Quick Identification and Localization of CpG Islands in Large Genomic Fragments by Partial Digestion with \textit{Hpa} II and \textit{Hha} I

Reiko \textsc{Kato} and Hiroyuki \textsc{Sasaki}\*  

\textit{Division of Disease Genes, Institute of Genetic Information, Kyushu University, 3-1-1 Maidashi, Higashi-ku, Fukuoka 812-8582, Japan}

(Received 12 July 1998)

Abstract

More than 50\% of mammalian genes are associated with CpG islands and thus they serve as a good gene marker. We have devised a simple method to scan large pieces of native or cloned genomic DNA for CpG islands. The method is based on the presence of multiple \textit{Hpa} II and \textit{Hha} I sites in CpG islands, at a frequency 30 times higher than in the rest of the genome. The steps include complete digestion of DNA with a rare-cutting restriction endonuclease (to produce large fragments with defined ends), partial digestion with \textit{Hpa} II and \textit{Hha} I, and subsequent Southern hybridization with an end probe. This identifies a CpG island as a cluster of sub-bands and, based on their electrophoretic mobility, one can immediately locate the island relative to the ends. For many vectors, universal probes flanking the cloning site are available, enabling the simultaneous analysis of a large number of samples. We demonstrated the usefulness of the method by analyzing known CpG islands in native genomic DNA and lambda, cosmid and P1 clones, and by isolating two novel transcribed islands from anonymous cosmid clones. Our method is quick, inexpensive, and can detect CpG islands with few or even no rare-cutter sites.

\textbf{Key words:} CpG islands; gene identification; DNA methylation; \textit{Hpa} II; \textit{Hha} I

1. Introduction

A haploid mammalian genome contains approximately $3 \times 10^9$ bp of DNA but only several percent of this codes for protein. Identification of new genes from a large region of DNA is therefore labor-intensive works. One way of isolating genes is to screen small DNA fragments prepared from the region of interest for spliced exons by the so-called exon trap assay.\textsuperscript{1} An alternative way is to look for CpG islands, which serve as a useful gene marker.\textsuperscript{2} CpG islands are genomic regions larger than 200 bp with a high G + C content (\textgreater 50\%) and a high ratio of observed versus expected CpG dinucleotides (\textgreater 0.6).\textsuperscript{3} Although cytosine residues of most CpGs in the genome are partially or completely methylated, those in CpG islands are normally free from methylation. In humans, all housekeeping genes and many tissue-specific genes, which together comprise 50–60\% of all genes, are associated with CpG islands.\textsuperscript{3–5} In particular, parentally imprinted genes on autosomes and X-linked genes that are subject to X chromosome inactivation are frequently associated with CpG islands. In most cases, the CpG island contains the promoter and the 5' portion of the gene.

\textsuperscript{1} Communicated by Yoshiyuki Sakaki
\* To whom correspondence should be addressed. Tel. +81-92-642-6168, Fax. +81-92-632-2375, E-mail: hsasaki@gen.kyushu-u.ac.jp

The most common approach used to identify CpG islands in a defined genomic region is to map the cleavage sites for rare-cutter restriction enzymes. Rare-cutter restriction endonucleases recognize 6- to 8-bp sequences containing at least one CpG dinucleotide. There are about eight such enzymes whose recognition sites are predominantly present in CpG islands (41–93\% of all sites in the genome).\textsuperscript{2} Therefore, if several sites for these enzymes are co-mapped in a defined region, the region is most probably a CpG island. However, the frequency of rare-cutter sites is not so high even in CpG islands (0.27–2.14 sites per island, based on an average island size of 1.4 kb)\textsuperscript{2} and there are many examples of islands with only few, or even no, rare-cutter sites.\textsuperscript{2} Therefore, one could overlook a proportion of islands by this strategy. Furthermore, most rare-cutter restriction enzymes are more expensive than the ordinary restriction enzymes.

We describe an alternative approach which uses two 4-base restriction enzymes \textit{Hpa} II (CCGG) and \textit{Hha} I (GCGC) instead of multiple rare-cutter enzymes. CpG islands contain multiple sites for both \textit{Hpa} II and \textit{Hha} I (20.35 and 21.86 sites per island, respectively)\textsuperscript{2} and their density is approximately one per 70 bp in islands and one per 2300 bp in the rest of the genome.\textsuperscript{2} Thus the frequency is approximately 30 times higher in islands than in non-island regions (Fig. 1A). This large difference in frequency enabled us to detect CpG islands as clusters.
2. Materials and Methods

2.1. Materials

Mouse genomic DNA was prepared from the spleen of C57BL/6 mice by a standard procedure. Mouse genomic clones carried by the lambda vector λFIXII were kindly provided by Drs. R. Kimura, K. Kawabe and K. Morohashi. The mouse cosmid K- and A-series were randomly picked up from a SuperCos 1 library (Stratagene, USA). The mouse cosmid clone cDH2 was previously iso-
lated from the same library. The human P1 clones D-26 and O-459Δ, carried by the pAd10 sacBI1 vector, were kindly gifts from Drs. M. Hattori and Y. Sakaki.

2.2. Digestion with rare-cutter restriction enzymes

Approximately 100 μg of cloned DNA (carried by either lambda, cosmid or P1 vectors) was cut to completion with appropriate rare-cutter restriction enzymes under the conditions recommended by the suppliers. DNA was extracted with phenol and chloroform, precipitated with ethanol, and resuspended in an appropriate volume of TE (pH 8.0) to a final concentration of approximately 200 ng/μl.

2.3. Partial digestion with Hpa II or Hha I

Mouse genomic DNA treated as above (100 μg) was brought into 1 mg of L buffer [10 mM Tris-HCl (pH 7.5), 10 mM MgCl₂, 1 mM dithiothreitol] or '1 × M buffer' [10 mM Tris-HCl (pH 7.5), 10 mM MgCl₂, 1 mM dithiothreitol, 50 mM NaCl] and partially digested with 5 units of Hpa II or 2.5 units of Hha I, respectively. Partial digestion was performed by limiting the time of digestion. In practice, aliquots of 100 μl were removed every 1.25 min after the start of the reaction. Each cloned DNA sample (1 μg) was brought into 100 μl of L or M buffer and digested with 5, 1.25 or 0.62 units of Hpa II (for lambda, cosmid or P1 clones, respectively), or 2.5, 0.62 or 0.31 units of Hha I (same as above), respectively. Aliquots of 10 μl were removed every 2 min. The removed reactions were immediately stopped by adding 1/20 volumes of 0.5 M EDTA (pH 8.0) and then precipitated with ethanol.

2.4. Gel electrophoresis

To separate high molecular weight DNA fragments efficiently, genomic, cosmid and P1 samples treated as above were subjected to field-inversion (biased sinusoidal field) gel electrophoresis by using the Genofield AE-8900 system (Atto, Japan). DNA was loaded on a 1% (w/v) agarose gel in 0.5 × TBE buffer and run at an appropriate DC/AC ratio and frequency. Lambda DNA was run through a conventional 0.4% (w/v) agarose gel in 1 × TBE buffer.

2.5. Southern blot analysis of native genomic DNA

After the gel was photographed, DNA was blotted onto a piece of Hybond-N+ filter (Amersham, UK) with alkaline transfer and then UV-crosslinked (Stratalinker 2400, Stratagene, USA). Labeling of the probe DNA fragment was carried out by incorporating [α-32P]dCTP by a random priming method (Megaprime DNA Labeling System, Amersham, UK). The filters were hybridized with the probe in Church's solution at 65 °C. Washing was performed in 0.2 × SSC, 0.1% SDS at 65 °C. The filters were then exposed to X-ray films (Fuji, Japan) at −80 °C.

2.6. Southern blot analysis of cloned DNA

DNA in the gel was transferred onto a piece of Hybond-N filter (Amersham, UK) with 20 × SSC. Hybridization and detection were carried out by using the Flash Non-radioactive Gene Mapping Kit (Stratagene, USA). This kit supplies T3 and T7 oligonucleotide probes conjugated with alkaline phosphatase. The filters were hybridized with either of the probes at a concentration of 2 nM in the supplied hybridization buffer. Hybridization was done at 37 °C for 60 min. The filters were washed in prewarmed 2 × SSC, 1% SDS and then in 2 × SSC, 0.1% Triton X-100 at 37 °C. The filters were further washed in 2 × SSC at room temperature and finally rinsed twice in the assay buffer (diethanolamine pH 10.0 and 1 mM MgCl₂). Visualization of bands was performed by reacting the filter with CSPD chemiluminescent substrate in the assay buffer at room temperature and then by exposing to X-ray films (Fuji, Japan).

2.7. DNA sequencing

Appropriate DNA fragments from the cosmids were subcloned into plasmids and sequenced by using the Dye Primer Cycle Sequence Kit (Perkin-Elmer, USA) and ABI PRISM 377 DNA sequencer. All sequence information was obtained on both strands.

2.8. Northern blot analysis

Total RNA was isolated from mouse embryos and adult mouse tissues by guanidium isothiocyanate extraction followed by CsCl centrifugation. RNA was treated with glyoxal, fractionated on an agarose gel, and transferred to Hybond-N+ membrane (Amersham, UK) according to standard procedures. RNA was UV-crosslinked and hybridized to 32P-labeled probes.

3. Results

3.1. Principles of the method

Our method is primarily based on the non-uniform distribution of Hpa II and Hha I sites in the genome: these sites appear 30 times more frequently in CpG islands than in the rest of the genome (Fig. 1A). A diagram in Fig. 1B summarizes the principles and experimental procedures. Both native and cloned genomic DNAs are first cut with an appropriate rare-cutter restriction enzyme to obtain defined ends, which serve as reference points for mapping islands. Most of the commonly used cloning vectors contain rare-cutter sites on both sides of the cloning site, and therefore these enzymes can be conveniently used to release the insert. The samples are then split into halves, and one is subjected to limited digestion with Hpa II and the other with Hha I. The condition of

No. 5| R. Kato and H. Sasaki 289
digestion is determined such that about half of the fragment of interest remains intact, but samples were usually collected at several time points during the reaction. After electrophoretic fractionation, the products are visualized by Southern blotting with an end probe, located close to either end of the fragment of interest. CpG islands then appear as clusters of sub-bands and one can immediately know their location by the electrophoretic mobility. A comparison of the results obtained with \textit{Hpa II} and \textit{Hha I} makes the identification and mapping easier and more convincing (Fig. 1B).

3.2. Detection of CpG islands in native genomic DNA

We first tested our method with genomic DNA from the mouse spleen. In this case, hypermethylation of the bulk genome should highlight non-methylated CpG islands since both \textit{Hpa II} and \textit{Hha I} are sensitive to CpG methylation. A 25-kb Sse8387I fragment within the mouse \textit{Ins2}/\textit{Igf2} region, which is known to contain two CpG islands (Fig. 2A),\(^8\) was analyzed. When the products of limited digestion by \textit{Hpa II} and \textit{Hha I} were probed with a 0.9-kb \textit{Bgl II} fragment located close to the \textit{Ins2} end of the Sse8387I fragment, the two CpG islands at about 3 kb upstream of \textit{Igf2} and at the second \textit{Igf2} promoter were clearly detected as two broad sub-bands in the expected positions (molecular sizes 15 and 20 kb, respectively) (Fig. 2B). Notably, the two enzymes gave very similar results. Additional bands were seen in the low molecular weight region, suggesting that some sites closer to \textit{Ins2} were not completely methylated, but these sub-bands were rather faint. Hybridization with a probe from the opposite end (0.74-kb \textit{BamHI-Pst I} fragment) confirmed these observations (data not shown). These results demonstrated that CpG islands are readily detectable by our method in native genomic DNA.

---

Figure 2. Detection of two CpG islands in the mouse \textit{Ins2}/\textit{Igf2} region on chromosome 7 by an analysis of native genomic DNA. A: Physical map of the region. Filled boxes represent the gene exons, and arrows above the boxes indicate the direction of transcription. Positions of the CpG islands, location of the hybridization probe, and interpretation of the results are shown below the map. B: Autoradiographs representing the results obtained with genomic DNA from the mouse spleen. The two CpG islands are detected as broad sub-bands at 15 and 20 kb.
3.3. Detection of CpG islands in P1 and lambda clones

We next tried to detect CpG islands in cloned DNA fragments, where the native CpG methylation was lost. Some commonly used vectors for large DNA fragments contain T3, T7 or SP6 bacteriophage promoters on both sides of their inserts, and the inserts can be released by digestion with rare-cutter restriction enzymes such as Not I, with the promoter sequences attached to both ends of the released inserts (Fig. 3). Thus, for the fragments conveyed by such vectors, T3, T7 and SP6 oligonucleotides hybridizing to the promoters could be used as 'universal' end probes. Also, in the subsequent experiments, we combined samples partially digested with different amounts of Hpa II or Hha I before loading onto the gel, which enabled simultaneous analysis of multiple clones in a single gel (see later for the analysis of cosmid clones).

Large human DNA fragments carried by the P1 vector pAd10sacBII were analyzed by the present method (Fig. 4). The P1 DNAs were completely digested with Not I prior to the partial digestion with Hpa II and Hha I, and a commercially available, alkaline phosphatase-conjugated T7 probe (Stratagene, USA) was used as a universal probe. Two human clones, O-495Δ and D-26, were analyzed: O-495Δ had a 35-kb insert from the APP region on human chromosome 21 and was known to contain a typical CpG island (this clone had undergone a deletion within a longer original insert); D-26 did not contain any CpG island. In the Hpa II and Hha I lanes

Figure 3. Structure of the lambda, cosmid and P1 vectors used in this study. Positions of the cloning site, T3, T7 and SP6 oligonucleotide probes, and rare-cutter sites for release of the insert are indicated.

Figure 4. Detection of CpG islands in the inserts of P1 clones. A CpG island from the human APP region on chromosome 21 was clearly detected as a thick sub-band in both the Hpa II and Hha I lanes of O-495Δ whereas no common signal was identified in D-26, which contained an adjacent genomic region.
Quick Detection of CpG Islands

Figure 5. Identification of CpG islands and possible candidate regions in cosmid clones. In cDH2, a CpG island associated with the mouse L23mrp gene on chromosome 7 was clearly detected as a cluster of sub-bands in both Hpa II and Hha I lanes. Seven mouse cosmid clones, randomly picked up from a SuperCos 1 library, were also analyzed, and the sub-bands marked a-d were selected as possible CpG islands.

3.4. Detection and isolation of CpG islands from cosmid clones

We next examined a cosmid clone (cDH2) known to contain a CpG island and seven randomly picked up clones (A- and K-series) with no information regarding the existence of islands. The inserts of these clones were released by Not I digestion, and both the T7 and T3 promoters were used as end probes. The clone cDH2 contained a CpG island from the mouse L23mrp region at about 3.5 kb from the "T7 end," and this was clearly seen as a bunch of sub-bands in both the Hpa II and Hha I lanes (Fig. 5). Although this clone contained some other regions rich in Hpa II or Hha I sites, their sub-bands did not appear in identical positions in the two lanes. Indeed, sequence analysis of the entire insert revealed that none of these regions qualified as an island.

Among the seven anonymous clones, both K4 and K5 contained a Not I site within their inserts, and the fragments from these inserts could be scanned by only one probe. In the case of K4, however, the T7 probe detected two clusters of sub-bands at 10 and 16 kb (designated regions a and b, respectively) upon both Hpa II and Hha I digestion. Two other possible CpG islands were identified in A6: one appeared at 21 kb from the T3 end (or 18 kb from the T7 end) (designated region c) and the other at 5 kb from the T3 end (or 34 kb from the T7 end) (designated region d). The rest of the clones did not contain regions rich in both Hpa II and Hha I sites.

To examine whether the four Hpa II/Hha I clusters identified above are CpG islands, we next attempted to clone them from the cosmids. The possible islands were conveniently located on the restriction maps by analyzing the cosmid inserts with some restriction enzymes appropriate for mapping, together with Hpa II and Hha I, us-
Fig. 6. An example of restriction mapping of the cosmid insert and simultaneous localization of the possible CpG islands on the map. A: A randomly picked up clone, designated A6, was analyzed by partial digestion with HpaII and HhaI, together with HpaII and HhaI, using the Flash Non-radioactive Gene Mapping Kit (Stratagene, USA) (left and middle). To produce a precise restriction map, the clone was also analyzed by complete digestion with BamHI, EcoRI, and by both together (right). B: The obtained restriction map and the localization of the possible islands.

We have described a simple and inexpensive method to identify and locate CpG islands in large DNA fragments from both native and cloned genomic DNA. This approach is based on the high density of HpaII and HhaI sites in CpG islands and does not involve any special techniques or equipment. The usefulness of the method has been demonstrated for DNA fragments as large as 70 kb in P1 clones. It worked particularly well with native genomic DNA, in which most non-island sites were methylated and thus insensitive to the enzymes. Furthermore, we isolated two novel CpG islands from randomly selected mouse cosmid clones.
Several methods have been developed to identify and isolate DNA fragments containing CpG islands. Cross et al. used a methyl-CpG binding protein to purify islands directly from total genomic DNA, based on both the high density and the unmethylated state of CpG dinucleotides within islands.\textsuperscript{10} To isolate CpG islands from cloned DNA, Shiraishi et al. devised a method that detects genomic sub-fragments with high G + C content using denaturing gradient gel electrophoresis.\textsuperscript{11} A drawback of this approach, however, is that such regions are not necessarily rich in CpG dinucleotide. A different approach reported by Valdes et al. is the ‘island rescue PCR,’ which recovers CpG islands with rare-cutter sites from cloned DNA using a PCR primer that matches the human \textit{Alu} repeat sequence.\textsuperscript{12} However, genes that lack an \textit{Alu} repeat within about 2 kb of a CpG island are less likely to be isolated. In all of these approaches, the location of the island has to be determined after its isolation, and this could be a laborious task.

So far, the most common approach used to identify and locate CpG islands in a defined region has been the long-range mapping by rare-cutter restriction enzymes. However, a serious problem is that not all CpG islands contain multiple rare-cutter sites.\textsuperscript{2} Our approach was designed to overcome this problem by using restriction endonucleases with 4-bp cleavage sites. Thus, one of the most important advantages of our method over the conventional approach is: (1) it can identify CpG islands with few or even no rare-cutter sites. Other advantages include the following: (2) by use of the universal oligonucleotide probes, one can handle many cloned DNA samples at a time; (3) since the same method can be used to construct a restriction map of a cosmid or a lambda clone, one can locate the CpG island on the map immediately and precisely; (4) upon the analysis of native genomic DNA, one can know the methylation status of these sites by, for example, comparing the profile with that obtained by \textit{Msp} I (a methylation-insensitive isoschizomer of \textit{Hpa} II) digestion. Thus, the method provides a convenient and quick approach for isolating CpG islands. Instead of using the T3 and T7 oligonucleotide probes, the ‘cos mapping’ technique may also be used for the analysis of cosmid and
lambda clones.\textsuperscript{13}

There are, however, two potential drawbacks to our approach. One is that the method requires at least one rare-cutter site as a reference point and an end probe located internal to it. This may pose a problem when analyzing native genomic DNA because one has to clone a probe fragment. A solution for this, however, may be the use of ‘jumping clones’ and ‘linking clones,’ which are short genomic clones flanking or containing a rare-cutter site.\textsuperscript{13} The other potential drawback is that it is sometimes difficult to judge from the blotting profile whether a region is a CpG island or not. An important guide is the co-localization of the $Hap\ II$ and $Hha\ I$ clusters, however, two out of the four candidate regions identified in this study were not qualified as islands. Just as with the other approaches, the final proof for an island is only obtained after cloning and sequencing. We should like to stress, however, that our method provides a more thorough screen for CpG islands.

In this study, we found that the two novel CpG islands were indeed transcribed. Surprisingly, however, we also found that the other two regions, which were relatively CpG-rich but did not meet the criteria for an island, contained transcribed sequences. Perhaps protein coding regions may preserve more CpG dinucleotides than the rest of the genome because of the structural and functional constraints imposed on the sequences during evolution. If this is the case, our method may work as a more general approach to isolate transcribed sequences.

Acknowledgments: We are grateful to Drs. R. Kimura, K. Kawabe, K. Morohashi, M. Zubair, M. Hattori, and Y. Sakaki for providing us with the genomic clones, Drs. A. Iwaki, Y. Fukumaki, K. Hayashi, and H. Furuumi for their excellent technical assistance. The work was supported by a Health Sciences Research Grant and encouragement. We especially thank N. Aoki and all members of our laboratory for their advice, interest and encouragement. We especially thank N. Aoki and H. Furuumi for their excellent technical assistance. The work was supported by a Health Sciences Research Grant on Human Genome and Gene Therapy from the Ministry of Health and Welfare of Japan, a Grant-in-Aid for Scientific Research from the Ministry of Education, Science, Sports and Culture of Japan, and a Research Grant from the Kato Memorial Bioscience Foundation.

References
