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

Citation for published version:

Digital Object Identifier (DOI):
10.1371/journal.pbio.0060022

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
PLoS Biology

Publisher Rights Statement:
RoMEO green

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
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 Bird¹

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. The Immobilised CXXC Domain Specifically Retains DNA Containing Clusters of Nonmethylated CpGs
(A) EMSA showing the CXXC complex with a DNA probe containing 27 nonmethylated CpG sites. Nonmethylated probe DNA (CG11) or methylated probe (MeCG11) was incubated with 0, 250, 500, 1,000, or 2,000 ng of recombinant CXXC protein.
(B) A typical elution profile of bulk genomic DNA (blue line) from a CXXC affinity chromatography column. Genomic DNA (100 μg) was applied to the CXXC affinity matrix (see Methods) in low salt (0.1 M NaCl) and eluted with a gradient of increasing NaCl (red line; see text). Eighteen fractions were interrogated by PCR (blue lines). The bracket above indicates fractions that were found to contain nonmethylated CGIs.
(C) Elution of specific CGI sequences of known methylation status. Methylated CGIs (NYESO and MAO in females) coelute with bulk genomic DNA (see bracket) whereas nonmethylated CGIs (P48 and MAO) elute at high NaCl concentration.

doi:10.1371/journal.pbio.0060022.g001
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.search*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 (4,082 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 the estimated number of CGIs in the human genome.
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 MseI 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
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) that allowed a comparison between male and female methylated CGI complements. As expected, these were positively correlated ($R = 0.865$ Pearson correlation) suggesting similar overall patterns. As the library comprises MseI fragments that sometimes overlap minimally with the cognate CpG-rich region, we chose to disregard data from spots that contained DNA with an average CpG frequency (observed/expected) of less than 0.5. Although the omitted fragments often denote CGIs, they include too little of the CpG-rich domain to be reliable for detection of MAP probes. This refinement reduced the number of analysable CGIs on the array to 14,318. To assess the relationship between hybridization signal relative to input and degree of enrichment by MAP, we measured a selection of CGIs in the probe by quantitative PCR and compared this data with the $M$ values ($\log_2$ [MAP signal]/[Input signal]) 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 = $1.213 \times 10^{-7}$) (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\textsuperscript{flaxed}</th>
<th>NCBI\textsuperscript{strict}</th>
<th>CGI</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCBI\textsuperscript{flaxed}</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\textsuperscript{strict}</td>
<td>24,163</td>
<td>24,163 (100%)</td>
<td>24,163</td>
<td>13,568 (56.2%)</td>
</tr>
<tr>
<td>CGI</td>
<td>17,387</td>
<td>15,799 (90%)</td>
<td>13,305 (76.5%)</td>
<td>17,387</td>
</tr>
</tbody>
</table>

\textsuperscript{a}All genes as classified on the ENSMBL genome browser including noncoding RNAs, pseudogenes, VDJ regions, etc.

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

### 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>43.5b</td>
</tr>
<tr>
<td></td>
<td>All genes\textsuperscript{a}</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\textsuperscript{a}</td>
<td>17,387</td>
<td>13,360</td>
<td>76.8</td>
<td>8,611</td>
<td>49.5</td>
</tr>
</tbody>
</table>

\textsuperscript{a}All genes as classified on the ENSMBL genome browser including noncoding RNAs, pseudogenes, VDJ regions, etc.

\textsuperscript{b}This fraction is less than the known fraction of genes with promoter CGIs (56%), because 31% of CGI inserts did not yield DNA sequence.
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 NYESO 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 (log$_2$[MBD/Input]) $\geq 1.5$ (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 disproportionately 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 (4 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 ($\pm$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 pre-mRNA transcript (Q59GD2) previously shown to contain methylated CpG sites [31], is specifically methylated in brain (Figure 5B), I11878 is a 3' CGI of ZNF649 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-Q59GD2 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.
that will be beneficial for studies of CGIs in normal human tissues and in disease settings. By focusing on CGIs alone, we excluded ~98% of the genome from our analysis. While it will ultimately be important to know in detail the methylation status of whole genomes, this currently represents a technical challenge that has been addressed only for the small-genomed plant Arabidopsis [33,34]. These studies used indirect micro-array-based methods for mapping DNA methylation that depend upon probes enriched in methylated domains. Current enrichment methods require clusters of CpG methylation, which are notably absent from the CpG-deficient majority of the mammalian genome. As a result, much bulk genomic DNA is beyond the resolution limit of this approach. Whole genome bisulfite sequencing, the most direct and reliable method for mapping methylated sites, has not yet been attempted in any organism. We therefore decided to study a discrete genomic fraction with evident biological relevance whose methylation status can be interrogated using microarray-based methods.

To isolate nonmethylated duplex CGIs from total genomic DNA, we harnessed the binding specificity of the CXXC protein domain. Extensive sequencing of the resulting library confirmed that CGIs represent a discrete fraction of the human genome with shared DNA sequence characteristics. The present CGI set supercedes a previous human CGI library that was prepared in our laboratory using an indirect affinity purification procedure [20]. The initial library was not comprehensive and appears to have acquired significant levels of non-CGI contamination following amplifications. We estimate that the new library represents ~25,000 CGIs, of which ~60% have been arrayed as full-length single strands on glass slides. Additional analysis of inserts that initially failed conventional sequencing strategies will generate an array that covers the great majority of CGIs that are nonmethylated in human blood. The choice of blood DNA as a starting material necessarily excludes from the set any CGIs that are nonmethylated in germ cells, but densely methylated in the soma [14]. In the future, it will be instructive to compare an exhaustive sequence analysis of this set with comparable sequences isolated by CAP from sperm DNA.

The library prepared using CAP defines CGIs based on the empirical criterion of clustered unmethylated CpGs, whereas criteria based purely on base sequence and composition necessarily ignore methylation status. Comparing our set with predicted CGIs on the NCBI database shows good overlap with predictions based on the “strict” algorithm. The CGI library did, however, identify 23% of CGIs that were negative by this criterion. This suggests that the software for DNA sequence-based CGI identification misses almost one in four CGIs that the more biological criterion of CAP is able to include. Recent CGI analyses identified large numbers of human CGI promoters that are enriched in methylation at lysine 4 of histone H3, a mark of transcriptional activity [14,35,36]. Since it has been proposed that hypomethylation is dependent on germ line and early embryonic transcription [3], we determined the overlap between our CGI set and the H3K4 sites in human embryonic stem cells [37]. We calculate that 90% of CGIs in the filtered set (14,318) coincide with H3K4 methylated promoters that were reported in the chromatin study. A better test of the relationship between CGIs and H3K4 methylation islands in ES cells is to exclude promoters of annotated genes and focus on intra- and intergenic CGIs. Here again, a high proportion (75%) of CGIs overlap with H3K4 methylation islands. These findings are compatible with the notion that the presence of CGIs is connected with specialised chromatin configurations in early embryonic cells. An intriguing proposal is that H3K4 methylation may be incompatible with docking of de novo methyltransferases [38]. This could in theory insure that these regions remain free of CpG methylation at a time when the rest of the embryonic genome is subject to global methylation.

We found that 49% of CGIs overlap the TSS of an annotated gene. In considering the function of the half of CGIs that are remote from an annotated TSS, it is noteworthy that several intragenic CGIs have been shown to coincide with previously unforeseen promoters that initiate bona fide transcripts [39,40]. This raises the possibility that all CGIs function as promoters and are therefore TSS-associated [40]. In this connection, it is of interest that genome-wide analysis by tiling arrays detected over 10,000 unanticipated human transcripts, many of which may represent noncoding RNAs [41]. It is conceivable that many inter- and intragenic CGIs mark promoters that drive the synthesis of these novel transcripts. The noncoding transcripts XIST and AIR, for example, whose RNA products play regulatory roles [42–44], both initiate within CGI promoters. The proximity of many methylated CGIs to developmentally important genes raises

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</td>
<td>149</td>
<td>135</td>
<td>214</td>
<td>247</td>
</tr>
<tr>
<td>Wilcoxon rank test; p-value:</td>
<td>50</td>
<td>35</td>
<td>178</td>
<td>140</td>
</tr>
<tr>
<td>Unclassified methylation</td>
<td>303</td>
<td>237</td>
<td>392</td>
<td>381</td>
</tr>
<tr>
<td>Total</td>
<td>910</td>
<td>815</td>
<td>1,192</td>
<td>1,176</td>
</tr>
<tr>
<td>CGIs</td>
<td>14,318</td>
<td>14,318</td>
<td>14,318</td>
<td>14,318</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 Chrs 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\[o/e\] was 0.75 between CGIs that become methylated and those that remain unmethylated (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>
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 CGIs themselves. These so-called “weak CpG islands” had an average CpG[oc] ratio intermediate between CGIs 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 CGIs 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 composite 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 EcoR1 sites of pet30b (Novagen) to generate a C terminally His-tagged pet30b-MeCP2_76–167. Primers: pet30bMeCP2_76–167Nde1 CGG TTC [C, followed by 44 cycles of 92

8C 2 min. (Figure 1B). Fractions retaining nonmethylated CpG-rich MseI fragments (Figure 1B) were pooled, digested with CA buffer, and re-chromatographed. The relevant fractions were precipitated with PGE/MS-Pr (Pronger).

CGLI 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 glycerol, Triton, EDTA (pH8) buffer plus NaOH and SDS and spun through Millipore Montage filter plates directly into propa-2-o1 to precipitate, followed by elution in water. In all, 172,800 clones were sequenced forward and reverse using T7 and SP6 primers and BigDye V3.1 chemistry, under the following conditions: 300 mM NaCl, 10% glycerol, 0.5 mM PMSF, 10 mM 2-Mercaptoethanol and 0.5% 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 ml of each was precipitated and resuspended in 40 ml x TE buffer. Aliquots were PCR- interrogated using Redhot taq polymerase (Axygen) and XASV (for CACGTGAAAGGGGAGGGGCGTCTGGTAACCATC, revCCACGGGACAGGTACCTC), (for CCGGATCGATGGAGTGACAAAAGCCTGTCG, rev GGTTAGCATGTTGTGTGAGC), (for CCCAGGTGCGTGATACCATC, revCCACGCGGACAGGTACCITT). (for CCGAGGTCTCTGTAACCATC, revCCACGCGGACAGGTACCITT). MGLC (MGLC) (for CCAATCCCGAGACAGGAGACCAGCATGCA, rev CTTCGAATCGAGGCTACACTACTA), P118 (for cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev). DNA was sequenced on a 1.5% TAE-agarose gel (Figure 1B). Fractions retaining nonmethylated CpG-rich MseI fragments (Figure 1B) were pooled, digested with CA buffer, and re-chromatographed. The relevant fractions were precipitated and ligated into the Ndel site of pGEM SmSf (Promega).

CGLI 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 alkali lysis method. Cells were lysed in glycerol, Triton, EDTA (pH8) buffer plus NaOH and SDS and spun through Millipore Montage filter plates directly into propa-2-o1 to precipitate, followed by elution in water. In all, 172,800 clones were sequenced forward and reverse using T7 and SP6 primers and BigDye V3.1 chemistry, under the following conditions: 300 mM NaCl, 10% glycerol, 0.5 mM PMSF, 10 mM 2-Mercaptoethanol and 0.5% 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 ml of each was precipitated and resuspended in 40 ml x TE buffer. Aliquots were PCR- interrogated using Redhot Taq polymerase (Axygen) and XASV (for CACGTGAAAGGGGAGGGGCGTCTGGTAACCATC, revCCACGGGACAGGTACCTC), MGLC (MGLC) (for CCAATCCCGAGACAGGAGACCAGCATGCA, rev CTTCGAATCGAGGCTACACTACTA), P118 (for cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev cagagagatgtcgctgca, rev). DNA was sequenced on a 1.5% TAE-agarose gel (Figure 1B). Fractions retaining nonmethylated CpG-rich MseI fragments (Figure 1B) were pooled, digested with CA buffer, and re-chromatographed. The relevant fractions were precipitated and ligated into the Ndel site of pGEM SmSf (Promega).
take into account mis-annotation of transcription start sites within poorly defined 5' UTRs.

**Microarray fabrication.** Amino-linked clone insert amplicons were generated by vector-specific PCR in 50 mM KCl, 5 mM Tris (pH 8.5), and 2.5 mM MgCl₂ including 1 M Betaine (10 min at 95°C), followed by 35 cycles of 95°C for 1 min, 50°C for 1 min, 72°C for 1 min; and a final extension of 92°C for 10 min; 5°C/10 s until reaching 3°C. The PCR products were ethanol-purified and resuspended in 1× microarray spotting buffer (250 mM sodium phosphate that has been adjusted to 0.1% sarkosyl, 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 denatured (95°C ddH₂O for 2 min), rinsed in ddH₂O, and dried by centrifugation for 5 min at 250×g.

**Microarray hybridization.** Human tissue DNA pooled from three individuals was digested with MseI, phosphatase-treated, and ligated to 5 μmol of phosphorylated catch-linkers (upper GGTT CCA TCC AACG CAG TCT and lower CCC AAGT GGG TGT CCT ATG ggg CATA TG g–3′ reverse primer 5′-CgC CAA GCT ATT TAG tTG AC-5′). PCR products were ethanol-purified and resuspended in 1× microarray spotting buffer (250 mM sodium phosphate that has been adjusted to 0.1% sarkosyl, 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 denatured (95°C ddH₂O for 2 min), rinsed in ddH₂O, and dried by centrifugation for 5 min at 250×g.

**Quantitative PCR.** Real-time PCR was carried out on the bisulfite-treated DNA using RedHot Taq DNA polymerase (Abgene) under the following conditions: 2 min at 94°C, followed by 40 cycles of 94°C for 30 s, 58°C for 30 s, 72°C for 1 min; and a final extension of 50°C for 3 min. PCR fragments were cloned using the Stratataclone PCR cloning system (Stratagene) and at least ten clones are amplified (as above) and sequenced (BigDye Terivdase, Applied Biosystems). Methylation status and experimental quality control was carried out with the aid of BiQAnalyzer [57].

**Supporting Information**

**Dataset S1.** Characterization of Human CGIs Identified as Being Methylated in One or More Somatic Tissues

Included are sequence characteristics (G+C and CpG[oe]) in gene association (Gene ID, ENSEMBL nomenclature), and the methylation profile in the four tissues tested (categories of methylation are colour-coded as for Figure 4B). The CGI island identifier (ID) corresponds to the arrayed CGI fragment. The genomic position of each CGI (Location) corresponds to the chromosomal coordinates derived from NCBI build 38.

Found at doi:10.1371/journal.pbio.0060022.sd001 (373 KB XLS).

**Figure S1.** Sequence Characteristics of CGIs Missed by NCBI-Strict

NCBI-strict relies on base composition to identify CGIs, utilising threshold values for CpG[oe] and G+C density (0.65 and 50%, respectively) as determined from CGIs retained by the CXXC affinity matrix but missed by NCBI-strict. The nonparametric distribution was determined using a Shapiro-Wilk test of normality and subsequent significance was determined using a two-sample Kolmogorov-Smirnov test (NCBI-missed n = 4,082 and all CGIs n = 13,305).

Found at doi:10.1371/journal.pbio.0060022.sg001 (48 KB DOC).

**Figure S2.** Sequence Properties of Methylated CGIs

Brock 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 of methylated 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 [58]) 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 EMBOSS 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).

Found at doi:10.1371/journal.pbio.0060022.sg002 (66 KB DOC).

**Table S1.** Gene Association: CpG Islands Missed by NCBI-Strict

All CGIs (n = 1,082) retained by the CXXC affinity matrix but not predicted by NCBI-strict were mapped relative to protein-coding genes. Gene overlap indicates the spatial association of CGIs relative to protein-coding genes.

Found at doi:10.1371/journal.pbio.0060022.sd001 (27 KB DOC).

**Table S2.** Methylated CGIs on Chromosome 16 and Chromosome X in Human Whole Blood DNA

CpG arrays hybridised with MBD probe from male and female human heparin DNA identify extensive methylation of X-linked CGIs. 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.

Found at doi:10.1371/journal.pbio.0060022.sd002 (30 KB DOC).

**Table S3.** Developmental Gene Categories Are Associated with Differentially Methylated CpG Islands

Ontology terms for gene-associated CGIs were compared with those for differentially methylated CGI-genes. Genes involved in developmental processes such as neurogenesis and segmentation are
significantly enriched and include transcriptional regulators such as homeobox genes. Significantly enriched biological processes and molecular functions were determined using the Web-based Panther classification system (http://www.pantherdb.org) [66].

Found at doi:10.1371/journal.pbio.0060092.s004 (31 KB DOC).

Table S4. Quantitative PCR Primers for Microarray Validation

Acknowledgments

We are grateful to Anne Evans and Alieen Greig for excellent technical support; to Jose de las Heras, Robert Andrews, lan Still, Michael Quail, and Carol Scott for help and advice; and to Miho Suzuki, Heather Owen, and Pete Skene for comments on the manuscript. We thank the Wellcome Trust Sanger Institute Microarray Facility and Mapping Core Group.

Author contributions. AB, RI, and HJ conceived and designed the experiments. RI and DD performed the experiments. RI, AK, JS, DJ, and TC analyzed the data. HJ contributed reagents/materials/analysis tools. PE, SH, and CL designed and printed the CGI microarrays. CC, RP, and JR performed DNA sequencing of the CGI library. RI and AB wrote the paper.

Funding. This work was supported by the Wellcome Trust and by a Medical Research Council (UK) studentship to RI.

Competing interests. The authors have declared that no competing interests exist.

References


