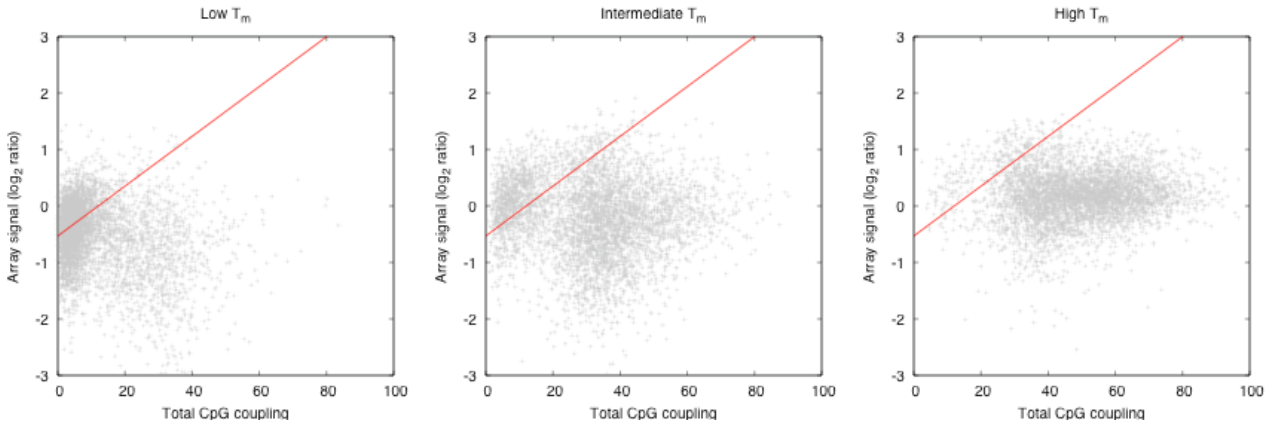
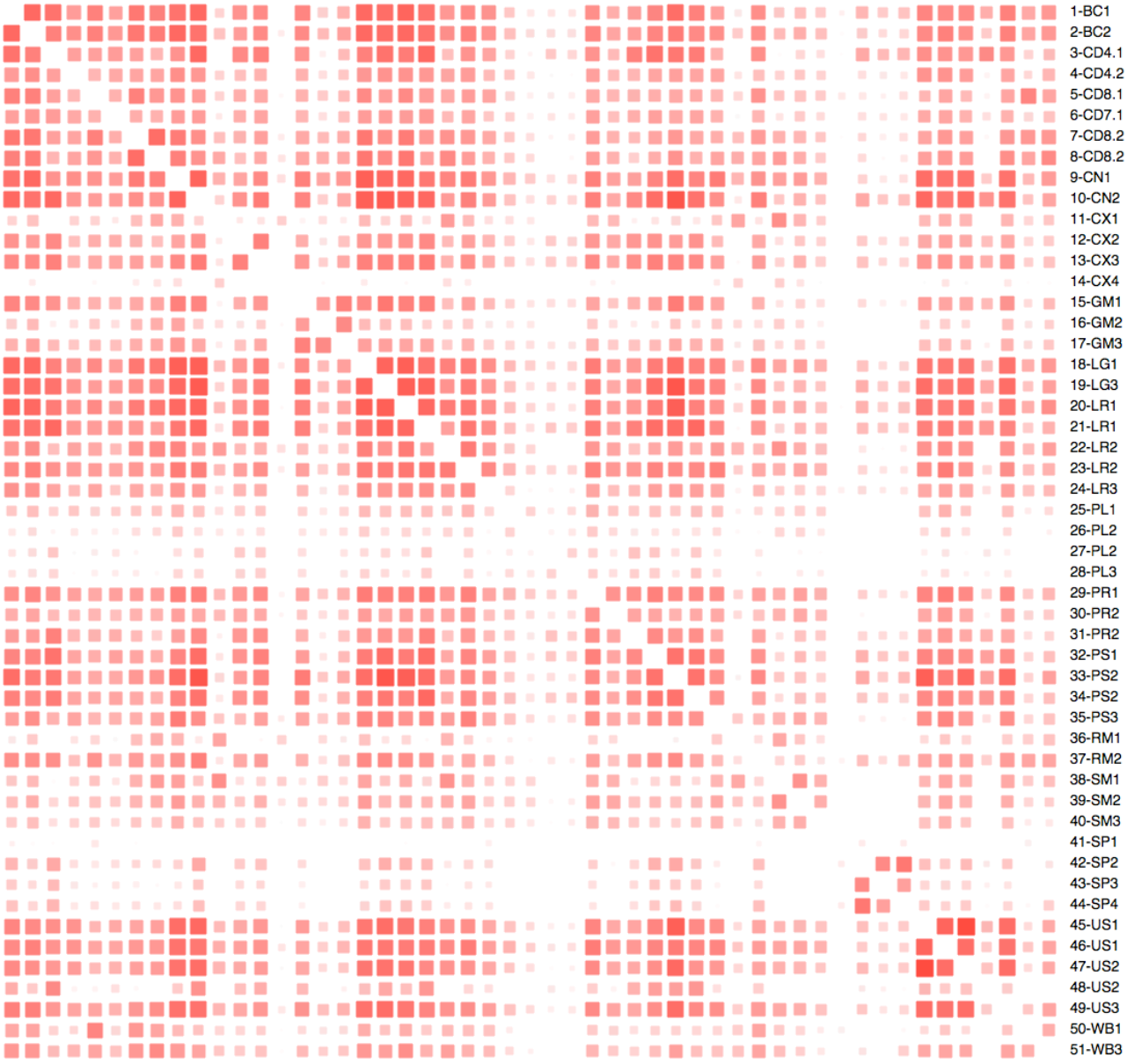
Supplementary information for "An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs)" – Rakyan et al.,

No.	Tissue	Biological replicate	Age (y)	Ethnic Ancestory	array_id	cy3	cy5	Baseline	Response	Notes
1	B-cells	1	45	African	76459	IP	Input	-0.65	17.88	
2	B-cells	2	40	European	69117	IP	Input	-0.48	25.55	
3	CD4 T-cells	1	22	European	66065	IP	Input	-0.36	24.90	
4	CD4 T-cells	2	31	European	82857	IP	Input	-0.23	49.24	
5	CD8 T-cells	1	41	European	62860	IP	Input	-0.58	24.55	
6	CD8 T-cells	1	41	European	78480	IP	Input	-0.32	40.01	Technical replicate of sample 5
7	CD8 T-cells	2	27	African	82312	IP	Input	-0.76	19.66	
8	CD8 T-cells	2	27	African	76460	IP	Input	-0.62	23.75	Technical replicate of sample 7
9	Colon	1	37	European	61257	IP	Input	-0.45	25.30	
10	Colon	2	38	European	61664	IP	Input	-0.49	19.76	
11	Cervix	1	44	European	62870	Input	IP	-0.53	33.87	
12	Cervix	2	50	European	75698	Input	IP	-0.47	26.92	
13	Cervix	2	50	European	72307	IP	Input	-0.49	24.38	Technical replicate of sample 12
14	Cervix	3	25	East Asian	76454	IP	Input	-0.40	44.04	
15	GM06990	1	41	European	86005	IP	Input	-0.67	15.28	
16	GM06990	2	41	European	86134	IP	Input	-0.62	24.08	
17	GM06990	3	41	European	86136	IP	Input	-0.52	25.70	
18	Lung	1	41	European	59181	IP	Input	-0.63	20.15	
19	Lung	3	36	East Asian	79060	IP	Input	-0.50	19.77	
20	Liver	1	37	European	82843	IP	Input	-0.42	23.97	
21	Liver	1	37	European	74196	Input	IP	-0.68	18.28	Technical replicate of sample 20
22	Liver	2	37	European	71622	IP	Input	-0.37	34.38	
23	Liver	2	37	European	74212	Input	IP	-0.52	29.16	Technical replicate of sample 22
24	Liver	3	26	East Asian	59504	IP	Input	-0.53	22.68	
25	Placenta	1	29 (mother)	European	83032	IP	Input	-0.34	42.38	
26	Placenta	2	31 (mother)	European	79812	Input	IP	-0.32	47.85	
27	Placenta	2	31 (mother)	European	81192	IP	Input	-0.24	49.11	Technical replicate of sample 26
28	Placenta	3	unknown	East Asian	83895	IP	Input	-0.34	37.51	
29	Prostate	1	51	European	71742	IP	Input	-0.47	28.31	
30	Prostate	2	46	European	77241	Input	IP	-0.41	35.40	
31	Prostate	2	46	European	71738	IP	Input	-0.28	39.02	Technical replicate of sample 30
32	Pancreas	1	37	European	79036	IP	Input	-0.44	24.25	
33	Pancreas	2	37	European	76450	Input	IP	-0.53	16.89	
34	Pancreas	2	37	European	82148	IP	Input	-0.39	25.68	Technical replicate of sample 33
35	Pancreas	3	33	East Asian	83896	IP	Input	-0.27	40.00	
36	Rectum	1	43	European	61631	IP	Input	-0.52	35.50	
37	Rectum	2	37	European	61622	IP	Input	-0.57	21.14	
38	Skeletal Muscle	1	37	European	74213	Input	IP	-0.67	25.88	
39	Skeletal Muscle	2	41	European	72308	IP	Input	-0.51	31.77	
40	Skeletal Muscle	3	26	East Asian	83891	IP	Input	-0.41	37.22	
41	Sperm	1	20-49	European	61246	IP	Input	-0.58	24.64	
42	Sperm	2	20-49	European	78923	IP	Input	-0.39	23.20	
43	Sperm	3	20-49	European	83890	IP	Input	-0.21	41.53	
44	Sperm	4	20-49	European	98489	Input	IP	-0.34	28.53	
45	Uterus	1	38	European	71619	IP	Input	-0.34	35.11	
46	Uterus	1	38	European	76451	Input	IP	-0.35	26.79	Technical replicate of sample 45
47	Uterus	2	41	European	82533	IP	Input	-0.24	14.56	
48	Uterus	2	41	European	76452	Input	IP	-0.41	25.27	Technical replicate of sample 47
49	Uterus	3	39	East Asian	82058	IP	Input	-0.58	19.91	
50	Whole Blood	1	26	European	76457	IP	Input	-0.61	25.59	
51	Whole Blood	3	44	East Asian	98131	Input	IP	-0.57	23.84	

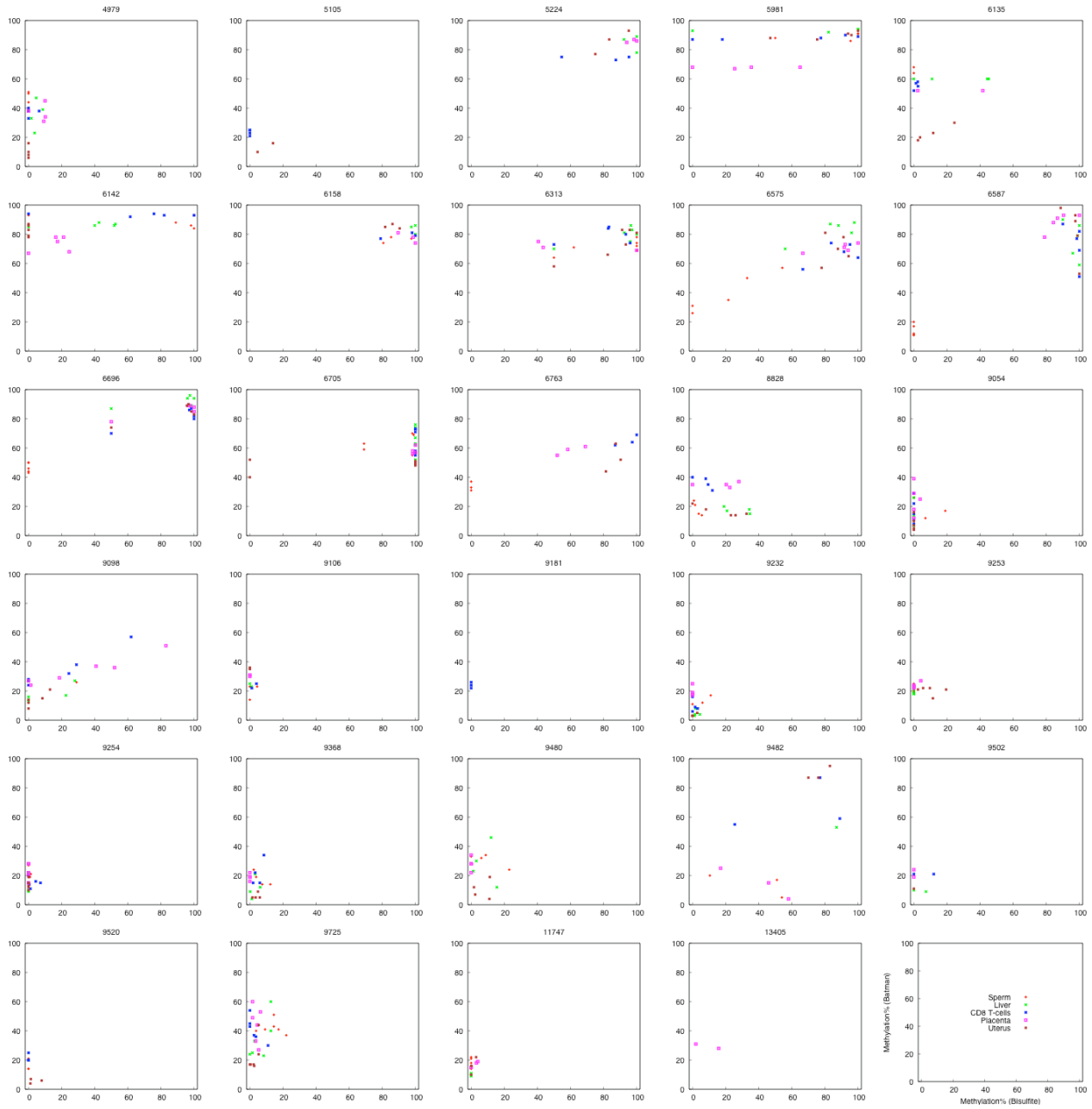
Supplementary Table 1. Tissue samples used in the study. The "Baseline" and "Response" parameters refer, respectively, to the intercept and inverse slope of a linear model fitted to the low-CpG portion of each array's data (refer to description of Batman). The "Response" parameter can be interpreted as the number of methylated cytosines in a region required to increase the observed array signal by one unit. Since the noise level of the arrays appears to be fairly uniform, this can be interpreted as a measure of the signal/noise ratio of the complete MeDIP-chip experiment. Data for the sperm samples have been previously described in Down et al., (in press).



Supplementary Figure 1. MeDIP-array data plotted against a measure of CpG density in the neighbourhood of the probe. Probes were sub-divided into three equal-sized sets according to probe melting temperatures calculated using Nimblegen's method (www.nimblegen.com). The red line shows the Batman calibration used for the complete dataset. As expected, most of the high Tm probes are in high-CpG regions. However, in regions of lower CpG density, the three populations of probes are similar, and in particular the linear model seems to fit all three populations reasonably well.



Supplementary Figure 2. Correlation coefficients between the 51 MeDIP-chip microarrays used in this study. Each array was analysed individually using the Batman method. Area of squares reflects the correlation coefficient (r) with an empty square indicating r <= 0.65 and a full square indicating r = 1.0. The overall correlation between arrays is high (with the great majority of pairwise comparisons showing r > 0.65), indicating a strong methylation pattern in common between most tissues. Some arrays show lower overall correlation than others: we believe that this reflects a slightly lower signal to noise ratio from these arrays, and note that it often corresponds with a relatively high Response parameter (see Sup. Table 1 and Sup. methods).

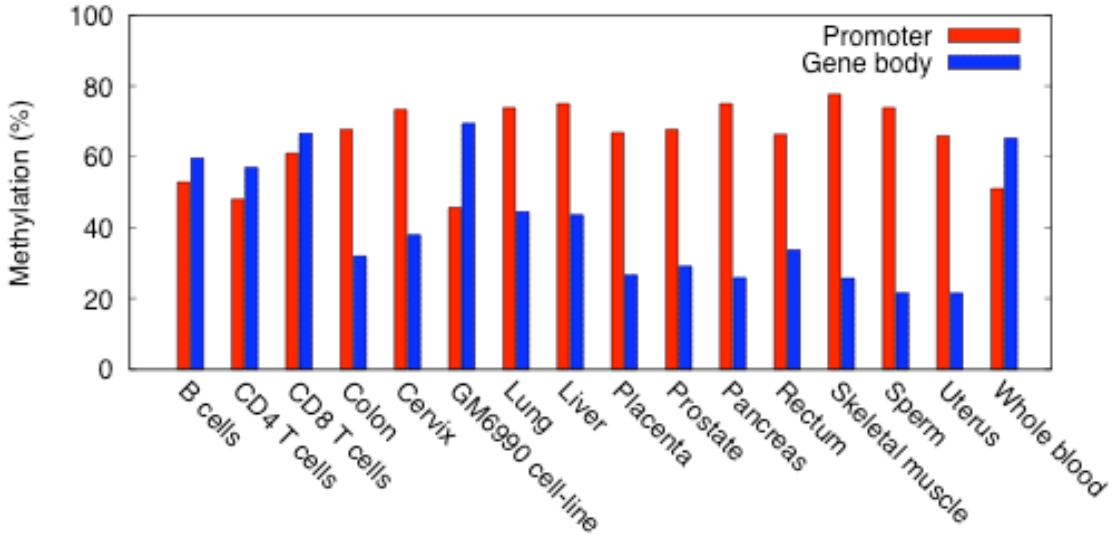


Supplementary Figure 3. Bisulfite-PCR validation of the Batman calls. Initially, 36 regions were chosen for bisulfite-PCR validation, spanning a range of CpG densitites, genomic locations, tDMRs and non-tDMRs (see Supplementary Table 2). However, PCR products could be obtained for only 29. The validation was performed for each of the same tissue samples analyzed on the arrays, resulting in >1,000 individual bisulfite-PCR sequences. For the sake of clarity, only 5 tissues are shown here. The bilsufite-PCR was performed as described previously[1], and then averaged across 100 bp tiles. DNA methylation data for the biological replicates for each tissue type were averaged. We classified both the bisulfite-PCR and Batman-called array data as unmethylated (< 40%) or methylated (> 60%). Based on this classification, only ROIs 5981 and 6142, are discordant between the bisulfite and array datasets.

Supplementary Table 2. Regions analyzed in Supplementary Figure 1.

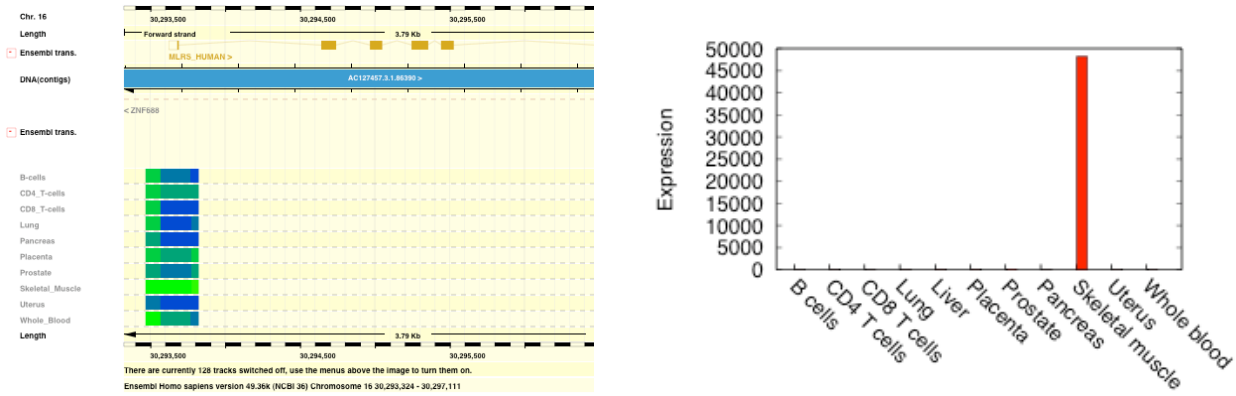
no.	Bisulfite-PCR amplicon ID	Chr	Amplicon start	Amplicon end	GC%	CpG%	Array ROI id	ROI start	ROI end
1	4979	22	29,938,315	29,938,715	67	6.2	27086	29,938,018	29,938,968
2	5105	22	17,545,712	17,546,102	77	11.8	26738	17,545,145	17,547,059
3	5224	22	20,130,777	20,131,135	64	6.7	26831	20,130,463	20,131,012
4	5981	22	38,296,618	38,297,072	67	3.5	27317	38,296,495	38,297,444
5	6135	22	35,970,069	35,970,424	67	4.5	27195	35,970,040	35,970,489
6	6142	22	35,938,425	35,938,903	67	3.8	27194	35,938,422	35,938,871
7	6158	22	49,216,499	49,216,926	61	4.9	27675	49,216,545	49,216,794
8	6313	22	18,510,668	18,511,158	70	8.4	26780	18,510,402	18,511,251
9	6575	22	49,334,595	49,335,038	67	7.4	27696	49,333,374	49,335,023
10	6587	22	29,281,454	29,281,947	63	8.3	27059	29,280,934	29,282,083
11	6696	22	41,419,027	41,419,524	66	8.2	27422	41,418,869	41,419,618
12	6705	22	45,453,093	45,453,592	59	5.6	27559	45,453,377	45,453,526
13	6763	22	39,964,476	39,964,879	70	6.4	27354	39,963,565	39,964,753
14	8828	6	101,018,918	101,019,406	64	6.5	34354	101,018,058	101,020,107
15	9054	6	139,136,417	139,136,888	60	5.3	34696	139,136,090	139,137,339
16	9098	6	46,811,222	46,811,720	51	3.8	34024	46,810,546	46,811,795
17	9106	6	53,322,056	53,322,372	42	2.5	34096	53,320,630	53,322,579
18	9181	6	150,963,434	150,963,699	64	9.8	34785	150,962,740	150,963,841
19	9232	6	28,475,339	28,475,830	59	6.3	33702	28,475,166	28,475,915
20	9253	6	126,111,195	126,111,673	53	4.8	34567	126,110,449	126,113,315
21	9254	6	153,346,203	153,346,693	70	8.0	34814	153,345,105	153,346,654
22	9368	6	170,735,558	170,735,982	51	3.4	35053	170,735,258	170,736,107
23	9480	6	33,787,387	33,787,734	63	8.9	33720	33,787,126	33,787,975
24	9482	6	54,281,191	54,281,533	41	1.2	34106	54,280,991	54,282,009
25	9502	6	154,872,494	154,872,915	67	6.9	34823	154,872,350	154,873,838
26	9520	6	76,368,619	76,368,942	74	13.0	34204	76,367,741	76,369,717
27	9725	6	37,774,253	37,774,700	68	8.3	33827	37,774,107	37,775,056
28	11747	20	2,801,490	2,801,889	57	4.3	24947	2,800,957	2,802,966
29	13405	22	35,777,451	35,777,920	73	10.6	27183	35,777,329	35,778,352

All co-ordinates are based on the NCBI36 version of the human genome
Primer sequences are available upon request

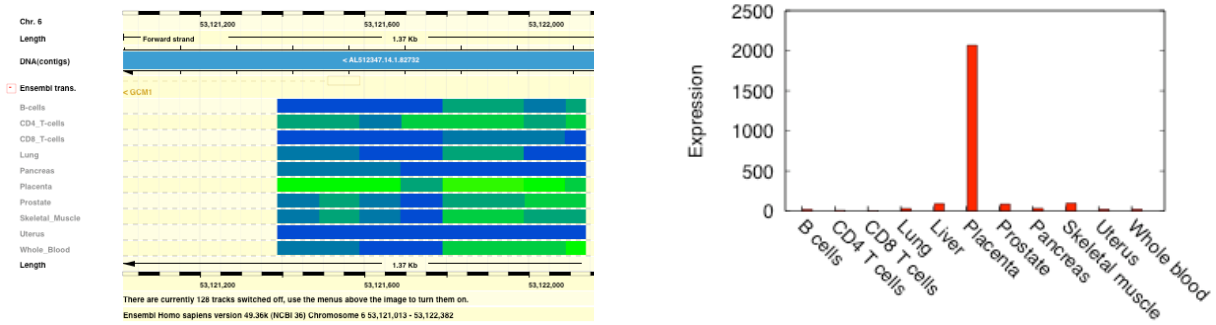


Supplementary Figure 4. DNA methylation status of the ICAM3 gene in a panel of tissues. Promoter methylation bars are based on a 500bp region upstream of the transcription start site (as annotated in Ensembl), while gene body bars show the mean of all available exonic and intronic data from the second exon onwards.

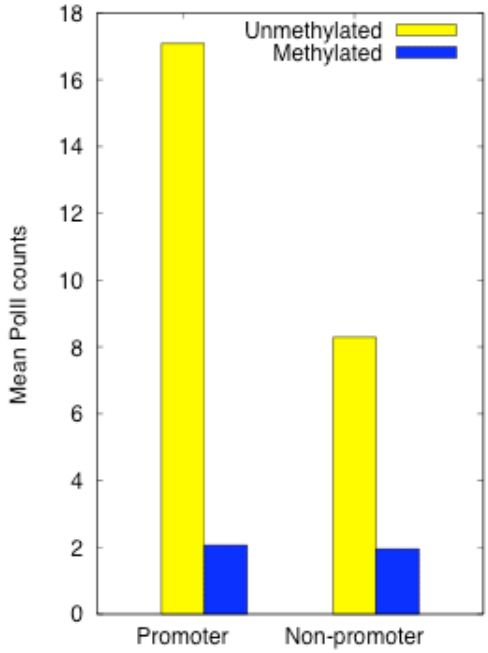
A



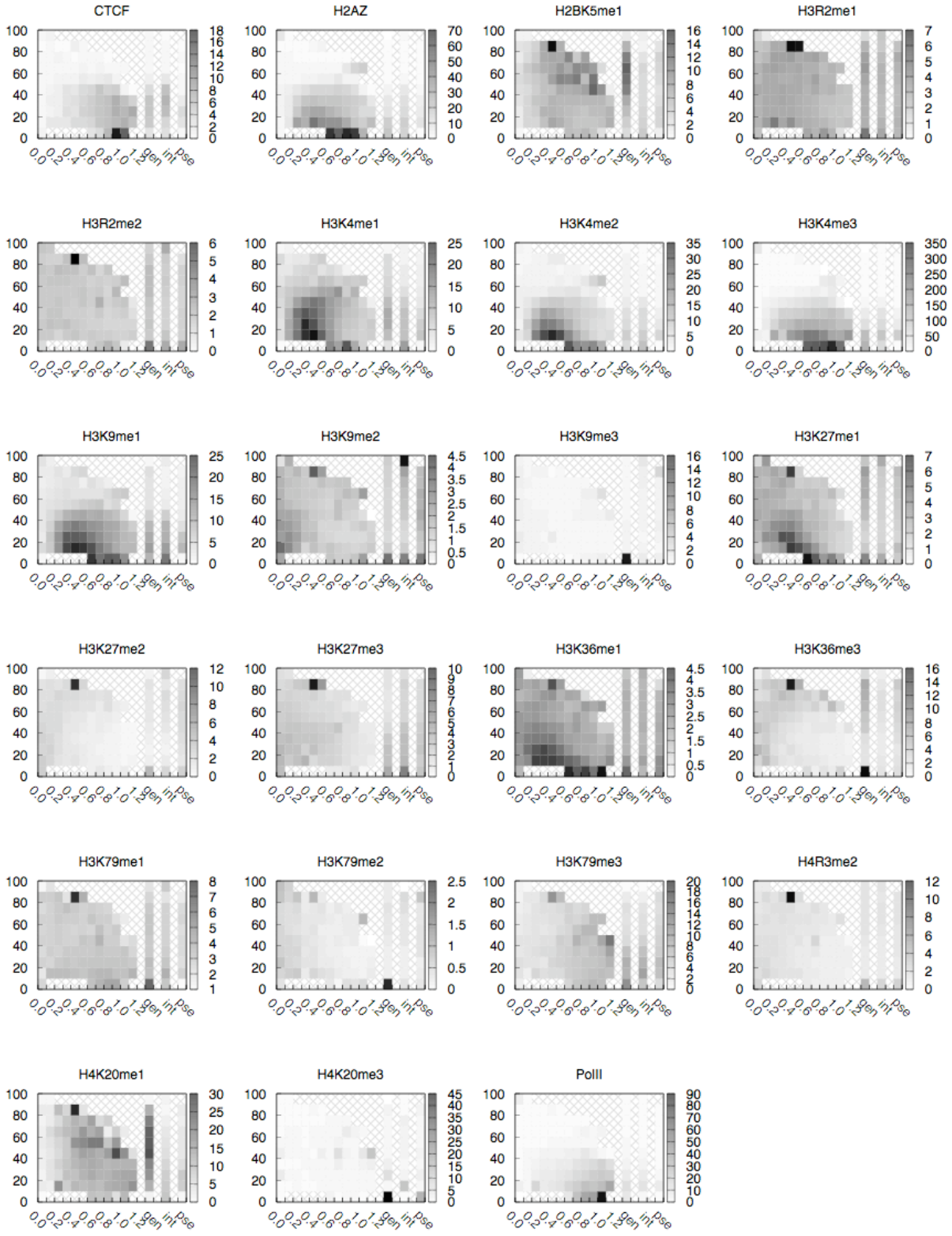
B



Supplementary Figure 5. Promoter DNA methylation and expression patterns for two tissue-specific genes. Gene expression data was plotted as in figure 2c. Expression data are from Su et al., (2004).



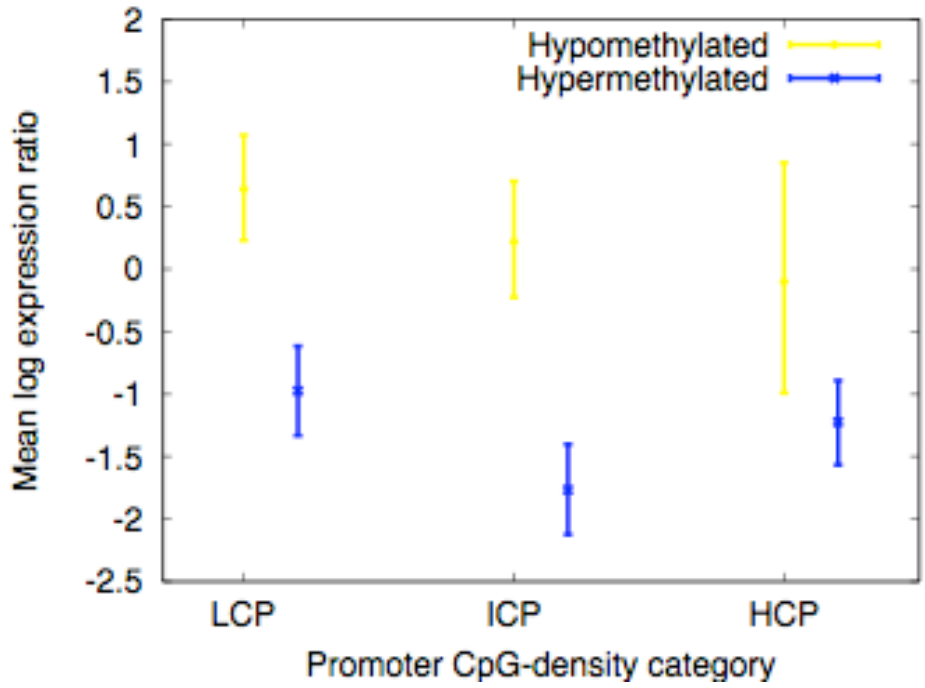
Supplementary Figure 6. DNA methylation data for promoter and non-promoter CpG islands from our study was correlated with genome-wide enrichment profiles RNA PolII generated by Barski et al., 2007 (ref. 2) using Solexa 1G sequencing technology. The y-axis DNA represents the average tag count for RNA PolII. Yellow is < 40% methylation and blue is >60% methylation.



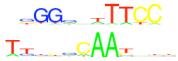
Supplementary Figure 7. DNA methylation data from our study was correlated with genome-wide enrichment profiles for 20 histone lysine and arginine methylations, H2A.Z, RNA PolII, and CTCF generated by Barski et al., 2007 (ref. 2) using Solexa 1G sequencing technology. The x-axes represent $CpG_{o/e}$ (there were insufficient data to stratify by $CpG_{o/e}$ in the non-promoter categories), the y-axes DNA methylation levels, and the grey-scale represents the average tag count for the histone modification or protein indicated. The exon and intron categories were combined into a single 'genic' category. Hatched regions indicate insufficient data were available.

Supplementary Table 3. Tissue-specific differentially methylated regions (tDMRs) were called in 500bp ROIs. To identify hypermethylated tDMRs in a given tissue, we looked for ROIs with a mean methylation of > 60% in the target tissues and < 40% in at least three somatic tissues (not including sperm, placenta, or cell line). Similarly, to identify hypomethylated tDMRs, we looked for ROIs with methylation <40% in the target tissue and > 60% in at least 3 somatic tissues.

	hypo-methylated tDMRs	hyper-methylated tDMRs
B-cells	613	531
CD4 T-cells	725	328
CD8 T-cells	555	908
Colon	579	392
cervix	473	593
GM06990 cells	1278	1667
Lung	439	472
Liver	520	670
Placenta	731	1192
Prostate	522	512
Pancreas	922	576
Rectum	676	554
Skeletal Muscle	625	1236
Sperm	4348	1030
Uterues	1822	1094
Whole Blood	739	861



Supplementary Figure 8. Comparison of tissue-specific DNA methylation and gene expression using a promoter classification similar to that previously used by Schubeler and colleagues[2]. 'Promoters' were defined as a 2,400 bp region centered on the TSS annotated in the Ensembl genome browser. High CpG density promoters (HCP) were defined as having at least one 500 bp window with $CpG_{o/e} > 0.75$ and GC% > 55%. Low CpG density promoters (LCP) were defined as having no 500 bp windows with $CpG_{o/e} > 0.48$ and GC% > 55%. All other promoters were classified as intermediate CpG density promoters (ICP). The figure shows a comparison of promoter-tDMR (located anywhere within 1.2 kb of the TSS) DNA methylation and gene expresson between whole blood and uterus. Gene expression data are from GNF SymAtlas database[3]. Yellow bars represent promoter-tDMRs that display <40% methylation in whole blood and > 60% methylation in uterus. Blue bars represent promoter-tDMRs that display > 60% methylation in whole blood and < 40% methylation in uterus. 95% confidence intervals for the mean log ratios were calculated by bootstraping.

	Motif	Tissue	Differential from somatic tissues	Significance
TFAP2A		Whole blood	+	0.00043
Pax5		Sperm	+	0.00001
		CD8+ T cells	+	0.00006
Evi1		Prostate	+	0.00057
FOXL1		CD8+ T cells	+	0.00016
		CD4+ T cells	+	0.00024
		Sperm	-	0.00075
Klf4		Sperm	+	<0.00001
		Skeletal muscle	+	0.00004
		Prostate	+	0.00012
		Placenta	+	0.00068
		CD4+ T cells	-	0.0008
FOXI1		B cells	-	0.0002
TCF1		Skeletal muscle	-	0.00063
Foxa2		Sperm	+	0.00031
NHLH1		Uterus	-	0.00004
IRF1		Sperm	+	0.00013
		Prostate	-	0.00086
MEF2A		Sperm	-	<0.00001
		Liver	-	0.00031
ZNF42_5-13		Uterus	-	0.00001
MAX		Sperm	+	0.00086
MYC-MAX		B cells	+	0.00047
Pax4		Cervix	+	0.00039
Pbx		Skeletal muscle	-	0.00019
		B cells	+	0.00039
		Uterus	-	0.0008
RXR-VDR		Lung	+	<0.00001
SP1		Uterus	-	0.00009
		Cervix	+	0.00011
		B cells	+	0.00012
		Pancreas	-	0.00025
SPIB		GM6990 cell line	-	<0.00001
		Skeletal muscle	+	0.00013
SRF		B cells	-	0.00098
SRY		Liver	-	0.00038
Staf		Sperm	+	0.00022
		Uterus	-	0.00024
TCF11-MafG		Uterus	+	0.00086
HAND1-TCF3		Liver	-	0.00002
USF1		GM6990 cell line	-	0.00052
REL		GM6990 cell line	-	0.00053
cEBP		GM6990 cell line	-	0.00038
NFKB1		Uterus	-	0.00012
		Cervix	+	0.0003
TBP		GM6990 cell line	-	0.00007
		Whole blood	-	0.00007
		CD4+ T cells	+	0.00016
		Colon	-	0.00044
		Prostate	+	0.00092
Spz1		Pancreas	-	0.00001
		Sperm	+	0.00072
		Prostate	-	0.00078
ESR1		Uterus	-	0.00022
HNF4		GM6990 cell line	-	0.00029
		Skeletal muscle	+	0.00041
		Pancreas	-	0.00075
MafB		Uterus	-	0.00049
Macho-1		CD8+ T cells	+	0.00051
Bapx1		Lung	-	0.00038

Supplementary Figure 9 (previous page). Complete list of motifs from the JASPAR CORE database which are significantly over-represented in promoter tDMRs ($p <= 0.001$, simulations indicate a false discovery rate <10% at this threshold). We compared hyper- and hypo-methylated promoter tDMRs from each tissue in this study with equal-sized sets of non-tDMR promoters with matching distributions of CpG dinucleotide frequencies. For each promoter, we scanned each of the JASPAR motifs using the nmscan algorithm from NestedMICA 0.8.0, and recorded the highest score for each motif in each promoter. For each motif, we then compared each tDMR set with its corresponding non-tDMR set, looking for significant differences in the distribution of motif scores. Significance was assessed empirically by randomly resampling promoters into the tDMR and non-tDMR categories.

References

1. Eckhardt, F. et al. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.* 38, 1378-1385 (2006).

2. Weber, M. et al. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.* 39, 457-466 (2007).

3. Su, A.I. et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* 101, 6062-6067 (2004).