding genes (y-axis); genes defined by RefSeq ID (top: 5-mC, bottom: 5-hmC). Regression line shown in red, P-values are for slope ($\alpha = 0.005$), with significant P-values shown in red. The inverse relationship of genic 5-mC and gene expression levels were consistently found in all tissues investigated. The relationship of 5-hmC and transcription levels was significant only in some non-neuronal tissues, although the slight upward trend is visible in all cases (also see Fig. 1c).



Supplementary Figure 4: 5-hmC in the adult mouse brain is higher in genes mapped to synapse-related categories, compared to that in genes outside these categories. (a) Probes in genes mapped to each Gene Ontology (GO) category (red) had greater cross-tissue differences (Brain-Other) than those in other genes (gray). The GO categories tested were the top three categories overrepresented in brain 5-hmC rich genes (Table 1); this result generalizes the observation in enriched genes to all genes in these categories. Each dot measures the difference in probe intensity between brain samples and samples from other tissues; probes were not averaged within genes. (P-values from two-tailed WMW test, $\alpha = 0.016$). (b) Probe-level differences persist even after probe stratification by GC content. This panel shows probes combined for all three GO terms tested in (a) (red), compared to other probes (gray). (Left): Increase in 5-hmC levels is evident, particularly in strata with most probes ($9 \leq GC \leq 16$). This increase is more pronounced for individual GO categories (not shown). (Right): Number of probes in each GC-stratum, (inset: probe proportions).

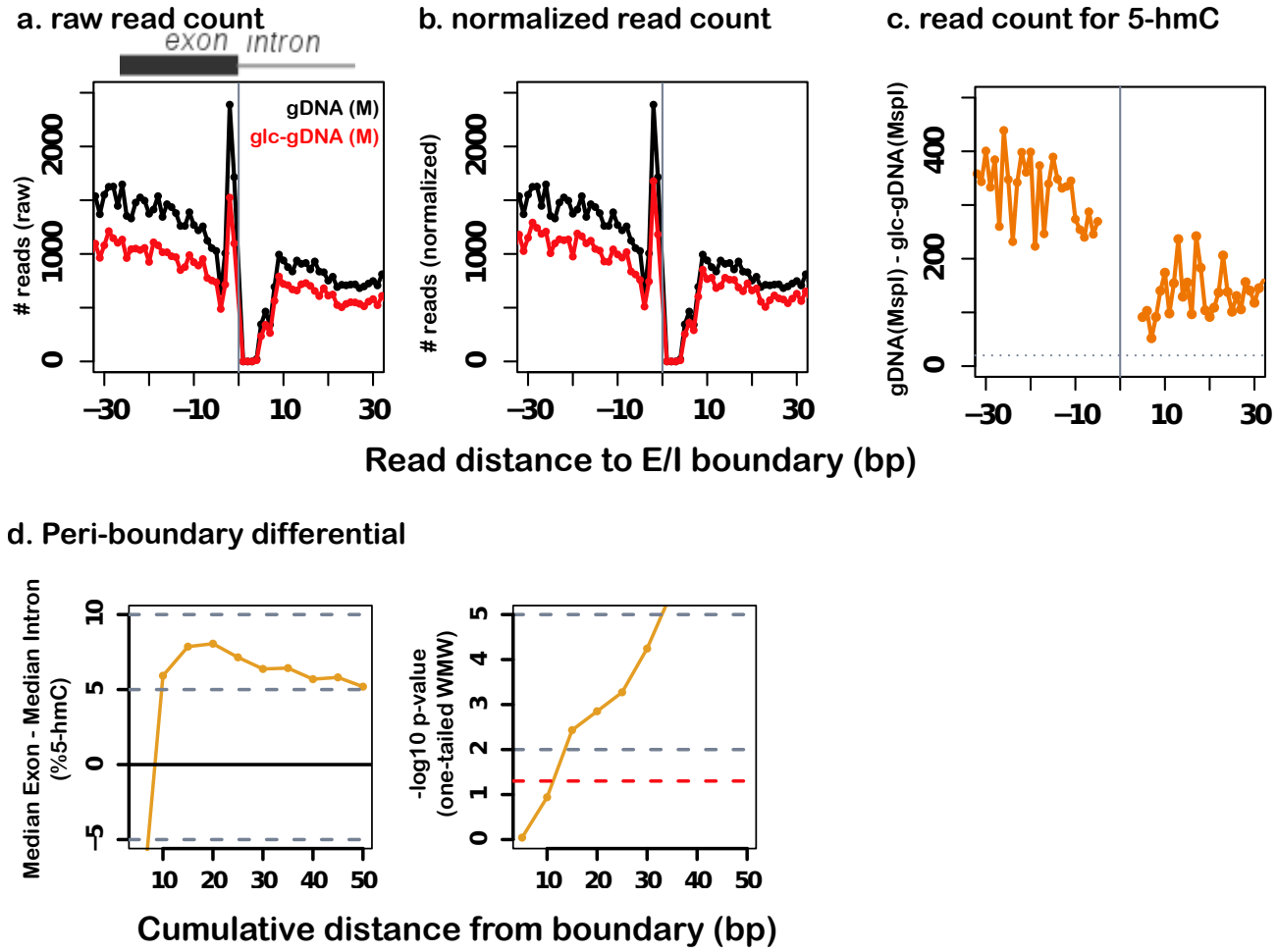


Supplementary Figure 5: 5-hmC in the adult human brain is higher in genes mapped to synapse-related categories, compared to that in genes outside these categories. (a) Probes in genes mapped to each GO category (red) had higher density of 5-hmC than those in other genes (gray). The GO categories tested were the same as those tested in the mouse (Supplementary Fig. 4). Each dot measures probe-level 5-hmC (sample-averaged). Probes were not averaged within genes. (b) Probe-level intensities with probe stratification by GC content. This panel shows probes combined for all three GO terms (red, "w/ GO"), compared to other probes (gray, "w/o GO"). (Left): 5-hmC intensity (Right): Number of probes in each GC-stratum (Inset: probe proportions).

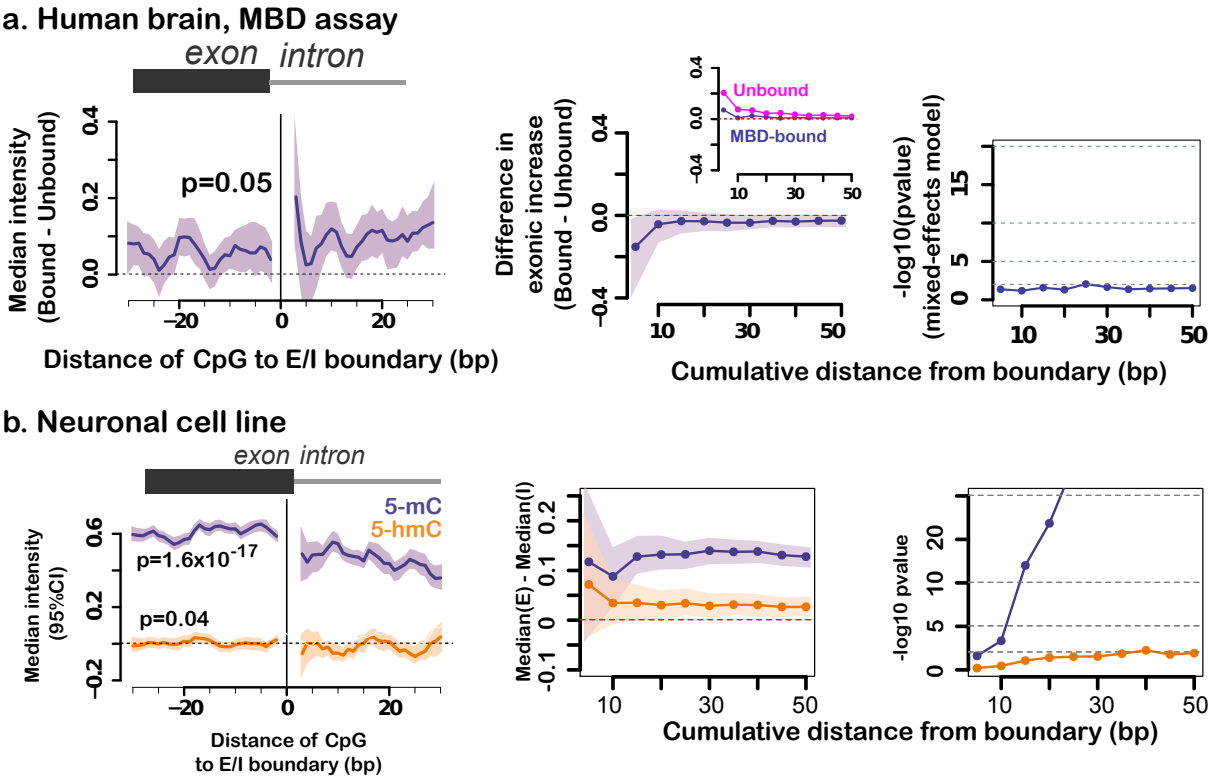


Supplementary Figure 6: Exonic increase in DNA modifications in human tissues. (a-d) show cross-boundary changes in DNA modifications in human tissues, for various cumulative distances (d = 5 – 50 bp). Data are shown for (a) human brains without diagnosis of mental illness (n = 28; 6 chromosomes), (b) human brains from individuals diagnosed with major psychosis (n = 54, 3 chromosomes), and for (c,d) an independent experiment on age- and sex-matched (c) brain and (d) liver samples. In each case, the top panel shows median exonic increase in DNA modifications at various cumulative distances from the exon-intron boundary, and the bottom panel shows corresponding informal P-values from statistical comparison of exonic and intronic probe intensities (linear mixed-effects model, see Online

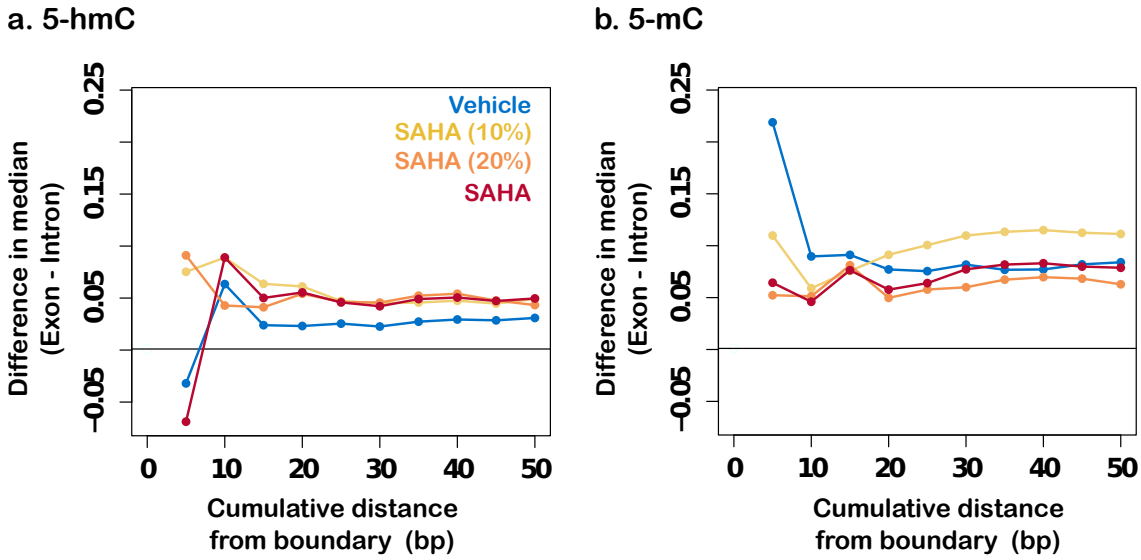
Methods). In brain samples (a-c), exonic increase in all DNA modifications (black) is predominantly mirrored by changes in 5-hmC (orange), and to a lesser extent in 5-mC (purple); in the liver (d), this pattern is reversed. (e,f) Exon-intron peri-boundary differential after probes on either side of the boundary are matched for GC content, at various cumulative distances from the boundary (100 iterations of matching; trendliens show median, shaded areas show the range between the 5th and 95th percentile of differences). (e) Following GC-matching, exonic increase in 5-hmC levels are notable at d = 5 and persist up to 20 bp in the peri-boundary region. (f) The relatively modest change in 5-mC persists after GC-matching. At d = 5, zero lies within the range of GC-matched values. Following GC-matching, 5-hmC and 5-mC values are similar for peri-boundary distances greater than 10bp. It is unclear at present whether this similarity is due to loss of statistical power from GC-matching.



Supplementary Figure 7: 5-hmC changes measured at exon-intron boundary using single molecule sequencing (SMS). One brain sample (cortex Brodmann Area 10; female, age at death 49 years, no diagnosis of brain disease) was analyzed (3 replicates) for DNA modifications on SMS (Supplementary Table 13). 5-hmC was estimated as the percent difference in read count of non-glucosylated DNA and that in glucosylated DNA, following MspI restriction digestion. In all panels, the x-axis shows the distance from the second cytosine in the target site (CCGG) of a read generated by a CCGG sequence, relative to an exon-intron boundary. The y-axis shows: (a) raw read count, (b) read count normalized by reads in unglucosylated channel, and (c) the difference in reads from restriction-digested DNA with and without glucosylation. (d) shows exonic increase in % 5-hmC at various cumulative distances from the boundary. The x-axis is the cumulative distance (in bp) from the second cytosine of a target read to an exon-intron boundary. Left: Percent difference in reads obtained with and without glucose protection of MspI sites. Right: P-values from comparison of exonic and intronic % 5-hmC at distances corresponding to the left graph (one-tailed Wilcoxon Mann Whitney test, *n* = Distance (in bp) from the boundary).



Supplementary Figure 8: Validation of cross-boundary change in brain samples as being due to 5-hmC. (a) Exon-intron boundary comparisons in DNA enriched for 5-mC, relative to that depleted in 5-mC. Sample genomic DNA (gDNA) from one human brain (n = 6 technical replicates) was separated into a 5-mC rich fraction ('bound', Methyl Miner Kit (Invitrogen)) and the rest ('unbound'). Both fractions were then analyzed for 5-mC following glucosylation and restriction enzyme treatment (as in Fig. 1a) on human tiling array chip E (chr 5,7,16). (Left): Junction detail showing the relative increase in 5-mC in the MBD-bound fraction. Compared to the unbound fraction, the bound fraction shows an intronic increase at the boundary. Shaded regions show bootstrapped (R = 1,000) 95 % CI. (Middle, Right): Changes in the bound fraction at various cumulative distances from the boundary (Inset: bound and unbound fractions). Middle: relative intensity; dip at 5 bp indicates intronic increase in 5-mC at the boundary. Right: informal p-values of cross-boundary comparisons. Shading shows 95 % CI from separately bootstrapping exonic and intronic values prior to subtraction (R = 1,000). (b) DNA modifications at exon-intron boundary in a mouse neuronal cell line with low global levels of 5-hmC. This cell line (mHypoA-2/24) has negligle amounts of 5-hmC (thin-layer chromatography, not shown; microarray, Supplementary Table 12) (n = 24, 3 chromosomes). Consistent with globally undetectable levels of 5-hmC, the main change at the exon-intron boundary was that in 5-mC rather than 5-hmC. (Left): Probe intensities in the region immediately around the exon-intron boundary. (Middle): median exonic increase at various cumulative distances from the boundary; (Right) informal p-values (linear mixed-effects model, Online Methods) for exon-intron comparisons.



Supplementary Figure 9: DNA modifications at the exon-intron boundary in B-lymphocytes following treatment with suberoylanilide hydroxamic acid (SAHA). Graphs show the exon-intron differential at various cumulative distances from the boundary. Treated cells showed an increase in cross-boundary differences that approximately corresponded to increasing doses of SAHA (10 % , 20 % full dose of SAHA (0.1µM)), relative to vehicle-treated control cells. Data for each trendline is an average of six technical replicates. (a) 5-hmC levels at the exon-intron boundary are higher in SAHA-treated samples (warm colors), relative to vehicle-treated samples (blue). (b) 5-mC levels show the opposite trend of decreasing with SAHA treatment. SAHA may also modulate splicing outcomes by changing the levels of DNA modifications at the exon-intron boundary.

Supplementary Tables

Dataset	Num. biological samples	Num. arrays (biological + technical replicates)	Array name, chroms	Num. CCGG probes [a]
Human				
Brain, BA10 (Stanley & Maclean, controls)	28	84	E: 5,7,16	69,252
Brain, BA10 (Stanley & Maclean, controls)	28 [b]	84	F: 8,11,12	67,225
Brain, BA10 (Stanley & Maclean, psychosis)	54	162	E: 5,7,16	69,252
Brain, BA10 (Stanley), for MBD	6 [b]	12	E: 5,7,16	69,252
Exp 2, Liver (Curline & Cambridge Biosci.)	13	39	E: 5,7,16	69,252
Exp 2, Brain (Stanley, controls)	12 [b]	36	E: 5,7,16	69,252
B-lymphocyte cell line	24	72	G: 10,13,14,17	76,102
Mouse				
Multiple organs (brain, liver, pancreas, kidney, heart)	32	141	A, G: 1,9,10,13,14,19	130,314 [c]
Neuronal cell line (mHypoA-2/24)	6	54	B: 2, X,Y	46,892
Frontal cortex	15	45	A: 1,9,19	69,052
Brain, non-frontal cortex	15	45	A: 1,9,19	69,052
Total	**187**	**774**		

a) After excluding repeat overlaps; b) Biological samples excluded from running total as already previously counted; c) this probe count is lower than that in the synapse-related analyses because this analysis has an additional filter: probes where the chromosomal and Affymetrix probe sequence did not both contain a target site were excluded.

Supplementary Table 1: Sample, array, and probe count for all datasets analyzed in the current study. Not included is the sample count for Helicos validation, which used 3 technical replicates of a single human brain.

Type of array measurement	**gDNA (MspI)**	**Glu-gDNA (MspI)**	**Combined**
# data points	*55*	*44*	*99*
Single probe	0.67	0.52	0.62
Probe-averaged window (340 bp)	0.02	– 0.29	0.09
Weighted window (340 bp)	0.20	– 0.15	0.27
Probe-averaged window (100bp)	0.42	0.03	0.39
Weighted window (100bp)	0.51	0.17	0.48

Supplementary Table 2: Correlation of sequence quantity at 11 loci, as measured by quantitative PCR and by microarrays. Correlation deteriorates dramatically for glucosylated DNA, when window-based averaging is used in arrays. The reason for this drop in correlation is not understood. Based on these results we decided to analyze arrays at the single probe level, without window-based averaging. qPCR coordinates provided in Supplementary Table 3.

F Primer	R Primer	qPCR start	qPCR stop	Amplicon Length (bp)
TGCTGCTCGATGCACAGGT	CATCTTCACCTTCCTGCTGAG	179497224	179497310	86
CAGTCTCTCCCTGCACACAG	GTGGCATGGTCTGGTTTCC	178618590	178618677	87
CCAAATTATAAGACAGATGCCTAG	GCCAACTTCTGTGAAACTACACT	53342655	53342775	120
CAGTGCCTTAGGCCTCTCTC	CTTGGTCTGCCATCTTCTGG	131634481	131634574	93
GAAAGGTGAGCTCCCTGAAG	AAGCAAACGCTGGCTGAG	176792842	176792935	93
TGAGTAGTCATGACCCCTTTC	CCAGGGTGTAACATGAATAGGA	161334060	161334151	91
CTCTTTGGTTCAACTGGTCCA	CTCTCAGAATCCCAACCAGGA	168636029	168636112	83
GCCCTTGACTGCCTCCTTAT	TTCCAGGACCCTAAAAAGCTC	16648811	16648944	133
CTTGCCTGGTCAGATGACAG	TGC CTC TCC ATC TAG CAT CC	111684652	111684729	77
TGGTTCTTTACCCCATTAGTCATA	CAGGGATCTGATGTGCCATAT	93076322	93076426	104
ACCCACCTGTGTAAGCCTGT	AGG AAC TCA GGA GAG CAG GT	169707109	169707244	135

Supplementary Table 3: Primers used for quantitative PCR experiments. These experiments were used to identify the optimal normalization algorithm. All probes are located on human chromosome 5 (*hg18*).

	5-hmC % (relative to 5-mC)		
	Sample	Mean	SD
Human	Brain	18.6	2.6
Mouse	Heart	5.1	0.4
	Kidney	6.0	0.1
	Brain	13.6	0.1
	Liver	8.8	0.3
	Pancreas	2.8	0.8

Supplementary Table 4: Thin layer chromatography quantification of 5-hmC**.** 5-hmC was estimated in 20 human *postmortem* brain gDNA samples (age range 34 – 85 years). Mouse: Brain, heart, kidney, and liver samples were obtained from a 24-month animal, and the pancreas sample from a 8-week old mouse (male). Standard deviation (SD) from 3 technical replicates per sample.

Array type	Animal num.	Age	Brain	Heart	Liver	Pancreas	Kidney
G	1	8wk	71B	72H	73L	74P	75K
G	2	24mos	136B	137H	138L	139P	140K
G	3	18 mos	121B	-	-	-	-
G	4	8wk	91B	1H	61L	121P	31K
G	5	8wk	92B	2H	62L	122P	32K
G	6	8wk	95B	-	-	-	-
A	5	8wk	92B	2H	62L	122P	32K
A	1	8wk	71B	72H	73L	74P	75K
A	4	8wk	91B	1H	61L	121P	31K
A	7	8wk	93B	3H	63L	123P	33K
A	8	8wk	51B	52H	53L	54P	55P

Supplementary Table 5: Sample information for the mouse tissue dataset. All animals were adult male inbred C57/BL6 strain mice; in several instances, multiple tissue samples were collected from the same animals (Animal num.). Not shown in table: An independent set of mouse brains was separated into frontal cortex and the remainder (including brain stem and cerebellum; 8-week old mice; $n = 15$). This set was used for the characterization of exon-intron boundaries in mouse brain. In this case, sample genomic DNA was processed for DNA modification profiling (Online Methods, Fig. 1a) and hybridized to mouse chip G.

Supplementary Table 6: List of genes with differential 5-hmC in the adult mouse brain, relative to other tissues.

Cnksr3	Pkib	BC030307	Rdh7
Oprm1	Edar	1700113H08Rik	Smarcc2
Rgs17	Sh3rf3	Pah	Pitrm1
Fbxo5	Sept10	4930547N16Rik	Pfkp
Syne1	Oit3	Dram1	Adarb2
Mthfd1l	Ccdc109a	Mybpc1	Dip2c
Plekhg1	Cbara1	Ano4	Chrm3
Ppp1r14c	Dnajb12	Anks1b	Ryr2
Ppil4	Ascc1	Elk3	Lyst
Tab2	Spock2	Ntn4	Hecw1
Ust	Cdh23	Vezt	Gli3
Sash1	Unc5b	Fgd6	Rala
Samd5	Adamts14	Plxnc1	Pou6f2
Grm1	X99384	Nudt4	Amph
Shprh	Lrrc20	Eea1	Elmo1
Epm2a	Col13a1	Mir297a-6	Aoah
Utrn	Supv3l1	Epyc	Scgn
Phactr2	Ddx21	Gm10754	Lrrc16a
Hivep2	Ccar1	Atp2b1	Dcdc2a
Nhsl1	Herc4	Tmtc3	Cdkal1
Pde7b	Ctnna3	Mgat4c	Mboat1
Ahi1	Ado	Nts	Irf4
Eya4	Zfp365	Slc6a15	Gmds
Moxd1	Rtkn2	Tmtc2	1700026J04Rik
Ctgf	Arid5b	BC067068	Fars2
Arg1	1700040L02Rik	Ppfia2	F13a1
Akap7	Ank3	Lin7a	Riok1
Samd3	Ccdc6	Ptprq	Bmp6
Lama2	Fam13c	Gm6924	Elovl2
Ptprk	Bicc1	Syt1	Gm10790
Themis	Tfam	Nav3	Phactr1
6330407J23Rik	Zwint	Osbpl8	Rnf182
Trmt11	Pcdh15	Kcnc2	Atxn1
Nkain2	Cabin1	Trhde	Rbm24
Nt5dc1	Gm5134	Tph2	Cap2
Hs3st5	Slc5a4b	Tmem19	Kif13a
Hdac2	Pcbp3	Lgr5	Shc3
Fyn	Col18a1	Tspan8	Sema4d
Rev3l	Adarb1	Ptprr	Sykb
Cdk19	D10Jhu81e	Kcnmb4	Ror2
Cdc40	Abca7	Cpm	4732471D19Rik
Sesn1	Tmprss9	Slc35e3	Cdhr2
Armc2	Zbtb7a	Grip1	Unc5a
Foxo3	Tjp3	Msrb3	Grk6
Nr2e1	Ncln	Lemd3	Pdlim7
Sobp	Gna11	Tbc1d30	H2afy
Bend3	BC025920	Gns	Slc25a48
Rtn4ip1	Nfyb	Rassf3	Tgfbi
Aim1	Chst11	Xpot	Smad5
Atg5	Slc41a2	Srgap1	Trpc7
Prep	A230046K03Rik	Ppm1h	Spock1
Grik2	Rfx4	Mon2	Gkap1
Ascc3	Ric8b	Fam19a2	Slc28a3
Ros1	Btbd11	Slc16a7	Ntrk2
Slc35f1	Syn3	Myo1a	Golm1

4921517D22Rik
2010111I01Rik
Cdc14b
BC018507
Adamts16
Lpcat1
Ahrr
Mctp1
9330111N05Rik
Gpr98
Edil3
Vcan
Xrcc4
Atp6ap1l
Atg10
Ssbp2
Rasgrf2
Msh3
Dhfr
Serinc5
Arsb
Lhfpl2
Scamp1
Ap3b1
Wdr41
Pde8b
Sv2c
Ankdd1b
Fam169a
Rgnef
Mrps27
Mtap1b
Ocln
Cdk7
Pik3r1
Mast4
Cwc27
Srek1ip1
Rgs7bp
Ndufaf2
Elovl7
Depdc1b
Pde4d
Rab3c
Skiv2l2
Ndufs4
Hcn1
Flnb
Pxk
Fam107a
4930452B06Rik
Fhit
Ptprg
Cadps
Synpr
Top2b
Thrb
Gng2
Myst4

1700112E06Rik
Gm10248
Kcnma1
Dlg5
Anxa11
Il17rd
Arhgef3
Erc2
Cacna2d3
Cacna1d
Itih1
Capn7
Mettl6
E130203B14Rik
Wdfy4
Ldb3
Grid1
Nrg3
Gpr137c
Fermt2
Samd4
Atg14
Peli2
6720456H20Rik
Slc35f4
Mettl3
Slc7a7
Dhrs2
Rabggta
Khnyn
Parp4
Zmym2
Cryl1
Atp8a2
Spata13
Tnfrsf19
Sacs
Gucy1b2
Blk
Xkr6
4930578I06Rik
Rp1l1
Msra
Hmbox1
Ints9
Fbxo16
Pnoc
Adam2
Ptk2b
Dpysl2
Bnip3l
Dock5
Chmp7
Xpo7
Gfra2
Lrch1
Lrrc63
Zc3h13
Siah3

Tsc22d1
Enox1
Tnfsf11
Akap11
Dgkh
1300010F03Rik
Diap3
Pcdh9
Klhl1
Dach1
Pibf1
Klf12
Mycbp2
Scel
Ednrb
Rnf219
D130009I18Rik
Gpc5
Gpc6
Abcc4
Uggt2
Hs6st3
Mbnl2
Farp1
Stk24
Dock9
Ubac2
Clybl
Pcca
Itgbl1
Fgf14
Xkr4
St18
Pcmtd1
Cspp1
Prex2
Slco5a1
Prdm14
Ncoa2
Trpa1
Kcnb2
Stau2
Ube2w
Tcfap2d
Pkhd1
Tram2
Rims1
Col19a1
Lmbrd1
Bai3
Khdrbs2
Prim2
Rab23
Prss39
Uggt1
Cnnm4
Actr1b
Tmem131
Inpp4a

2010300C02Rik
Eif5b
Aff3
Chst10
Npas2
Rfx8
Il1r2
Mfsd9
Tmem182
Mrps9
Col3a1
Tmeff2
Hecw2
Ankrd44
Plcl1
1700066M21Rik
9430016H08Rik
Spats2l
Gm973
Sumo1
Carf
Abi2
Icos
Pard3b
Ccnyl1
Plekhm3
Unc80
Rpe
Erbb4
Ikzf2
Mreg
March4
Ttll4
Acsl3
Serpine2
Dock10
9430031J16Rik
Rhbdd1
Col4a4
Sphkap
Dner
Trip12
Dis3l2
Inpp5d
Sag
Heatr7b1
Trpm8
Agap1
Iqca
Hdac4
Neu4
D1Ertd622e
Pam
Cntnap5b
Cdh20
Phlpp1
Bcl2
Cntnap5a
Tcfcp2l1

Gli2
Marco
Dpp10
Gpr39
Nckap5
Mgat5
Tmem163
AA986860
Pigr
Il19
Slc41a1
Slc45a3
Lemd1
Tmcc2
Dstyk
Cntn2
Nfasc
Atp2b4
Optc
Ppfia4
Ppp1r12b
Pkp1
Dennd1b
Kcnt2
Glrx2
Hmcn1
1700025G04Rik
Apobec4
Arpc5
Smg7
Nmnat2
Lamc1
Rgs8
Cacna1e
Acbd6
Lhx4
Cep350
Fam163a
Tdrd5
Fam20b
Rasal2
Fam5b
Astn1
Pappa2
Tnr
Rabgap1l
Dnm3
Prrx1
Kifap3
Nme7
Dpt
Pou2f1
Gpa33
Fam78b
Uck2
Rxrg
Lmx1a
Pbx1
Sdhc
Ncstn
Ccdc19
Fmn2
Grem2
Rgs7
Wdr64
Pld5
Sdccag8
1700016C15Rik
Adss
Efcab2
Kif26b
Smyd3
Cnst
Cabc1
Itpkb
Lin9
Trp53bp2
Capn2
Mark1
Tgfb2
Spata17
Gpatch2
Esrrg
Kcnk2
Smyd2
Prox1
Vash2
Batf3
Kcnh1
AA408296
Gucy1a2
Gria4
Birc3
Arhgap42
Cntn5
Sesn3
Folr4
Ccdc67
Fat3
Pde4a
Cdkn2d
Bmper
Dpy19l1
Eepd1
Ncapd3
Jam3
Spata19
Opcml
Ntm
Ets1
Kirrel3
Cdon
Fez1
BC024479
Gramd1b
Ubash3b
Sorl1
Sc5d
Grik4
Arhgef12
Tmem136
Trim29
Arcn1
Tmprss13
Dscaml1
Cadm1
Htr3a
Ncam1
Ppp2r1b
Layn
Arhgap20
Crabp1
Ube2q2
Nrg4
Scaper
Pstpip1
Lingo1
1600029O15Rik
6030419C18Rik
Hcn4
Neo1
Thsd4
Itga11
Map2k5
Megf11
Clpx
Pdcd7
Zfp609
Rab8b
Tln2
Rora
Sltm
Prtg
Unc13c
Wdr72
Myo5a
Mapk6
Lysmd2
Bmp5
E330016A19Rik
Slc17a5
Impg1
4930486G11Rik
Bckdhb
Pgm3
Zic4
Slc9a9
Pcolce2
Acpl2
Trim42
Clstn2
Pik3cb
Ephb1
Slco2a1
Rab6b
Tmem108
Nphp3
Dnajc13
Acpp
Cpne4
Aste1
Gm7455
Rpl29
Abhd14a
Dock3
Cacna2d2
Traip
Qrich1
Prkar2a
Col7a1
Dhx30
Klhl18
Arpp21
2900079G21Rik
Cnot10
Cmtm8
Tgfbr2
Rbms3
Itga9
Wdr48
Myrip
Ulk4
D9Ertd402e
Kif15
Mtl5
Suv420h1
Rps6kb2
Kdm2a
Pacs1
Catsper1
Mus81
Pola2
Slc22a12
2700081O15Rik
Slc3a2
Cd6
Mpeg1
Tle4
Gnaq
Gna14
Prune2
Pcsk5
Tmc1
Trpm3
Smc5
Mamdc2
1700028P14Rik
Apba1
Fam189a2
Pip5k1b
Pgm5
Dmrt1
Vldlr
Rfx3
Glis3
Ermp1

9930021J03Rik	Myof	Arl3	Acsl5
Prkg1	Plce1	Cnnm2	Tcf7l2
Sgms1	Sorbs1	Nt5c2	Afap1l2
Rnls	Slit1	Col17a1	Trub1
Ifit1	Loxl4	Itprip	Atrnl1
Pank1	Hpse2	Ccdc147	Gfra1
Htr7	Abcc2	Sorcs3	Slc18a2
Rpp30	Cyp2c44	Sorcs1	E330013P04Rik
Btaf1	Btrc	Xpnpep1	
Cpeb3	Gbf1	Gpam	

Supplementary Table 7: Coordinates for intergenic CCGG probes with differential 5-hmC in adult mouse brain relative to other tissues (build *mm8;* half-open start).

Chr	Probe start coordinate	Brain avg	Other tissue avg	*P*-value	*Q*-value
chr1	78846561	2.45	8.64×10^{-5}	2.91×10^{-7}	0.009
chr9	92416795	1.77	-0.15	2.96×10^{-7}	0.009
chr1	51146331	1.90	-0.04	1.02×10^{-6}	0.019
chr1	72493877	1.76	0.10	1.87×10^{-6}	0.019
chr10	86917636	1.50	-0.25	1.63×10^{-6}	0.019
chr13	76268792	2.36	0.75	1.31×10^{-6}	0.019
chr1	139633267	1.88	0.04	2.63×10^{-6}	0.023
chr14	119838901	1.35	-0.14	3.23×10^{-6}	0.025
chr1	156603214	2.21	0.24	4.09×10^{-6}	0.026
chr10	89066698	1.80	0.02	5.51×10^{-6}	0.026
chr10	91257722	1.89	-0.05	5.26×10^{-6}	0.026
chr14	83544790	1.36	-0.17	4.70×10^{-6}	0.026
chr14	117226832	1.45	-0.32	4.38×10^{-6}	0.026
chr9	43709321	1.77	-0.07	5.98×10^{-6}	0.026
chr13	112370877	1.61	0.01	6.45×10^{-6}	0.026
chr1	77902064	1.95	0.28	7.87×10^{-6}	0.026
chr1	98379668	1.27	-0.38	8.70×10^{-6}	0.026
chr10	83795353	1.20	-0.13	8.43×10^{-6}	0.026
chr13	114468819	1.38	-0.41	8.68×10^{-6}	0.026
chr9	25761651	1.93	-0.22	7.17×10^{-6}	0.026
chr14	45383301	1.83	-0.17	1.07×10^{-5}	0.031
chr9	15660435	1.57	0.07	1.12×10^{-5}	0.031
chr1	24884965	2.04	0.28	1.18×10^{-5}	0.031
chr1	96939020	1.62	-0.11	1.28×10^{-5}	0.032
chr1	3135214	1.07	-0.19	1.65×10^{-5}	0.035
chr1	57156744	1.52	-0.03	1.66×10^{-5}	0.035
chr1	192199530	1.54	-0.07	1.54×10^{-5}	0.035
chr10	118203110	1.39	0.18	1.68×10^{-5}	0.035
chr13	44607871	1.80	0.25	1.73×10^{-5}	0.035
chr13	72140572	1.51	-0.07	1.52×10^{-5}	0.035
chr1	81617348	1.28	-0.39	1.83×10^{-5}	0.036
chr1	88168252	1.69	0.21	2.30×10^{-5}	0.036
chr1	133416455	1.42	0.08	2.31×10^{-5}	0.036
chr10	9980174	1.90	0.01	2.14×10^{-5}	0.036
chr10	35925246	1.43	-0.04	2.09×10^{-5}	0.036
chr10	68187278	1.19	-0.23	1.98×10^{-5}	0.036
chr13	97283307	1.39	-0.09	2.21×10^{-5}	0.036
chr14	13921293	1.38	-0.13	1.94×10^{-5}	0.036
chr19	36396799	1.50	-0.17	2.05×10^{-5}	0.036
chr1	13793147	1.06	-0.33	2.53×10^{-5}	0.037
chr1	159585964	1.53	0.05	2.58×10^{-5}	0.037
chr13	67687899	1.68	0.20	2.46×10^{-5}	0.037
chr1	144913497	1.26	-0.12	2.82×10^{-5}	0.039
chr19	20712929	1.34	0.04	2.84×10^{-5}	0.039
chr9	71004321	1.54	0.02	2.87×10^{-5}	0.039
chr1	159841792	1.11	-0.35	3.04×10^{-5}	0.039
chr1	189092861	1.31	-0.19	2.96×10^{-5}	0.039
chr13	49971195	1.70	0.25	3.25×10^{-5}	0.039
chr13	85162768	1.37	-0.02	3.12×10^{-5}	0.039
chr14	101564081	1.28	-0.14	3.15×10^{-5}	0.039

chr9	111681105	1.33	-0.08	3.26×10^{-5}	0.039
chr1	83732759	1.60	0.12	3.71×10^{-5}	0.039
chr1	97353272	1.48	0.00	3.50×10^{-5}	0.039
chr1	132346528	1.70	0.05	3.96×10^{-5}	0.039
chr1	138266914	1.58	0.02	3.77×10^{-5}	0.039
chr1	154066489	1.83	-0.21	3.83×10^{-5}	0.039
chr1	165409207	1.00	-0.45	3.91×10^{-5}	0.039
chr10	12826411	0.85	-0.34	3.63×10^{-5}	0.039
chr14	55672471	1.57	0.09	3.51×10^{-5}	0.039
chr19	26121407	1.07	-0.20	3.62×10^{-5}	0.039
chr9	118303616	1.63	0.16	3.87×10^{-5}	0.039
chr14	97724075	1.35	-0.14	4.06×10^{-5}	0.04
chr1	56395976	1.74	0.50	4.71×10^{-5}	0.045
chr1	172940703	1.39	-0.17	4.78×10^{-5}	0.045
chr14	100205839	1.40	-0.05	4.66×10^{-5}	0.045
chr14	20249274	1.17	-0.06	4.89×10^{-5}	0.045
chr13	37675381	1.39	0.03	5.04×10^{-5}	0.046
chr1	168823169	1.45	0.00	5.35×10^{-5}	0.047
chr19	53141154	1.97	0.00	5.38×10^{-5}	0.047
chr9	103237221	1.88	-0.04	5.33×10^{-5}	0.047
chr14	50591504	1.48	0.07	5.47×10^{-5}	0.047
chr10	13887553	2.26	0.62	5.67×10^{-5}	0.047
chr13	98890154	1.71	0.23	5.72×10^{-5}	0.047
chr14	113911315	1.22	-0.27	5.79×10^{-5}	0.047
chr13	30027675	1.33	-0.13	6.18×10^{-5}	0.049
chr13	43877899	1.37	-0.11	6.25×10^{-5}	0.049
chr14	101966649	1.01	-0.08	6.15×10^{-5}	0.049
chr19	21183149	1.62	-0.15	6.13×10^{-5}	0.049
chr1	195292433	1.26	-0.10	6.49×10^{-5}	0.049
chr10	91740445	1.29	-0.09	6.52×10^{-5}	0.049
chr10	129068102	1.41	0.28	6.52×10^{-5}	0.049
chr14	77939470	1.65	0.28	6.60×10^{-5}	0.049

Functional annotation category	Annotation terms in cluster	*P* of term	*Q* of term
Annotation Cluster 1	**Enrichment Score: 2.79**		
GOTERM_MF_FAT	GO:0005216~ion channel activity	5.4×10^{-4}	0.16
GOTERM_MF_FAT	GO:0022838~substrate specific channel activity	8.4×10^{-4}	0.16
GOTERM_MF_FAT	GO:0015267~channel activity	1.5×10^{-3}	0.18
GOTERM_MF_FAT	GO:0022803~passive transmembrane transporter activity	1.5×10^{-3}	0.18
GOTERM_MF_FAT	GO:0022836~gated channel activity	2.6×10^{-3}	0.21
GOTERM_MF_FAT	GO:0005261~cation channel activity	6.1×10^{-3}	0.32
Annotation Cluster 2	**Enrichment Score: 2.42**		
GOTERM_MF_FAT	GO:0005089~Rho guanyl-nucleotide exchange factor activity	1.0×10^{-3}	0.15
GOTERM_MF_FAT	GO:0005088~Ras guanyl-nucleotide exchange factor activity	2.2×10^{-3}	0.21
GOTERM_BP_FAT	GO:0035023~regulation of Rho protein signal transduction	2.4×10^{-2}	0.78
Annotation Cluster 3	**Enrichment Score: 2.05**		
GOTERM_BP_FAT	GO:0048666~neuron development	4.3×10^{-3}	0.48
GOTERM_BP_FAT	GO:0031175~neuron projection development	1.2×10^{-2}	0.65
GOTERM_BP_FAT	GO:0030030~cell projection organization	1.4×10^{-2}	0.68

Supplementary Table 8: Functional annotation clusters for 5-hmC enriched brain genes (DAVID). Each cluster represents a group of genes with significant overlap in annotation terms. The Enrichment Score of a cluster is the geometric mean of the exponents of the *P*-values associated with all the member terms in a cluster. The low *P*-values of individual GO terms are a trade-off for identifying clusters where genes had greater overlap in annotation terms (DAVID classification stringency = "High"). Using the default setting would have identified clusters with higher enrichment scores but lower overlap.

Supplementary Table 9: Genes enriched for 5-hmC in mouse brain, which also have steady-state mRNA levels enriched in particular cell types within brain tissue. The list of genes with statistically enriched steady-state mRNA levels was obtained from [1]. Genes with fold-enrichment > 5.0 were considered enriched in particular cell types.

Astrocyte-enriched	Neuron-enriched	Oligodendrocyte-enriched
Bcl2	2010300C02Rik	Ank3
Bmper	Arhgap20	Ccnyl1
Ednrb	Arpp21	Cnksr3
Elovl2	Atp2b1	Cpm
Eya4	Btbd11	Dock10
Fgd6	Cacna1e	Elovl7
Gli3	Cacna2d3	Jam3
Glis3	Cadps	Nfasc
Gpam	Clstn2	Opcml
Gpc5	Col19a1	Pcdh9
Gpc6	Cpne4	Serinc5
Il17rd	Dnm3	Slc45a3
Mamdc2	Dpp10	St18
Nhsl1	Erc2	Tmcc2
Nr2e1	Fgf14	Tmeff2
Ntrk2	Gfra2	Tmem108
Peli2	Gng2	Tmem163
Rfx4	Gpr39	Tmem182
Rnf182	Grem2	Unc5b
Slc41a1	Gria4	Ust
Sorbs1	Grip1	Zfp365
Tcfcp2l1	Hcn1	
Tgfb2	Hecw1	
Tnfrsf19	Hivep2	
Tph2	Hs3st5	
	Kcnc2	
	Kcnh1	
	Kcnma1	
	Kcnt2	
	Lin7a	
	Lingo1	
	Mctp1	
	Nav3	
	Nrg3	
	Ntrk2	
	Nts	
	Pfkp	

Pkib
Ppfia2
Ptk2b
Ptprk
Rab3c
Rgs17
Rgs7bp
Rgs8
Rims1
Ryr2
Slc35f4
Slc6a15
Slco5a1
Spock1
Ssbp2
Stau2
Syne1
Synpr
Syt1
Trhde

		Brain, BA10 (Control)	Brain, BA10 (Bipolar)	Brain, BA10 (Schizophrenia)	Liver
Number of Samples		28	28	30	13
Mean Age in years (range)		49 (32 – 80)	49 (29 – 75.5)	49.5 (24 – 75.5)	54.31 (31 – 75)
Sex	Male	13	15	15	6
	Female	15	13	15	7
Post Mortem Interval in hours, (std. dev.)		30 (12.4)	32.6 (19.9)	32.4 (15.6)	5 (2.2)

Supplementary Table 10: Demographic information for human samples.

				5-hydroxymethylcytosine (5-hmC)			5-methylcytosine (5-mC)		
Biological context	***n*** [1]	**# E** [2]	**# I** [2]	**E – I** [3]	**95%CI**	***P*** [4]	**E – I** [3]	**95%CI**	***P*** [4]
d = 5 bp from exon-intron boundary									
Human									
Brain BA10, Controls	28	762	43	0.12	[0.07, 0.16]	5.7×10^{-8}	-0.01	[-.06, 0.04]	0.85
Brain BA10, Psychosis	54	239	26	0.11	[0.06, 0.14]	6.8×10^{-9}	- 0.01	[-0.05, 0.05]	0.06
Brain BA10, Controls (exp2)	12	339	26	0.06	[-0.04, 0.17]	1.4×10^{-2}	0.13	[0.05, 0.24]	2.60×10^{-4}
Liver (exp2)	13	339	26	0.02	[-0.05, 0.12]	0.30	0.16	[0.09, 0.26]	4.80×10^{-6}
Mouse									
Frontal cortex	15	393	11	0.00	[-0.12, 0.18]	0.35	0.22	[0.14, 0.34]	9.20×10^{-4}
Rest of the brain	15	393	11	0.05	[-0.07, 0.22]	0.60	0.11	[-0.04, 0.10]	6.40×10^{-2}
Liver, Pancreas, Heart, Kidney	24	504	15	-0.01	[-0.13, 0.08]	0.25	0.11	[0.04, 0.24]	3.10×10^{-2}
Neuronal cell line (mouse)	18	285	10	0.07	[-0.04, 0.21]	0.72	0.12	[-0.05, 0.28]	2.80×10^{-2}
B-lymphocyte cell line (human)	6	400	13						
Vehicle- treated (0 μM SAHA)	6	400	13	-0.03	[-0.11, 0.06]	0.95	0.21	[0.02, 0.45]	0.40
$1/10^{th}$ of IC-10 (0.01 μM SAHA)	6	400	13	0.07	[-0.13, 0.24]	0.35	0.10	[0.02, 0.45]	0.27
$1/5^{th}$ of IC-10 (0.02 μM SAHA)	6	400	13	0.09	[-0.11, 0.20]	0.97	0.05	[-0.26, 0.24]	0.95
IC-10 (0.10 μM SAHA)	6	400	13	-0.07	[-0.26, 0.08]	0.44	0.06	[-0.27, 0.26]	0.84
d= 20 bp from exon-intron boundary									
Human									
Brain BA10, Controls	28	2,445	811	0.05	[0.04, 0.06]	4.30×10^{-23}	0.02	[0.01, 0.03]	1.70×10^{-4}
Brain BA10, Psychosis	54	1,148	424	0.04	[0.03, 0.06]	9.20×10^{-20}	0.01	[0.00, 0.03]	2.60×10^{-2}
Brain BA10, Controls (exp2)	12	1,148	424	0.07	[0.05, 0.08]	1.50×10^{-12}	0.05	[0.03, 0.07]	4.60×10^{-10}
Liver (exp2)	13	1,148	424	0.01	[-0.01, 0.03]	3.10×10^{-2}	0.05	[0.03, 0.07]	4.80×10^{-12}
Mouse									
Frontal cortex	15	1,244	309	0.06	[0.04, 0.09]	7.50×10^{-6}	-0.01	[-0.03, 0.01]	0.84
Rest of the brain	15	1,244	309	0.06	[0.04, 0.09]	3.00×10^{-8}	0.01	[-0.02, 0.04]	0.33
Liver, Pancreas, Heart, Kidney	24	1,613	422	0.03	[0.01, 0.04]	5.70×10^{-3}	0.08	[0.06, 0.10]	3.40×10^{-19}

Neuronal cell line (mouse)	18	857	203	0.03	[-0.00, 0.06]	4.50×10^{-2}	0.13	[0.10, 0.17]	1.60×10^{-17}
B-lymphocyte cell line (human)									
Vehicle- treated (0 μM SAHA)	6	1,301	418	0.02	[-0.01, 0.05]	0.275	0.08	[0.01, 0.12]	2.70×10^{-5}
$1/10^{th}$ of IC-10 (0.01 μM SAHA)	6	1,301	418	0.06	[-0.01, 0.08]	0.0144	0.09	[0.05, 0.15]	6.10×10^{-8}
$1/5^{th}$ of IC-10 (0.02 μM SAHA)	6	1,301	418	0.05	[0.02, 0.10]	8.8×10^{-4}	0.05	[0.02, 0.11]	4.00×10^{-2}
IC-10 (0.10 μM SAHA)	6	1,301	418	0.05	[0.02, 0.09]	3.6×10^{-4}	0.06	[0.00, 0.09]	7.20×10^{-4}

[1] Number of biological samples; [2] Number of measurements for boundary side = (# probes * *n*); [3] Median increase in exons, relative to introns (value > 0 indicate higher exonic levels); [4] P-values from linear mixed-effects model (Online Methods)

Supplementary Table 11: Statistics on DNA modification changes at cumulative distances (d = 5 and d = 20) from the exon-intron boundary.

	Num. arrays	probes per array	5-hmC + 5-mC		5-hmC		5-mC	
			Median	IQR	Median	IQR	Median	IQR
Human								
Controls	56	136,477	0.26	1.02	0.11	0.99	0.14	0.97
Psychosis	54	69,252	0.34	1.00	0.11	1.01	0.23	0.97
Exp2, Brain	12	69,252	0.45	0.97	0.23	0.86	0.23	0.81
Exp2, Liver	13	69,252	0.40	0.98	0.04	0.84	0.37	0.90
Mouse								
Frontal cortex	15	69,052	0.56	1.02	0.19	0.91	0.37	0.90
Brain, non-frontal cortex	15	69,052	0.59	1.02	0.18	0.90	0.41	0.93
Non-brain organs	36	130,314	0.57	1.10	0.05	0.89	0.51	1.05
Cell lines								
Human B-lymphocyte	24	76,102	0.36	0.99	-0.07	0.88	0.44	0.98
Mouse neuronal	18	46,892	0.47	1.05	0.02	0.93	0.47	1.06

Supplementary Table 12: Median probe intensities for DNA modifications for all datasets used in the exon-intron boundary analysis**.** Probes were pooled across all samples and arrays without aggregation before median and interquantile range (IQR) were computed. Consistent with previous literature, 5-hmC intensities are higher in tissues sampled from the brain, relative to those sampled from other tissues (e.g. human liver, mouse non-brain organs). Negligible 5-hmC was detected in the two cell lines tested.

	Channel	Aligned	Target (T)	Non-target (NT)	(T / (T+NT)) * 100
Replicate 1	Undigested	408,987	1,341	276,142	0.5
	gDNA (MspI)	2,761,844	1,083,483	662,416	62.1
	gDNA (HpaII)	1,407,105	444,298	511,772	46.5
	glc-gDNA (MspI)	1,231,589	421,646	356,355	54.2
Replicate 2	Undigested	1,301,480	4,766	850,546	0.6
	gDNA (MspI)	2,224,007	829,158	593,030	58.3
	gDNA (HpaII)	2,840,191	918,979	1,004,470	47.8
	glc-gDNA (MspI)	3,166,749	1,185,353	833,898	58.7
Replicate 3	Undigested	2,395,206	5,758	1,663,120	0.3
	gDNA (MspI)	2,166,860	646,267	620,031	51.0
	gDNA (HpaII)	2,115,006	662,905	534,374	55.4
	glc-gDNA (MspI)	1,487,507	374,791	469,684	44.4

Supplementary Table 13: Read counts from Helicos single-molecule sequencing. Target reads are reads where the 5' end lies within ± 3bp of a CCGG site. Non-target reads are reads where the 5' end lies outside ±200bp of a CCGG site.

	Brain (six chromosomes)		Liver (three chromosomes)	
	Constitutive exons	**Alternative exons**	**Constitutive exons**	**Alternative exons**
# exons on array chroms	5,862	980	1,048	118
Whole exon				
exons with probes	1,010	95	177	15
probes	1,234	199	224	18
d<=20 from boundary				
exons with probes	349	30	71	4
probes	358	31	73	4

Supplementary Table 14: Exonic probe count of RNAseq data from human liver and brain (cortex). RNAseq data was obtained from Brawand *et a*l., 2011 [2]

Adaptor end	Adaptor sequence (5'-3')
A1	5’ AGT TAC ATC TTG TAG TCA GTC TCC A 3’
A25	5’ TGG AGA CTG ACT ACA AGA T 3’

Supplementary Table 15: Adaptor primer sequence for blunt-ended adaptors ligated to sheared genomic DNA. Adaptors are prepared by mixing equal molar amounts (100 μM) of complementary primers annealed in 10 mM TrisHCl (pH 8.0), heating at 95 ^{o}C for 5 minutes followed by slow cooling (1 ^{o}C/minute) to room temperature.

Supplementary Note 1

Production of a 31-mer DNA duplex containing modified cytosines at a CCGG target site

Equal molar amounts (150μM) of complementary single-stranded oligonucleotides (5'-tgacccacgctcgcc and 3'-actgggtgcgagcgggcctctatttaataca) were annealed in water by heating at 95°C for 5 minutes, followed by slow cooling to room temperature. Annealed DNA (5 μM) was supplemented with dGTP, dTTP, dATP and dCTP, dmCTP or dhmCTP (Bioline, USA) (1mM each) and Klenow Fragment (0.15 U / μl, Fermentas), and incubated in Klenow reaction buffer at 37°C for 40 minutes to produce duplexes containing cytosine (C), 5-mC or 5-hmC at the target site, respectively. 1 μM of duplex oligo with 5-hmC, 200 μM UDP-Glc (Sigma) and 0.04 μg BGT were incubated for 1.5 hrs at 37°C in buffer (15 μl, 100 mM Tris-HCl pH 8.0, 25 mM magnesium chloride). Then, 2 μl of Tango buffer, 1 μl (10 U) of MspI (Fermentas) and 2 μl of water was added to the reaction, and incubation was continued for 1.5 hrs. Samples were supplemented with 1/6 of 6x Loading Dye Solution and analyzed by 15% polyacrylamide gel electrophoresis.

Thin-layer chromatography quantification of total genomic 5-mC, 5-hmC and C at CCGG sites

Genomic DNA (40 ng) was digested with MspI (Fermentas) endonuclease for 2 hrs at 37°C and dephosphorylated with 0.1 U of FastAp (Fermentas) for 1 h at 37°C. Enzymes were inactivated by heating at 75°C for 10 minutes. 5'-end-labelling of DNA fragments was carried out with 4 U T4 Polynucleotide Kinase (Fermentas) in the presence of 3.3 μCi of [$^{\gamma}$33-P]-ATP (Hartmann Analytic) at 37°C for 10 minutes in T4 Polynucleotide Kinase reaction buffer, followed by enzyme inactivation at 90°C for 3 minutes. Labelled fragments were ethanol-precipitated using sodium acetate pH 7.0 (3 M) as part of a standard protocol. Air-dried pellets were dissolved in 4 μL Lambda Exonuclease buffer and incubated with 2.5 U Lambda Exonuclease at 37°C for 2 hours (Fermentas). Aliquots of hydrolysate (3 replicates) were spotted on PEI cellulose plates (PEI Cellulose F, 20 x 20 cm, Merck) and chromatographed by eluting with isobutyric acid/water/conc. ammonia (66:17:4, vol/vol/vol). Plates were dried, autoradiographed to phosphorimager screens and analyzed with a FLA-5100 scanner and MultiGauge software (Fujifilm). Ratios of C, 5-mC and 5-hmC were calculated after subtracting corresponding gel density values from control experiments. Note that methylation of repetitive elements was quantified by TLC, while repeat-overlapping probes were excluded from the microarray data analysis; this difference could partially account for the discrepancy between these two methods.

Quantitative Polymerase Chain Reaction

BGT-treated and -untreated DNA was subjected to MspI digestion. In addition, DNA was digested by HpaII and an undigested control was used (Online Methods). Locus-specific real-time PCR was performed using 10 ng genomic DNA and SYBR® Green PCR Master Mix (Applied Biosystems) on the 7500 Real-Time PCR System (Applied Biosystems), according to the manufacturer's recommendations (melting temperature of 60ºC). Primer sequences and genomic coordinates

(UCSC genome build *hg18*) tested for qPCR are listed in Supplmentary Table 3. Each sample was performed in duplicate and the corresponding *Ct* values were obtained from the 7500 System SDS Software v1.3.1 (Applied Biosystems). All primer pairs flanked either one or two MspI target sites (CCGG) (target primer pairs). One primer pair did not flank an MspI target site and was used as an internal control (reference primer pair, *ref*). The efficiency (*E*) of each primer pair was calculated from the slope of regression line obtained by plotting *Ct* values against varying DNA concentration [3]. ΔCt and percent modification values were calculated from the formula:

$$\Delta Ct = Ct_{meanUndigested} - Ct_{meanDigested}$$

$$\% \text{ modification} = \frac{(E_{target})^{\Delta Ct(target)}}{(E_{ref})^{\Delta Ct(ref)}} \times 100$$

Adaptor PCR amplification for Affymetrix tiling arrays

Restriction enzyme-digested DNA fragments were amplified with an adaptor primer (5'-agttacatcttgtagtcagtctcca-3'), and dUTP was included in the dNTP mix as specified by Affymetrix. Two rounds of PCR amplifications were performed to achieve the required DNA amount for tiling array hybridization. PCR cycling for the 1st round of amplification was performed on the restriction enzyme-digested gDNA sample. The second round of amplification was done on 1/10th of the 1st PCR template; both rounds of amplification used the same PCR cycling conditions (i.e. 95ºC for 1 minute, followed by 15 cycles of 94ºC for 15 seconds, 65ºC for 30 seconds and 1 minute at 72ºC, with an extension of 5 second in each subsequent cycle). The amplicons were then purified using QIAquick 96 PCR Purification Kit (Qiagen) and checked for quality and quantity on a NanoDrop 2000 spectrophotometer (Thermo Scientific). Nine micrograms of PCR amplicons were fragmented to 50–100 bp using uracil DNA glucosylase enzyme, which cleaves DNA at incorporated dUTP (GeneChip® WT Double-Stranded DNA Terminal LabelingKit, Affymetrix). Fragments were end-labeled according to the manufacturer's instructions. Prior to labelling, 1 µL of fragmented DNA was analyzed on a Bioanalyzer using the DNA1000Chip (Agilent Technologies) to check the uniformity of the fragmented products. Individual samples were hybridized on a separate Gene Chip of Human or Mouse Tiling 2.0R array for 16 hrs at 45°C.

Selection of array normalization algorithm

We first investigated various methods of array preprocessing to identify the algorithm best suited to analyze DNA modification data on tiling arrays. We considered quantile normalization and two variants of probe-sequence based normalization. Quantile normalization, a conventional choice, results in every microarray having the same overall intensity distribution, an assumption that may be invalid in cases where microarrays represent different tissues and interrogate modifications that may vary several-fold in magnitude among them [4] (e.g. 5-hmC is higher in brain than in other tissues [5-7]). Moreover, it does not explicitly correct for probe sequence-based affinity bias, a known issue in tiling arrays [8]. We considered MAT (model-based analysis of tiling arrays, [9]) and an alternative

sequence-based normalization scheme with fewer parameters (the "Potter" algorithm; [10]). We then correlated single-probe intensities normalized using each algorithm with 11 arbitrary loci on which we performed quantitative PCR. The Potter algorithm showed the highest correlation with qPCR estimates (Supplementary Table 2), so we chose this algorithm. We also determined that fitting the sequence-based model (equation 1 in Supplementary Note 1) to non-target probes with the same GC-composition as the target-probes, rather than to all non-target probes, resulted in a more uniform baseline for the non-target probes (not shown, for definition of target and non-target probes, see Online Methods).

It was originally unclear if targets analyzed at the single-probe level had a smaller measurement bias than those analyzed by averaging probe intensities in a window surrounding the target. We therefore correlated digestion efficacies from qPCR experiments with microarray intensities measured at the single-probe level, and using rectangular or distance-weighted windows (Supplementary Table 2). Both types of windows were tested at longer (~340 bp, microarray amplicon size) and shorter (~100 bp, average size of qPCR amplicon) lengths. Single-probe intensities showed the strongest correlation with qPCR estimates (Supplementary Table 2, Supplementary Fig. 2a). Windowed probe averages in glucosylated samples had dramatically lower correlations with qPCR estimates, relative to single probe measurements (Correlations: Single probe = 0.52, 100 bp rectangular window = 0.03, 100 bp distance-weighted window = 0.17). We concluded that single-probe estimates provided the best balance between bias and precision for these data, and analyzed our data at the single probe level.

Array normalization

Non-target probes were first trimmed to proportionally match target probes in GC content. The probe sequence-based affinity model (equation (1), the "Potter" model) was applied to non-target probes. The fitted value was subtracted from raw intensities of all probes, resulting in normally-distributed probe-level intensities. In equation (1), α corrects for baseline chip-level intensity differences, β represents the total number of each nucleotide, γ and θ for position of each nucleotide. Each chip was individually normalized. All downstream analyses were carried out at the single-probe level (i.e. without windowing or peak-calling) and exclusively on target probes (henceforth referred to simply as 'probes').

$$\hat{y} = \alpha + \sum_{j\epsilon\{A,C,G,T\}} \beta_j n_j + \sum_{j\epsilon\{A,C,G\}} \gamma_j \sum_{k=1}^{25} I(b_k = j) + \sum_{j\epsilon\{A,C,G\}} \theta_j \Big(\sum_{k=1}^{25} I(b_k = j) \Big)^2 \quad (1)$$

Values for various DNA modifications were generated by computing the log-ratios of base channels of a given biological sample (restriction enzyme treatments are indicated by corresponding names in parenthesis; all values are log2-transformed):

$$5hmC = log_2(gDNA_{glu}(MspI)) - log_2(gDNA_{ref}(MspI)) \quad (2)$$

$$5mC = log_2(gDNA_{ref}(HpaII)) - log_2(gDNA_{glu}(MspI)) \quad (3)$$

$$5hmC + 5mC = log_2(gDNA_{ref}(HpaII)) - log_2(gDNA_{ref}(MspI)) \quad (4)$$

Identification of differentially enriched 5-hmC intergenic regions in the mouse brain

We identified differential 5-hmC in intergenic regions using probe-wise linear regression. Intergenic probes were defined as probes which did not overlap any RefSeq genes on either strand; 60,721 probes met this criterion. A probe-wise linear regression was conducted, with the regressor being an indicator variable of tissue type 'Brain' or 'Other' (*lmFit* from the R package limma). The fit was first moderated using Empirical Bayes shrinkage (eBayes), and nominal p-values were adjusted using Benjamini-Hochberg FDR. Eighty-three probes had *Q*-values < 5 % and were called 'differential'. All 83 probes were enriched in the brain, relative to other tissues.

Functional annotation analysis of 5-hmC rich genes

Gene Ontology (GO) overrepresentation analysis (ORA) was done using DAVID (Database for Annotation, Visualization and Integrated Discovery [11]); for the background gene set, we used the 5,925 RefSeq IDs associated with the 4,357 genes (defined by MGI symbols) for which tests were performed. The foreground consisted of genes (MGI symbols) identified as enriched based on gene-wise tests. GO-related databases (GOTERM_CC_FAT, GOTERM_BP_FAT, and GOTERM_MF_FAT) were chosen for annotation databases.

DAVID also identifies 'clusters of annotation terms' with member genes that share annotation terms more than expected by chance. In part, this aggregation serves to combine individual terms into groups potentially representing biological pathways. The 'Classification Stringency' parameter was set to "High" (Default = "Medium") to create smaller clusters with greater overlaps in annotation terms. The Enrichment Score (ES) of an annotation cluster is the geometric mean of the exponents of *P*-values associated with individual member annotation terms in the cluster [11].

Categorization of genes by brain cell type

The list of genes with cell-type specific enrichment scores was downloaded from the Supplementary Online Material accompanying a dataset of steady-state mRNA levels in FACS-sorted brain cell populations [1]. Genes with relative mRNA enrichment > 5.0 were called as being enriched in a particular cell-type. Genes with > 20.0 enrichment were deemed to be cell-type specific (after analyses and threshold set in the source paper).

Analysis of genes with particular GO terms, for mouse and human brain

The list of all mouse (or human) genes mapped to specific GO terms was downloaded from the AmiGO Gene Ontology browser (release date 2011-05-07, AmiGO version 1.8, download date 2011-05-13 (mouse), 2011-05-15 (human)). Gene association files were downloaded for GO:0045202 ("synapse"), GO:0044456 ("synapse part"), GO:0007155 ("cell adhesion"), and GO:0005886 ("plasma membrane") (filter for species *Mus musculus* (or *Homo sapiens*); GO evidence codes not filtered). Genic probes were defined as those that overlapped RefSeq genes on at least one strand (*refGene* table from UCSC genome browser, *hg18* for human, *mm8* for mouse). Genes on interrogated tiling arrays were divided into those that were mapped to the GO term being analyzed, and those that were not. Within each group of genes, individual probes were first averaged (mean) across samples in the tissue group (e.g. brain). Probes were not averaged across a gene. GC content of each probe was obtained using the probe sequence provided in the Affymetrix array annotation (bpmap) file.

Calculation of exon-intron boundary differential

A linear mixed-effects model [12] was used to test probe intensity differences between the exonic and intronic side of the junction, using junction side (*junctionSide*='Exon' or 'Intron') as the fixed-effects term, and sample (*Sample* in eqn 5,6) as random-effects terms (*lmer4* package in R). For datasets that used multiple array types, array type (*Array* in eqn. 5,6) was used as an additional random-effects term.

ANOVA was used to determine whether the data better fit the null model:

$$Intensity = 1 + (Array + Sample) + residual \qquad (5)$$

or the alternative model, which took into account the side of the junction on which the probe occurred (*junctionSide*)):

$$Intensity = 1 + junctionSide + (Array + Sample) + residual \qquad (6)$$

Tests with P-value < 0.025 were deemed significant.

The Wilcoxon-Mann-Whitney test (WMW test), a more common choice for testing difference in medians, would have been an inappropriate choice to compare exonic and intronic intensities. Our data contained multiple measurements per sample, violating the assumption of independence required by the WMW.

Relating DNA modifications to mRNA levels with transcriptomic data

We used a previously-published dataset (GSE10246 [13]) that measured steady-state mRNA levels in a variety of adult mouse tissues. Normalized expression values were downloaded in series matrix format from the Gene Expression Omnibus [14], and analyzed in R using the BioConductor package GEOquery [15] . Array annotation was downloaded from Bioconductor ("mouse4302.db"). Probes were averaged across samples within a tissue, and then averaged within RefSeq IDs.

The transcriptomic dataset was validated prior to use. Samples were subjected to unsupervised hierarchical clustering (distance = Pearson's correlation, clustering method = "ward"), and the cluster heatmap was manually examined to establish that tissues with similar developmental origin were grouped into closer subtrees than tissues from different cellular lineages (heatmap visualization done in Seurat [16]). Further spot checks were done for individual genes with a known characteristic expression pattern (e.g. *Nanog*, a transcription factor expressed in embryonic stem cells, is expected to be relatively overexpressed in ES cell lines and underexpressed in differentiated tissues). RNA samples were separated into brain ("cerebral_cortex_prefrontal", "cerebral_cortex", or "cerebellum", n = 6 arrays), liver, heart, kidney and pancreas (2 arrays each). For each tissue, genes (RefSeq IDs) were stratified into deciles, based on mRNA level.

Separately, in our dataset of DNA modification estimates, samples were grouped by tissue (brain = 11; liver, kidney, heart, pancreas = 9 arrays each). For each tissue, probes were first averaged across samples and then within a gene, resulting in one value per RefSeq ID. Genes were binned according to their mRNA expression decile (previous paragraph), and the average quantity of 5-mC or 5-hmC in each decile was computed.

Helicos single molecule sequencing (SMS) and analysis

Micrococcus nuclease digestion was used to fragment genomic DNA to a median size of 500 bp and to reduce 3' hydroxyl end at the DNA fragments, where the latter served as the starting end for SMS. 5 µg of genomic DNA was treated with 1 U of micrococcal nuclease enzyme (NEB) and the reaction was stopped by adding 10 µl of 0.5M EDTA (in excess) in a time series. A small aliquot was then checked on 1% agarose gel and samples with median fragment size of 500 bp were column purified with buffer PN (QIAquick Nucleotide Removal Kit columns, Qiagen). Glucosylation and control treatments were performed as described before (Online Methods), and 200ng of each glucosylated or non-glucosylated treated DNA was subjected to 10 U of restriction enzyme digestions respectively at 37 ^{o}C for 8h, and inactivated at 80^{o}C for 20 minutes.

10 ng of each digested product, quantified by Quant-iTTM PicoGreen dsDNA Reagent Kit (Invitrogen), was then processed for Helicos sequencing. In brief, 10ng of DNA was heat denatured at 95^{o}C for 5 minutes prior to 3' end labeling with 5 U of terminal transferase (NEB) in presence of 200 µmoles of dATP (Roche) and 5 mmoles of $CoCl_2$ (NEB) in 20 µl reaction volume at 37^{o}C for 1 h, and then inactivated at 70^{o}C for 10 minutes. Fragments were biotinylated by repeating the terminal transferase enzymatic reaction step in the presence of 100 µmoles of biotin labeled ddATP (Perkin Elmer) instead of dATP in a reaction volume of 30 µl. These processed samples were then sent to the Helicos sequencing service facility (www.helicosbio.com; USA).

Three technical replicates of the same human brain DNA sample were processed for glucosylation and respective restriction digestion with MspI enzyme with or without glucosylation treatment, and

with HpaII enzyme on non-glucosylated gDNA. Data from all three runs were pooled for analysis, after each run had been separately normalized using the corresponding number of non-target reads (see below). SMS reads were trimmed for leading "T" homopolymers, filtered for reads with a minimal length of 25 bases after trimming as well as for other standard Helicos quality metrics using a suite of Helicos tools available at: http://open.helicosbio.com/mwiki/index.php/Releases. Alignments to the *hg18* version of the human genome were conducted with indexDPgenomic software freely available on the Helicos website (http://open.helicosbio.com/mwiki/index.php/Releases). The sequence reads were aligned using a minimum normalized score of 4.3. Only uniquely-mapped reads were considered for the present analysis (Supplementary Table 13 for read counts).

Reads with a 5' coordinate $\leq$ 3 bp from a target sequence (CCGG) were defined as target reads. Reads with a 5' coordinate $\geq$ 200 bp away from a CCGG sequence were used to normalize the read count. Junction distance of target reads were computed as for the microarray analysis, using the coordinate of modifiable cytosine (underlined "C" in "C<u>C</u>GG") of the read-associated target site. Raw reads were first aggregated by junction distance (e.g. distance to exon start/end or intron start/end) respectively for both channels (unglucosylated or glucosylated DNA samples). Aggregated reads were normalized by non-target read count and scaled relative to the number of reads in the channel with non-glucosylated DNA. Percent 5-hmC was computed as the fold-difference in reads from the glucosylated channel, relative to those in the non-glucosylated channel. The proportion of reads arising from CCGG target sequences was greater in the non-glucosylated DNA sample, compared to the glucosylated DNA sample. This is expected since higher levels of digestion will generate more DNA fragments with 3' hydroxyl ends, a prerequisite for Helicos single molecule sequencing (Supplementary Table 13).

HpaII digestion resulted in more reads than expected from previous estimates of total DNA modification in the average mammalian cell. (Supplementary Table 13, [17]). One possibility is that the HpaII enzyme generates single strand nicks in the modified DNA, which remained undetected in earlier studies that estimated total DNA methylation within the genome; this observation requires further investigation. For this study, only MspI digestion was taken in account, as it has identical restriction conditions for glucosylated and for non-glucosylated DNA.

Identification of cassette exons for exon inclusion analysis

To identify cassette exons, first, all available human expressed sequence tags (ESTs) and mRNA sequences were mapped to the human genome (*hg19*) using SIM4. The information on intron-exon structures was then merged with Ensembl annotation (release 65). From this database, a bowtie library of exon-exon junction (EEJ) sequences was generated by combining every possible splicing donor and acceptor within each gene. RNAseq from liver and cortex [2] was mapped to this library using bowtie with –m 1 –v 2 parameters. Reads were trimmed to 50 nucleotides and reads mapping to the genome were previously discarded (since EEJs must not exist in the genome). A minimum of eight mapped nucleotides were required at each of the two exons in a given EEJ. The outputs were

then parsed to identify cassette exons (exons that are either included or fully excluded from the transcripts), by examining exons that have associated reads maps to (i) both EEJs, supporting the inclusion of the exon (constitutive upstream exon (C1)-cassette exon (A) or A-constitutive downstream exon (C2)) and (ii) the EEJ for the exclusion of the exon (C1-C2). Genome coordinates were converted to build *hg18* (liftOver, UCSC genome browser) prior to the analysis with DNA modification arrays.

Treatment of suberoylanilide hydroxamic acid (SAHA) on human B-lymphocyte cells

Transformed human B-lymphocyte cells (GM10851, Coriell Cell Repositories) were treated with the histone deacteylase inhibitor SAHA. Prior to the experiment, a cell viability assay (ATP luminescence assay; Cell Titer-Glo; Promega) for SAHA was conducted by titrating different SAHA concentrations. The maximum concentration of SAHA that induced minimal cytotoxicity (e.g., not more than a 10 % decrease in ATP levels on the cytotoxicity concentration response curve) is referred to as IC10 (0.1 μM), while the other concentrations used were 1/5th (0.02 μM) and 1/10th (0.01 μM) of the maximum concentration. SAHA concentrations were dissolved in DMSO (Fisher Scientific). To assess the influence of SAHA on 5-hmC DNA modification, B-lymphocytes cells cultured at 37^0C in 6-well plates were exposed to SAHA for 30 or 72 hrs. A comparable cell confluence was attained for each time point by plating 1×10^6 cells in 4 mL of culture media (RPMI 1640 with 1 % l-glutamine (Invitrogen) supplemented with 15 % FCS (USDA tested (Hyclone)) for the 30 hrs time point and 0.3×10^6 cells in 2.4 mL for the 72 hrs time point. For the 30 hrs time point cells, each of the 3 compound concentrations or vehicle (DMSO, with less than 0.4 % DMSO/well) were added at 5X in 1 mL culture media, while for the 72 hrs time point each of 3 compound concentrations or vehicle were initially added at 5X in 0.6 mL culture media and then at 24 and 48 hrs time points, 1X compound concentration or vehicle in 1 mL media was added to each well. Triplicates were performed for each respective treatment and cells were harvested for gDNA extraction. Genomic DNA was isolated with phenol chloroform and isopropanol precipitation and glucosylation, restriction enzyme digestion and analysis on tiling microarray were performed as described before.

Supplementary Note 2

Verification of glucosyltransferase-based quantification of 5-hmC

We performed three groups of control experiments to demonstrate the validity of using T4 β-glucosyltransferase (BGT) to estimate the quantity of 5-hmC (Online Methods). First, glucosylation treatment was investigated on a 31-mer DNA duplex (see below) that contained 5-hmC modification (Supplementary Fig. 1a). Second, we determined the influence of the glucosylation treatment on unmethylated (C) and on methylated cytosines (5-mC). This was performed on whole genome PCR-amplified DNA that had lost all genomic modifications. The glucosylation and restriction digestion procedure was then applied to either whole-genome amplified (WGA, unmethylated genome) or SssI methyltransferase treated WGA DNA (fully methylated genome). Real time qPCR was used to estimate the % modifications (5-mC or 5-hmC) present at specific loci (n = 3). These two control experiments showed that there is no influence of the glucosylation procedure on 5-mC or on unmethylated cytosines, and that it is specific for 5-hmC. As a third control, we evaluated the linearity of the measure of 5-hmC by employing the BGT-based procedure in a model system (Supplementary Fig. 1b). A 200 bp DNA fragment containing one MspI/HpaII site for qPCR analysis was generated by PCR from mouse genomic DNA with primers 5'-gcatcctggagattgtgggcaacatc^{hm}cgg (IBA, Germany) and 5'-gcccatgtcgctgtg (Metabion, Germany). Enzymatic BGT glucosylation of the PCR product was perfomed in the presence of UDP-G (Online Methods) and PCR products were subsequently subjected to MspI restriction hydrolysis for 16 hrs. Real-time PCR experiments were performed with a Rotor-Gene™6000 real-time PCR system (Corbett Research) using Maxima™SYBR Green qPCR Master Mix (Fermentas); 0.3 mM primers were used in each reaction in a final volume of 25 µl. The amplification program was set as: 95°C for 10 minutes, 40 cycles for 15 s, 60°C for 1 minutes, and a melt curve analysis step at the end to check the specificity of the PCR product. Data were analyzed by Rotor-Gene™6000 real-time PCR software.

Comparison of biological versus technical variability

Genomic DNA from two human brain samples (Stanley Medical Research Institute (SMRI) [18] was used to create two sets of technical replicates. Each DNA sample was split six ways, and six technical replicates were generated for MspI-treated genomic DNA (MspI-gDNA). These technical replicates were compared to six biological replicates, using MspI-treated genomic DNA from six individual human brain samples (SMRI; samples randomly chosen in R from full set of 28 used in the study). DNA was hybridized onto Affymetrix 2.0R human whole-genome tiling arrays (Array E: chr 5,7,16), generating a total of 24 arrays. Arrays were normalized using the Potter algorithm (Supplementary Note 1) and target probes were extracted for chromosome 5 (27,546 probes). For each of the three sets (two technical replicate sets, and one set of biological replicates), we computed the sample range of individual probe intensities. The probe-wise range in technical replicates (presumably owing to technical variation) was subtracted from that in biological replicates (Supplementary Fig. 2b), and the shift in range was tested using a one-sample t-test (α =

0.05)

References

1. Cahoy, J.D. et al. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. *J Neurosci* **28**, 264-78 (2008).
2. Brawand, D. et al. The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343-8 (2011).
3. Pfaffl, M. Real-time PCR. Vol. 63 (ed. MT, D.) 63--82 (Taylor and Francis Group, 2006).
4. Laird, P.W. Principles and challenges of genome-wide DNA methylation analysis. *Nat Rev Genet* **11**, 191-203 (2010).
5. Globisch, D. et al. Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates. *PLoS One* **5**, e15367 (2010).
6. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science* **324**, 929--930 (2009).
7. Nestor, C.E. et al. Tissue-type is a major modifier of the 5-hydroxymethylcytosine content of human genes. *Genome Res* (2011).
8. Liu, X.S. Getting started in tiling microarray analysis. *PLoS Comput Biol* **3**, 1842-4 (2007).
9. Johnson, W.E. et al. Model-based analysis of tiling-arrays for ChIP-chip. *Proc Natl Acad Sci U S A* **103**, 12457--62 (2006).
10. Potter, D.P., Yan, P., Huang, T.H.M. & Lin, S. Probe signal correction for differential methylation hybridization experiments. *BMC Bioinformatics* **9**, 453 (2008).
11. Huang, D.W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44--57 (2009).
12. Vittinghoff E., G.D.V., Shiboski S.C., McCulloch C.E. Regression Methods in Biostatistics. 253-290 (2005).
13. Lattin, J.E. et al. Expression analysis of G Protein-Coupled Receptors in mouse macrophages. *Immunome Res* **4**, 5 (2008).
14. Edgar, R., Domrachev, M. & Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30**, 207-10 (2002).
15. Sean, D. & Meltzer, P.S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **23**, 1846--7 (2007).
16. Gribov, A. et al. SEURAT: visual analytics for the integrated analysis of microarray data. *BMC Med Genomics* **3**, 21 (2011).
17. Robertson, K.D. & Jones, P.A. DNA methylation: past, present and future directions. *Carcinogenesis* **21**, 461-7 (2000).
18. Torrey, E.F., Webster, M., Knable, M., Johnston, N. & Yolken, R.H. The stanley foundation brain collection and neuropathology consortium. *Schizophr Res* **44**, 151-5 (2000).