Nucleic Acids Research, 1999, Vol. 27, No. 10 **2099–2107**

Isolation of CpG islands from large genomic clones

Sally H. Cross*, Victoria H. Clark and Adrian P. Bird

Institute of Cell and Molecular Biology, University of Edinburgh, Darwin Building, King's Buildings, Mayfield Road, Edinburgh, EH9 3JR, UK

Received February 10, 1999; Revised and Accepted March 26, 1999

DDBJ/EMBL/GenBank accession nos AJ132338–AJ132341

ABSTRACT

Positional cloning is a powerful method for the identification of genes. Using genetic and physical mapping methods the genomic region within which a particular gene is located can relatively easily be narrowed down to a comparatively small area contained within cosmid, PAC or BAC clones. It is then a matter of identifying genes within these clones. Here we describe the application of a technique, which has been successfully used for the bulk purification of CpG islands from whole genomes, to the isolation of CpG island sequences from such clones. As CpG islands overlap transcription units they can be used to isolate full-length cDNAs for associated genes, either by probing cDNA libraries or by searching databases. CpG islands are linked with ~60% of human genes and because their isolation is independent of the expression profile of these genes this approach would complement other expression-based methods of gene identification. By applying this technique to a cosmid clone known to contain the *PAX6* gene we successfully isolated the CpG island for this gene along with other CpG island-like sequences. Closer examination revealed that an extensive genomic region around the 5′-end of *PAX6* is unusual with regard to methylation and GC content. CpG island sequences were also successfully isolated from a PAC clone carrying the *MBD1* gene. These included the complete CpG island containing the first exon and regulatory sequences from *MBD1*.

INTRODUCTION

In the human genome there are estimated to be 45 000 CpG islands (CGIs) which co-localise with the 5′-ends of ~60% of human genes (1). CGIs are distinctive patches of genomic DNA which are GC-rich and do not exhibit suppression of the dinucleotide CpG. They are unmethylated, regardless of the activity status of the associated gene, with the exception of CGIs on the inactive X chromosome and those associated with some imprinted genes. CGIs are found dispersed throughout otherwise heavily methylated, comparatively GC-poor and CpG-suppressed vertebrate genomes and are, on average, between 0.5 and 2 kb in size. In humans they account for between 1 and 2% of the genome (reviewed in 2).

Largely intact CGIs have been purified from total genomic DNA using a method which takes advantage of their unusual base composition and methylation status in combination with a technique by which DNA is separated according to its level of methylation (3). To date total genome CGI libraries have been prepared for human (3), chicken (4), mouse (5) and pig (6) and these libraries have been used to examine the gross distribution of CGIs in these genomes. In every case the distribution of CGIs has been found to be non-random, such that CGIs are concentrated in particular regions. This is most extreme in the chicken genome, where CGIs appear to be clustered on the microchromosomes and to be relatively scarce on the macrochromosomes (4). In the human genome, the distribution of CGIs closely parallels that of R bands (7). Moreover, CGIs have been shown to co-localise with early replicating, highly acetylated genomic regions and it is thought that areas rich in CpG islands are also generally gene dense (8,9).

The human genome can be thought of as containing two kinds of domain which differ in the frequency with which the overall GC-poor, CpG-depleted DNA is interspersed with CGIs. Restriction maps of two contrasting genomic regions of 85 kb from human chromosomes 18 and 19, which illustrate these two kinds of domain, are shown in Figure 1. The clustering of sites for restriction enzymes which recognise GC-rich sequences, infrequent in the rest of the genome, indicate the location of CGIs in these sequences. The chromosome 18 sequence is an example of a region in which CGIs are comparatively rare (Fig. 1A). The single CGI present is associated with the *ST8SIA V* gene (10). In contrast, the region from chromosome 19 is very different and contains up to six CGIs (Fig. 1B, boxes I–VI). The 5′-end of *EAAT4* (11), the only gene mapped within this region, is found within box II. Part of box IV matches clone cpg83b2 from the total CGI library (3; Sanger Centre CpG island sequence database, http://www.sanger.ac.uk).

In a positional cloning project the task is usually to detect genes within clones containing between 35 kb (cosmids) and 300 kb (BACs) of genomic DNA, i.e. from sequences of a similar size order to that shown in Figure 1. Several different methods, such as exon trapping (12) and direct cDNA selection (13,14), have been devised to allow the detection of putative gene sequences. Each has been successfully used, although clones isolated have to be carefully analysed as both methods are prone to generate false positives. Some gene sequences may be missed, either because the exons are too small to be trapped or, for genes with restricted

*To whom correspondence should be addressed at present address: MRC Human Genetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK. Tel: +44 131 332 2471; Fax: +44 131 343 2620; Email: sally.cross@hgu.mrc.ac.uk

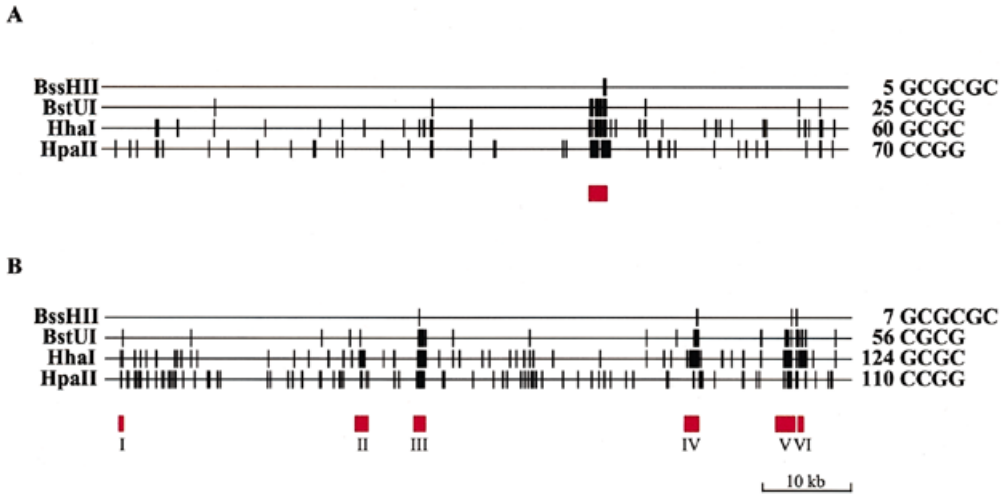

Figure 1. Diagram showing the positions of CGIs in contrasting chromosomal domains. Positions of sites for the rare-cutting restriction enzymes *Bss*HII, *Bst*UI, *Hha*I and *Hpa*II within 85 kb of sequence from each domain are indicated by vertical lines. The name of each restriction enzyme is shown to the left and the number of sites present and the recognition sequence for each are shown to the right. Red boxes underneath each sequence show where potential CGIs are located (labelled I–VI in B). (**A**) Sequence from chromosome 18 (AC003971). (**B**) Sequence from chromosome 19 (AC004659).

patterns of expression, they would only be isolated in a cDNA selection experiment if the appropriate cDNA was used. The method we describe here is an alternative, complementary approach aimed at isolating largely intact CGIs by exploiting their sequence characteristics. It depends on the differential affinity of DNA fragments containing different numbers of methyl-CpGs for a methyl-CpG binding domain (MBD) column (3). Other techniques aimed at isolating such sequences have been developed. The first is a PCR-based method which is dependent on an Alu sequence being present close to a CGI (15). This has been useful for YAC clones as the Alu-specific primer excludes products being generated from contaminating yeast DNA, however it does depend on the presence of an Alu sequence close to a CGI. In the second method, fragments from CGIs are selected by the retention of partly melted DNA fragments in a denaturing gradient gel (16). In both these methods only parts of CGIs would be recovered. Recently a method by which genomic clones can be scanned for the presence of CGIs has been reported (17).

If the sequence of a genomic clone is available it is easy to see where CGIs, and therefore the 5′-end of potential candidate genes, are located (Fig. 1). In the absence of sequence information the presence of CGIs can be detected by restriction analysis with rare-cutting enzymes, although these enzymes are not useful for CGI isolation because by their use CGIs are fragmented. The dinucleotide CpG occurs, on average, every 100 bp outside of CGIs and is usually methylated at such sites. Within CGIs, CpG occurs once every 10 bp and is unmethylated (Fig. 1). On cloning, methylation is erased from genomic sequences but the difference in number of CpGs between CGIs and other sequences is unaltered, which enables CGIs from cloned DNA to be purified using the MBD column in a way similar to that used to purify CGIs from genomic DNA. First, genomic clones are treated with the restriction enzyme *Mse*I, which fragments bulk genomic DNA but leaves CGIs largely intact. *Mse*I fragments derived from CGIs are then selected on the MBD column. The number of methylated CpGs present on an *Mse*I fragment determines the strength of binding to the MBD column (3); large *Mse*I fragments outside CGIs do not bind but small fragments from CGIs, which contain many CpGs, do bind. It should be noted that if, as an alternative way of isolating CGI *Mse*I fragments, selection was based solely on large size, many of the selected fragments would not be from CGIs and small CGI fragments would be missed. For example, in the chromosome 18 region (Fig. 1A) there are 40 *Mse*I fragments of $\geq$0.5 kb, of which only two are from the only CGI present. This CGI contains, in addition, two smaller *Mse*I fragments. In the chromosome 19 region (Fig. 1B) there are 50 *Mse*I fragments of $\geq$0.5 kb. Four of these are derived from four of the CGIs present, but the other two CGIs are represented by *Mse*I fragments <0.5 kb. Once potential CGI *Mse*I fragments have been isolated it can easily be established if they are genuine CGIs by testing their methylation status in the genome. Then, because CGIs are usually single copy, they can be used to isolate the associated full-length transcripts, either by screening cDNA libraries or by searching databases. In this paper we show that it is possible to successfully isolate CGIs from both a cosmid and a PAC clone.

MATERIALS AND METHODS

Sequence names and accession numbers

The sequences of the clones P1–P6 reported here are available under accession nos AJ132338–AJ132341.

Isolation of CGIs from genomic clones

The MBD column was prepared and tested essentially as described (3,18). Briefly, 30 mg of histidine-tagged methyl-CpG binding domain protein, purified from crude bacterial extracts, was coupled to 1 ml of Ni^{2+}–NTA–agarose (Qiagen) and packed into a HR 5/5 column (Pharmacia). DNAs were loaded onto, washed and eluted from the column in 20 mM HEPES (pH 7.9), 10% glycerol, 0.1% Triton X-100, 0.5 mM PMSF and NaCl concentrations varying between 0.5 and 1 M. Cosmid FAT5 DNA was prepared by standard alkaline lysis followed by equilibrium

centrifugation in a CsCl/EtBr gradient. PAC 286-e7 DNA was prepared using Qiagen columns according to the manufacturer's directions. DNAs were digested to completion with *Mse*I and then methylated at all CpGs using CpG methylase (NEB). Methylated *Mse*I-digested FAT5 DNA (20 µg) was loaded twice onto the MBD column at 0.5 M NaCl, washed first with 0.5 M buffer up to a volume of 10 ml to elute unbound fragments, then with a 30 ml gradient of 0.5–1 M NaCl to elute bound fragments and finally with 8 ml of 1 M NaCl buffer. Fractions of 2 ml were collected. DNA from 1/5 of each fraction was precipitated, separated on a 1.2% agarose gel and transferred to Hybond-N^+ (Amersham). By probing this filter with a 2.2 kb *Pst*I fragment (positions 13 883–16 107 in FAT5) overlapping the *PAX6* CGI, fractions expected to contain CGI fragments were identified. For a second round of purification these fractions, eluting between 0.8 and 0.9 M NaCl, were pooled, diluted back to 0.5 M NaCl, loaded twice onto the MBD column, washed with 6 ml 0.5 M buffer and subsequently treated as for the first round of purification. A third round of purification was performed as for the second round and DNA was precipitated from fractions eluting between 0.8 and 0.9 M NaCl expected to contain CGIs. Fragments were ligated into the *Nde*I sites of pGEM-5Zf(–) (Promega) or pBS-ANA, which has *Afl*II sites flanking the *Nde*I site (a gift of Dr W. Rideout III) and transformed by electroporation into the SURE bacterial strain (Stratagene). The purification and cloning scheme used for PAC 286-e7 was the same with the following changes. Methylated *Mse*I-digested DNA (45 µg) was loaded onto the column for the first round of purification and fractions eluting between 0.74 and 1 M NaCl were selected for the second and third rounds.

PCR

Inserts cloned into pBS-ANA were amplified using primers 5′-CGATAAGCTTGATCTTAAGC-3′ and 5′-GCAGGAATTC-GATCTTAAGC-3′ that flank the *Nde*I cloning site. Otherwise inserts were amplified as described (3) except that 1 U Prozyme (Bioline) with the supplied buffer was used and 5% DMSO was included in the reaction mix for some inserts. Reactions were heated to 95°C for 3 min followed by 30 cycles of 95°C for 1 min, 55°C for 1 min, 72°C for 3 min and a final extension of 72°C for 10 min. For the PCR reactions to determine if any of the fragments cloned from 286-e7 were adjacent the following primers were used for each clone: P1, 5′-GAATTCCGTACGTT-CCTAGGC-3′ and 5′-GTGGTGAGCCATAACCGGAG-3′; P2, 5′-TCTGTTTCTCCGGTTCTCCC-3′ and 5′-TCACAGAAGA-GTCGTGTGGC-3′; P3, 5′-GTGTCACCACACTGAAGGCG-3′ and 5′- CTGTCGTTGAACGTCAGCAC-3′; P4, 5′-AGAGCC-AGACCCTGTCTCAA-3′ and 5′-CATGGGGACTCTAATGG-CAG-3′; P5, 5′-GCAAGACCCTGAGATTTTTCC-3′ and 5′-TC-AGCCCGAAAAGTGGAGAC-3′; P6, 5′-ACTCTAGGCCCG-TGGACC-3′. Using 200 ng of 286-e7 as a template the reactions were carried out as for the clone inserts using the following protocol: initial incubation at 95°C for 3 min followed by 25 cycles of 95°C for 30 s, 58°C for 30 s, 72°C for 2 min and a final extension of 72°C for 10 min. In all cases parallel reactions without DNA template were performed as controls.

Sequence and database analysis

Plasmid DNAs were prepared using Qiagen columns. Sequencing was performed on a Perkin Elmer-Applied Biosystems 373 Stretch DNA sequencer. The FAT5 cosmid was end-sequenced using SP6 and T7 primers. Sequences were analysed using Gene Jockey (Biosoft), the GCG Wisconsin Package v.9.1 (Genetics Computer Group, Madison, WI), BLAST (19), the Lasergene DNA analysis software package (DNASTAR, Madison, WI) and custom written programs. Database searches were carried out at NCBI via Email (http://www.ncbi.nlm.nih.gov/). To establish if the chromosomal location and gene identity of sequences was known Unigene (http://www.ncbi.nlm.nih.gov/UniGene/) was searched with the accession numbers of database matches. Restriction digests, Southern blotting and hybridisations were carried out using standard protocols. All Southern blots were washed at high stringency in 0.2× SSC, 0.1% SDS at 65°C.

RESULTS

Isolation of CGIs from a cosmid clone

First we wished to determine if CGIs could be efficiently isolated from cloned genomic DNA within a cosmid. We chose to use a cosmid containing a known gene with a CGI so that we could track this CGI during the purification procedure. Cosmid FAT5 (20) contains the *PAX6* gene from human chromosome 11. Mutations of *PAX6* result in aniridia and other eye development disorders (21,22). The sequence of FAT5 is contained in two overlapping entries in the database (accession nos Z95332 and Z83307). Positions 1–20 770 correspond to positions 1–20 770 of Z95332 and positions 20 771–36 111 to positions 1–15 341 of Z83307. Figure 2A shows a plot of the distribution of CpGs and selected restriction sites in FAT5 with the location of exons from *PAX6* shown underneath.

To isolate potential CGIs from FAT5, methylated, *Mse*I-digested cosmid DNA was passed over a MBD column as described in Materials and Methods. DNA fragments which bound tightly were selected and cloned. Thirty-two clones were picked and the inserts analysed by restriction digestion and sequencing which showed that some were identical. In total six different fragments, named C1–C6, were isolated. Each fragment possessed the typical sequence characteristics of CGI DNA, i.e. a high GC content and close to the expected number of CpGs (Table 1). To test if the fragments were derived from unmethylated regions of the genome, as expected for CGIs, they were used to probe Southern blots of human blood DNA digested with *Mse*I, alone or in combination with *Msp*I (methylation insensitive), *Hpa*II or *Bst*UI (methylation sensitive). In each case the genomic DNA was found to be unmethylated (Table 1 and data not shown). In addition, this experiment showed that the probes were single copy, except for C2 which was slightly repetitive. The locations of C1–C6 within FAT5 were determined by sequence comparison and are shown in Figure 2A. Clones C2 and C3 are derived from the *PAX6* CGI and the other fragments are found elsewhere within the cosmid. C4 overlaps exon 4 of *PAX6*. No matches with any other gene sequences were found on searching the databases with the sequences of the other clones. However, C5 is 600 bp proximal to a proposed neuroretina-specific enhancer in intron 4 of *PAX6* (23) and C6 lies in a non-coding region which has a high degree of conservation between man and the distantly related vertebrate *Fugu rubripes* (24).

Inspection of the sequence of FAT5 revealed that it is atypical (Fig. 2A and B). Generally human DNA has a GC content of ~40% and the occurrence of the dinucleotide CpG is suppressed to ~25% of the number expected from base composition (2). In contrast to this, the majority of the DNA within FAT5 is more like

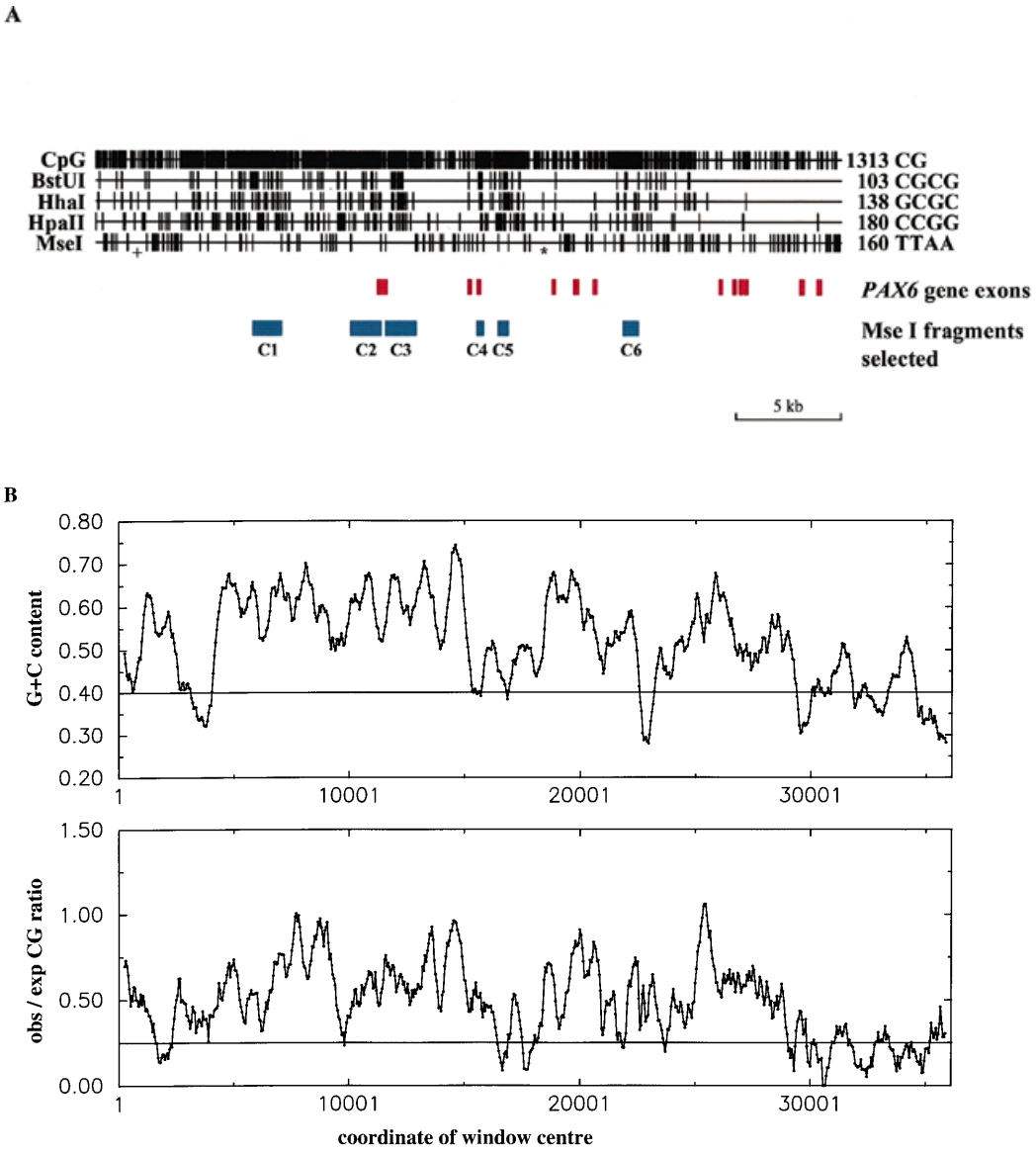

Figure 2. Diagram showing the distribution of CpGs and the positions of *PAX6* exons and cloned *Mse*I fragments in FAT5. (**A**) Positions of CpGs and sites for the restriction enzymes *Bst*UI, *Hha*I, *Hpa*II and *Mse*I are indicated by vertical lines. The name of each restriction enzyme is shown to the left and the number of sites present and the recognition sequence for each are shown to the right. The *Mse*I fragments denoted by a + and an * are referred to in the text. Shown underneath the restriction plot are the positions of *PAX6* gene exons (red boxes) and the cloned *Mse*I fragments (blue boxes, labelled C1–C6). (**B**) Per cent G+C content and CG observed/expected values plotted across the FAT5 sequence in steps of 50 bp with a window size of 500 bp. The horizontal lines in each plot show the average value for non-CGI human DNA (%G+C content = 0.40 and CG observed/expected value = 0.25).

CGI DNA, i.e. a higher GC content and closer to the expected number of CpGs (Fig. 2A and B). It is quite striking that only six *Mse*I fragments are in the cloned set of fragments prepared from the fraction which bound tightly to the MBD column, demonstrating the power of the MBD column in discriminating between genuine CGI-like fragments and other GC-rich DNA fragments. Where did the largest *Mse*I fragment in FAT5, which was not present in the clone collection, elute from the column? To determine this the fragment, denoted by an * in Figure 2A, was gel purified and used to probe a Southern blot of DNA isolated from the fractions collected during the first round of purification (see Materials and Methods). This showed that it eluted principally between 0.67 and 0.74 M NaCl, before the fractions selected for further purification (data not shown). The size of this *Mse*I fragment was 1497 bp and it had a %G+C content of 53.5% and a CpG observed/expected value of 0.42. Therefore, whilst being fairly GC-rich, the number of CpGs present is diminished, although not to the extent generally seen with bulk DNA. It was also found to be unmethylated in blood DNA by using the same method as that used to test the methylation status of clones C1–C6 (see above; data not shown). Recently exon 5 of the *PAX6* gene, which lies within this *Mse*I fragment, has been shown to be unmethylated in a wide variety of normal somatic tissues but to be hypermethylated in many tumours (25).

Table 1. Summary of *Mse*I fragments purified from cosmid FAT5 and PAC 286-e7

Name[a]	Size (bp)	%G+C[b]	CpG O/E[c]	Number[d]	Methylated?[e]	Comments
C1	1430	62.4	0.84	9	U	
C2	1469	62.0	0.67	12	U	*PAX6* CGI
C3	1485	58.5	0.80	1	U	*PAX6* CGI
C4	309	67.6	0.79	6	U	Overlaps *PAX6* exon 4
C5	492	65.2	0.82	3	U	
C6	712	64.6	0.67	1	U	
P1	832	59.9	1.03	3	U	EST matches (AA325016, AA349398)
P2	1102	59.7	0.58	1	M	Adjacent to *MBD1* CGI
P3	757	67.0	0.73	2	U	*MBD1* CGI
P4	1248	59.2	0.47	1	P	Alu repeat[f]
P5	525	53.2	0.70	1	U	
P6	1088	56.3	0.51	1	U	*MBD1* CGI

[a]C1–C6 were isolated from FAT5 and P1–P6 were isolated from 286-e7.
[b]%G+C content of the clone.
[c]Number of CpGs observed/expected.
[d]Number of clones of those analysed which contained this *Mse*I fragment.
[e]Methylation status of the corresponding genomic fragment in blood DNA. U, unmethylated; M, methylated; P, partially methylated.
[f]Alu repeat contained within the first 400 bp of the clone.

Most of the sequence in FAT5, essentially the first ~30 kb, is CGI-like in sequence composition and only a small part, approximately the final 6 kb, resembles bulk genomic DNA (Fig. 2A and B). To find such a high incidence of CGI-like sequence contiguous over a large genomic region of ~30 kb is very uncommon (contrast with the sequence composition of the genomic segments shown in Fig. 1). As another feature of CGIs is that they are unmethylated, regardless of the transcriptional state of their associated genes, we tested the genomic methylation status across the cosmid to determine if any regions lacked methylation in a similar fashion. Sequences located at the right-hand end of FAT5 were shown to be methylated in blood DNA by probing a Southern blot of DNA cleaved with *Eco*RI, alone or in combination with *Msp*I (methylation insensitive), *Hpa*II or *Hha*I (methylation sensitive), with a 3 kb *Sca*I–*Eco*RI fragment (positions 31 032–33 990) (data not shown). This was expected as this part of FAT5 has a typical bulk DNA-like sequence composition (Fig. 2A and B). The same Southern blot was stripped and probed with a 2 kb *Sau*3A fragment (positions 202–2163) which showed that a large 8–9 kb *Eco*RI fragment overlapping the left-hand end of FAT5 (first *Eco*RI site in FAT5 at position 3187) was unmethylated in blood DNA (data not shown). However, although the first 4 kb of sequence was fairly GC-rich (47.0%) it did exhibit some reduction in the number of CpGs present (CpG observed/expected value of 0.42), suggesting that C→T transitions resulting from the spontaneous deamination of 5-methylcytosine had taken place within this segment. As the consequence of such mutations can only become fixed if they occur in the germline, we compared the methylation status of the

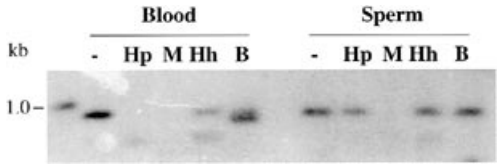

Figure 3. Sequences from the left-hand end of FAT5 are unmethylated in blood DNA but methylated in sperm DNA. Blood and sperm DNAs from the same human male digested with *Mse*I alone (–) or in combination with *Hpa*II (Hp), *Msp*I (M), *Hha*I (Hh) or *Bst*UI (B) were separated on a 1.2% agarose gel, transferred to Hybond-N$^+$ (Amersham) and probed with an *Mse*I–*Bst*UI fragment from FAT5 (positions 1541–2515) which detects the *Mse*I fragment denoted with a + in Figure 2A. Sizes are in kb.

*Mse*I fragment denoted by a + in Figure 2A in blood and sperm DNA, which showed that whilst it was largely unmethylated in blood DNA it was methylated in sperm (Fig. 3). It is likely that the *Mse*I fragment denoted by an * in Figure 2A is also methylated in sperm as it exhibits a similar suppression of CpG although unmethylated in blood DNA (see above). Evidence that other regions of FAT5 which are also CpG-suppressed are likely to be methylated in the germline comes from mutational analysis that has been carried out for the *PAX6* gene (22; I.Hanson, personal communication). In a database of mutations found in the *PAX6* gene (26) all the mutations caused by CpG→TpG transitions, and therefore likely to have resulted from deamination of 5-methylcytosine, are found within regions of comparatively low GC content (≤50%) which exhibit CpG suppression (Table 2).

Table 2. *PAX6* mutations resulting from CpG to TpG or CpA transitions

Position in FAT5	Position in PAX6	Codon no.	Domain[a]	Mutation[b]	Sequence change	%G+C[c]	CpG O/E[c]	No. found[d]
22269	exon 5	44	PD	R→ter	CGA→TGA	46.5	0.48	1
23373	exon 6	103	PD	R→ter	CGA→TGA	42.6	0.41	2
24152	exon 7	128	PD	R→C	CGC→TGC	50.2	0.42	1
30279	exon 8	203	HD	R→ter	CGA→TGA	38.7	0.21	5
30294	exon 8	208	HD	R→W	CGG→CAG	38.7	0.21	1
				R→Q	CGG→TGG			1
30905	exon 9	240	HD	R→ter	CGA→TGA	43.4	0.22	12
31197	exon 10	261	HD	R→ter	CGA→TGA	43.4	0.22	1
31463	exon 11	317	PST	R→ter	CGA→TGA	43.3	0.20	9
23423	intron 6			splice	CG→CA	42.1	0.42	4

Note: The mutation information shown was extracted from the *PAX6* mutation database (http://www.hgu.mrc.ac.uk/Softdata/PAX6/).
[a]Domain of protein in which mutation occurs. PD, paired domain; HD, homeodomain; PST, Pro/Ser/Thr-rich domain.
[b]Mutational change. The single amino acid code is used. ter, stop codon.
[c]%G+C content and CpG observed/expected values of a 2001 bp window centred on the CpG mutated.
[d]Number of independent occurrences of this mutational change listed in the database.

Isolation of CGIs from a PAC clone

The above experiments demonstrate that it is possible to isolate CGIs efficiently from a cosmid clone. To determine if CGI isolation was similarly successful with larger genomic clones we applied the method to PAC clone 286-e7 from the RPCI1 PAC library obtained from the UK HGMP Resource Centre (27). Clone 286-e7 was selected because it contained the *MBD1* gene which encodes a methyl-CpG binding protein mapping to chromosome 18 (28; B.Hendrich *et al*., submitted for publication). Although the available *MBD1* cDNAs did not extend to the 5′-end of the gene, the first 250 bases of available sequence were CGI-like in that they were both GC-rich and contained over half the expected number of CpGs. *MBD1* would be expected to have a CGI as it has a housekeeping pattern of expression (28). Consistent with this, on probing Southern blots of genomic DNA digests with a *Not*I–*Bam*HI fragment containing bases 1–264 of the longest available cDNA (accession no. Y10746), two genomic *Mse*I fragments were detected, the larger of which was unmethylated (data not shown). This suggests that the cDNA fragment contains portions of the first and second exons of the gene, the first exon being part of a CGI.

In order to isolate this and any other CGIs from 286-e7, *Mse*I-digested, methylated DNA was passed over the MBD column as described in Materials and Methods. Figure 4A shows a Southern blot of fractions from the first round of purification probed with the 5′-end of the *MBD1* gene cDNA. It can be seen that the larger fragment, likely to be part of the CGI, is bound on the column and elutes at high salt in fractions 11–14 (0.67–0.8 M NaCl), whereas the smaller *Mse*I fragment elutes at low salt. To determine what proportion of *Mse*I fragments from this PAC are tightly bound by the MBD column the load DNA and bound fractions from the third round of purification were compared. Figure 4B shows that only a small proportion of fragments are retained on the column. These are potential CGIs along with vector fragments. Of the vector pCYPAC2, 17 kb is present in the recombinant, which contains four *Mse*I fragments >500 bp which would be expected to bind to the MBD column as they are CGI-like in nature.

DNA fragments present in these fractions, 12–14 from the third round of purification, were cloned and clones with inserts >500 bp analysed. Sequencing of these 18 clones showed that 10 were *Escherichia coli* fragments. The remaining eight represented five different human fragments, named P1–P5 (Table 1). Sequence comparison with the *MBD1* cDNA showed that none of these contained the 5′-end of the *MBD1* cDNA. It was found that although the *Mse*I fragment which did was present in the selected fractions its peak of elution was earlier, with the result that it might be expected to be under-represented in the clone collection (Fig. 2A and data not shown). To isolate this fragment 96 clones were screened and one which was positive was found (clone P6). All the clones possessed the typical sequence characteristics of CGI DNA, i.e. a high GC content and close to the expected number of CpGs (Table 1). The fragments were shown to be derived from largely unmethylated regions of the genome, as described for the clones isolated from the cosmid, with the exception of P2, which was methylated (Table 1 and data not shown). This analysis also showed that all were single copy sequences except for part of P4, which contained an Alu repeat (a single copy 860 bp *Tsp*509I fragment from P4 was used for the methylation analysis). By using P2 as a probe it was shown that the corresponding genomic *Mse*I fragment binds tightly to the MBD column and so, whilst not being a genuine CGI fragment, it would be expected to be isolated in this experiment (data not shown).

Database searching revealed that P1 had matches with two expressed sequence tags (ESTs) which overlap (Table 1). P1 contains 414 bp of sequence upstream of the most 5′ EST (AA325016), providing additional information about transcribed and promoter sequences of this gene. The homology with both ESTs ends at a putative splice site junction (position 647 in P1). The ESTs are part of an EST clone contig which has weak similarity to human transcription factor *TFIIS* (entry number

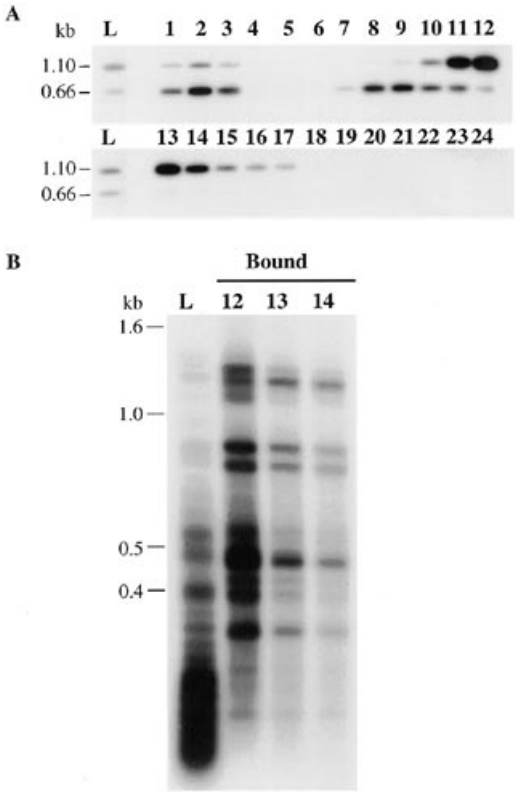

Figure 4. Purification of CGIs from PAC clone 286-e7 using a MBD column. (**A**) Aliquots of the load DNA (L) and fractions 1–24 collected after a first round of purification of methylated *Mse*I-digested 286-e7 DNA on the MBD column were separated on a 1.5% agarose gel, transferred to Hybond-N^+ (Amersham) and probed with the *Not*I–*Bam*HI fragment, bases 1–264, of the MBD1 cDNA. (**B**) Aliquots of DNA from the load (L) and the bound fractions eluting between 0.8 and 0.9 M NaCl (12–14) in the third round of purification of methylated *Mse*I-digested 286-e7 DNA using the MBD column were end-labelled, separated on a 1.5% agarose gel and the dried gel autoradiographed. Sizes are in kb.

Hs.9571 in Unigene) and, interestingly, the protein encoded by this gene potentially contains a cysteine-rich region which is highly related to three such regions found in *MBD1* (28,29). There is no other homology between the two genes. The ESTs forming the contig are derived from a wide range of cDNA sources, suggesting that the gene is ubiquitously expressed. It therefore would be expected to have a CGI. The contig has been mapped to chromosome 18 between D18S472 and D18S835, in agreement with the map position found for *MBD1*, to which it must be very closely linked (B.Hendrich *et al*., submitted for publication).

No significant matches to any other sequences, apart from *MBD1* in the case of P6, were found, but sequence comparison with the mouse *Mbd1* gene promoter (B.Hendrich, submitted for publication) suggested strongly that clone P3 contained part of the *MBD1* CGI (Table 1). To determine if P3 and P6, both part of the *MBD1* CGI, and any of the other *Mse*I fragments were adjacent to each other PCR was carried out using primers from the ends of the clones in pairwise combinations on 286-e7 DNA and PCR products of the expected predicted size were sequenced to confirm the join. This analysis revealed that clones P2, P3 and P6 were adjacent and formed a large CGI covering the 5′-end and promoter region of the *MBD1* gene (Fig. 5). By comparing the sequence of this with that of the *MBD1* cDNA and the mouse *Mbd1* gene CGI (B.Hendrich *et al*., submitted for publication) the position of the first exon was deduced (Fig. 5). Furthermore, this result suggested that the other three clones were derived from independent CGIs found on this PAC. This conclusion was supported by hybridisation analysis which found that on probing Southern blots of 286-e7 DNA digests with these clones different restriction digest patterns were detected (data not shown). In conclusion, by using the MBD column the entire *MBD1* CGI was isolated from PAC 286-e7 along with three other CGIs. One of these is the CGI associated with a gene previously identified as an EST clone contig and the other two are potentially part of two more CGIs found in the genomic region contained in this clone.

DISCUSSION

In this paper we have shown that the method we originally developed for the isolation of CpG island sequences from genomic DNAs can successfully be applied to large genomic clones. From the cosmid FAT5 six fragments were repeatedly isolated and, although no genes apart from *PAX6* were found associated with these, they all possessed the canonical features of CGIs. Three fragments overlapped the *PAX6* transcript, C2 and C3 were part of the CGI and C4 overlapped exon 4. Sequence and methylation analysis of the genomic DNA cloned in FAT5 revealed that an unusually large expansive CGI-like domain of ~30 kb is present which includes all the upstream sequences of *PAX6* found in FAT5 and extends to include most of intron 7 of the gene. The cloned fragments C4 and C5 lie in intron 4 close to the site of a proposed neuroretina-specific enhancer (23). C6 lies in intron 7 and is part of a CGI which has been suggested to be an alternative start site of transcription for *PAX6* which would result in an isoform lacking a paired domain (24). Alternatively, the CGI of which C6 is a part could denote the start of a different gene because an EST clone contig, which has a *Fugu* homolog (24), is found 6 kb 3′ to *PAX6*. It is striking that when the human and *Fugu* sequences between *WT1* and *PAX6* were compared the only significant regions of homology without coding sequences were found around the *PAX6* locus. Of these the most notable were two clusters found >5 kb 5′ of exon 1 and in introns 4 and 7 (24). In human all these sequences are CGI-like and fragments from these clusters were selected (clones C1, C4, C5 and C6). This degree of conservation in the sequence of these non-coding regions between two such distantly related species implies that they are of regulatory significance and this could account for the unusually large CGI-like domain found associated with *PAX6*.

Whilst unmethylated in somatic DNA, it is likely that the parts of the extended CGI-like domain not bound by the MBD column are methylated in the germline. Figure 3 shows that this is the case for the *Mse*I fragment denoted by a + in Figure 2A. Suppression of CpGs, consistent with germline methylation, is also observed, although it is interesting to note that over much of the region the suppression seen is not as marked as is usually found in bulk DNA (Fig. 2A and B and Table 2). The high incidence of mutations in the *PAX6* gene probably caused by deamination of 5-methylcytosine to thymidine also supports the argument that much of this region is methylated in the germline. Of the 149 mutations recorded in the *PAX6* mutation database, 37 (25%) are CpG→TpG transitions and these are found at nine positions in the *PAX6* gene (Table 2). All of these are probably loss of function mutations. Six are mutations leading to premature termination of translation and, of

Figure 5. The structure of the *MBD1* CGI assembled from clones P2, P3 and P6. P2, P3 and P6 are the first, second and third *Mse*I fragments, respectively. Positions of CpGs, GpCs and sites for the restriction enzymes *Bst*UI, *Hha*I, *Hpa*II, *Mse*I and *Not*I are indicated by vertical lines. The name of each restriction enzyme is shown to the left and the number of sites present and the recognition sequence for each are shown to the right. Shown underneath the restriction plot is the position of the first exon of the *MBD1* gene (black box) and the line with the arrowhead indicates the direction of transcription.

the three missense mutations, the two found at codon 208 result in the change of an amino acid found immediately before the homeodomain, which is invariant in all known pax proteins with an intact homeodomain (30). At the other 36 CpGs in the coding sequence of *PAX6* a C→T transition would result in a missense mutation. As missense mutations have only rarely been documented for *PAX6* (22) it is not surprising that mutations have not been identified involving these CpGs in spite of their being potentially as prone to mutation as those at which mutations have been found.

The size of the PAC clone 286-e7 is not known, but it is probably between 100 and 150 kb, as the average insert size of clones from the RPCI1 library is 110 kb (27). Six CGI-like fragments were isolated from this PAC, some more than once, and no non-CGI-like fragments were isolated (Table 1). Three of the fragments form a large CGI associated with the *MBD1* gene. Clone P1 is from a CGI associated with a gene of unknown function represented as an EST clone contig. This finding anchors the 5′-end of this gene close to the *MBD1* gene on chromosome 18. The other two are probably derived from independent CGIs found elsewhere on the PAC. The genes associated with these have not been identified to date. The major contaminant in the clone set was *E.coli* DNA (10/18 clones analysed). As *E.coli* DNA resembles CGI DNA in sequence composition it will co-purify with CGIs. More recent experiments have found that by following the PAC DNA preparation protocol outlined on the Sanger Centre web site it is possible to prepare PAC DNA largely free of *E.coli* DNA (http://www.sanger.ac.uk/Teams/Team53/PAC.shtml).

The cosmid and the PAC clone used here were chosen solely because they were known to contain at least one CGI. Only fragments with the sequence composition of CGIs were isolated from both clones using the MBD column. In both cases the CGI of the known gene was found together with others. In the case of the cosmid FAT5, only genuine CGI fragments were isolated even though these lie in an unusually GC-rich domain. The same number of CGI-like fragments were isolated from both FAT5 and 286-e7, even though the PAC clone is at least three times larger than the cosmid clone. The unusual nature of the region surrounding the *PAX6* gene probably accounts for the high incidence of CGIs in FAT5. The generally CGI-poor nature of chromosome 18 is probably the reason why a low number of CGI sequences were found on PAC 286-e7 (7). The success of the method described here in efficiently enabling the purification of CGIs from two such different chromosome domains implies that it should be generally applicable and should prove useful for isolating CGIs from large genomic clones, facilitating rapid gene identification. Indeed, the method described in this paper has been used to estimate the number of CGIs present in chicken DNA cloned into cosmids (9). For the genomic clones used here probes from either a known CGI or a known gene were used to monitor which fractions to select during the purification procedure. In both cases these fractions eluted at the same salt concentration as the methylated plasmid used to calibrate the column. Therefore, this technique can be applied to clones where there are no such probes available.

CGIs can be used to isolate the associated genes because they overlap transcripts (Figs 2A and 5). Either full-length cDNA libraries can be screened or the genes can be identified by searching databases. Approximately a third of the clones analysed from two CGI libraries prepared from human chromosomes 18 and 22 match with sequences present in the EST database (S.H.Cross *et al*., submitted for publication). As the amount of sequence data available increases this figure should rise. CGIs are associated with ~60% of human genes (1). Therefore this method will be complementary to other methods of gene detection, principally exon trapping and cDNA selection (12–14). However, this method does have the advantage that it depends only on sequence composition and is unaffected by gene expression patterns. In addition, because CGIs are overwhelming single copy they can be used for mapping, as they mark the 5′-ends of genes.

ACKNOWLEDGEMENTS

We thank V. van Heyningen for the FAT5 cosmid, Heather McQueen for the 286-e7 PAC clone, William Rideout III for the pBS-ANA plasmid vector, Brian Hendrich for sharing data prior to publication, Martin Simmen for help with computer analyses and Isabel Hanson and Heather McQueen for useful discussions. We thank Aileen Greig and Joan Davidson for excellent technical assistance. We would also like to thank Ian Jackson and Martin Simmen for critical reading of the manuscript. S.H.C. and V.H.C. were supported by a grant from the UK Medical Research Council. A.P.B. was a Howard Hughes International Scholar.

REFERENCES

1 Antequera,F. and Bird,A. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 11995–11999.
2 Cross,S.H. and Bird,A.P. (1995) *Curr. Opin. Genet. Dev.*, **5**, 309–314.
3 Cross,S.H., Charlton,J.A., Nan,X. and Bird,A.P. (1994) *Nature Genet.*, **6**, 236–244.

4 McQueen,H.A., Fantes,J., Cross,S.H., Clark,V.H., Archibald,A.L. and Bird,A.P. (1996) *Nature Genet*., **12**, 321–324.
5 Cross,S.H., Lee,M., Clark,V.H., Craig,J.M., Bird,A.P. and Bickmore,W.A. (1997) *Genomics*, **40**, 454–461.
6 McQueen,H.A., Clark,V.H., Bird,A.P., Yerle,M. and Archibald,A.L. (1997) *Genome Res*., **7**, 924–931.
7 Craig,J.M. and Bickmore,W.A. (1994) *Nature Genet*., **7**, 376–382.
8 Craig,J.M. and Bickmore,W.A. (1993) *Bioessays*, **15**, 349–354.
9 McQueen,H.A., Siriaco,G. and Bird,A.P. (1998) *Genome Res*., **8**, 621–630.
10 Kim,Y.J., Kim,K.S., Do,S., Kim,C.H., Kim,S.K. and Lee,Y.C. (1997) *Biochem. Biophys. Res. Commun*., **235**, 327–330.
11 Fairman,W.A., Vandenberg,R.J., Arriza,J.L., Kavanaugh,M.P. and Amara,S.G. (1995) *Nature*, **375**, 599–603.
12 Krizman,D.B. (1997) In Birren,B., Green,E.D., Klapholz,S., Myers,R.M. and Roskams,J. (eds), *Genome Analysis: A Laboratory Manual, Detecting Genes*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, Vol. 2, pp. 191–216.
13 Lovett,M., Kere,J.H. and Hinton,L.M. (1991) *Proc. Natl Acad. Sci. USA*, **88**, 9628–9632.
14 Parimoo,S., Patanjali,S.R., Shukla,H., Chaplin,D.D. and Weissman,S.M. (1991) *Proc. Natl Acad. Sci. USA*, **88**, 9623–9627.
15 Valdes,J.M., Tagle,D.A. and Collins,F.S. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 5377–5381.
16 Shiraishi,M., Oates,A.J., Li,X., Hosoda,F., Ohki,M., Alitalo,T., Lerman,L.S. and Sekiya,T. (1998) *Nucleic Acids Res*., **26**, 5544–5550.
17 Kato,R. and Sasaki,H. (1998) *DNA Res*., **5**, 287–295.
18 John,R.M. and Cross,S.H. (1997) In Birren,B., Green,E.D., Klapholz,S., Myers,R.M. and Roskams,J. (eds), *Genome Analysis: A Laboratory Manual, Detecting Genes*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, Vol. 2, pp. 217–285.
19 Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J. Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res*., **25**, 3389–3402.
20 Fantes,J.A., Bickmore,W.A., Fletcher,J.M., Ballesta,F., Hanson,I.M. and van Heyningen,V. (1992) *Am. J. Hum. Genet*., **51**, 1286–1294.
21 Glaser,T., Walton,D.S. and Maas,R.L. (1992) *Nature Genet*., **2**, 232–239.
22 Prosser,J. and van Heyningen,V. (1998) *Hum. Mutat*., **11**, 93–108.
23 Plaza,S., Dozier,C., Langlois,M.-C. and Saule,S. (1995) *Mol. Cell. Biol*., **15**, 892–903.
24 Miles,C., Elgar,G., Coles,E., Kleinjan,D.-J., van Heyningen,V. and Hastie,N.D. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 13068–13072.
25 Liang,G., Salem,C.E., Yu,M.C., Nguyen,H.D., Gonzales,F.A., Nguyen,T.T., Nichols,P.W. and Jones,P.A. (1998) *Genomics*, **53**, 260–268.
26 Brown,A., McKie,M., van Heyningen,V. and Prosser,J. (1998) *Nucleic Acids Res*., **26**, 259–264.
27 Ioannou,P.A. and de Jong,P.J. (1996) In Dracopoli,N.C., Haines,J.L., Korf,B.R., Moir,D.T., Morton,C.C., Seidman,C.E., Seidman,J.G. and Smith,D.R. (eds), *Current Protocols in Human Genetics*. John Wiley & Sons, New York, NY, pp. 5.15.1–5.15.24.
28 Cross,S.H., Meehan,R.R., Nan,X. and Bird,A. (1997) *Nature Genet*., **16**, 256–259.
29 Hendrich,B.H. and Bird,A. (1998) *Mol. Cell. Biol*., **18**, 6538–6547.
30 Hanson,I., Seawright,A., Hardman,K., Hodgson,S., Zaletayev,D., Fekete,G. and van Heyningen,V. (1993) *Hum. Mol. Genet*., **2**, 915–920.