A Novel CpG Island Set Identifies Tissue-Specific Methylation at Developmental Gene Loci

Robert Illingworth¹, Alastair Kerr¹, Dina DeSousa¹, Helle Jørgensen², Peter Ellis³, Jim Stalker³, David Jackson³, Chris Clee³, Robert Plumb³, Jane Rogers³, Sean Humphray³, Tony Cox³, Cordelia Langford³, Adrian Bird†

¹Wellcome Trust Centre for Cell Biology, University of Edinburgh, Edinburgh, United Kingdom, ²Lymphocyte Development Group, MRC Clinical Sciences Centre, Imperial College School of Medicine, London, United Kingdom, ³Wellcome Trust Sanger Centre, Hinxton, Cambridge, United Kingdom

CpG islands (CGIs) are dense clusters of CpG sequences that punctuate the CpG-deficient human genome and associate with many gene promoters. As CGIs also differ from bulk chromosomal DNA by their frequent lack of cytosine methylation, we devised a CGI enrichment method based on nonmethylated CpG affinity chromatography. The resulting library was sequenced to define a novel human blood CGI set that includes many that are not detected by current algorithms. Approximately half of CGIs were associated with annotated gene transcription start sites, the remainder being intra- or intergenic. Using an array representing over 17,000 CGIs, we established that 6%–8% of CGIs are methylated in genomic DNA of human blood, brain, muscle, and spleen. Inter- and intragenic CGIs are preferentially susceptible to methylation. CGIs showing tissue-specific methylation were overrepresented at numerous genetic loci that are essential for development, including HOX and PAX family members. The findings enable a comprehensive analysis of the roles played by CGI methylation in normal and diseased human tissues.


Introduction

DNA methylation in the mammalian genome arises due to covalent addition of a methyl group to the 5′ position of cytosine in the context of the palindromic dinucleotide, CpG. This modification is established and maintained by a family of DNA methyltransferases that are essential for development and viability [1,2]. The pattern of CpG methylation in the human genome distinguishes two fractions with distinct properties: a major fraction (~98%), in which CpGs are relatively infrequent (on average 1 per 100 bp) but highly methylated (approximately 80% of all CpG sites), and a minor fraction (<2%) that comprises short stretches of DNA (~1,000 bp) in which CpG is frequent (~1 per 10 bp) and methylation-free. The latter are known as CpG islands (CGIs) and they frequently colocalise with the transcription start sites (TSSs) of genes [3,4].

Although CGIs are often free of methylation, there are circumstances in which they become heavily methylated, and this invariably correlates with silencing of any promoter within the CGI. Artificial methylation of CGI promoters has long been known to extinguish transcription when the constructs are introduced into living cells [5]. Moreover, demethylation of endogenous methylated CGIs using DNA methyltransferase inhibitors can restore expression of the gene [6]. These findings demonstrate that dense CpG methylation prevents expression of CGI promoters. Because of this biological consequence, it is important to know the extent of CGI methylation in both normal and diseased tissue states. The classical example is X chromosome inactivation in placental mammals, during which hundreds of CGI promoters become methylated and contribute to the stability of gene inactivation on this chromosome [7,8]. Genomic imprinting can also depend upon differential CGI methylation between maternal and paternal alleles [9]. Certain “testis-specific antigen” genes possess CGIs that are methylated in all somatic tissues, but not in testis, where the genes are expressed [10]. Several additional candidates for CGI methylation in normal tissues have been reported [11,12], and the number of cases has recently grown due to large-scale bisulfite sequencing [13] and analysis of promoter methylation using microarrays [14].

In the cases of X chromosome inactivation and genomic imprinting, the biological processes were described initially, and CpG methylation was subsequently implicated through mechanistic studies. To uncover new biological roles for CGI methylation in hitherto undiscovered biological processes, it would be advantageous to comprehensively screen genomic DNA for methylated CGIs in normal or diseased cell types. A persistent limitation affecting this kind of approach has been uncertainty concerning CGI identification [15]. The criteria for designating a sequence as CGI-like are currently exclusively bioinformatic in nature, relying on the differences in the base composition and CpG frequencies (observed/expected) between bulk genomic DNA and CGIs [16,17]. In an
attempt to address this limitation and create a resource for future analysis, we developed a method for CGI identification and purification based on their lack of CpG methylation in an otherwise highly methylated genome.

Our method utilised a protein domain with a specific affinity for clustered nonmethylated CpG sites [18,19]. Using this reagent we physically purified DNA sequences that contain clusters of nonmethylated CpG-rich DNA from human blood DNA. Large-scale sequencing of the fraction identified a CGI set that was annotated on the ENSEMBL database. We found that many CGIs in the set were not associated with promoters of annotated genes, but were either within transcription units or between genes. By arraying the intact CGI sequences, we were able to interrogate genomic DNA fractions from several human tissues in order to identify methylated CGIs. The results revealed large numbers of CGIs that are methylated in normal human tissues, many of which showed tissue-specific methylation.

Results

A Novel Technique for Purification of CpG Islands

To enrich for nonmethylated CpG-rich DNA (CpG islands), we developed the technique of CXXC affinity purification (CAP). This uses the cysteine-rich CXXC3 domain that has a high affinity for nonmethylated CpG sites [18,19]. A recombinant CXXC domain from mouse Mbd1 [19] was expressed in bacteria, and its binding specificity for nonmethylated CpG sites was confirmed (Figure 1A). The CXXC domain had no detectable affinity for DNA containing only methylated CpGs or for DNA lacking CpGs altogether. We linked the CXXC domain to a sepharose matrix and confirmed that this fractionated DNA fragments according to CpG density and methylation status (unpublished data). All DNA bound to the

![Figure 1](https://example.com/figure1.png)
CGIs are identified bioinformatically as DNA sequences with a base composition greater than 50% G+C and a CpG[o/e] of more than 0.6 [23]. The DNA length over which this condition applies is critical. Initially the threshold most often used was 200 bp, whereas 500 bp is now more commonly applied [17]. These two criteria are formalised as “NCBI-relaxed” and “NCBI-strict,” respectively (http://www.ncbi.nlm.nih.gov/mapview/static/humansearch.html#cpg). The relaxed algorithm predicts 307,193 CGIs in the human genome, which includes many repeated sequences and gene exons. Over 90% of NCBI-relaxed CGIs are not represented in either our library or the set predicted by the NCBI-strict. This and other arguments suggest that the great majority (>90%) are false positives. On the other hand, 77% of clones in the CGI library match CGIs predicted by the “NCBI-strict” algorithm (Table 1). Examples of the coincidence of NCBI-strict predicted CGIs and sequenced CGI clones are illustrated for the three typical regions of the human genome (Figure 2D).

Altogether, NCBI-strict identifies 24,163 CGIs in the human genome, which accords with the adjusted CGI library estimate of 25,200. The coincidence of these numbers masks significant differences, however, as 23% of CGIs in the library are not detected by the NCBI-strict algorithm (4,092 out of 17,387; Table 1). Four randomly selected examples of library CGIs not detected by NCBI-strict (Figure 2D and 2E, numbered) gave CpG maps resembling CGIs; three of these coincided with the promoters of annotated protein-coding genes (Figure 2D and 2E). The presence of bioinformatically predicted CGIs that are missing from the CGI library is most probably due to sequence failure of ~31% of library inserts. Analysis of the CGIs missed by the NCBI-strict algorithm shows them to be, as expected, significantly weaker with respect to CpG and G+C content than the total set (Figure S1). It was not obvious, however, that the algorithm could be easily improved based on this information. Relaxation of the sequence parameters reduces the number of false negatives, but leads to increased numbers of false positives. We suggest that CAP identifies islands that fail the NCBI criteria, but reduces the false discovery rate by excluding spurious methylated CpG-rich sequences. Like the majority of CGIs, most NCBI-missed islands are gene-associated, although with an increased incidence of intragenic islands (Table S1). The CGI library therefore includes a significant fraction of bona fide CGIs that are missed by one of the best available algorithms.

CAP defines a set of CGIs that is coherent with respect to clustering of nonmethylated CpG sites. The genomic distribution of these CGI sequences correlates strongly with gene density (Figure 2F). For example, gene-rich Chromosome 19 is also CGI-rich, whereas gene-poor Chromosome 18 is correspondingly CGI-poor. With respect to annotated protein-coding genes, we found that 76% of CGIs are within 1.5 kb of a transcription unit, but only 49% overlap with the TSS (Table 2). It follows that half of CGIs are not TSS-associated, but are either within downstream regions of transcription units (22%) or located in intergenic DNA. Previous studies have detected CGIs at the TSS of 56% of human protein-coding genes [24]. As 43.5% of TSSs overlap sequenced CGIs, we calculate that the sequenced set of 17,387 CGIs represents 78% of the CGI complement. According to this calculation, the total CGI number would be 22,400, somewhat less than...
Figure 2. A Library of DNA Sequences that Bind Tightly to the CXXC Column Represents a Comprehensive Set of CGIs

(A and B) Plots of fragment length versus G+C content (A) and CpG(o/e) (B) for 28,013 unique Mse1 inserts. Fragments shorter than 512 bp with a G+C content \( < 50\% \) and a CpG(o/e) \( < 0.6 \) (grey dots) were filtered out as contamination. The dashed line indicates the base composition (A) and CpG o/e (B) of bulk genomic DNA.

(C) A filtered insert set representing 17,387 CGIs shows a discrete distribution that is distant from bulk genomic DNA (black dot).

(D) Three random chromosomal regions showing CGI sequences mapped by ENSEMBL (green bars). Also shown are CGIs predicted by the NCBI-strict and NCBI-relaxed algorithms (blue bars). The directions of transcription of coding sequences (yellow bars) are arrowed. Numbered CGIs (1–4) represent sequences not detected by the NCBI-strict algorithm.

(E) CpG maps of the four CGI clones not predicted by NCBI-strict. Transcription start sites in examples 1, 3, and 4 are indicated by arrows. Sequenced MseI fragments are denoted by dashed lines and CpG sites by vertical black strokes.

(F) The distribution of cloned CGIs (red strokes) on human chromosomes. The number of CGIs on each chromosome is shown (right) and centromeres are denoted by blue dots.

doi:10.1371/journal.pbio.0060022.g002

Methylation of Human CpG Islands
the figure of 25,200 deduced from the fraction of sequenced inserts.

**MBD Affinity Purification and Blood CGI Methylation**

CAP selects CGIs from blood DNA based on their lack of methylation and therefore excludes the small fraction of CGIs (\(<3\%\)) that are fully methylated in somatic cells from the set [14]. Indeed, CGIs associated with the human testis-specific antigen genes [10], which are methylated in somatic tissues, were not enriched by CAP (Figure 1C) or present in the library (unpublished data). Despite the absence of these fully methylated CGIs, we reasoned that the blood CGI library provides an opportunity to screen for methylation that affects a fraction of all copies of a specific CGI in whole blood DNA. Also, it permits a screen for differential methylation of CGIs in tissues and cell types other than blood. To investigate CGI methylation in normal human tissues, we constructed an array of sequenced CGIs from the library by immobilising single-stranded PCR-amplified inserts on glass slides using 5’-aminolink chemistry as described (http://www.sanger.ac.uk/Projects/Microarrays/arraylab/methods.shtml). As probes for the array, methylated CGIs were enriched from genomic DNA using MBD affinity purification (MAP), which was shown previously to efficiently bind methylated CGIs [20] (Figure 3A and 3B). Human male and female blood DNA was MseI-digested and ligated to universal catch linkers. We verified by PCR that affinity fractionation using MAP effectively separated known methylated CGIs (XIST on the active X chromosome and NYES0) from bulk genomic DNA and nonmethylated CGIs (P48 and XIST on the inactive X chromosome; see Figure 3B). Male and female DNA fractions were pooled after two rounds of MAP, amplified by linker-mediated PCR, cyanine labeled, and hybridized to the CGI microarray. Quadruplicate hybridisations (inclusive of cyanine dye swaps) gave mean enrichment values (MAP/Input) for those sequences (Figure 3C). The results established that M values greater than 1.5 denote CGIs that are significantly enriched by MAP and therefore methylated. CGIs of the BEST1 and R4RL1 genes were predicted to be nonmethylated (\(M = 0.2–0.4\)) and methylated (\(M = 2.2–2.8\)), respectively, based on the array data. Bisulfite genomic sequencing confirmed this expectation (Figure 3G and 3H).

The major difference in CGI methylation between male and female DNA was expected to be due to X chromosome inactivation (see also [25]). We therefore compared the methylation status of CGIs on Chr 16 and Chr X in male versus female DNA. Chr 16 CGIs did not vary between males and females, whereas Chr X CGIs were significantly enriched in female DNA as predicted (Figure 3D–3F; Table S2). Studies of human X chromosome inactivation have indicated that a proportion of genes escape inactivation and are therefore expressed from both chromosomes [26,27]. By comparing the microarray data for a set of inactivated and escaping CGIs, we found that inactivated genes had significantly higher M values (\(p\)-value = 1.213 × 10^{-7}) (Figure 3I). This finding affirms the long-standing link between CGI methylation and gene silencing and validates the present experimental system as a means of detecting genes that are shut down in this way.

**Table 1.** Comparison of Human Blood CGI Set with Bioinformatic Prediction

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Number of CGIs</th>
<th>NCBI&lt;sub&gt;fl&lt;/sub&gt;</th>
<th>NCBI&lt;sub&gt;strict&lt;/sub&gt;</th>
<th>CGI</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCBI&lt;sub&gt;fl&lt;/sub&gt;</td>
<td>307,193</td>
<td>307,193</td>
<td>24,163 (7.9%)</td>
<td>16,072 (5.2%)</td>
</tr>
<tr>
<td>NCBI&lt;sub&gt;strict&lt;/sub&gt;</td>
<td>24,163</td>
<td>24,163</td>
<td>24,163</td>
<td>13,568 (56.2%)</td>
</tr>
<tr>
<td>CGI</td>
<td>17,387</td>
<td>15,799</td>
<td>13,305 (76.5%)</td>
<td>17,387</td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pbio.0060022.s001

**Table 2.** Relationship between CGI Library Inserts and Protein-Coding Genes

<table>
<thead>
<tr>
<th>Type of Overlap</th>
<th>Gene Type</th>
<th>Total Genes (CGIs)</th>
<th>Overlap with CGI (Gene)</th>
<th>Percentage</th>
<th>Overlap with CGI (TSS)</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Overlap: CGI</td>
<td>Protein</td>
<td>21,384</td>
<td>15,118</td>
<td>70.7</td>
<td>9,312</td>
<td>43.5*</td>
</tr>
<tr>
<td></td>
<td>All genes</td>
<td>31,524</td>
<td>15,433</td>
<td>49.0</td>
<td>9,529</td>
<td>30.2</td>
</tr>
<tr>
<td>CGI Overlap: Genes</td>
<td>Protein</td>
<td>17,387</td>
<td>13,271</td>
<td>76.3</td>
<td>8,491</td>
<td>48.8</td>
</tr>
<tr>
<td></td>
<td>All genes</td>
<td>17,387</td>
<td>13,360</td>
<td>76.8</td>
<td>8,611</td>
<td>49.5</td>
</tr>
</tbody>
</table>

*This fraction is less than the known fraction of genes with promoter CGIs (56%), because 31% of CGI inserts did not yield DNA sequence.

*All genes as classified on the ENSEMBL genome browser including noncoding RNAs, pseudogenes, VDJ regions, etc.
doi:10.1371/journal.pbio.0060022.s002
Figure 3. Use of an Arrayed CGI Library to Detect CGI Methylation in Human Blood DNA

(A) Schematic showing isolation of densely methylated CGIs using MBD affinity purification based on reference [20]. Open and filled circles represent nonmethylated and methylated CpG sites, respectively.

(B) Examples of retention of known methylated CGIs by MBD affinity chromatography. Methylated XIST and NYESC CGIs elute at high salt concentration, whereas nonmethylated P48 and female XIST co-elute with bulk genomic DNA (blue line) at low salt concentration (red line).

(C) M values (log2 [MBD/Input]) ≥1.5 (dashed vertical arrow) denote DNA fragments enriched by MAP. M values are plotted against the ratio of fragment abundance in the MAP probe versus input DNA as determined by quantitative PCR. Error bars represent ± standard deviation.

(D–F) MAP CGI array hybridization identifies CGIs that are methylated on the inactive X chromosome. (D) Probes isolated by MAP from male and female

G

H

I

11_61475084-61485083
BEST1
1kb
Mse I3878
Female (0.2)
Male (0.4)

17_17823965-1792394
R4RL1
1kb
Mse I9112
Female (2.2)
Male (2.8)

Methylated CpG rich DNA
DNA (Absorbance)
NaCl conc (M)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

Female
Male
Female
Male
Female
Male

Figure 3. Use of an Arrayed CGI Library to Detect CGI Methylation in Human Blood DNA

(A) Schematic showing isolation of densely methylated CGIs using MBD affinity purification based on reference [20]. Open and filled circles represent nonmethylated and methylated CpG sites, respectively.

(B) Examples of retention of known methylated CGIs by MBD affinity chromatography. Methylated XIST and NYESC CGIs elute at high salt concentration, whereas nonmethylated P48 and female XIST co-elute with bulk genomic DNA (blue line) at low salt concentration (red line).

(C) M values (log2 [MBD/Input]) ≥1.5 (dashed vertical arrow) denote DNA fragments enriched by MAP. M values are plotted against the ratio of fragment abundance in the MAP probe versus input DNA as determined by quantitative PCR. Error bars represent ± standard deviation.

(D–F) MAP CGI array hybridization identifies CGIs that are methylated on the inactive X chromosome. (D) Probes isolated by MAP from male and female
Differential CGI Methylation in Human Tissues

Methylation of CGIs on the inactive X chromosome and at imprinted genes is well known, but CGI methylation at other chromosomal loci in normal cells and tissues is incompletely characterized [12,13,28,29,30]. To investigate this issue on a large scale, we probed CGI arrays with MAP fractions from genomic DNA (three individuals per pool) of brain, muscle, spleen, and sperm in addition to blood (Figure 4A). MAP enrichment of methylated CGIs in sperm DNA consistently failed to generate enough DNA for labeling using our standard PCR amplification conditions and was therefore not analysed further. We conclude that the level of CGI methylation in sperm is far lower than in any of the somatic tissues. Taking M values greater than 1.5 to signify methylation, we observed between 5.7% and 8.3% of CGIs methylated in the somatic tissues that were tested (Figure 4B; Table 3; Dataset S1). Some CGIs were methylated in common between all the tested somatic tissues, whereas others were methylated in only one or a subset of the tissues. We noted that methylated CGIs disproportionally involved those that are remote from the TSS of an annotated gene. In the dataset as a whole, only 8% of TSS CGIs showed evidence of methylation in at least one tissue, whereas 22% of 3' CGIs were methylated (Table 4). Do the methylated CGIs differ in sequence characteristics from CGIs that remain methylation-free? We plotted the CpG[o/e] frequencies of 1,657 CGIs that acquired methylation in one or more tissues and found a mean CpG[o/e] of 0.77 compared with 0.75 for methylated CGIs (Figure 4C). Though statistically significant (p-value = 1.413e-10) the biological significance of this small difference is unclear.

We checked by bisulfite sequence analysis a panel of seven CGIs with M values suggestive of tissue-specific methylation (M values differing between tissues by >0.75). In each case, bisulfite data confirmed the microarray predictions. CGI I11878 is not associated with an annotated gene (±1.5 kb) and is methylated exclusively in muscle and brain (Figure 4D). CGI I2985 spans the transcription start site of the SEC31B gene, whose product is implicated in vesicular trafficking, and is compositely methylated only in blood and spleen (Figure 4E). CGIs I13406 (Figure 4F) and I12175 (Figure 5A) are methylated specifically in muscle. These overlap the predicted gene 67313 and the 3' end of OSR1. CGI I13654, which is associated with the promoter region of an annotated PAX6 transcript (Q59GKD2), previously shown to contain methylated CpG sites [31], is specifically methylated in brain (Figure 5B). I11878 is a 3' CGI of ZN649 and is only methylated in spleen (Figure 4G).

Many methylated CGIs were associated with genes that are essential for development (Figure 5). This was confirmed by analysis of gene ontology, which showed significant over-representation of genes whose products are involved in developmental processes, including ectoderm and mesoderm development, neurogenesis, and segment specification (Table S3). Transcription factors, including homeobox family members and other DNA binding proteins, were twice as abundant as expected by chance. Other gene categories did not show significant enrichment. Among the CGIs whose methylation status was confirmed by bisulfite sequencing, PAX6 is involved in eye development and neurogenesis [32], the HOXC cluster lays down the embryonic body plan, and OSR1 is related to a gene involved in Drosophila gut development. We examined the extended HOXC and PAX6 loci for CGI methylation status using the MAP-CGI array data. Our library identified 19 CGIs within the 150-kb HOXC gene cluster of which eight were methylated differentially in blood, muscle, and spleen (Figure 5C). Brain was the only tissue that lacked obvious HOXC CGI methylation. Of nine CGIs near PAX6, two showed differential methylation. In addition to brain-specific methylation of the PAX6-Q59GKD2 CGI (see Figure 5B), we observed methylation of a CGI upstream of the major PAX6 promoter in muscle and brain (Figure 5D).

The majority of CGIs identified as methylated by MAP-CGI array hybridization display composite methylation (Figures 3, 4, and 5), whereby DNA strands at a specific locus were either heavily methylated or essentially nonmethylated. This can explain why CGIs that were initially selected by being nonmethylated in blood DNA (by CAP) nevertheless register as methylated by MAP-CGI array analysis. One potential explanation for composite CGI methylation is that different individuals within the tissue pools exhibit different CGI methylation. To look for such “polymorphism,” we examined CGI I5134, which is within the HOXC cluster and shows composite methylation by bisulfite genomic sequencing. Analysis of individuals by MAP-CGI arrays showed highly significant differences between individual C and individuals A and B (Figure 5E). This strikingly confirms individual variability in methylation at this CGI.

Another potential explanation for composite CGI methylation is that cell types within the tissue sample possess different CGI methylation profiles. Blood, for example, consists of monocytes and granulocytes, each of which is subdivided into other cell types. As CGI I2985 was methylated at about half of DNA strands in blood, we tested the level of CGI methylation in DNA from monocytes and granulocytes separately. The results showed that monocytes had high methylation levels at this CGI, whereas granulocytes had very low methylation (Figure 5F). These findings indicate a developmental origin for cell type-specific methylation at this genomic CGI.

Discussion

A Comprehensive CGI Set

We describe the characterisation of a comprehensive, verified CGI set derived from human blood genomic DNA
Figure 4. Tissue-Specific CGI Methylation in a Panel of Human Tissues

(A) Examples of pairwise comparisons using MAP CGI probes derived from blood, brain, muscle, and spleen. Broken red lines indicate threshold M values used to determine differential CGI methylation.

(B) Frequencies of methylated CGIs in blood, brain, muscle, and spleen. The following categories are represented: CGIs methylated in all tested tissues (black); CGIs methylated in more than one tissue tested but not all (green); CGIs methylated in one tissue only (blue); CGIs methylated in one tissue tested but unclassified in other tissues (white).

(C) Somatically methylated CGIs display a very small but significant reduction in CpG(o/e) (0.75) relative to the whole CGI set (0.77; n = 1,657 and 12,661, Wilcoxon rank test: p-value: 1.0226 **). The histogram shows the CpG(o/e) profile for the total CGI set (white bars) overlaid with the CpG(o/e) profile for methylated CGIs (red line).

(D–G) Confirmation of candidate CGIs showing evidence of tissue specific methylation by bisulfite genomic sequencing. Layout is as for Figure 3G.

Table 3. CGI Methylation in Human Tissues

<table>
<thead>
<tr>
<th>Methylation Status</th>
<th>Blood</th>
<th>Brain</th>
<th>Muscle</th>
<th>Spleen</th>
</tr>
</thead>
<tbody>
<tr>
<td>Differentially methylated (multiple*)</td>
<td>408</td>
<td>408</td>
<td>408</td>
<td>408</td>
</tr>
<tr>
<td>Wilcoxon rank test: p-value:</td>
<td>149</td>
<td>135</td>
<td>214</td>
<td>247</td>
</tr>
<tr>
<td>Unclassified methylation</td>
<td>50</td>
<td>35</td>
<td>178</td>
<td>140</td>
</tr>
<tr>
<td>Total</td>
<td>303</td>
<td>237</td>
<td>392</td>
<td>381</td>
</tr>
<tr>
<td>CGIs</td>
<td>910</td>
<td>815</td>
<td>1,192</td>
<td>1,176</td>
</tr>
<tr>
<td>Percentage methylated</td>
<td>6.4</td>
<td>5.7</td>
<td>8.3</td>
<td>8.2</td>
</tr>
</tbody>
</table>

*Refers to number of tissues tested.

doi:10.1371/journal.pbio.0060022.t003
the possibility that putative CGI transcripts play regulatory roles during development. Recent analyses of the human HOX gene cluster highlight the functional importance of noncoding RNAs [45]. Large numbers of potential CGI promoters within HOX gene loci may therefore contribute to the regulation of these complex loci.

**CGI Methylation in Normal Tissues**

CGI methylation has been extensively studied in cancers and their derivative cell lines, but relatively less attention has been paid to the phenomenon in normal tissues. Several studies have reported somatic CGI methylation, but in early examples the bioinformatics procedure used to identify these sequences was often equivalent to the NCBI-relaxed algorithm, which generates a large excess of questionable CGI candidates. The MASPIN gene, for example, scores as a methylated CGI promoter by the relaxed criterion [28], but it is not detected as such either by the NCBI-strict algorithm or by CAP (unpublished data). A recent report addressing the methylation status of 16,000 human promoters identified that 3% of TSS-associated CGIs are normally methylated in somatic tissue [14], which is somewhat below the levels observed in our study (7.8%; Table 4). We detect a much higher frequency of methylation at nonpromoter CGIs (average = 16%), which are obviously absent from promoter arrays. In particular, 22% of CGIs near the 3′ ends of genes are methylated. Extensive bisulfite sequence analysis [13] surveyed 512 CGIs on Chr 6, 20, and 21 and reported 9.2% to be methylated in somatic tissues. This is similar to the overall level of 11.6% methylation among 14,318 CGIs detected by our study (Table 4).

Our findings raise important questions about the relationship of CGI methylation to gene expression. On the X chromosome, it is clear that methylated CGIs correlate with inactivated genes whereas unmethylated CGIs correlate with genes known to escape inactivation. The generalisation that CGI methylation silences promoters is therefore supported (see also [25]). The relevance to gene expression of the autosomal methylated CGIs identified here is complicated by the frequent presence of both methylated and nonmethylated alleles in a specific tissue (see below). This means that even if CGI methylation silences a promoter completely, large changes in gene expression are not to be expected. Also, many CGIs are not at promoters of annotated genes, but are within or between transcription units. Their function with respect to transcription, if any, may be positive or negative. Finally, any transcripts originating from these “orphan” CGIs have yet to be identified and cannot be tested. For these reasons, it is difficult to make predictions about the effect of CGI methylation on global transcription levels. We nevertheless mined published expression microarray data to determine whether tissues in which a specific set of promoter CGIs was methylated expressed the associated genes at a different level from tissues where the same CGI was unmethylated. The results showed no obvious correlation between CGI methylation and expression. This, therefore, remains an open question that demands detailed analysis of specific cases.

Genes that play an important role in development were prominent among the set of methylated CGIs identified by MAP-CGI array hybridization. Out of 109 CGI-associated genes that contain homeobox-like domains, 27 (~25%) were unmethylated in at least one tissue compared with ~11% of all CGI-associated genes (see Table 4). Specifically, we identified 79 CGIs in the four human HOX gene clusters A–D, of which 22 were methylated in at least one of the tissues that we tested. Given the relatively small selection of tissues analysed in the study, the actual frequency of HOX CGI methylation in all human tissues is likely to be higher than one in four. Interestingly, methylation of HOX gene CGIs is also reported in cancers [46], raising the possibility that cancer CGI methylation patterns mimic patterns that arise during development. A potential link between normal development and cancer is suggested by the finding that CGIs methylated in cancer preferentially include promoters that are marked by association with polycomb group proteins in embryonic stem cells [47–49]. In contrast, we found little difference between the fractions of all CGIs (5.9% = 845/14,318) and of methylated CGIs (7.7% = 127/1,657) that were polycomb-associated in embryonic cells [37]. The origins of CGI methylation in cancer may be distinct from the mechanisms that lead to CGI methylation in normal tissues.

It was reported that the most CpG-rich CGIs among 512 analysed on Chr 6, Chr 20, and Chr 22 were never methylated, suggesting that the CpG-richness may protect from methylation [13]. In a larger CGI set, we detected a very small, but statistically significant, difference in sequence properties between CGIs that become methylated and those that remain immune in the tested cell types. The mean CpG[oc] was 0.75 for methylated CGIs compared with 0.77 for bulk CGIs (Figure 4C). Bock and colleagues [50] identified sequence features that were predictive for CGI methylation, including specific repeats, sequence patterns, and DNA structure. Contrary to predictions of this method, methylated CGIs were significantly depleted in repetitive elements and showed no difference in predicted base twist. We did, however, observe small, but statistically significant, increases in simple

---

**Table 4. Location of Methylated CGI Relative to Protein-Coding Genes**

<table>
<thead>
<tr>
<th>CGI Gene Association</th>
<th>All CGIs</th>
<th>Methylated</th>
<th>Methylated (%)</th>
<th>Differentially Methylated</th>
<th>Differentially Methylated (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All CGIs</td>
<td>14,318</td>
<td>1,657</td>
<td>11.6</td>
<td>711</td>
<td>5</td>
</tr>
<tr>
<td>Genes</td>
<td>11,383</td>
<td>1,209</td>
<td>10.6</td>
<td>501</td>
<td>4.4</td>
</tr>
<tr>
<td>5′</td>
<td>7,915</td>
<td>620</td>
<td>7.8</td>
<td>243</td>
<td>3.1</td>
</tr>
<tr>
<td>3′</td>
<td>765</td>
<td>166</td>
<td>21.7</td>
<td>77</td>
<td>10.1</td>
</tr>
<tr>
<td>Intragenic</td>
<td>3,478</td>
<td>536</td>
<td>15.4</td>
<td>230</td>
<td>6.6</td>
</tr>
<tr>
<td>Intergenic</td>
<td>2,863</td>
<td>435</td>
<td>15.2</td>
<td>203</td>
<td>7.1</td>
</tr>
</tbody>
</table>

---

**Figure 3. CGI Methylation and Gene Expression.**

Methylation of Human CpG Islands

---

[Table 4. Location of Methylated CGI Relative to Protein-Coding Genes]

[Figure 3. CGI Methylation and Gene Expression]
Methylation of Human CpG Islands

A

B

C

D

E

F
sequence elements (TGTG(CACA) and base-stacking energy (see Figure S2)). The biological relevance of these minimal differences is uncertain.

Weber and coworkers [14] identified ~2,000 promoters out of 16,000 that were more susceptible to methylation than CGGs themselves. These so-called “weak CpG islands” had an average CpG[occ] ratio intermediate between CGGs and bulk genomic DNA. We have determined that 75% of weak CpG islands reported by Weber et al. are absent from the CGI library. Weak CGGs may be depleted because they are heavily methylated and therefore not enriched by CAP. Indeed, 22 methylated weak CpG islands [14] were not detected in our library. Alternatively, their relatively low CpG density and somewhat elevated frequency of Mse1 sites may result in too few CpGs per fragment for efficient retention by the CXXC matrix.

Composite Methylation of CGIs

Those CGIs that were methylated often showed a mixture of heavily methylated and nonmethylated strands by bisulfite analysis. There are several possible explanations for composite methylation patterns. Firstly, at the highest level, it is possible that different individuals contributing to the DNA pool are polymorphic with respect to this epigenetic mark. We analysed specific CGIs in muscle DNA from three individuals and found evidence of individual variation of this kind. A large-scale survey would be required to determine the extent of inter-individual variability. A second possibility is that cells within the analysed tissue are heterogeneous with respect to CGI methylation. Each of the analysed tissues consists of multiple differentiated cell types that should be analysed separately to address this possibility. Analysis of three compositely methylated CGIs in blood showed one that was highly methylated in monocytes, but weakly methylated in granulocytes, indicating that cell-type specific CGI methylation underlay heterogeneous DNA methylation. A third possible explanation for composite methylation is monoallelic methylation. A previous study of 149 CGIs on Chr 21q detected three that were mono-allelically methylated, indicating that this explanation also accounts for some cases of composite CGI methylation [12].

Methods

Plasmid cloning and recombinant protein purification. Cloning of the His-CXXC construct from murine Mbd1a was described previously [19]. The MBD construct was subcloned from pet30bMeCP2 [51]. A fragment of human MeCP2 corresponding to amino acids 76–167 was PCR-amplified and ligated into the NdeI and EcoRI sites of pet30b (Novagen) to generate a C terminally His-tagged pet30b-MeCP2_76–167. Primers: pet30bMeCP2_76–167Nde1 CGG TTC AAT ACC ATA TGG CCT CTG CCT CCC CCA ACG GCC and pet30bMeCP2_76–167EcoRI CGG AAG TCA AAG AAT TCT CAT (for cagaaggtcatcatctgcca, rev tgagttgtttttcatccacctgtccagtcc , MAO (for CGGTATCAGATTGAAACAT, rev CTCTAAGCATGGC-CCCAGCGTCTGGTAACCATC, revCCACGGGACAGGTACCTC ), NYS/E (for CCCACGGTCTGGTAACCATC, revCCACGGGACAGGTACCTC ).

Methylation of Human CpG Islands

Tissue, Cell-Type, and Individual-Specific CGI Methylation at Developmental Gene Loci

(A–B and E–F) Bisulfite genomic sequencing confirmed tissue-specific CGI methylation associated with the developmental genes OSR1 (A) and PAX6 (B). Multiple CGIs (red boxes) span the HOXC (C) and PAK6 (D) gene loci. Plots of the MAP-CGI array profiles for blood, brain, muscle, and spleen identify tissue-specific CGI methylation (vertical black bars extending above M = 1.5). Gray bars extending downwards below M = 1.5 (broken blue line) represent nonmethylated CGIs. The region of PAK6 analysed by bisulfite genomic sequencing (see Figure SB) is indicated (asterisk in panel D). Tick marks on the y-axis are spaced at intervals of 1 M value unit. Coding sequences are diagrammed as yellow bars.

(E) Individual-specific CGI methylation internal to the HOXC cluster in muscle DNA.

(F) Cell type–specific methylation is seen at the SEC31B promoter CGI in monocytes and granulocytes derived from whole human blood. Bisulfite genomic sequencing results (A–B and E–F) are diagrammed as in Figure 3G.

Methylation of Human CpG Islands

exchange using Sp-Sepharose (GE Healthcare) cation exchange as previously described [51]. Recombinant protein was bound to Nickel sepharose prior to longer term storage.

Electrophoretic mobility shift assay (EMSA). CXXC-EMSA was carried out essentially as described in [19]. Briefly, binding reactions including 0, 250, 500, 1,000, or 2,000 ng of purified recombinant His-CXXC were preincubated in 1× binding buffer (5 × binding buffer: 30 mM Tris-HCl [pH8], 750 mM NaCl, 5 mM DTT, 30 mM MgCl2, 15% Glycerol, 50 mM BSA, and 0.05 μg/μl of poly(dIdT)) (Amersham). End-labeled CGI[12] probe (1 ng) was added to each reaction and incubated for a further 25 min. Complexes between probe DNA and the CXXC domain were resolved on a 1.3% agarose Triton-borate-EDTA gel and imaged by Phosphor Imager (Molecular Devices).

Human DNA samples. Whole blood was collected from voluntary donors and used in anonymized pools. Donors were aware of, and consented to, its use for preparation of DNA. Monocyte and granulocyte cells were prepared from whole human blood using Ficoll gradient centrifugation. Whole blood (3 ml) was layered onto an equivalent volume of Histopaque-1077 (Sigma-Aldrich) and sedimented, according to the manufacturer’s instructions. Mononucleated cells were recovered from the plasma-ficoll interphase and granulocytes from the cell pellet. Whole human blood, monocyte, and granulocyte DNA was extracted using the Genomic-tip 5000 G (Qiagen 10302) genomic DNA purification kit. Sperm DNA was prepared as described [53]. Human skeletal muscle, spleen, and brain genomic DNAs were purchased from Ambion.

CXXC affinity purification. 50–60 mg of recombinant CXXC was dialysed into W1 buffer (50 mM sodium phosphate buffer [pH8], 300 mM NaCl, 10% glycerol, 15 mM β-mercaptoethanol, 0.5 mM PMSF), bound to nickel-charged sepharose, and then washed with 10 column volumes (CVs) of W1, 10 CVs of W2 (W1 + 10 mM Imadazole), and 10 CVs of W1. Beads were packed onto a 1-ml Tricorn chromatography column (GE Healthcare). Mse1 digested male DNA (100 μg) pooled from three individuals was bound to the CXXC column in 90% CA buffer (20 mM Hepes [pH7.9], 0.1% Triton X-100, 10% glycerol, 0.3 mM PMSF, 10 mM 2-Mercaptoethanol) and 10% CB buffer (CA + 1 M NaCl). Equilibrated DNA was then eluted over an increasing NaCl gradient of 10%–100% CB buffer (Figure 1B). Fractions (3 ml) were collected and 200 μl of each was precipitated and resuspended in 40 μl 1× TE buffer. Aliquots were PCR- interrogated using Redhot taq polymerase (Abgene) for probe XIST (for CAGCTGACAAAGG- CATTG, rev GGTAGCATGTTGTTTGAC). NYS/E (for CCCACGGTCTGGTAACCATC, revCCACGGGACAGGTACCTC ), MAB (for CGGTATCAGATTGAAACAT, rev CTCTAAGCATGGCC- TACATCATA), P48 (for cagaaggtcatcatctgcca, rev tgagttgtttttcatccacctgtccagtcc ) under the following conditions: 2 min at 94 °C, followed by 30 cycles of 94 °C for 50 s, 55 °C for 50 s, 72 °C for 50 s; 72 °C for 1 min; a final extension of 72 °C for 7 min. PCR products were resolved on a 1.5% TAE-agarose gel (Figure 1B). Fractions retaining nonmethylated CpG-rich Mse1 fragments (Figure 1B) were pooled, digested with CA buffer, and re-chromatographed. The relevant fractions were precipitated and ligated into the Ndel site of pGEM5zf– (Promega).

CGI library sequencing. The clone set was arrayed into 384-well plates in glycerol for long-term storage. Copies were taken and DNA prepared for sequencing using a modified alkaline lysis method. Cells were lysed in glucose, Tris, EDTA gel and imaged by Phosphor Imager (Molecular Devices). The relevant fractions were precipitated and ligated into the Ndel site of pGEM5zf– (Promega).
GenePix Pro 6.0 (Axon) software package. All subsequent analysis was performed in ddH2O, before being dried by centrifugation (500 g). The slides were blocked (1% ammonium hydroxide for 5 min, followed by 0.1% SDS for 5 min) and then denatured (55°C ddH2O for 2 min), rinsed in ddH2O, and dried by centrifugation for 5 min at 250 g. MAP, labeling, and microarray hybridization. Human tissue DNA pooled from three individuals was digested with MseI, phosphatase-treated, and ligated to 5′-CgC AAG TCT ATT g−3′ reverse primer 5′-CgC CAA CGT ATT Tag tGt AG-3′. PCR products were ethanol-purified and resuspended in 1 × microarray spotting buffer (250 mM sodium phosphate (pH 7.5), 0.1% sodium azide). Arrays were spotted onto amine-binding slides at 20–25°C, 40–50% relative humidity. After an overnight incubation in a humid chamber, the slides were blocked (1% ammonium hydroxide for 5 min, followed by 0.1% SDS for 5 min) and then denatured (55°C ddH2O for 2 min), rinsed in ddH2O, and dried by centrifugation for 5 min at 250 g.

Microarray scanning and data analysis. Arrays were scanned with a GenePix Pro 6.0 (Axon) software package. All subsequent analysis was performed using the ImageJ software. Features with poor signal-to-noise ratios were stabilized using a base value of 1,000 for background-subtracted intensities. Cy3 and Cy5 signals ( Mayor-Tyndall and Perry [46], Mayor-Tyndall and Smith [47]) were normalized by print-tip loess. Each tissue analysis is represented by four microarrays comprising two independent replicates with respective dye swaps. Processed values were averaged through linear modeling and used to determine the relative enrichment of MAP DNA relative to Input. An M value of >1.5 was designated as the threshold for hypermethylation as determined by quantitative PCR (Figure 3C) and bisulfite genomic sequencing (Figures 3G and 3H, 4D–4G, and 5A and 5B). This threshold was confirmed as significant (p-value = 0.001, see text for discussion). In contrast, the percentage of X-linked CGIs methylated on the single male X chromosome was comparable to the levels found on the human autosomes, as illustrated for Chr16.

Bisulfite genomic sequencing. Bisulfite treatment of genomic DNA was carried out on DNA purified from human blood (Promega). Bisulfite converted DNA was cloned using the Strataclone PCR cloning system (Stratagene) and at least five clones were sequenced for each CGI. Methylation levels were calculated for each CG of the CGIs using the MethPrimer software [3] (sequences are available on request). There are no significant differences between the DNA sequences from different individuals (data not shown). Bisulfite sequencing of specific repeats (TGTG/CACA) show small but significant increase in stacking energy relative to all CGIs (p-value < 0.001). In contrast we found no significant difference in the base twist for differentially methylated CGI-genes. Genes involved in developmental processes such as neurogenesis and segmentation are significantly under-represented (p-value < 0.001).

Figure S2. Sequence Properties of Methylated CGIs

Bock and colleagues determined a number of DNA sequence features that are correlated with DNA methylation at CGIs [50]. Here we compare the sequence attributes of the methylated and total CGI sets with respect to DNA structure (stacking energy and base twist) and specific repeats (TGTG/CACA). Methylated CGIs show small but significant increase in stacking energy relative to all CGIs (p-value < 0.001). In contrast we found no significant difference in the base twist between methylated and unmethylated CGIs. TGTG/CACA specific repeats were found to be significantly enriched in methylated CGIs (p-value < 0.001; see text for discussion). In contrast, all repetitive elements (as outlined in Repbase [38]) were found to be marginally depleted in methylated CGIs (p-value < 0.01, Wilcoxon rank sum test, n = 4,082 and 10,236). Stacking energy and base twist were calculated using the EMSOSS b-twisted program with default settings [59]. All distributions were tested for parametric distribution by the Shapiro-Wilk test of normality. Nonparametric significance values were determined using the Wilcoxon rank sum test (n = 4,082 and 10,236).

Bisulfite sequencing of specific repeats (TGTG/CACA) show small but significant increase in stacking energy relative to all CGIs (p-value < 0.001). In contrast we found no significant difference in the base twist between methylated and unmethylated CGIs. TGTG/CACA specific repeats were found to be significantly enriched in methylated CGIs (p-value < 0.001; see text for discussion). In contrast, all repetitive elements (as outlined in Repbase [38]) were found to be marginally depleted in methylated CGIs (p-value < 0.01, Wilcoxon rank sum test, n = 4,082 and 10,236). Stacking energy and base twist were calculated using the EMSOSS b-twisted program with default settings [59]. All distributions were tested for parametric distribution by the Shapiro-Wilk test of normality. Nonparametric significance values were determined using the Wilcoxon rank sum test (n = 4,082 and 10,236).

Figure S2. Sequence Properties of Methylated CGIs

Bock and colleagues determined a number of DNA sequence features that are correlated with DNA methylation at CGIs [50]. Here we compare the sequence attributes of the methylated and total CGI sets with respect to DNA structure (stacking energy and base twist) and specific repeats (TGTG/CACA). Methylated CGIs show small but significant increase in stacking energy relative to all CGIs (p-value < 0.001). In contrast we found no significant difference in the base twist between methylated and unmethylated CGIs. TGTG/CACA specific repeats were found to be significantly enriched in methylated CGIs (p-value < 0.001; see text for discussion). In contrast, all repetitive elements (as outlined in Repbase [38]) were found to be marginally depleted in methylated CGIs (p-value < 0.01, Wilcoxon rank sum test, n = 4,082 and 10,236). Stacking energy and base twist were calculated using the EMSOSS b-twisted program with default settings [59]. All distributions were tested for parametric distribution by the Shapiro-Wilk test of normality. Nonparametric significance values were determined using the Wilcoxon rank sum test (n = 4,082 and 10,236).
References


