GENOME RESEARCH

An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs)

Vardhman K. Rakyan, Thomas A. Down, Natalie P. Thorne, *et al.*

Genome Res. 2008 18: 1518-1529; originally published online Jun 24, 2008;
Access the most recent version at doi:10.1101/gr.077479.108

An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs)

Vardhman K. Rakyan,[1,9,10] Thomas A. Down,[2,9] Natalie P. Thorne,[3,9] Paul Flicek,[4,9] Eugene Kulesha,[4] Stefan Gräf,[4] Eleni M. Tomazou,[5] Liselotte Bäckdahl,[6] Nathan Johnson,[4] Marlis Herberth,[7] Kevin L. Howe,[3] David K. Jackson,[5] Marcos M. Miretti,[5] Heike Fiegler,[5,8] John C. Marioni,[3] Ewan Birney,[4] Tim J.P. Hubbard,[5] Nigel P. Carter,[5] Simon Tavaré,[3] and Stephan Beck[6,10]

[1]Institute of Cell and Molecular Science, Barts and the London, London E1 2AT, United Kingdom; [2]Wellcome Trust Cancer Research UK Gurdon Institute, University of Cambridge, Cambridge CB2 1QR, United Kingdom; [3]Department of Oncology, University of Cambridge, Cancer Research UK Cambridge Research Institute, Cambridge CB2 0RE, United Kingdom; [4]European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, United Kingdom; [5]Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, United Kingdom; [6]UCL Cancer Institute, University College London, London WC1E 6DD, United Kingdom; [7]Institute of Biotechnology, University of Cambridge, Cambridge CB2 1QT, United Kingdom

We report a novel resource (methylation profiles of DNA, or mPod) for human genome-wide tissue-specific DNA methylation profiles. mPod consists of three fully integrated parts, genome-wide DNA methylation reference profiles of 13 normal somatic tissues, placenta, sperm, and an immortalized cell line, a visualization tool that has been integrated with the Ensembl genome browser and a new algorithm for the analysis of immunoprecipitation-based DNA methylation profiles. We demonstrate the utility of our resource by identifying the first comprehensive genome-wide set of tissue-specific differentially methylated regions (tDMRs) that may play a role in cellular identity and the regulation of tissue-specific genome function. We also discuss the implications of our findings with respect to the regulatory potential of regions with varied CpG density, gene expression, transcription factor motifs, gene ontology, and correlation with other epigenetic marks such as histone modifications.

[Supplemental material is available online at www.genome.org. The array data from this study have been submitted to ArrayExpress under accession no. E-TABM-445.]

DNA methylation is indispensable for genome function in mammals. It is the only known epigenetic modification of mammalian DNA and plays critical roles in transcriptional regulation, chromosomal stability, genomic imprinting, and X-inactivation (for review, see Bird 2002). Its importance is further underlined by observations that various complex diseases such as cancer are associated with perturbed DNA methylation profiles (Laird 2003). Surprisingly, the role of DNA methylation in regulating normal tissue-specific genome function is still poorly understood, even though this is one of the functions originally postulated for this epigenetic modification (Bird 2002). Several recent genome-wide studies show that DNA methylation profiles in mammals are tissue specific (Rakyan et al. 2004; Eckhardt et al. 2006; Khulan et al. 2006; Kitamura et al. 2007; Illingworth et al. 2008). However, our understanding of the role of tissue-specific DNA methylation is still limited, and many questions remain open, including the genomic distribution of tissue-specific DNA methylation profiles, the relative impact of tissue-specific methylation at CpG-island versus non-CpG-island promoters, and the role of tissue-specific methylation in nonpromoter regions, including nonpromoter CpG islands. Comprehensive genome-wide profiles would signficantly improve our ability to address these questions and to better understand the role DNA methylation plays in tissue-specific genome function.

As a resource for the scientific community, we have performed the most comprehensive genome-wide study of human tissue-specific differentially methylated regions (tDMRs), representing the largest available data set for DNA methylation in any organism. Using a combination of methylated DNA immunoprecipitation (MeDIP) (Weber et al. 2005; Keshet et al. 2006), custom high-density microarrays, and novel bioinformatic analytical tools, we have generated reference human genome-wide DNA methylation profiles for 13 normal somatic tissues, placenta, sperm, and the GM06990 immortalized cell line that was used in the ENCODE pilot study (The ENCODE Project Consortium 2007). This work represents a valuable resource for researchers seeking to understand the role of mammalian tissue-specific DNA methylation. Using a newly developed visualization tool, all of our data have been integrated into the Ensembl genome browser

[8]Present address: Roche Diagnostics Limited, West Sussex RH15 9RY, UK.
[9]These authors contributed equally to this work.
[10]Corresponding authors.
E-mail v.rakyan@qmul.ac.uk; fax 44-020-78822186.
E-mail s.beck@ucl.ac.uk; fax 44-020-76790815.

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.077479.108. Freely available online through the *Genome Research* Open Access option.

(Flicek et al. 2008), and are the first genome-wide DNA methylation data to be included in any genome browser. The final part of our integrated system consists of a novel bioinformatic algorithm we have recently developed, Bayesian tool for methylation analysis (Batman), that enables the estimation of absolute methylation levels from immunoprecipitation-based DNA methylation profiles (Down et al. 2008). Bioinformatic analyses of our data confirm some conclusions from previous smaller studies and also suggest several novel roles for DNA methylation. A negative correlation between DNA methylation and gene expression is observed at high-, medium-, and contrary to previous notions, at even some low-CpG density promoters. On the other hand, gene-body methylation positively correlates with gene expression. Furthermore, in addition to the study by Illingworth et al. (2008), our study represents one of the first systematic genome-wide efforts to characterize nonpromoter CpG islands, and we propose that only a fraction of these are likely to be functional regulatory elements. Overall, this work represents an important contribution to current efforts in understanding the epigenetic code, and its role in tissue-specific genome function in mammals.

Results

Genome-wide mapping of human tissue-specific DNA methylation profiles

We based our DNA methylation profiling strategy on a recently developed technique—methylated DNA immunoprecipitation (MeDIP)—which utilizes a monoclonal antibody against 5-methylcytosine to enrich for the methylated fraction of a genomic DNA sample (Weber et al. 2005; Keshet et al. 2006). MeDIP combined with microarrays is a powerful approach for DNA methylation profiling (Weber et al. 2005, 2007; Keshet et al. 2006; Zhang et al. 2006; Zilberman et al. 2006). We designed a custom high-density oligonucleotide array that encompassed all known promoters and CpG islands (both promoter- and nonpromoter-CpG islands) in the human genome based on the Ensembl genome browser (*Homo sapiens* release 45.36g based on NCBI_36). To cover these regions, we chose regions of interest (ROIs) that were 500 bp in length, typically containing 5 × 50-mer probes. Most promoters/CpG islands were represented by multiple ROIs. Repetitive elements were not represented on the array. The final array design included ROIs that overlapped 82% of all known autosomal transcriptional start sites (TSSs) in Ensembl, which we used as a proxy for promoters, 72% of autosomal nonpromoter CpG islands (for additional information about the ROIs and array design, see Table 1 and Methods), and also some randomly selected CpG-poor nonpromoter regions. For technical reasons, probes could not be designed against the remaining TSSs and nonpromoter CpG islands. Data were obtained for several biological and technical replicates (dye-swaps) for each of 13 different normal human somatic tissues, placenta, sperm, and the GM06990 EBV-transformed lymphoblastoid cell line, resulting in 51 genome-wide DNA methylation profiles (Fig. 1; Supplemental Figs. 1, 2; Supplemental Table 1). The sperm data are from a recent study that we published (Down et al. 2008).

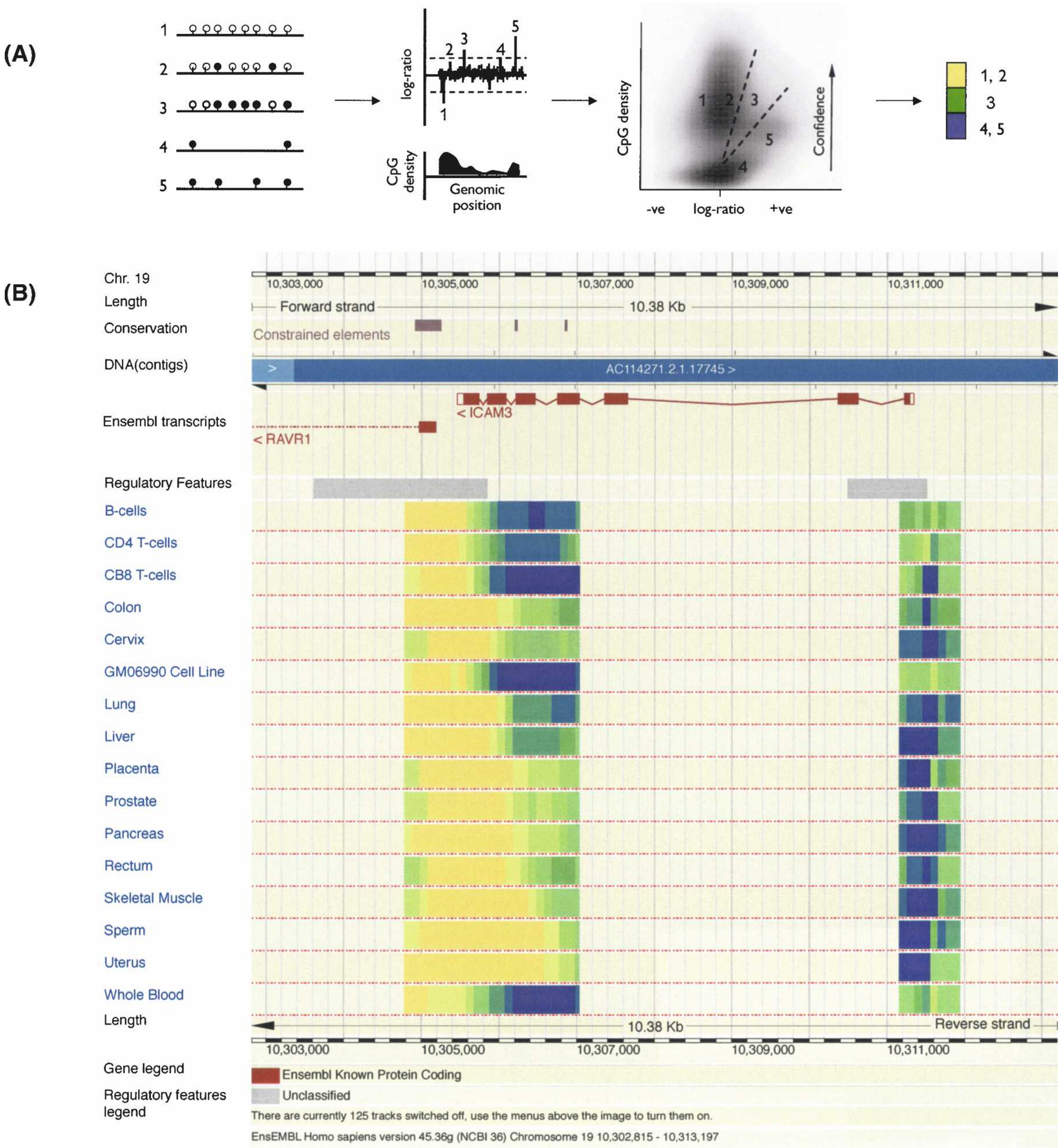
Quantitation of DNA methylation levels and integration into the Ensembl genome browser

Until now, it has not been possible to transform MeDIP enrichment ratios into absolute methylation values. This is because MeDIP enrichment depends on the density of methylated cytosines (Weber et al. 2005; Keshet et al. 2006), which varies greatly within the human genome (DNA methylation in mammals occurs almost exclusively at CpG dinucleotides). Any attempt to correct for this CpG density effect at the level of the array design or with experimental constraints dramatically lowers the amount of the genome that can be assessed. To overcome this constraint, we used a novel algorithm that we recently developed—Bayesian tool for methylation analysis (Batman)—based on a Bayesian deconvolution strategy similar to joint binding deconvolution (Qi et al. 2006) to assign MeDIP signal to CpG dinucleotides in the sequence (Down et al. 2008; schematically shown in Fig. 1A). Briefly, Batman corrects for the observation that methylated sequences with higher CpG densities will have stronger MeDIP enrichment, thereby allowing estimation of absolute methylation levels. Comparison of Batman-called methylation values with bisulfite–PCR sequencing data from the Human Epigenome Project (Eckhardt et al. 2006) and 29 random regions represented on the array used here demonstrates that the

Table 1. Description of the genomic regions represented on the arrays

Genomic category[a]	Description[b]	No. of ROIs	No. of genes represented[c]	No. of CpG islands represented[d]	Modal $\text{CpG}_{o/e}$ in this study[e]	Modal $\text{CpG}_{o/e}$ in genome
Promoter	ROI located within 1.5 kb upstream of or downstream from the TSS of a protein-coding gene[f]	44,337	17,271	11,202	Bimodal, 0.2 and 0.8	Bimodal, 0.2 and 0.8
Exon	>50% of the ROI overlaps any exon except 1st or last exons	7,104	3,705	1,831	0.55	0.2
Intron	>50% of the ROI overlaps intron except 1st or last introns	5,132	2,457	1,089	0.65	0.15
Pseudogene	ROI located within 1.5 kb upstream of or downstream from a pseudogene	3,033	2,143	406	0.2	0.15
Intergenic	ROI not classified in any of the above categories	9,904	NA	3,028	0.75	0.15

[a] Categories are mutually exclusive.
[b] ROI, region of interest. Each ROI was 500 bp, typically containing 5 × 50-mer probes.
[c] Most genes were represented by multiple ROIs. All genome annotations were from Ensembl genome browser (*Homo sapiens* release 45.36g based on NCBI 36). Pseudogene annotations were obtained from www.pseudogene.org.
[d] ROIs in nonpromoter categories were biased toward CpG islands annotated in the Ensembl genome browser.
[e] $\text{CpG}_{o/e}$ was calculated as (no. of CpGs × sequence length)/(no. of Cs × no. of Gs).
[f] 94% of promoter-ROIs are located within 800 bp of the annotated TSS.
NA, Not available.



There are currently 125 tracks switched off, use the menus above the image to turn them on.

EnsEMBL Homo sapiens version 45.36g (NCBI 36) Chromosome 19 10,302,815 - 10,313,197

Figure 1. (*A*) Schematic description of Batman. The *left* panel shows five hypothetical genomic regions of varying CpG densities and number of methylated CpG sites (filled and empty circles represent methylated and unmethylated CpG sites, respectively). As MeDIP enrichment is proportional to the number of methylated CpG sites, the normalized enrichment ratios of these five hypothetical regions, shown in the *second* panel, will not accurately reflect the absolute methylation levels at the genomic region of interest (ROI). Batman is based on the observation that the log-ratio MeDIP signal of methylated DNA scales linearly with the number of methylated CpG sites in a sequence. We use a Bayesian deconvolution strategy, taking into account the estimated distribution of DNA fragment lengths, to find the most likely configurations of methylated and unmethylated CpGs in a sequence that explains the observed MeDIP signals. This allows estimation of absolute methylation levels at the ROI. Yellow, green, and blue represent unmethylated, intermediately methylated, and methylated regions, respectively. Batman is described in detail in Down et al. (2008). (*B*) Integration of Batman-called methylation values into the Ensembl genome browser—screenshot of the data integrated into the Ensembl genome browser (www.ensembl.org). The web display uses a color gradient to show the Batman methylation score for each of the probes in the ROI. The color represents the value of the probe on a sliding scale from 20 (bright yellow) to 80 (dark blue). Probes with Batman values of less than 20 or greater than 80 are colored with the maximum and minimum shades to increase the contrast in the display. Each tissue-type is configured as a dedicated DAS source, allowing the user to select any possible subset of tissues for viewing. Users clicking on a probe will see a small pop-up window, which displays the exact chromosome position of the probe and the Batman score.

two methods correlate very well (Supplemental Fig. 3; $R = 0.84$ in Down et al. 2008**)**.

All of the Batman-analyzed data from this study can be visualized as a set of extra tracks within the Ensembl genome browser (Flicek et al. 2008) (Fig 1B). The web display uses a color gradient to show the Batman methylation score for each of the probes in the ROI. These data represent the first genome-wide DNA methylation data to be included in any genome browser, providing a valuable resource for the scientific community. Furthermore, we have set up another browser (http://td-blade.gurdon.cam.ac.uk/hepscape/) to allow direct comparison between the MeDIP-array data from this study and the Human Epigenome Project (Eckhardt et al. 2006).

Canonical somatic DNA methylation profiles of human promoters

Consistent with previous findings, we see a bimodal distribution of observed/expected CpG densities ($CpG_{o/e}$) in promoter regions (Fig. 2A) (Takai and Jones 2002; Saxonov et al. 2006; Weber et al. 2007). The CpG-dense population (modal $CpG_{o/e} \sim 0.8$) corresponds to CpG islands (CGIs)—regions where the $CpG_{o/e}$ greatly exceeds the genome average of ~0.2. CGIs are considered to be important regulatory elements, as they are generally unmethylated and ~60% of all known human genes contain CGIs at their 5′-end. Several methods have been proposed for classifying CGIs, varying in their use of cut-offs for length, GC%, and CpG density (Bird et al. 1985; Gardiner-Garden and Frommer 1987; Takai and Jones 2002; Saxonov et al. 2006; Glass et al. 2007). We used the CGI definition of the Ensembl genome browser, (length > 400 bases and $CpG_{o/e} > 0.6$), which results in the exclusion of most small, CpG-rich repetitive sequences in the human genome. As expected, in a typical somatic tissue ~90% of CGI-associated "regions of interest" (ROIs, defined in Table 1) within the promoter category display low levels of methylation (<40% methylation) and hereafter are operationally termed "unmethylated". Validation by bisulfite sequencing confirmed that use of this threshold minimizes false positives (Down et al. 2008). In fact, significant numbers of unmethylated ROIs (Table 2) were observed across the entire range of $CpG_{o/e}$, raising the possibility that maintenance of an unmethylated state is also important for the activity of non-CGI promoters. This is somewhat in contrast to the recent study by Weber et al. (2007), who concluded that most low CpG density promoters (LCPs) were methylated. Comparison of our data with genome-wide expression profiles from a public database (Su et al. 2004) revealed a small but significant negative correlation between promoter DNA methylation and gene expression across a broad range of $CpG_{o/e}$ ($P < 10^{-5}$) (Fig. 2B). We tested this for eight tissues where suitable data were available from the GNF expression database, and observed a significant negative correlation between methylation and expression ($P < 0.05$)

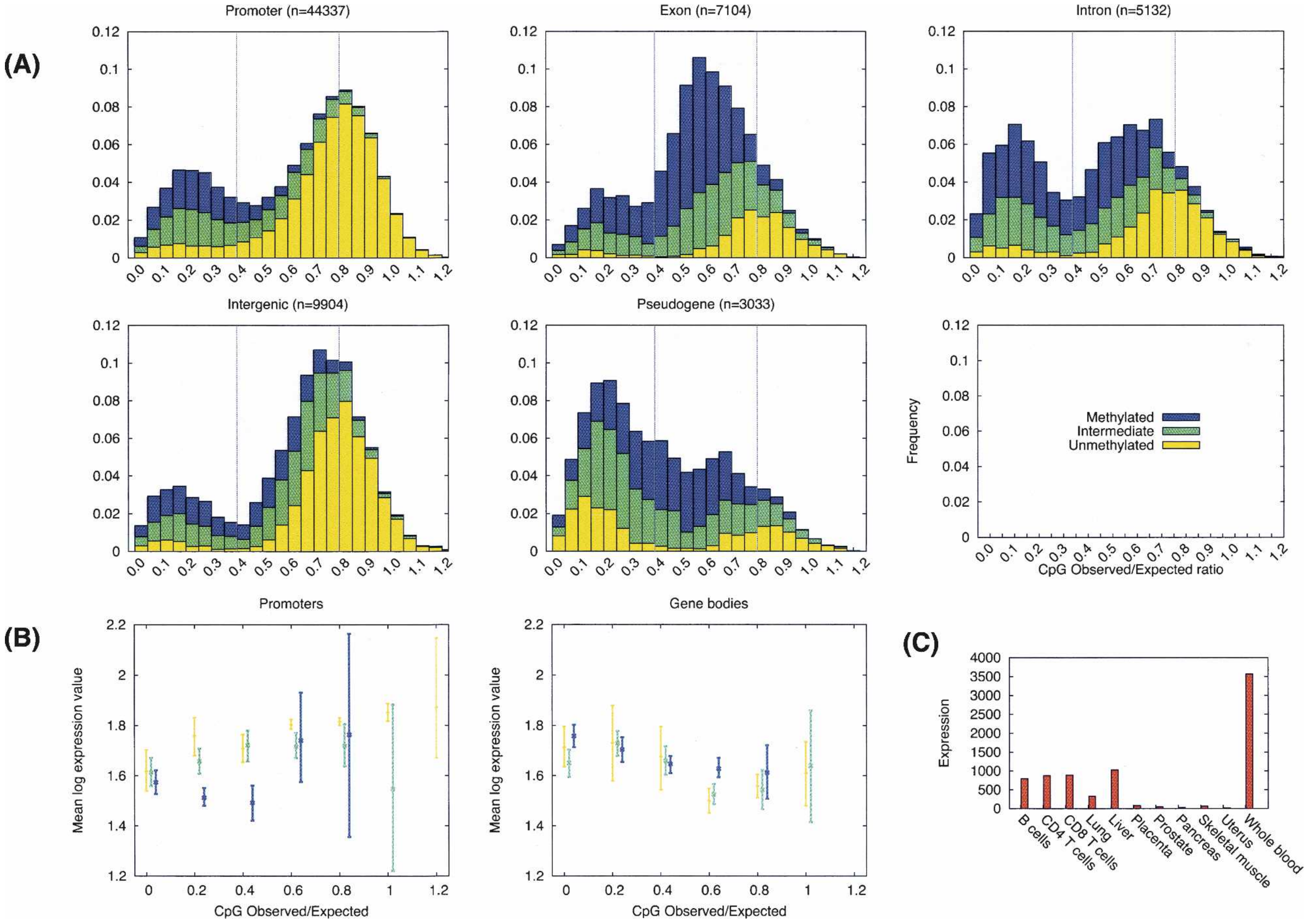


Figure 2. (Continued on next page)

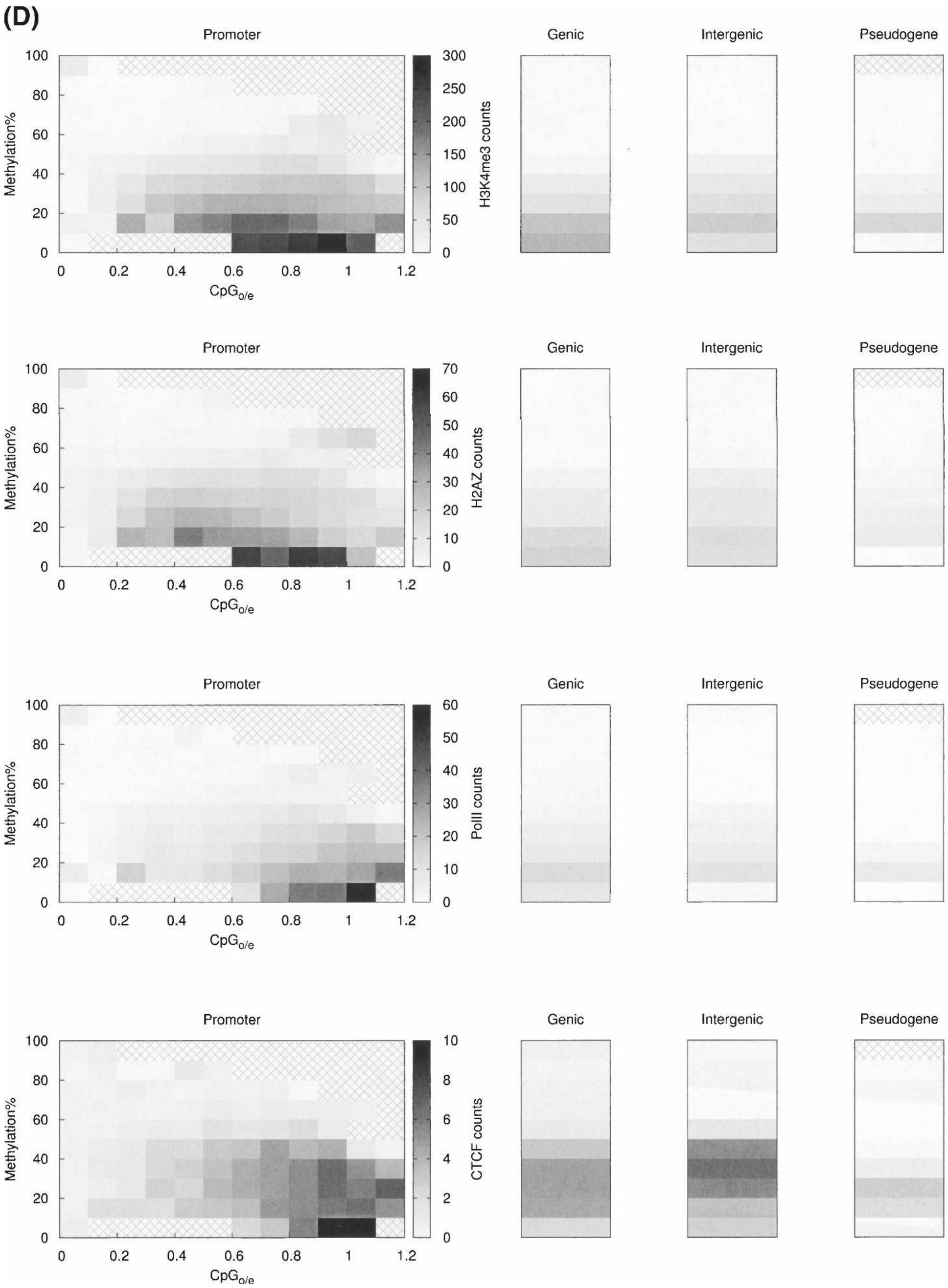


Figure 2. Analysis of somatic DNA methylation profiles. (*A*) Distribution of data with respect to CpG$_{o/e}$ for the different genome feature categories. The data were operationally categorized into unmethylated (<40%), intermediate (40%–60%), and methylated (>60%). Within the nonpromoter categories, we focused predominantly on CpG islands as annotated in the Ensembl genome browser (NCBI_36). Therefore, the average CpG$_{o/e}$ within the nonpromoter categories in our data set was higher than the genome average (refer to Table 1). However, because probes could not be chosen for all nonpromoter CpG islands, we also randomly selected some CpG-poor nonpromoter regions, and hence, a bimodality of CpG$_{o/e}$ is also observed in some of the nonpromoter categories. Methylation data used in these plots are from whole blood. (*B*) Comparison of promoter DNA methylation with gene expression across a range of promoter CpG$_{o/e}$. Whole-blood DNA methylation data (only ROIs overlapping the TSS were used) was correlated with whole-blood genome-wide expression profiles obtained from the GNF SymAtlas database (Su et al. 2004). There were insufficient data for intermediately methylated promoters in the CpG$_{o/e}$ = 1.2 category, and methylated promoters in the CpG$_{o/e}$ ≥ 1 categories. The color code is the same as in *A*, and error bars represent 95% confidence intervals. (*C*) Gene expression levels for *ICAM3* were obtained from a public database (Su et al. 2004). Expression values represent average difference values computed by Affymetrix software. These values are proportional to mRNA content in the sample. (*D*) Correlation of DNA methlyation with H3K4me3, H2A.Z, RNA PolII, and CTCF enrichment. DNA methylation data (500 bp ROIs) from our study were 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) using Illumina 1G sequencing technology. The remaining 19 comparisons are presented in the Supplementary section. The *X*-axes represent CpG$_{o/e}$ (there were insufficient data to stratify by CpG$_{o/e}$ in the nonpromoter categories), the *Y*-axes DNA methylation levels, and the grayscale 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 that insufficient data were available.

Table 2. Number of unmethylated ROIs in each $CpG_{o/e}$ category

CpGo/e	ROIs with <40% methylation
0.0–0.2	1,030 (19.2%)
0.2–0.4	1,139 (16.0%)
0.4–0.6	2,415 (42.9%)
0.6–0.8	9,358 (77.7%)
0.8–1.0	11,639 (84.1%)
1.0–1.0	17,29 (96.8%)

for every bin between 0.2 and 1.0 $CpG_{o/e}$, that is, even for CpG-poor promoter ROIs ($CpG_{o/e} < 0.4$), which corresponds to a previously defined "low-CpG" promoter category (Saxonov et al. 2006; Weber et al. 2007). For example, the ROI associated with the *ICAM3* TSS has a $CpG_{o/e}$ of 0.29 and is hypomethylated only in the tissues in which this gene is expressed (Figs. 1, 2C; Supplemental Fig. 4). Additional examples are shown in Supplemental Figure 5. However, our data also suggest that many promoters can be silent, irrespective of the DNA methylation status.

Repression of CGI-promoters by DNA methylation is well documented (Eckhardt et al. 2006; Khulan et al. 2006; Estecio et al. 2007; Jones and Baylin 2007; Weber et al. 2007). Indeed, we found 5%–10% of CpG island promoters to be predominantly methylated in any given tissue, consistent with emerging evidence that methylation of CpG islands in normal cell function is more common than previously appreciated (Eckhardt et al. 2006; Shen et al. 2007; Weber et al. 2007; Illingworth et al. 2008). With regard to CpG-poor promoters, the recent study by Weber et al. (2007) shows a rather complex correlation between CpG-poor promoter methylation and gene expression—certain promoters with few CpGs were shown to be active and methylated, whereas other promoters of that group can be unmethylated when active. Overall, our data suggest that DNA methylation is involved in regulating the activity of a small but significant number of promoters over a broad range of $CpG_{o/e}$, including CpG-poor promoters. Further studies will be required to determine the exact number of such promoters, establish causality, and understand why some promoters use DNA methylation for regulating their activity, whereas many other promoters do not seem to require DNA methylation to be silenced.

DNA methylation profiles of nonpromoter CpG islands

There are 8449 autosomal nonpromoter CGIs annotated in the Ensembl genome browser, but their function remains poorly understood. In addition to the recent report by Illingworth et al. (2008), our study represents one of first systematic genome-wide effort to characterize nonpromoter CGIs. We observed that the populations of unmethylated nonpromoter CGIs ($CpG_{o/e} > 0.6$) in the various nonpromoter categories have a strikingly similar "bell-shaped" distribution to the unmethylated CGI-promoter population (modal $CpG_{o/e} = 0.8$ in all categories) (Fig. 2A). However, the unmethylated nonpromoter CGIs represent only a small proportion of currently annotated nonpromoter CGIs: in a typical somatic tissue, 20% of exonic CGIs, 39% of intronic CGIs, 48% of intergenic CGIs, and 23% of pseudogenic CGIs are unmethylated. Only 29% of nonpromoter CGIs were found to be unmethylated in all 16 tissues tested, compared with 67% of promoter-CGIs that are constitutively unmethylated. Comparison of DNA methylation data (from $CD4^{+}$ cells) of promoter- and nonpromoter-CGIs with RNA polymerase II binding profiles, generated by Barski et al. (2007), for human $CD4^{+}$ T-cells using the Illumina 1G sequencing (formerly known as Solexa sequencing technology), revealed RNA polymerase II levels at unmethylated nonpromoter CGIs to be approximately half of those observed at promoter-CGIs. (Supplemental Fig. 6). Consequently, we propose that approximately only half of nonpromoter CGIs (as classified in the Ensembl Genome Browser) are likely to be functional in the sense that promoter CGIs are thought to be. A number of definitions for CGIs—based on CpG density and local GC%—have been proposed over the last 20 yr (Bird et al. 1985; Gardiner-Garden and Frommer 1987; Takai and Jones 2002; Saxonov et al. 2006; Glass et al. 2007). The incorporation of experimental data, such as those presented here, will greatly assist in refining CGI definition and thereby help in understanding their function in the context of both promoter and nonpromoter regions of mammalian genomes.

Association between DNA methylation and chromatin signatures

It is known that active regulatory elements bear distinctive chromatin "signatures" (The ENCODE Project Consortium 2007) and that DNA methylation interacts with the chromatin regulatory machinery (Bird 2002). To better understand the regulatory potential of the regions analyzed in our study, we compared our DNA methylation data from $CD4^{+}$ T-cells with genome-wide profiles for 20 histone lysine and arginine methylations, histone variant H2A.Z, RNA polymerase II, and the insulator binding protein CTCF (Barski et al. 2007). These profiles were generated for human $CD4^{+}$ T-cells using the Illumina 1G sequencing. We found that unmethylated promoter-ROIs are strongly associated ($P < 10^{-5}$, nonparametric empirical test; refer to the Methods section) with signatures of active chromatin such as H3K4me3, H2A.Z, and RNA polymerase II (Fig. 2D; Supplemental Fig. 3). Although these associations were more pronounced at high $CpG_{o/e}$, there was clear enrichment across the entire range of $CpG_{o/e}$, including CpG-poor promoter-ROIs. Hypermethylated promoter-ROIs, across a range of CpG densities, did not show clear associations with either H3K27me3 or H3K9me3, two well-established "repressive" histone modifications (Supplemental Fig. 7). However, even in the original study by Barski et al. (2007) these modifications showed only a modest correlation with inactive promoters.

Enrichment for H3K4me3, H2A.Z, and RNA polymerase II was also observed at unmethylated nonpromoter regions (which are mostly CpG-rich as a result of our array design), albeit at relatively lower levels compared with promoter regions (Fig. 2C). This would suggest that at least a subset of nonpromoter CGIs are unannotated TSSs, consistent with recent evidence suggesting that there are many more TSSs in the human genome than has previously been appreciated (The ENCODE Project Consortium 2007). Alternatively, these unmethylated regions could function as other types of regulatory elements such as insulators that restrict transcriptional enhancers from activating unrelated promoters, i.e., "enhancer blockers" (West and Fraser 2005). All known vertebrate enhancer blockers interact with the CTCF protein, and it has been shown that CTCF preferentially binds to unmethylated sites (Mukhopadhyay et al. 2004). Indeed, we observed a strong correlation ($P < 10^{-5}$) between unmethylated domains (over a range of $CpG_{o/e}$) and CTCF binding at promoter and nonpromoter regions. Overall, this analysis reinforces the idea that a significant proportion of unmethylated nonpromoter CGIs identified in our study are functional regulatory elements.

Tissue-specific differentially methylated regions (tDMRs)

DNA methylation profiles are known to be tissue specific, but the role of this epigenetic modification in controlling tissue-specific transcriptional programs remains controversial (Walsh and Bestor 1999; Weber et al. 2007). Approximately 18% of the genomic regions in our study were classified as tissue-specific differentially methylated regions (tDMRs) (Khulan et al. 2006), i.e., regions of the genome that display significant differences in DNA methylation levels among the 16 tissues analyzed (see Methods, for a detailed description of the tDMR calling strategy). Comparison with data from the Human Epigenome Project (Eckhardt et al. 2006) revealed that our tDMR classification approach had a positive predictive value of 78% and sensitivity of 61%. tDMRs were found across a range of CpG densities in all genomic feature categories, although in the promoter category, tDMRs were relatively more common at low-to intermediate CpG-density promoters (Fig. 3A). Consistent with results from the Human Epigenome Project (Eckhardt et al. 2006), mature sperm, the product of the male germ line, was significantly hypomethylated relative to all other tissues in all the genomic categories; in fact, 27% of all tDMRs are sperm specific (Supplemental Table 3). DNA methylation patterns for the GM06990 EBV-transformed lymphoblastoid cell line were substantially divergent (both hypo- and hypermethylated loci) from all the other tissues (Supplemental Table 3). Although we profiled only a single cell line (three biological replicates), our results further support the idea that epigenetic profiles obtained from transformed cells are not representative of primary tissues (Liu et al. 2005), and future studies aiming to elucidate disease-specific epigenetic variants should take this into account.

Correlation of promoter tDMRs with gene expression

To investigate the role of DNA methylation in tissue-specific transcriptional programs, we compared tissue-specific promoter methylation and gene expression profiles (Fig. 3B, left). The important aspect of this analysis, compared with that described above in the section "Comparison of genome-wide DNA methylation and gene expression profiles" is that, here, we are comparing the expression of the same gene in different tissues. A significant negative correlation ($P < 10^{-5}$) was observed between tissue-specific promoter methylation and gene expression across a broad range of $\mathrm{CpG}_{o/e}$, including CpG-poor promoter-tDMRs. This is in contrast to the recent proposal by Weber et al. (2007) that DNA methylation is unlikely to play a significant role in regulating CpG-poor promoters. They defined low-CpG promoters (LCPs) as being a region that spans 900 bp upstream to 400 bp downstream relative to the TSS and does not contain any 500-bp windows with $\mathrm{CpG}_{o/e} > 0.48$. We reanalyzed our data using this promoter classification system, but using an even bigger promoter region (e.g., LCP was defined as a 2400-bp region, centered on the TSS, that does not contain any 500-bp windows with a $\mathrm{CpG}_{o/e} > 0.48$), and again observed a statistically significant negative correlation ($P < 10^{-5}$) across the entire range of $\mathrm{CpG}_{o/e}$ (Supplemental Fig. 8). Overall, our analysis suggests that some promoter-tDMRs, including CpG-poor promoter-tDMRs, are involved in regulating tissue-specific gene expression. There are a number of possible reasons for the difference between our results and those of Weber et al. (2007): (1) we analyzed more tissues and, hence, have increased the power to detect such differences, and (2) they used RNA Pol II binding as a proxy for gene expression. However, it has been shown recently that significant RNA Pol II binding is observed even at promoters associated with non-expressed genes (Guenther et al. 2007). It is therefore worth considering the possibility that DNA methylation actually influences the binding of proteins involved in elongation of transcription, and not necessarily RNA polymerase II binding.

Correlation of gene-body tDMRs with gene expression

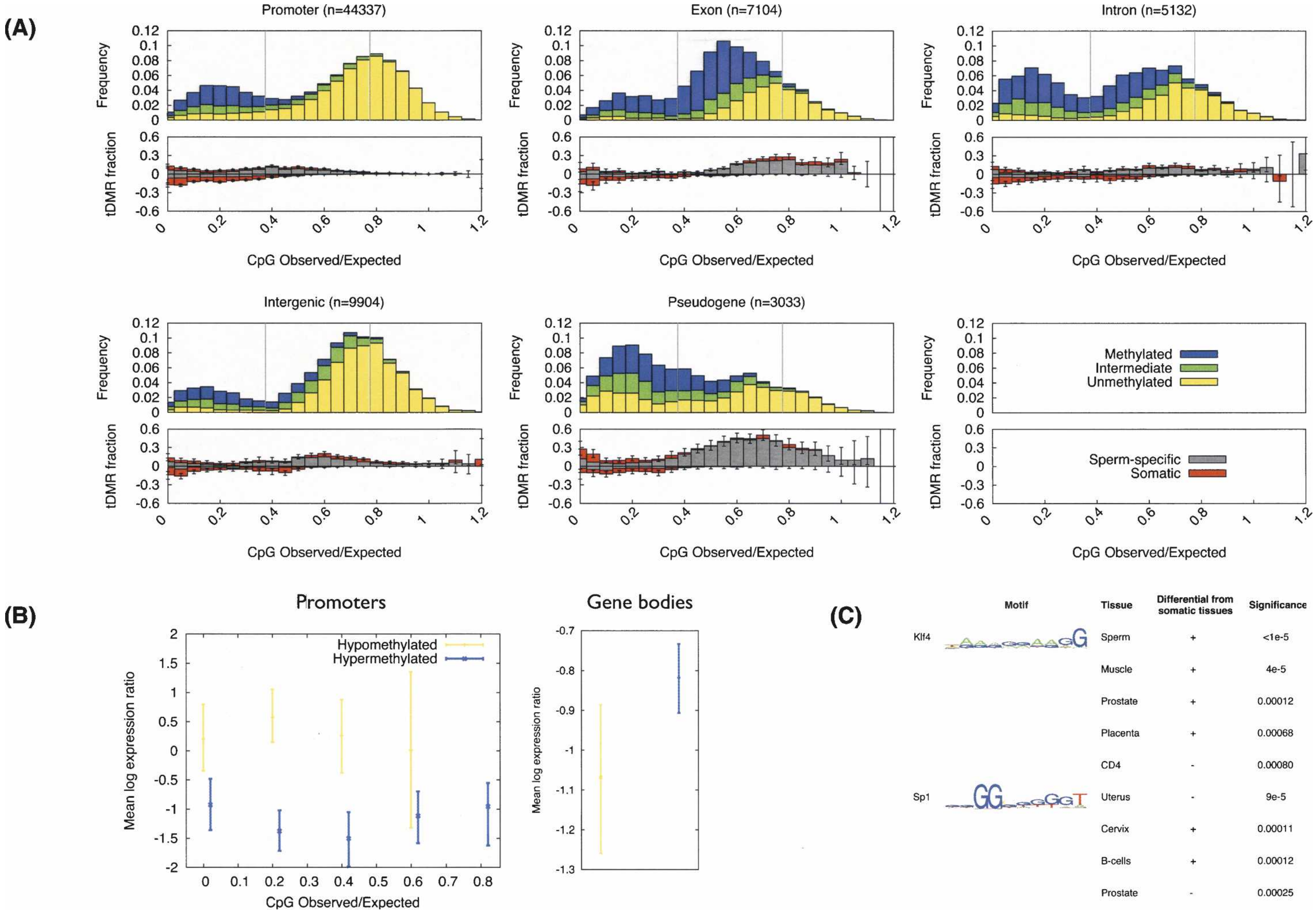
Surprisingly, gene-body tDMRs showed a small, but significant, positive correlation between DNA methylation and gene expression ($P = 0.024$; Fig. 3B, right panel). This is reminiscent of a recent report of hypomethylation at gene promoters and hypermethylation of gene bodies on the active X chromosome in humans (Hellman and Chess 2007). This type of phenomenon is exemplified by the *ICAM3* gene, which displays promoter hypomethylation and gene-body hypermethylation in the tissues in which it is expressed (Fig. 2C). The functional relevance of tissue-specific gene-body methylation is unclear. It may be associated with expression potentiality or act to suppress spurious transcriptional initiation within actively transcribed genes (Zilberman et al. 2006; Hellman and Chess 2007; Suzuki et al. 2007). Elucidating the role of gene-body methylation represents an important area of investigation for the future.

Motif analysis of promoter tDMRs

To further explore the tissue-specific regulatory potential of tDMRs we used the JASPAR database (http://jaspar.genereg.net; Vlieghe et al. 2006) to search for over-represented transcription factor binding sites in tDMRs. Promoter-tDMRs were enriched for motifs associated with various tissue-specific transcription factors (Supplemental Fig. 9). Motifs for SP1 and Krüppel-like factor 4 (KLF4) were significantly over-represented in tDMRs associated with multiple tissues (Fig. 3C). There is extensive evidence that DNA methylation can modulate SP1 binding, and, consequently, tissue-specific gene expression (Li et al. 2004). KLF4 is known to regulate numerous biological processes including differentiation and development, and recently it has been shown that combined ectopic expression of KLF4, POU5F1 (formerly known as OCT4), SOX2, and MYC can induce fibroblasts to revert to a pluripotent state in vitro (Takahashi and Yamanaka 2006) with concomitant reprogramming of DNA methylation, gene expression, and chromatin states (Wernig et al. 2007). Furthermore, since KLF4 and SP1 can act synergistically to regulate gene expression (Sze et al. 2007), it is possible that these two transcription factors are required for tDMR-promoter function in multiple tissues. We saw much less evidence for over-representation of transcription factor binding motifs from the JASPAR database in intergenic tDMRs (data not shown), suggesting that such tDMRs are subject to a different set of tissue-specific DNA–protein interactions. This is consistent with observations from the ENCODE pilot project (The ENCODE Project Consortium 2007) that many sequence-specific factors show differential occupancy at TSSs compared with distal DNase I hypersensitivity sites (that are assumed to contain regulatory information). Motif analysis was not performed for gene-body tDMRs due to the confounding effect of sequence constraints associated with protein-coding regions.

Gene Ontology analysis of genes associated with promoter-tDMRs

The functional relevance of promoter-tDMRs was also investigated by an analysis of Gene Ontology (GO) terms (http://www.geneontology.org/) (Table 3). Previous analyses by others



Motif	Tissue	Differential from somatic tissues	Significance
Klf4	Sperm	+	<1e-5
	Muscle	+	4e-5
	Prostate	+	0.00012
	Placenta	+	0.00068
	CD4	-	0.00080
Sp1	Uterus	-	9e-5
	Cervix	+	0.00011
	B-cells	+	0.00012
	Prostate	-	0.00025

Figure 3. Analysis of tissue-specific differentially methylated regions (tDMRs). (*A*) The *top* half of each panel shows the DNA methylation profiles observed in sperm. In the *bottom* half of each panel, bars above the line represent the proportion of ROIs in each CpGo/e category that display <40% methylation in sperm, but >60% methylation in all somatic tissues (gray) or >60% methylation in one or more (but not all) somatic tissues (red, i.e., these are somatic tDMRs). Bars below the line (i.e., negative values) represent the proportion of ROIs in each CpGo/e category that display >60% methylation in sperm, but <40% methylation in all somatic tissues (gray) or <40% methylation in one or more (but not all) somatic tissues (red, i.e., these are also somatic tDMRs). (*B*) Comparison of tissue-specific DNA methylation and gene expression. (*Left*) Comparison of promoter DNA methylation (only ROIs overlapping the TSS were used) and gene expresson between whole blood and uterus. Gene expression data are from GNF SymAtlas database (Su et al. 2004). Insufficient data were available for $\text{CpG}_{o/e} > 0.8$. Yellow bars represent genomic regions that display <40% methylation in whole blood and >60% methylation in uterus. Blue bars represent genomic regions that display >60% methylation in whole blood and <40% methylation in uterus. (*Right*) Comparison of gene-body methylation with gene expression of the associated gene. Intronic and exonic methylation data were combined for this analysis. All ROIs in these categories were used, not just CpG-dense ROIs, but we did not stratify by $\text{CpG}_{o/e}$ since there were insufficient numbers of exonic/intronic CpG-poor ROIs. Both figures show 95% confidence interval. The color code is the same as at *left* (promoter). Pairwise comparisons of other tissues showed similar results (data not shown). (*C*) Known transcription factor motif analysis of tDMR promoters. We used the JASPAR database (http://jaspar.genereg.net) to search for over-represented transcription factor binding sites in tDMRs. (+) tDMRs that are hypermethylated in the tissue of interest; (−) tDMRs that are hypomethylated in the tissue of interest.

Table 3. Comparison of Gene Ontology (GO) terms (http://www.geneontology.org/) associated with tDMR and non-tDMR promoters analyzed in this study

	Low CpG	High CpG
Constitutively methylated	Serine-type endopeptidase activity Membrane	
Somatic tDMR	Olfactory receptor activity Sensor perception of smell Response to stimulus G-protein coupled receptor protein signaling pathway Receptor activity Signal transduction Immune response Integral to membrane Extracellular region Inflammatory response	Multicellular organismal development Nucleus
Constitutively unmethylated	Intracellular Mitochondrion Nucleic acid binding	Nucleus Transcription RNA binding Nucleotide binding Protein binding RNA splicing DNA binding mRNA processing Regulation of transcription, DNA-dependent ATP binding

Promoters were defined as 500 bp upstream of the annotated TSS in the Ensembl genome browser. Only signficant over-representations ($P < 10^{-5}$) are shown.

(Saxonov et al. 2006) have shown that CGI-promoters are strongly associated with "house-keeping" genes, and CpG-poor promoters with tissue-specific genes. Our analysis shows that constitutively unmethylated CGI-promoters (unmethylated in all tissues including sperm) can be distinguished from tDMR CGI-promoters, as the latter show a relative enrichment for tissue-specific functions, in particular, neural processes. CpG-poor promoters that are either constitutively methylated or associated with tDMRs were found to be strongly associated with tissue-specific functions. However, constitutively unmethylated CpG-poor promoters, although rare (599), are associated with housekeeping genes. Therefore, promoter-tDMRs, across a broad range of CpG densities, are associated with genes that are thought to function in a tissue-specific manner. Furthermore, this analysis shows that it is constitutively unmethylated promoters, and not CGI-promoters per se, that are associated with housekeeping genes.

DNA methylation profiles of mature spermatozoa

Mammalian genomes undergo genome-wide epigenetic reprogramming during gametogenesis (Bird 2002), presumably to restore totipotency. Consistent with the recent study by Weber et al. (2007), we found that within the promoter category, 94% of CGI-associated ROIs and 62% of intermediate CpG-density ROIs ($CpG_{o/e}$ = 0.4–0.6) are unmethylated in sperm, with the latter more likely to undergo de novo methylation in one or more somatic tissues (Fig. 3A). (Note: we first reported our data for sperm in Down et al. 2008.) Among the CpG-poor ROIs, 13% are unmethylated in sperm and undergo de novo methylation in somatic tissues, and 9% are methylated in sperm and undergo de novo demethylation in one or more somatic tissues. The various nonpromoter CGI categories also displayed hypomethylation in sperm (Fig. 3A). However, the major difference in DNA methylation dynamics between promoter and nonpromoter CGIs is that the latter are more likely to undergo soma-wide de novo methylation in somatic tissues.

Discussion

Here we report a novel integrated resource (methylation profiles of DNA, or mPod) for genome-wide human tissue-specific DNA methylation profiles. Firstly, the tissue-specific genome-wide DNA methylation profiles of 16 different human tissues represent the largest and most comprehensive available data set for this epigenetic modification. Second, all of our data are displayed via a novel visualization tool within the Ensembl genome browser, making the data accessible to the wider scientific community. Finally, the development of the Batman algorithm allowed us to estimate absolute methylation levels from MeDIP, a technique that is readily applied to genome-wide DNA methylation profiling. Batman can also be used to analyze genome-wide MeDIP data generated from other array platforms and next-generation sequencing technologies such as the Illumina Genome Analyzer (Down et al. 2008).

Our study—which includes a range of comparative analyses with independent genome-wide data-sets of gene expression, histone modifications and other regulatory proteins, transcription factor binding preferences, and gene-ontology terms—suggests DNA methylation is involved in regulating at least some promoters over a wide range of CpG densities in the context of cell- and tissue-specific transcriptional programs. Furthermore, we propose that only a fraction of the nonpromoter CGIs predicted by previous bioinformatic approaches are likely to be regulatory elements in the same sense that promoter-CGIs are thought to be.

Obviously, many questions regarding the role of DNA methylation remain to be answered, including how tDMRs are established, maintained, and function. Although it is easy to postulate how tissue-specific methylation at promoters could influence gene expression, the role of gene-body tDMRs is less clear. Understanding the role of tissue-specific differential methylation in the context of nongenic regions, including repetitive sequences that we did not study, will also be critical, especially in light of recent genome-wide association studies of complex diseases that have revealed many putative causative variants to be located within nongenic regions. Future studies, such as the recently proposed International Human Epigenome Project (Jones and Martienssen 2005) will undoubtedly address many of these important questions concerning the role of DNA methylation in genome function.

Methods

Samples

Sixteen different tissue types were analyzed: B-cells, CD8 T-cells, CD4 T-cells, cervix, colon, liver, lung, rectum, pancreas, prostate, placenta, skeletal muscle, sperm, uterus, whole blood, and the EBV-transformed GM06690 cell line. Individual sample information is listed in Supplemental Table 1. Tissue samples were obtained from AMS Biotechnology or Analytical Biological Services. The GM06990 cell line was a gift from Dr Ian Dunham

(Wellcome Trust Sanger Institute). (Note: we first reported our data for sperm in Down et al. 2008).

Immunoprecipitation of methylated DNA

Methylated DNA Immunoprecipitation (MeDIP) was based on a previously published protocol, but we also included a ligatin-mediated PCR (LM-PCR) step (Oberley et al. 2004) to amplify the material. Array hybridizations performed before and after LM-PCR showed that the LM-PCR did not introduce significant amplification bias (Down et al. 2008). A total of 2.5 µg of genomic DNA was sheared to a size range of from 300 to 800 bp. The resulting fragments were blunt-ended by incubation for 20 min at 12°C in a 120-µL reaction containing the DNA sample, 1× Buffer 2 (NEB), 10× BSA (NEB), 100 µM dNTP mix, and T4 DNA polymerase (NEB). The reaction was purified using a Zymo-5 kit (Genetix) according to the manufacturer's instructions, but the final elution was done in 30 µL of TE buffer (pH 8.5). Ligation of the adaptors was performed by incubating overnight at 16°C in a final volume of 100 µL containing the DNA sample, 40 µL adaptors, T4 DNA ligase 10× buffer, 5 µL of T4 DNA ligase (NEB). The reactions were purified using a Zymo-5 kit as described above. To fill in the overhangs, the sample DNA was incubated at 72°C for 10 min in a reaction containing the DNA, 100 µM dNTPs, 1× AmpliTaq Gold PCR buffer (Applied Biosystems), 1.5 mM $MgCl_2$, 5 U AmpliTaq Polymerase. The DNA was purified using a Zymo-5 kit as described above. A total of 50 ng of the ligated sample was set aside as the input fraction; 1.2 µg of the ligated DNA sample was subjected to MeDIP as described previously, after scaling down the reaction accordingly. The immunoprecipitated (IP) sample was purified using Zymo-5 kit (using 700 µL of binding buffer) according to the manufacturer's instructions. Ten nanograms of each IP and input fraction for each sample were subjected to PCR (20 cycles) using the Advantage-GC genomic PCR kit (Clontech). PCR cycling conditions are available upon request. After the LM-PCR, the duplicate reactions were combined, purified using a Qiagen PCR-clean up kit (Qiagen), and eluted with 50 µL of water. The MeDIP and input fractions were sent to Nimblegen for hybridization.

Array design

Our microarray consists of 382,178 50-bp probes. Although we aimed to target all annotated TSSs and nonpromoter CGIs, we were unable to design enough suitable unique probes for 18% of the TSSs and 28% of nonpromoter CGIs, largely due to the presence of repeat elements. In addition to the regions described in Table 1, the array contained 50-bp probes tiled at ~100 bp density across the entire human Major Histocompatibility Complex, and promoters and nonpromoter CpG islands on the X and Y chromosomes. Analyses of these regions will be presented elsewhere. The array was originally designed using the NCBI build 35 version of the human genome assembly, but then mapped to NCBI build 36 using Exonerate (Slater and Birney 2005). To be mapped, probes were required to align full-length and without gaps or mismatches. Probes that aligned more than once to the NCBI36 sequence were removed from the analysis. Tiled regions were defined by clustering uniquely mapped probes within 200 bp of one another. Singleton probes were discarded. The tiled regions were then divided into 500-bp ROIs. Following hybridization, arrays were LOESS normalized using custom R-scripts prior to Batman analysis of the resulting $\log_2$ ratios.

Bayesian tool for methylation analysis (Batman)

We model the MeDIP-array experiment by assuming that the observed array signal is proportional to the density of methylated CpG dinucleotides. We can then use Bayesian inference to determine the actual methylation state of CpGs. Batman consists of a suite of Java programs that implement this inference process using the Nested Sampling strategy (http://www.inference.phy.cam.ac.uk/bayesys/). Refer to Down et al. (2008) for a detailed description of Batman.

Bioinformatic analyses

Methylation data were compared with genomic features obtained from Ensembl genome browser (*Homo sapiens* release 45.36g based on NCBI_36). Pseudogene annotation is from http://www.pseudogene.org/. All analyses were performed using a series of custom Java, Perl, and R scripts, which are available upon request. All analyses were performed at the level of ROIs (500-bp intervals) unless otherwise stated.

For the expression analyses, Affymetrix expression data were downloaded from the Gene Expression Omnibus (accession no. GSE1133). Assignments of Affymetrix probe-sets to Ensembl transcripts were extracted from the Ensembl core database version homo_sapiens_core_44_36f. When more than one probe-set was mapped to a given transcript, we used the median expression score for all available probe-sets. For the expression plots, we used means of log-expression-scores, or log-expression-ratios for the tissue-specific expression analyses. The 95% confidence intervals were calculated by bootstrapped difference-of-means tests.

Data localizing histone modifications, CTCF binding sites, histone variant H2A.Z, and RNA polymerase II was obtained in the form of sequencing read alignments from http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcell.html, and originally generated by Barski et al. (2007) For each ROI, the number of overlapping reads was determined using a custom Perl script. The over-representation tests were bootstrapped difference-of-means tests.

Associations of GO terms to Ensembl gene IDs were extracted from the Ensembl core database version homo_sapiens_core_44_36f. Core promoter regions for all Ensembl transcripts were defined as the region 500 bp upstream of the annotated transcript start site. All promoters with available methylation data were then classified according to CpG density and methylation state across this 500-bp window. For each of the five promoter classes shown in Table 2, we performed a hypergeometric test to check for over-representation of each term in the GO vocabulary for genes with promoters of that class compared with the complete set of genes with available promoter methylation data. We report significant over-representations ($P < 10^{-5}$).

Tissue-specific differentially methylated regions (tDMRs) were called in 500-bp 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 (i.e., not sperm, placenta, or the cell line). Although the 40% and 60% cut-offs are arbitrary, they were chosen in an informed manner after looking at the distribution of scores round different features. More stringent thresholds did not materially affect our conclusions (data not shown). To identify hypomethylated tDMRs, we looked for ROIs with methylation <40% in the target tissue and >60% in at least three somatic tissues. There are six tissues in common between our study and the Human Epigenome Project (HEP): CD4 T-cells, CD8 T-cells, liver, placenta, skeletal muscle, and sperm, and 850 genomic regions in common between the two data sets. We called tDMRs in the HEP data set using the same strategy as for the MeDIP-array data. Only those genomic regions for which DNA methylation data were available for all six tissues in both data sets were used to calculate the positive predictive values (PPV, i.e., true positives/[true positives + false positives]) and sensitivity of the tDMR calls for the MeDIP-array data. The PPV is the fraction of MeDIP-array tDMRs that were also called as tDMRs in the HEP study—78%. Sensitivity (true positives/[true positives + false negatives], or the fraction of HEP tDMRs that are also classified as tDMRs in the MeDIP-array data), was 61%.

Statistics

All credible intervals were estimated by bootstrapping unless otherwise stated. Statistical testing for the GO analysis was performed with hypergeometric tests. *P*-values for significant association between methylation state and gene expression or ChIP data were all calculated using a nonparametric empirical test: briefly, the data were divided into bins (typically high and low methylation) and the mean expression was calculated for each bin. The data were then repeatedly resampled, and an empirical *P*-value was calculated by counting the number of times an equal or greater difference of means was seen in the resampled data compared with the original data. A similar empirical test was used for the motif analysis, except that the area under an ROC curve was used as the test statistic.

Acknowledgments

V.K.R. was supported by a C.J. Martin Fellowship from the National Health and Medical Research Council of Australia. T.A.D., E.M.T., L.B., K.L.H., D.K.J., M.M.M., T.J.P.H., N.P.C., and S.B. were supported by the Wellcome Trust. S.G., N.J., and M.H. were supported by a grant (Highthroughput Epigenetic Regulatory Organization in Chromatin [HEROIC], LSHG-CT-2005-018883) from the European Union under the 6th Framework Programme to S.B. (M.H.) and E.B. (S.G., N.L.). N.P.T., J.C.M., and S.T. were supported by grant C14303/A8646 from Cancer Research UK.

References

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129:** 823–837.

Bird, A. 2002. DNA methylation patterns and epigenetic memory. *Genes & Dev.* **16:** 6–21.

Bird, A., Taggart, M., Frommer, M., Miller, O.J., and Macleod, D. 1985. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **40:** 91–99.

Down, T.A., Rakyan, V.K., Turner, D.J., Flicek, P., Li, H., Kulesha, E., Graf, S., Johnson, N., Herrero, J., Tomazou, E.M., et al. 2008. A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat. Biotechnol.* 26: 779–785.

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V.K., Attwood, J., Burger, M., Burton, J., Cox, T.V., Davies, R., Down, T.A., et al. 2006. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.* **38:** 1378–1385.

The ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

Estecio, M.R., Yan, P.S., Ibrahim, A.E., Tellez, C.S., Shen, L., Huang, T.H., and Issa, J.P. 2007. High-throughput methylation profiling by MCA coupled to CpG island microarray. *Genome Res.* **17:** 1529–1536.

Flicek, P., Aken, B.L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cummingham, F., Cutts, T., et al. 2008. Ensembl 2008. *Nucleic Acids Res.* **36:** D707–D714.

Gardiner-Garden, M. and Frommer, M. 1987. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196:** 261–265.

Glass, J.L., Thompson, R.F., Khulan, B., Figueroa, M.E., Olivier, E.N., Oakley, E.J., Van Zant, G., Bouhassira, E.E., Melnick, A., Golden, A., et al. 2007. CG dinucleotide clustering is a species-specific property of the genome. *Nucleic Acids Res.* **35:** 6798–6807.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130:** 77–88.

Hellman, A. and Chess, A. 2007. Gene body-specific methylation on the active X chromosome. *Science* **315:** 1141–1143.

Illingworth, R., Kerr, A., DeSousa, D., Jørgensen, H., Ellis, P., Stalker, J., Jackson, D., Clee, C., Plumb, R., Rogers, J., et al. 2008. A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS Biol.* **6:** e22. doi: 10.1371/journal.pbio.0060022.

Jones, P.A. and Baylin, S.B. 2007. The epigenomics of cancer. *Cell* **128:** 683–692.

Jones, P.A. and Martienssen, R. 2005. A blueprint for a Human Epigenome Project: The AACR Human Epigenome Workshop. *Cancer Res.* **65:** 11241–11246.

Keshet, I., Schiesinhger, Y., Farkash, S.W., Rand, E., Hecht, M., Segal, E., Pikarski, E., Young, R.A., Niveleau, A., Cedar, H., et al. 2006. Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nat. Genet.* **38:** 149–153.

Khulan, B., Thompson, R.F., Ye, K., Fazzari, M.J., Suzuki, M., Stassiek, E., Figueroa, M.E., Glass, J.L., Chen, Q., Montagna, C., et al. 2006. Comparative isoschizomer profiling of cytosine methylation: The HELP assay. *Genome Res.* **16:** 1046–1055.

Kitamura, E., Igarashi, J., Morohashi, A., Hida, N., Oinuma, T., Nemoto, N., Song, F., Ghosh, S., Held, W.A., Yoshida-Noro, C., et al. 2007. Analysis of tissue-specific differentially methylated regions (TDMs) in humans. *Genomics* **89:** 326–337.

Laird, P.W. 2003. The power and promise of DNA methylation markers. *Nat. Rev. Cancer* **3:** 253–266.

Li, L., He, S., Sun, J.M., and Davie, J.R. 2004. Gene regulation by Sp1 and Sp3. *Biochem. Cell Biol.* **82:** 460–471.

Liu, L., Zhang, J., Bates, S., Li, J.-J., Peehl, D.M., Rhim, J.S., and Pfeifer, G.P. 2005. A methylation profile of in vitro immortalized human cell lines. *Int. J. Oncol.* **26:** 275–285.

Mukhopadhyay, R., Yu, W., Whitehead, J., Xu, J., Lezcano, M., Pack, S., Kanduri, C., Kanduri, M., Ginjala, V., Vostrov, A., et al. 2004. The binding sites for the chromatin insulator protein CTCF map to DNA methylation-free domains genome-wide. *Genome Res.* **14:** 1594–1602.

Oberley, M.J., Tsao, J., Yau, P., and Farnham, P.J. 2004. High-throughput screening of chromatin immunoprecipitates using CpG-island microarrays. *Methods Enzymol.* **376:** 315–334.

Qi, Y., Rolfe, A., MacIsaac, K.D., Gerber, G.K., Pokholok, D., Zeitlinger, J., Danford, T., Dowell, R.D., Fraenkel, E., Jaakkola, T.A., et al. 2006. High-resolution computational models of genome binding events. *Nat. Biotechnol.* **24:** 963–970.

Rakyan, V.K., Hildmann, T., Novik, K.L., Lewin, J., Tost, J., Cox, A.V., Andrews, T.D., Howe, K.L., Otto, T., Olek, A., et al. 2004. DNA methylation profiling of the human major histocompatibility complex: A pilot study for the human epigenome project. *PLoS Biol.* **2:** e405. doi: 10.1371/journal.pbio.0020405.

Saxonov, S., Berg, P., and Brutlag, D.L. 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci.* **103:** 1412–1417.

Shen, L., Kondo, Y., Guo, Y., Zhang, J., Zhang, L., Ahmed, S., Shu, J., Chen, X., Waterland, R.A., and Issa, J.P. 2007. Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. *PLoS Genet.* **3:** 2023–2036.

Slater, G.S. and Birney, E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **15:** 31–34.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci.* **101:** 6062–6067.

Suzuki, M.M., Kerr, A.R., De Sousa, D., and Bird, A. 2007. CpG methylation is targeted to transcription units in an invertebrate genome. *Genome Res.* **17:** 625–631.

Sze, K.L., Lee, W.M., and Lui, W.Y. 2007. Expression of CLMP, a novel tight junction protein, is mediated via the interaction of GATA with the kruppel family proteins, KLF4 and Sp1, in mouse TM4 sertoli cells. *J. Cell. Physiol.* **214:** 334–344.

Takahashi, K. and Yamanaka, S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126:** 663–676.

Takai, D. and Jones, P.A. 2002. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci.* **99:** 3740–3745.

Vlieghe, D., Sandelin, A., De Biesder, P.J., Vieminckx, K., Wasserman, W.W., van Roy, F., and Lenhard, B. 2006. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res.* **34:** D95–D97.

Walsh, C.P. and Bestor, T.H. 1999. Cytosine methylation and mammalian development. *Genes & Dev.* **13:** 26–34.

Weber, M., Davies, J.J., Wittig, D., Oakeley, E.J., Haase, M., Lam, W.L., and Schubeier, D. 2005. Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.* **37:** 853–862.

Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Paabo, S., Rebhan, M., and Schubeler, D. 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat. Genet.* **39:** 457–466.

Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., Hochedlinger, K., Bernstein, B.E., and Jaenisch, R. 2007. In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* **448:** 318–324.

West, A.G. and Fraser, P. 2005. Remote control of gene transcription. *Hum. Mol. Genet.* **14:** R101–R111.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S., Chen, H., Henderson, I., Shinn, P., Pellegrini, M., and Jacobsen, S. 2006. Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. *Cell* **126:** 1189–1201.

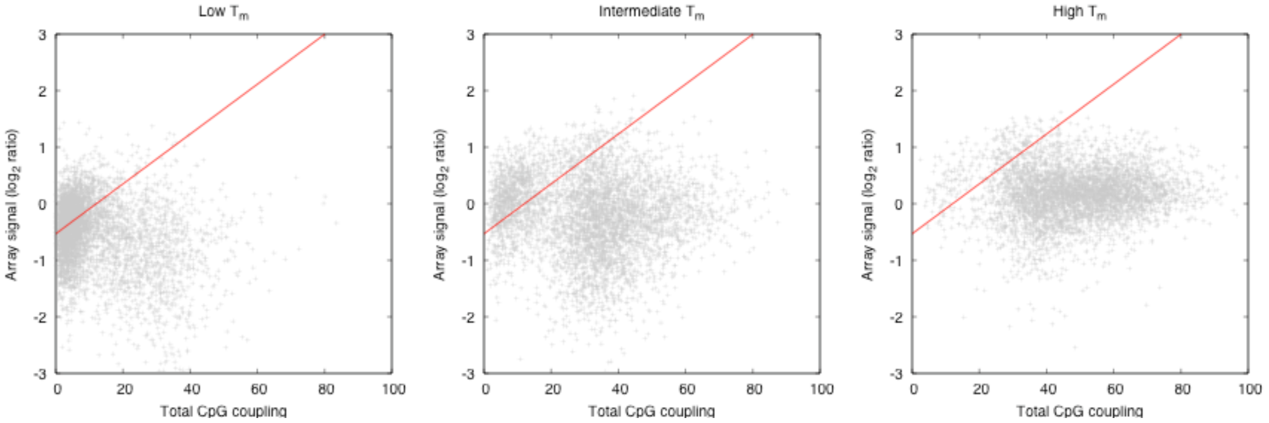
Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T., and Henikoff, S. 2006. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nat. Genet.* **39:** 61–69.

Received February 13, 2008; accepted in revised form June 15, 2008.

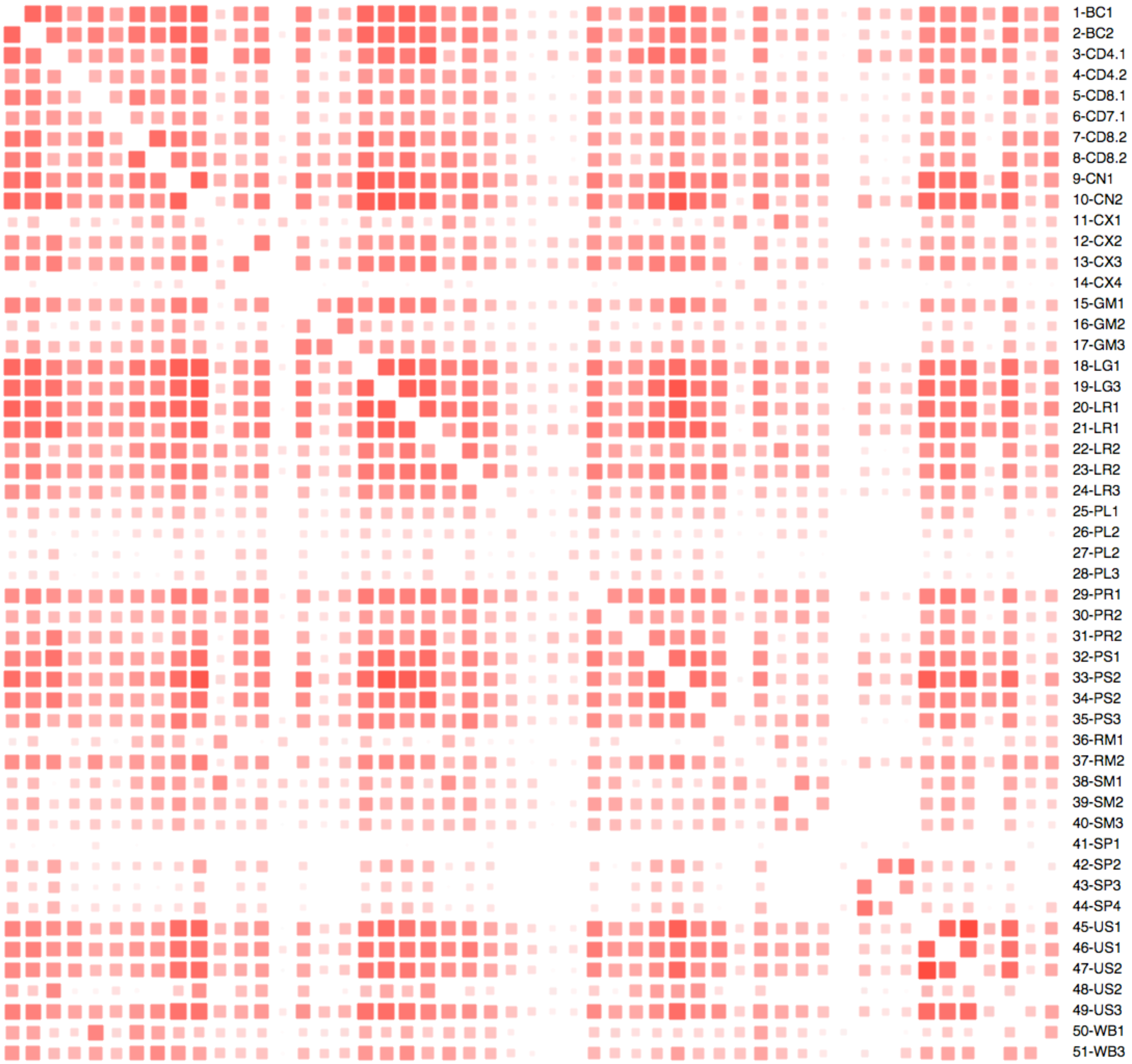
Supplementary information for "An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs)" – Rakyan et al.,

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	

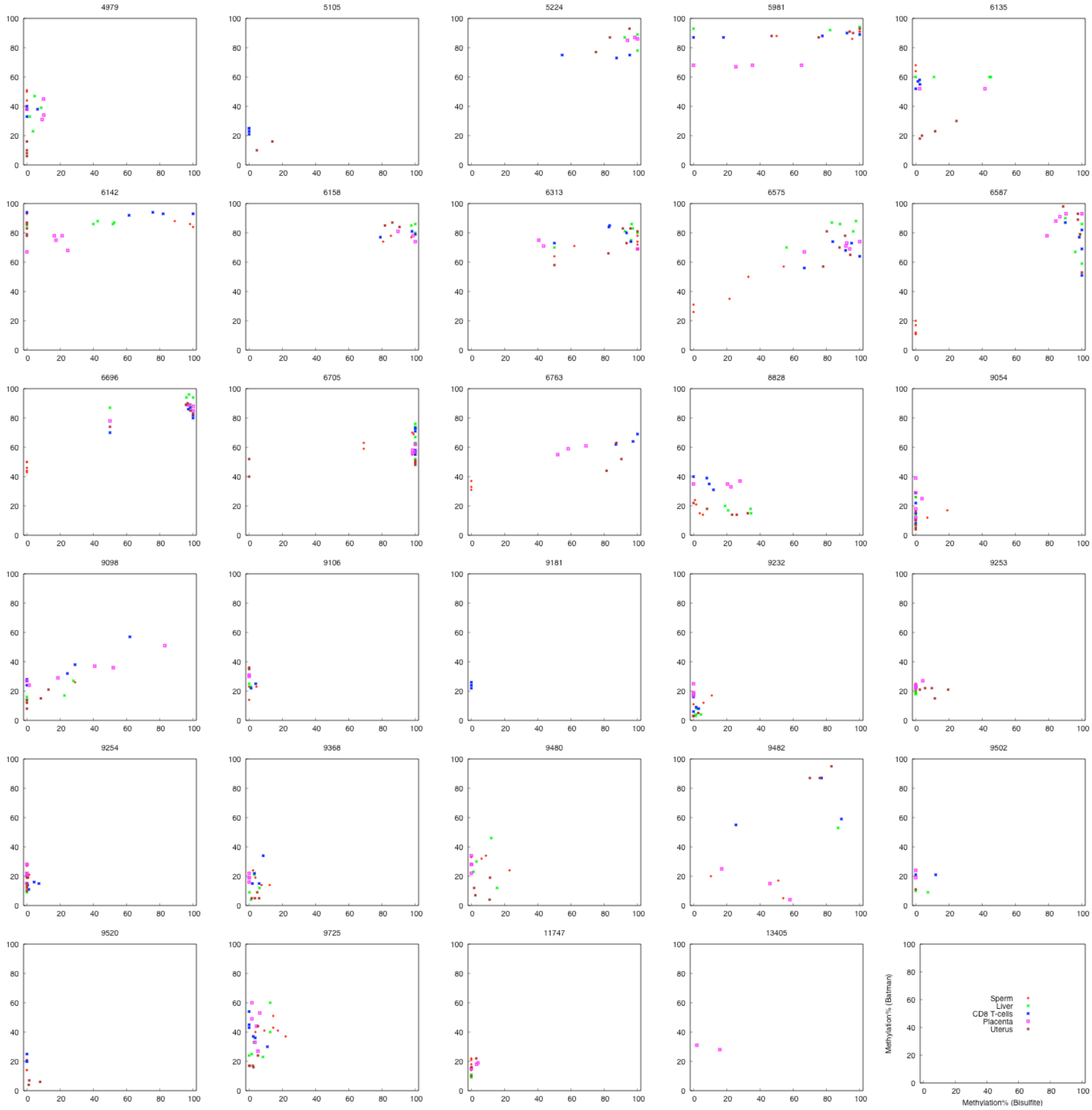
Supplementary Table 1. Tissue samples used in the study. The "Baseline" and "Response" parameters refer, respectively, to the intercept and inverse slope of a linear model fitted to the low-CpG portion of each array's data (refer to description of Batman). The "Response" parameter can be interpreted 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).

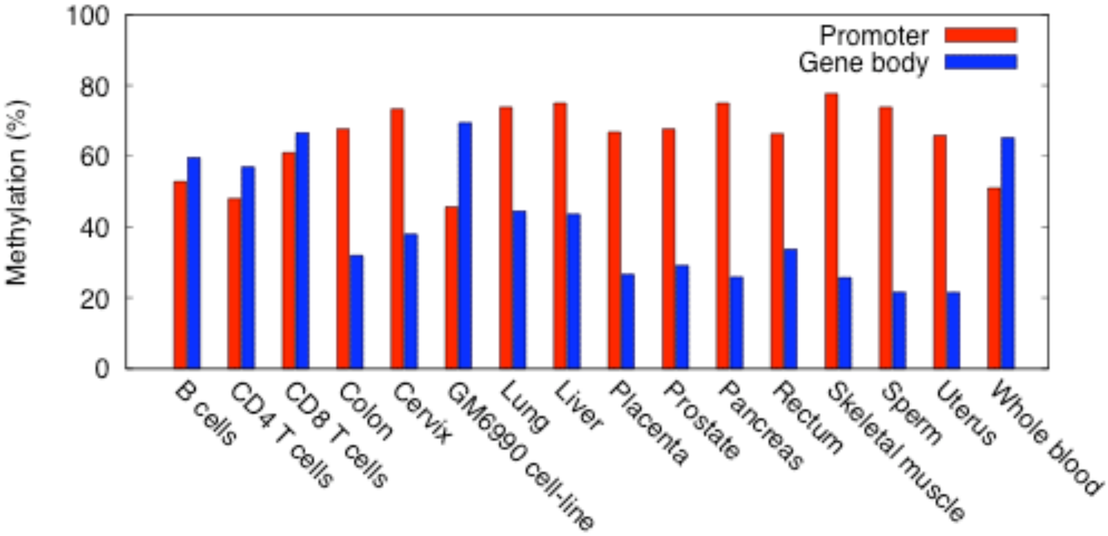


Supplementary Figure 3. Bisulfite-PCR validation of the Batman calls. Initially, 36 regions were chosen for bisulfite-PCR validation, spanning a range of CpG densitites, genomic locations, tDMRs and non-tDMRs (see Supplementary Table 2). However, PCR products could be obtained for only 29. The validation 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.

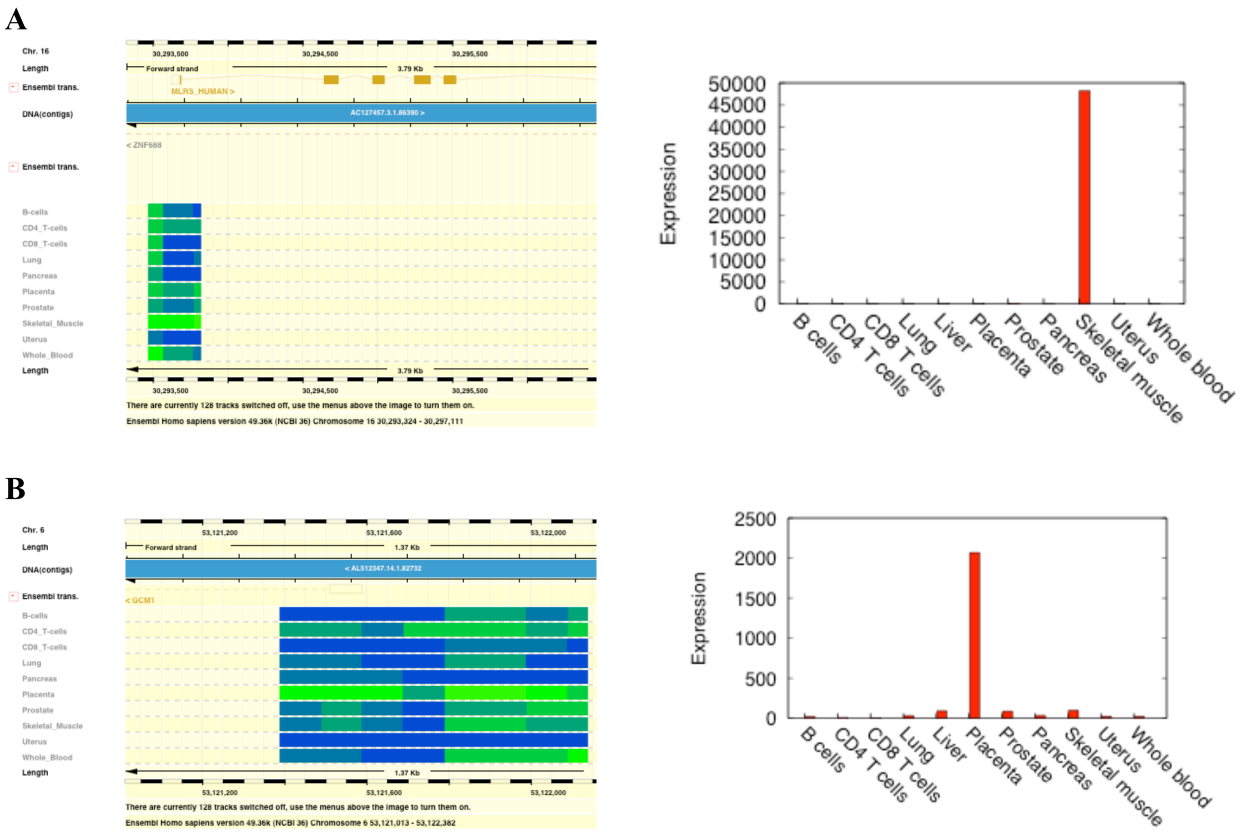
Supplementary Table 2. Regions analyzed in Supplementary Figure 1.

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.



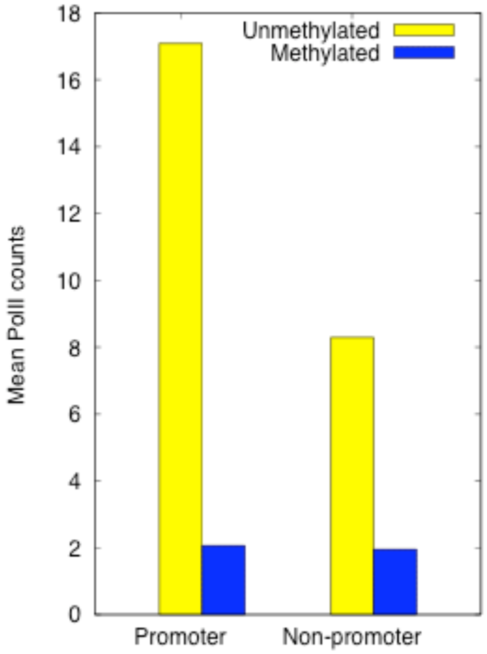
There are currently 128 tracks switched off, use the menus above the image to turn them on.

Ensembl Homo sapiens version 49.36k (NCBI 36) Chromosome 16 30,293,324 - 30,297,111

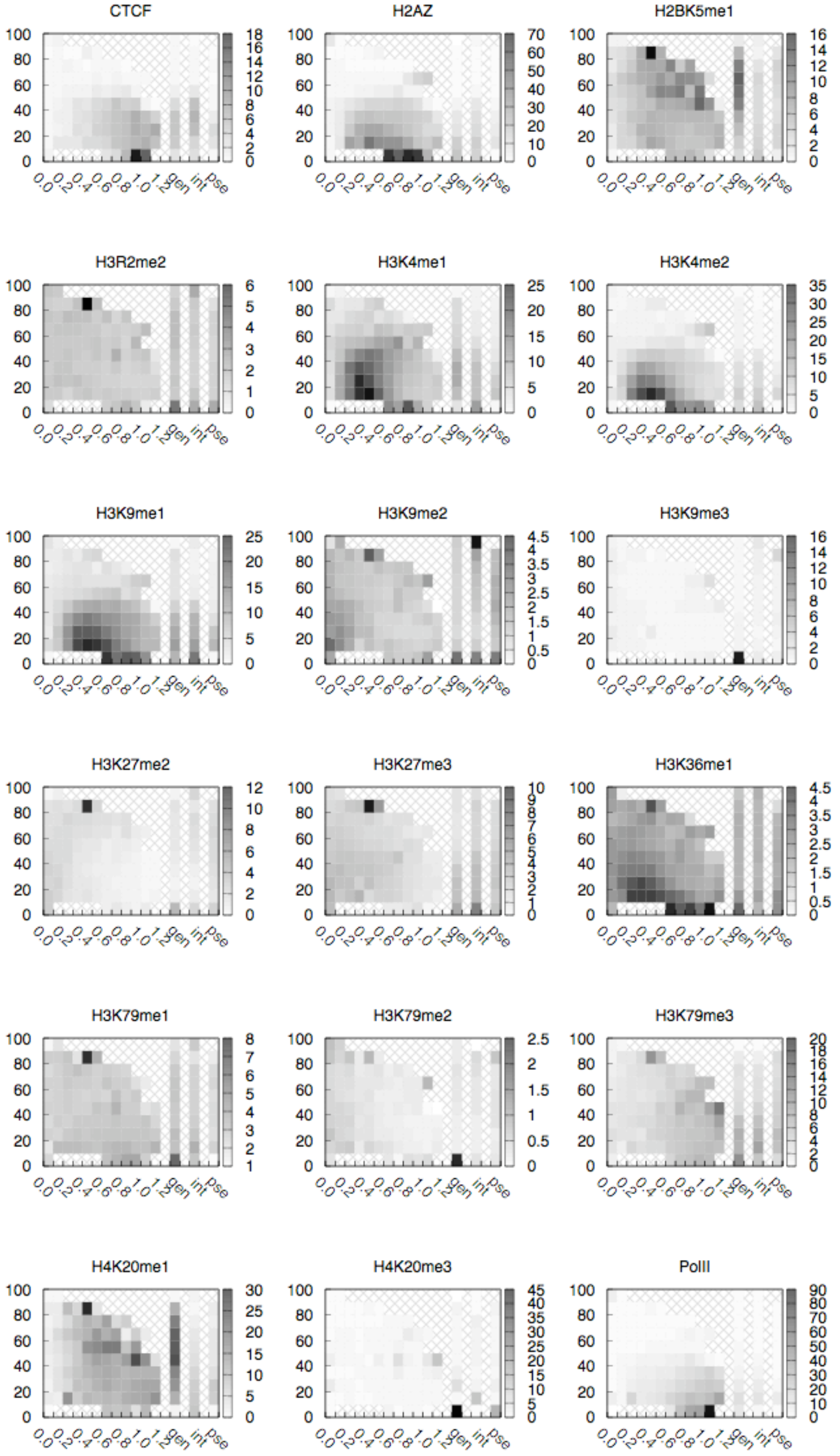
There are currently 128 tracks switched off, use the menus above the image to turn them on.

Ensembl Homo sapiens version 49.36k (NCBI 36) Chromosome 6 53,121,013 - 53,122,382

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).



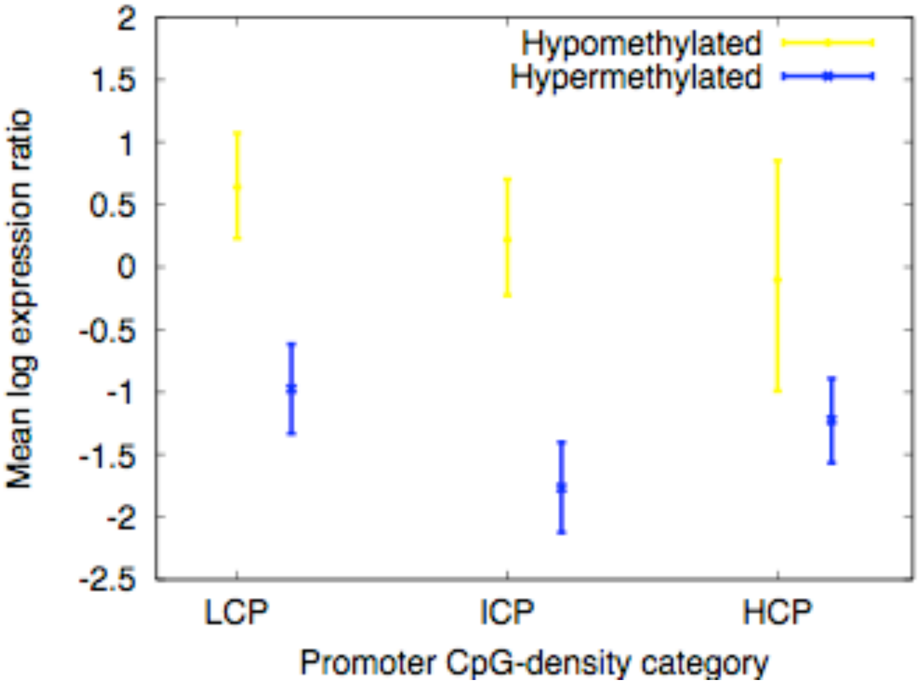
Supplementary Figure 6. DNA methylation data for promoter and non-promoter CpG islands from our study was correlated with genome-wide enrichment profiles RNA PolII generated by Barski et al., 2007 (ref. 2) using Solexa 1G sequencing technology. The y-axis DNA represents the average tag count for RNA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



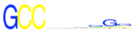
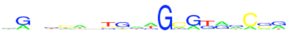
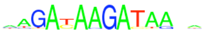
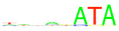
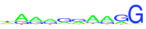
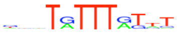
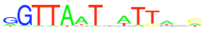
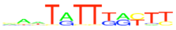
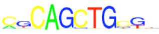
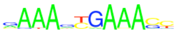
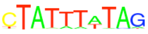
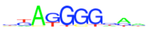
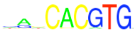
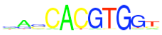
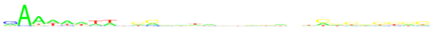
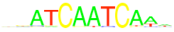
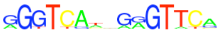
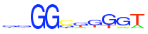
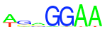
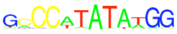
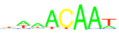
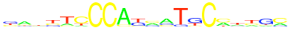
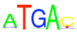
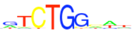
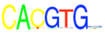
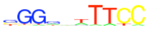
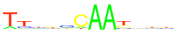
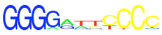
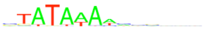
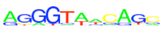
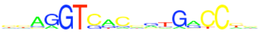
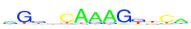
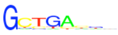
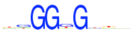
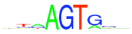
Supplementary Figure 7. DNA methylation data from our study was correlated with genome-wide enrichment profiles for 20 histone lysine and arginine methylations, H2A.Z, RNA PolII, and CTCF generated by Barski et al., 2007 (ref. 2) using Solexa 1G sequencing technology. The x-axes represent $CpG_{o/e}$ (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).