Methylation and sequence analysis around Eagi sites: identification of 28 new CpG islands in XQ24–XQ28

Carla Tribioli, Filippo Tamanini, Cristina Patrosso[1], Luciano Milanesi[1], Anna Villa, Rossana Pergolizzi[1], Elena Maestrini, Stefano Rivella, Silvia Bione, Mita Mancini, Paolo Vezzoni[1] and Daniela Toniolo*

Istituto di Genetica Biochimica ed Evoluzionistica, CNR, Via Abbiategrasso 207, 27100 Pavia and [1]Istituto di Tecnologie Biomediche Avanzate, CNR, Via Ampere 56, Milan, Italy

Received November 26, 1991; Revised and Accepted January 16, 1992 EMBL accession nos X62354–X62365 (incl.)

ABSTRACT

Thirtytwo probes for CpG islands of the distal long arm of the human X chromosome have been identified. From a genomic library of DNA of the hamster-human cell hybrid X3000.1 digested with the rare cutter restriction enzyme EagI, 53 different human clones have been isolated and characterized by methylation and sequence analysis. The characteristic pattern of DNA methylation of CpG islands at the 5′ end of genes of the X chromosome has been used to distinguish between EagI sites in CpG islands versus isolated EagI sites. The sequence analysis has confirmed and completed the characterization showing that sequences at the 5′ end of known genes were among the clones defined CpG islands and that the non-CpG islands clones were mostly repetitive sequences with a non-methylated or variably methylated EagI site. Thus, since clones corresponding to repetitive sequences can be easily identified by sequencing, such libraries are a very good source of CpG islands. The methylation analysis of 28 different new probes allows to state that demethylation of CpG islands of the active X and methylation of those on the inactive X chromosome are the general rule. Moreover, the finding, in all instances, of methylation differences between male and female DNA is in very strong support of the notion that most genes of the distal long arm of the X chromosome are subject to X inactivation.

INTRODUCTION

Between 50.000 and 100.000 genes may be present in the human genome but only a few thousand have been cloned and sequenced. The majority have been identified upon prior information on their protein product. For a large number of genes however their identification is based on the description of an inherited disorder and the biochemical defect is unknown. The cloning of such genes may be successfully achieved using approaches dependent on high resolution genetic and physical mapping. The precise identification of a genomic region corresponding to a gene may in fact bring to the isolation of coding sequences by screening for cDNAs or evolutionary conserved sequences and by analyzing sequenced DNA for the presence of ORFs.

It is likely that large scale mapping and sequencing of the human genome will also allow the prediction of coding regions from the nucleotide sequence of large segments of DNA and eventually the construction of a detailed physical map of human genes and of human inherited disorders. However this is a formidable task that will not be accomplished very soon, and different methodologies to directly identify human genes and genomic regions corresponding to genes are being tested. Automated partial sequencing of cDNAs will generate expressed sequence tags which may help in the identification of new genes and of coding regions in genomic sequences (1). This simple approach, in connection with the use of PCR and panels of somatic cell hybrids to localize the cDNAs, will provide a high resolution map of the genes along the chromosomes and will identify gene regions to sequence. This is however a random approach that does not allow the direct identification of genes from specific portions of the genome.

Different methods have to be designed to isolate gene sequences from specific chromosomal regions. The one we have used and we report here is based on the use of 'rare cutter' restriction enzymes to identify CpG islands. Most mammalian genes are in fact tagged by the presence at their 5′ end of a CpG island, unmethylated G+C and CpG rich region (2,3). Since about 30.000 CpG islands may be present in the mammalian genome, roughly 1/2 of the genes may be recognized from the CpG island surrounding their 5′ end. CpG islands are preferential sites of cleavage of a class of restriction enzymes, the so called 'rare cutter restriction enzymes', whose recognition sequence is CpG rich and methylation sensitive (4). The presence of a cluster of sites for such restriction enzymes is a very strong indication of the presence of a CpG island and of a gene (2). We report now that genomic libraries constructed digesting with a rare cutter restriction enzyme DNA of human-rodent hybrids containing specific portions of human chromosomes may be an enriched source of CpG islands.

* To whom correspondence should be addressed

We are interested in establishing a detailed physical map of the genes of the distal long arm of the human X chromosome. This is among the most gene dense portions of the human genome, where many inherited disorders have been mapped (5). Recently a physical map of a total of 12 Mb of DNA at the end of the long arm of the human X chromosome from Xq27.2 to the telomere has been described (6). This map establishes the order and the orientation of most loci and genes in Xq28. YAC clones from a library specific for the Xq24-Xq28 region have also been isolated and large ordered YAC contigs covering most of the distal portion of the long arm of the human X chromosome are available (7). A detailed physical map of the genes in this part of the human genome will be therefore immediately useful to the isolation and sequencing of candidate genes of the many inherited disorders mapped in the region.

As a first step in the construction of the complete gene map of this chromosomal region a genomic library of the DNA of the hamster human cell hybrid X3000.1, digested with the rare cutter enzyme EagI, was prepared (8). The hybrid carries the Xq24-Xqter portion of the human X chromosome as the only human DNA (9). In this paper we report the characterization of the 53 human clones obtained to determine how many are actually CpG islands, and, therefore, how useful may be our approach

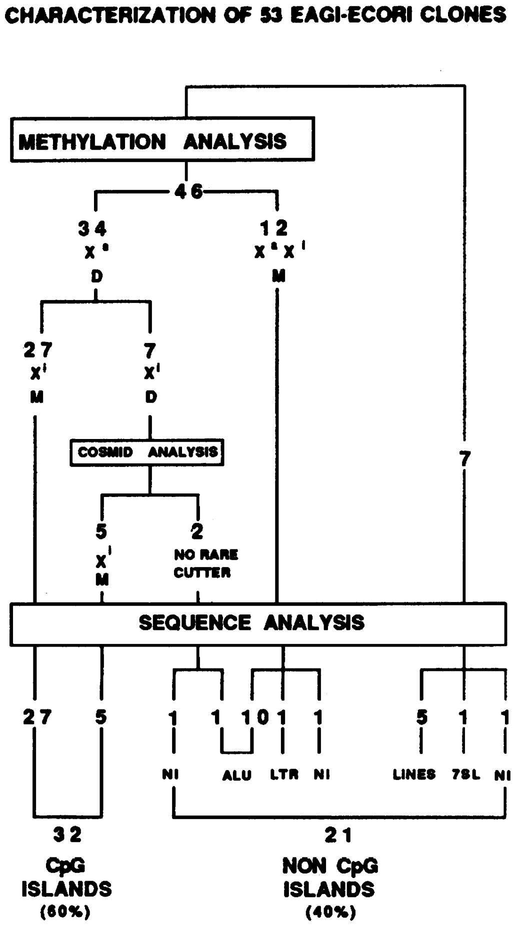

Fig. 1. Scheme of the characterization of 53 possible probes for CpG islands. X^a= active X; X^i= inactive X. D: demethylation in genomic DNA; M: methylation in genomic DNA. NI: non identified.

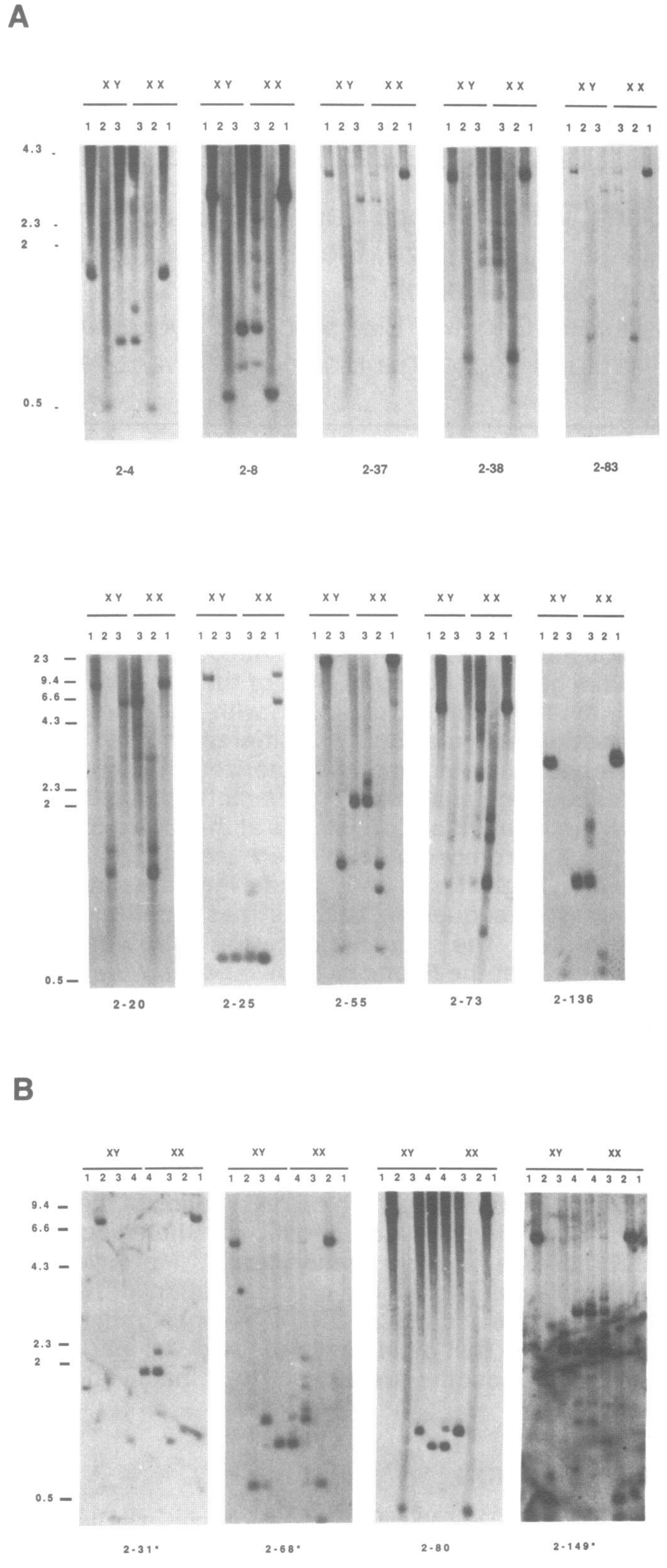

Fig. 2. Methylation analysis of genomic DNA surrounding the EagI sites. DNA from male (XY) or female (XX) WBC was digested with the following restriction enzymes. Panel A: EcoRI:1; EcoRI+ MspI:2; EcoRI+HpaII:3. Panel B: EcoRI:1; EcoRI+MspI:2; EcoRI+ HpaII:3; EcoRI+HhaI:4. After Southern Blotting, filters were sequentially hybridized to the probes indicated below. 2-31*, 2-68* and 2-149* are fragments prepared from cosmids and corresponding to the other side of the EagI site, in the original clone. Molecular weight markers are in kb. Probe 2-55 demonstrates a MspI polymorphism, previously described (31)

to obtain region specific probes for CpG islands and to establish a detailed map of the CpG islands of the distal long arm of the human X chromosome.

MATERIALS AND METHODS

DNA methylation analysis

Genomic DNA was prepared from white blood cells (WBC) and it was digested and blotted to nylon filters as previously described (8). Probes from large clones (more than 3 kb) were prepared for hybridization by restriction enzyme digestion and purification of the fragment indicated from low melting agarose gels (10). Clones with an insert size of 3 kb or less were prepared by PCR from the T3 and T7 promoter region of the vector (Bluescript). All probes still contained repetitive sequences and cold human placenta DNA was used in the hybridization, as described (8).

Nucleotide sequence analysis

Clones were sequenced for an average of 300 bp using an automated sequencing apparatus (EMBL, Heidelberg), from both ends of the insert. Sequencing was performed with the dideoxy chain termination method of Sanger using primers labelled with a single fluorescent dye (11).

Table I. Methylation state of genomic DNA near EagI sites

			Size of restriction fragments (kb)					Methylation state	
Clone	Rare cutter	R.E.	R.E.	R.E. +EagI	R.E. +MspI*	R.E. +HpaII*	R.E. +HhaI*	Xa°	Xiˆ
2-4	−	EcoRI	1.5	1.2	0.5	1	nd	D	M
2-5	−	EcoRI	2	1.4	1.4/0.2	2	nd	M	M
2-6	−	EcoRI	10	3.5	2/1.6	10	nd	M	M
2-7	+	EcoRI	15	1.5	0.8	0.85/0.8	0.75/0.6	D	M
2-8	−	EcoRI	3	1.6	0.6	1/0.8	nd	D	M
2-18	+	EcoRI	3.8	2	0.8	1.6	nd	D	M
2-19	+	EcoRI	6.5	6	0.8/0.6	3.6/3	nd	D	M
2-20	−	EcoRI	7.5	2.3	1.45/1.2	6	nd	D	M
2-22	+	HindIII	23/15§	4	0.2	1	nd	D	M
2-23	−	EcoRI	5	3.5	1.8	5	nd	M	M
2-25	+	HindIII	7.5	1	0.7	0.7	nd	D	M
2-28	−	HindIII	15	2	1.45	1.45	0.95	D	M
2-31	+	EcoRI	6	4	0.9	2.3	2.6	D	D
2-34	−	EcoRI	5	1.6	2.5−0.3	5	nd	M	M
2-37	+	SacI	4	4	2-1.8	3	nd	D	M
2-38	+	EcoRI	4	2.3	0.9	2/1.85	nd	D	M
2-39	+	EcoRI	12	2	1.4	1.7	nd	D	M
2-40	+	EcoRI	5.5	1.8	0.9/0.3	1.8/1.2	1.7/1.2	D	D
2-46	−	EcoRI	3	1.2	1.2	3	nd	M	M
2-48	−	EcoRI	7	2	1.6	2	7	M	M
2-53	−	XbaI	2	1.25	1.2	2	2	M	M
2-55	+	EcoRI	18	2.5	1.2/0.85	2	2.5	D	M
2-59	+	EcoRI	4	1.5	0.6/0.2	0.6	nd	D	M
2-62	−	EcoRI	9	1.5	0.9	1.2/0.9	1.2	D	M
2-63	+	EcoRI	4.3	2.5	1.8−0.5	2.3	2.7	D	M
2-65	−	EcoRI	6	3	3	3	2.8	D	D
2-68	+	EcoRI	5.5	4	2/1.45	3.7	4	D	D
2-71	−	EcoRI	4	3	3/2.5	4	nd	M	M
2-72	+	EcoRI	10	2.1	1.4/1.2	1.7/1.4	nd	D	M
2-73	+	EcoRI	5	3.5	1.9−0.65	2.5−1.1	nd	D	M
2-77	−	BglII	6	1.6	1.45	6	nd	M	M
2-78	+	EcoRI	8	3.5	1.6/1.5	3	nd	D	M
2-80	+	EcoRI	8	1.3	0.6/0.4	1.3	1.2	D	M
2-81	+	EcoRI	5	2.3	0.7	2.4/1.65	nd	D	M
2-83	+	SacI	4	3.5	1.5/1.1	3.2	nd	D	M
2-104	−	EcoRI	8	4.5	3.2	4.5	4.5	D	D
2-108	+	SacI	6.5	1.1	0.5−0.2	0.8−0.2	nd	D	M
2-110	−	BglII	3.3	2	1.6	1.9/1.6	nd	D	M
2-128	−	EcoRI	20	3.4	1.4/1.2	9.4	nd	M	M
2-132	−	HindIII	5	0.85	4	4	nd	M	M
2-136	+	BglII	3.5	2.2	0.5−0.2	1.2/0.2	nd	D	M
2-140	+	EcoRI	15	2.8	1.65−0.2	2.5	nd	D	M
2-142	+	EcoRI	10	6	1.4	6	nd	D	D
2-147	−	EcoRI	10	1.4	0.7	1.2/0.9	nd	D	M
2-149	+	EcoRI	6	3	0.9/0.6	1.9	2/1.8	D	D
2-151	−	EcoRI	4.8	1.4	1.1	4.8	nd	M	M

*: Fragment size is from digests of male DNA;

°: Methylation state of the active X chromosome in male DNA;

ˆ: Methylation state of the inactive X chromosome in female DNA;

D: Demethylated;

M: Methylated;

§: This probe hybridizes to two different HindIII and EcoRI fragments in male DNA.

Table II. GC conent of CPG island sequences

Clone	Homologies	Base pairs sequenced*	% G+C	CpG	GpC	CpG/GpC
2-4		358	48	10	27	0.37
2-7	LAMP2	350	64	29	38	0.76
2-8		341	72	32	51	0.63
2-18		393	71	36	40	0.9
2-19		251	73	24	36	0.67
2-20		339	41	6	11	0.55
2-22		346	61	19	27	0.7
2-25		309	44	9	10	0.9
2-28		123	60	6	13	0.46
2-31		241	68	19	18	1
2-37†		372	76	48	56	0.86
2-38		312	76	54	48	1.13
2-39	G6PD	469	65	50	43	1.16
2-40		429	55	19	34	0.56
2-55		435	69	45	62	0.73
2-59		495	60	39	53	0.74
2-62		506	55	16	32	0.5
2-63		335	67	28	30	0.93
2-68		298	76	34	44	0.78
2-72		200	35	2	12	0.17
2-73		362	57	18	20	0.9
2-78		375	60	27	30	0.9
2-80		379	72	17	20	0.85
2-81		330	64	18	25	0.72
2-83		304	63	28	35	0.8
2-108		315	66	17	32	0.53
2-110		305	49	14	12	1.17
2-136		119	92	24	21	1.14
2-140		381	73	45	48	0.94
2-142		137	82	20	27	0.74
2-147	HPRT	382	53	12	22	0.55
2-149		143	82	21	21	1

* From the EagI site

† Previously identified CpG island at the 3'end of GdX gene (21,22): the sequence from the EcoRI site is identical to the 3'end of the GdX gene and the size of the EagI-EcoRI insert is identical to the distance between the GdX gene and the known CpG island.

Table III. Sequence homologies

Clone	Sequenced from	Identical to	Region of homology
2-39	EagI	G6PD	5' flanking DNA
2-147	EagI	HPRT	5' flanking DNA
2-71	EagI	LAMP-2	5'end, first exon and intron
2-37	EcoRI	GdX	3'end

Sequence analysis was performed with a GCG (Genetic Computer Group, Inc.) package version 7. To search for homology the FASTA algorithm was used (12).

RESULTS

The clones from the EagI-EcoRI library (8) were analyzed as outlined in Fig.1.

Methylation of genomic DNA

In vertebrates, CpG islands are unmethylated regions of DNA: the only exception, so far, are the methylated CpG islands of the inactive X chromosome. Southern blots of female DNA digested with HpaII and hybridized to probes for CpG islands of the X chromosome, in addition to bands in common with male DNA, show higher molecular weight bands corresponding to methylation of HpaII sites on the inactive X chromosome (13). To determine the methylation state of the genomic DNA in the vicinity of each EagI site, the human clones from the EagI-EcoRI library were used as probes in Southern blots of genomic DNA of male or female individuals. As it has been done previously for 38 clones (8), the DNA was digested with EcoRI or with a more appropriate restriction enzyme (indicated in Table I:RE), distant no more than 1–3 kb from the EagI site, as well as with the same RE + MspI or HpaII or Hha. Only 8 additional clones could be analyzed, since 7 did not demonstrate single bands in genomic DNA. A total of 46 clones was studied.

Results of the hybridizations of 14 clones are in Fig.2 and a summary of the methylation studies is in Table I. In 34 clones, the HpaII+ RE or HhaI+RE (or both) digestions of male DNA produced a fragment smaller than the EagI-RE distance, indicating demethylation of HpaII or HhaI sites in the vicinity of the EagI site. In 27/34 clones additional higher molecular weight bands which could be caused by methylation of the same and additional HpaII or HhaI sites in the vicinity of the EagI site, were present only in female DNA (Table I). The remaining 7 clones did not show methylation differences between male and female DNAs with neither of the two methylation sensitive restriction enzymes.

A) Alu SEQUENCES

```
2-151                CGGCCG...ACA............-.................-..........-...............--..A.........A.........T......T....
2-5                  CGGCCG..T.CT.........A..C.................T........T.-.A....C......T..AAAAA..........GG..................
2-34                 CGGCCG..A........G...C---.A...............-..........-.AC.......C...--...-.........A..G......T......T....
2-48                 CGGCCG..CAC.....T.......-...........T.....-.......T.......A.........--.--....................A.........G..
Alu CONSENSUS         GGCTGGGCGTGGTGGCTCACGCC-TGTAATCCCAGCACTTT-GGGAGGCCGA-GGTGGTGGATCACCTGA--GGTCAGGAGTTCAAGACCAGCCTGGCCAACAT  100
2-132             TTTT.......A............T..-.................-........T.-..CA..GA.....T...--.....G...................A......
2-71              TCAA...CA.....C..-.........-.A...............-........AA-..CGGCCG

2-151 .....-......A................A(7) ..GT...............G.T.......G...........G.....................G..G.................C-
2-5   .T...-.....T.................A(14)A.....A..........T.G.T...................................................T.A....C..G..
2-34  .....-....-.A................A(7) ......A.T..........G..A......G...........................A.....G..A.......A.....C...C.
2-48  .....T.................................A..........A.CG..A..................TA..............-.....T.G.......A......C.....
AluC  GGTGA-AACCCCGTCTCTACTAAAAATACAAAAATTAGCCGGGCGTGGTGGCGCGCGCCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATCGCTTGAACCCGGGAGGTGGAGG 219
2-132 ..A..-.....T.................A(4) .....CGGCCG

2-151 ...........T................T................-..................A_{13}CAGA_4GA_5GA_4CATTGGTCATCTTTAACAAAAGCTAC
2-5   ................A...T.T.T..A..............A..A....T..A..........A_8GA_{13}GA_3GA_3GA_3GGAAGGAAG
2-34  ...........A.......T.........................-.....A.......A....A_3TA_4TA_4TA_4TA_4TA_4TA_3TA_4CTGGTCCCATCTGGCCTTTGAGGTTT
2-48  ...........T.......T......................A..-.....A............A_8GA_{10}CA
AluC  TTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGAC-AGAGCGAGACTCCGTCTCA_n
```

B) LTR LIKE SEQUENCES

```
2-77          CGGCCG......A...CC.........AA.......C.....AC..CAT.....TT.TG..TC...CT.C......C..AC.................
HuERS-P2 GAGCCAGGTCACTGAGCTGCCGGACTCCAGGG-GAAAACCATCTCCC-TTCTGGCTCCC--CCATCCACTG--A-GAGCTATTTCTACTCAATAAAACCTTGC 506

2-77  .C..........A......G.......CA..............C...........T.....G.............A.C..C........GT....C..G.A......TA......A...
HuERS AGTCATTCTCCAGGCCCACATGTGATCTGATTCTTCTGGTACATCAAGGCAAGAAACCCCAGGATACAGAAAGCCCTCTGTCCTTGTGACAAGGTAGAGGGTCTAACTGAGCTGGTTAA 625

2-77  .......CA...AC..GC...A..............A...........C..T......G....G..C..........CC..................G...TG.CA...CTCATGCT...
HuERS CACAAGCTGCCTGTAGATGGC-AAACTAAAAGAGCACCCTGTAACACATGCCCACTGGAGCTTCAGGAGCTGTAAACATTCACCCCTAGACACTGCCATGGGATTGGAG--------CCC 736

2-77  .G..A.C....CA..TGCC...A...T..A..G.T.....A....T......CA
HuERS CACAGCTTGCCTGTCCAAATGCTCCC CT AGAGGTTTGAGCAGCAGGGCACTGAAGAAGTGAGCCACACCCCCATCACATCCCCTGTGAGGGGAACAAGGAACCTTTCCTATTTCAAC 854
```

C) LINE SEQUENCES

```
2-33                     CGGCCG......C.........CTT.............A...A..........................................
LINE CONSENSUS GGGGGAGGGGTGCCCACCATTGTCCAGGCTTGAGCAGGTAAACAAAGCCGCCTGGAAGCTCGAACTGGGTGGAGCCCACCACAGCTCAAGGAGGC 530

2-33   ...........C...............G........................A....G..GTA...T..................-......................C...A......
LINE   CTGCCTGCCTCTGTAGGCTCCACCTCTAGGGGCAGGGCACAGACAAACAAAA-GACAACAAG-ACC-CTGCAGACTTAAATGTCCCTGTCTGACAGCTTTGAAGAGAGTAGTGGTTCTCC 646

2-33   ......GC....GG............G..........-.......................T..CCCCCGA.C.................C..G....C....G.A..............
LINE   CAGCACATAGCTTCAGATCTGAGAACAGGCAGACTGCCTCCTCAAGTGGGTCCCTGACCCCCGA-------GTAGCCTAACTGGGAGGCATCCCCCAGTAGGG-CGGACTGACACCTCAC 758

2-33   .....A...TA..
LINE   ATGGCTGGTACTCCTCTAAGACAAAACTTCCAGAGGAATGATCAGGCAGCAGCATTTGCGGTTCACCAATATCCACTGTTCTGCAGCCACCGCTGCTGATACCCAGGAAAACAGCATCTG 878
2-138  CGGCCG...........G....................C......A..........A..A......G..A.....................A......G........C.......GG...
2-101  CGGCCG...........G...TG................T...........A.....T.T..............G........G....T.T................C.......GG...
2-106  CGGCCG...........G........T...........AT...........A.......T.........C..T...............T.T................C.......GG...
2-58    CGGCCG...______________________________________________________________________________________________________________

LINE   GAGTGGACCTCCAGTAAACTCCAACAGACCTGCAGCTGAGGGTCCTGACTGTTAGAAGGAAAACTAACAAACAGAAAGGACATCCACACCAAAAACCCATCTGTACATCACCATCATCAA 998
2-138  .....A.....T..C....G...........................T..............................................-.........................
2-101  ........A.....C.........A..................................-....C.........-......T-...........C..........TG.............
2-106  .............AC...........................A.......................G............................C..........TG.............
2-58   ___________________.....................T......T.................G......................................................

LINE   AGACCAAAGGTAGATAAAACCATAAAGATGGGGAAAAAGCAGAGCAGAAAAACTGGACACTCTAAAAATGAGAGTGCCTCTCCTTCTCCAAAGTAACGCAGCTCCTCACCAGCAATGGA  1117
2-138  ........A....G........C...........G...A..................A..........G...............C........G.....................C...
2-101  ......................C...............A..................A...........C....C.........C........G.........................
2-106  ......................C...............A..................A.T......C..C....C.........C........G.T...................C...
2-58   ........A.............C...............A..................A..........GC....CA........C........G...A...T.............C...
```

D) 7SL RNA

```
2-19B           CGGCCGCC..G..C.T..................AC...............................ACA.......
7SL    GTAGCTTTTCGCAGCGTCT-CCGACCGCCGGGCGCGGTGGCGCGTGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGTGGGAGGAT 110

2-19B  ........T................................................G......................................G.......................
7SL    CGCTTGAGCCCAGGAGTTCTGGGCTGTAGTGCGCTATGCCGATCGGGTGTCCGCACTAAGTTCGGCATCAATATGGTGACCTCCCGGGAGCGGGGGACCACCAGGTTGCCTAAGGAGGGG 300

2-19B  .-....T.G.....................................................................................-.G.......................
7SL    TGAACCGGCCCAGGTCGGAAACGGAGCAGGTCAAAACTCCCGTGCTGATCAGTAGTGGGATCGCGCCTGTGAATAGCCACTGCACTCCAGCCTGAGCAACATAGCGAGACCCCGTCTCTT 420

2-19B  ..TG.ATT..TT.CTT.....ACT.TA..T.TTT...T.GC.CA.G.--.T.A.T.C.
7SL    TTGCCCCCCTCCCTACTTAAGGGATCTTTGTAAGTAAATAGTGTCTTTGAAGTTAAGAAGTTTGCT CTTTC.GTTCATACGTATTAAGAAACATACTAATGTGCACATTTAAAGACGAG 539
```

Fig. 3. Nucleotide sequence from the EagI site of clones corresponding to repetitive sequences. A: some of the Alus are aligned with the Alu consensus sequence (24). B: clone 2-77 is aligned to the LTR sequence HuERS.P2 (26). C: examples of Lines aligned to the Line consensus sequence (16). Underlined is an insertion of 130 bp not present in all Lines; D: alignment of clone 2-19B to a 7SL sequence (25). In bold are the EagI sites. The sequences have been submitted to the EMBL Data Library. Accession numbers are: 2-151: X62354; 2-5: X62355; 2-34: X62356; 2-48: X62357; 2-132: X62358; 2-33: X62359; 2-58: X62360; 2-101: X65361; 2-106: X62362; 2-138: X62363; 2-196: X62364; 2-77: X62365.

To study DNA methylation of the DNA identified by the remaining 7 clones, cosmids were isolated either from a X chromosome specific (14) or a Xq28 specific library (15): the presence of a cluster of rare cutter restriction enzymes like SacII, BssHII or NruI within 1−2 kb from the EagI site demonstrated the presence of a CpG island in all clones but 2 (not shown): 2-65 and 2-104 did not appear to correspond to CpG islands and their methylation has not been studied further. From the remaining cosmids, fragments flanking the CpG island were prepared and used to probe the same HpaII+RE and HhaI+ RE digestions described above. In all instances we were able to show the methylation differences between male and female DNA typical of CpG islands of the X chromosome (probes 2-31*, 2-68* and 2-149* in Fig.2B). From this analysis we can conclude that 32 of the 46 clones analyzed have the characteristics of the CpG islands of the human X chromosome (Fig.1).

12 clones corresponded to methylated regions of DNA (Fig.1 and Table I): none of them also corresponds to a cluster of rare cutter enzymes (8).

Sequence analysis of the CpG islands

All the 32 clones defined as DNA fragments flanking CpG islands, were sequenced for 300−400 nucleotides from the EagI site. 30 DNAs were also sequenced from the EcoRI site. Analysis of the G+C and CpG content of the nucleotide sequence flanking the Eag I site revealed in most cases a G+C content greater than 60% and a CpG content close to the GpCs (Table II). The nucleotide sequences were compared to the GeneBank (Release 68). This analysis has shown that four of the clones were identical to known CpG islands or to the 5'end of known genes (Table III). The remaining 28 clones did not show any identity to known sequences.

Nucleotide sequence of the clones non-corresponding to CpG islands

Also the 14 clones classified as non-CpG islands (Fig.1) were sequenced from the EagI site. The majority resulted to be known repetitive sequences, containing a EagI site: eight were Alu sequences with a EagI site at their 5' end (Fig.2A), two (2-132 and 2-71) were also Alu sequences but the Eag I site was elsewhere in the sequence (Fig.3A). Clone 2-77 was a sequence in the middle of a viral LTR (Fig.3B).

In keeping with this finding, among the 7 clone predicted to contain almost exclusively repetitive sequences as they did not show hybridization to single bands in Southern Blots (Fig.1), 5 were the 5' region of Line sequences (Fig.3C). They were all about 3 kb in length, in agreement with the presence of a conserved EcoRI site in the middle of many Lines (16). Clone 2−19b was 98% identical to 7SL repetitive sequences with an EagI site at the 5' end, in the same position where it is found in the Alu sequence (Fig.3D).

Only 4 clones (2-21,2-65 2-128 and 2-104) did not correspond to a known repetitive sequence, but did not have any of the characteristics of a CpG island as well. Cosmids hybridizing to 2-104 have been isolated and the DNA flanking the EagI site on the opposite direction from 2-104 has been sequenced: it was an Alu sequence with a EagI site at the 5' end (not shown). The same may be therefore true for the remaining 3 unidentified clones.

DISCUSSION

Clusters of rare cutter restriction enzymes have been extremely useful to identify genes in cloned genomic DNA (17,18,19,20). In this paper we now show that using rare cutter restriction enzymes it is also possible to construct genomic libraries enriched in fragments flanking genes: we have analyzed in detail a genomic library of EagI-EcoRI fragments of the Xq24-Xqter portion of the human genome. We show that 60% of the EagI clones correspond to CpG islands. The remaining clones were almost exclusively repetitive sequences containing the rare cutter site. We had previously shown that the inserts of the library were enriched in sites for rare cutter restriction enzyme (8). We have now studied the methylation state of the genomic DNA surrounding the EagI site identified by our clones. The characteristic pattern of DNA methylation of the CpG islands at the 5' end of genes of the human X chromosome has enabled us to clearly distinguish between EagI sites in a CpG island versus isolated EagI sites. In most cases the genomic DNA surrounding the EagI site was demethylated in male DNA. In females, the pattern of DNA methylation also showed higher molecular weight bands and was compatible with the presence of a unmethylated allele on the active X and methylation of the allele on the inactive X chromosome. Seven clones did not show such pattern but only DNA demethylation. The isolation of cosmid clones and of probes at the opposite side from EagI has demonstrated that also in 5 of them methylation differences between male and female DNA can be demonstrated. Thus, 32 clones have the characteristics of CpG islands of the X chromosome.

In keeping with our definition of CpG islands is the finding that in most cases the base composition of the DNA flanking the EagI site is G+C and CpG rich and that the nucleotide sequence of four of the clones corresponds to known CpG islands (HPRT, G6PD, and a previously described one 3' from the GdX gene) (21,22) or to the 5' of known genes (LAMP2) (23). Moreover, the remaining clones, which were defined as non-CpG islands, were known repetitive sequences containing a EagI site: Alu (24), Lines (16) and other less abundant sequences: 7SL (25) and LTR (26). In most cases they appeared methylated in leukocyte DNA. Alu sequences and the 5'end of the Lines are known as highly methylated CpG rich sequences (27, 28). However, the methylation status of such DNAs must be not highly conserved since they must be at least in part unmethylated in the X3000-1 DNA used to make the library. The X3000-1 DNA, however, was not analyzed.

The high percentage of known repetitive sequences containing a EagI site found in our library can be advantageously used to identify CpG islands. To isolate the remaining CpG islands from this region of the genome, we are constructing new libraries using EagI as well as other restriction enzymes with similar characteristics. Sequence or PCR analysis with repetitive sequence specific primers will be the first screening step. Once repetitive sequences flanking the rare cutter site have been identified more than 80% of the clones will be CpG islands. It has to be pointed out that our results also indicate that the usefulness of such libraries mainly depends from the ratio between CpG islands and unmethylated (or variably methylated) rare cutter containing repetitive sequences in the DNA to be cloned. This may be a very important parameter, since CpG islands are not uniformly distributed along chromosomes (29, 8) and it has been critical in the cloning of most of the CpG islands of Xq28 (Maestrini et al., in preparation).

The methylation analysis of such a large number of CpG islands of the X chromosome definitely demonstrates that demethylation of the active and methylation of the inactive X chromosome are a general rule. In addition, methylation of CpG islands has been shown to be responsible of the transcriptionally inactive state of X chromosome genes (13). Thus, the finding, in all instances, of methylation differences between male and female DNA is in very strong support of the notion that most genes of the distal long arm of the X chromosome are subject to X inactivation. This is in contrast with the situation in the short arm or proximal long arm where many genes remain active also on the inactive X chromosome (30).

In conclusion we have now 28 new probes for CpG islands of the X chromosome, mainly localized in Xq24–25 and Xq28 (8). The precise physical mapping of these and of new gene tags will be the next step in the construction of the complete map of CpG islands of this chromosomal region.

ACKNOWLEDGEMENTS

We thank Dr.A.Lisa for help with the computers and Dr.G.Camerino for critical reading of the manuscript. This work was supported by Progetti Finalizzati CNR 'Ingegneria Genetica' e 'Biotecnologie e Biostrumentazione' to D.T. and P.V. and by Telethon Italy and the Associazione Italiana per la Ricerca sul Cancro to D.T.

REFERENCES

1. Adams,M.D., Kelley,J.M., Gocayne,J.D., Dubnick,M., Polymeropoulos,M.H., Xiao,H., Merril,C.R., Wu,A., Olde,B., Moreno R.F., Kerlavage,A.R., McCombie,W.R., Venter,J.C. (1991) Science, 252, 1651–1656.
2. Bird,A.P. (1987) Trends Genet., 3, 342–347.
3. Gardiner-Garden,M., Frommer,M. (1987) J.Mol.Biol., 196, 261–282.
4. Bird,A.P. (1989) Nucleic Acids Res., 17, 9485.
5. McKusick,V.A. (1990) Mendelian inheritance in man. Ninth Edition. Johns Hopkins University Press.
6. Poustka,A., Dietrich,A., Langenstein,G., Toniolo,D., Warren,S.T., and Lehrach,H. (1991) Proc. Natl. Acad. Sci. USA., 88, 8302–8306.
7. Schlessinger,D., Little,R.D., Freije,D., Abidi,F., Zucchi,I., Porta,G., Pilia,G., Nagaraja,R., Johnson,S.K., Yoon,J.Y., Srivastava,A., Kere,J., Palmieri,G., Ciccodicola,A., Montanaro,V., Romano,G., Casamassimi,A., and D'Urso,M. (1991) Genomics, 11, in press.
8. Maestrini,E., Rivella,S., Tribioli,C., Purtilo,D., Rocchi,M., Archidiacono,N., and Toniolo,D. (1990) Genomics, 8, 664–670.
9. Nussbaum,R.L., Airhart,S.D., and Ledbetter,D.H. (1986) Amer. J. Med. Genet., 23, 457–466.
10. Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Molecular cloning, a laboratory manual. Second Edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
11. Ansorge,W., Sproat,B. (1987) Nucleic Acids Res., 15, 4593–4602.
12. Pearson and Lipman (1988) Proc. Natl. Acad. Sci. USA., 85, 2444–2448.
13. Grant,S.G., Chapman,V.M. (1988) Ann. Rev. Genet., 22, 199–233.
14. Lehrach,H., et al. (1990) In Davies,K.E. & Tilghman,S.M. (eds.), Genome Analysis. Genetic and Physical Mapping. Cold Spring Harbor Laboratory Press. Cold Spring Harbor. Vol.1. pp 39–81.
15. Dietrich,A., Kioschis,P., Monaco,A.P., Gross,B., Korn,B., Williams,S.V., Sheer,D., Heitz,D., Oberle,I.,Toniolo,D., Warren,S.T., Lehrach,H., and Poustka,A. (1991) Nucleic Acids Res., 19, 2567–2572.
16. Crowther,P.J., Doherty,J.P., Linsenmeyer,M.E., Williamson,M.R., and Woodcock,D.M. (1991) Nucleic Acids Res., 19, 2395–2401.
17. Estivill,X., Farrall,M., Scambler,P.J., Bell,G.M., Hawley,K.M.F., Lench,N.J., Bates,G.P., Kruyer,H.C., Frederick,P.A., Stanier,P., Watson,E.K., Williamson,R., & Wainwright,B.J. (1987) Nature, 326, 840–845.
18. Bates,G.P., MacDonald,M.E., Baxendale,S., Youngman,S., Lin,C., Whaley,W.L., Wasmuth,J.J., Gusella,J.F., and Lehrach,H. (1991) Am. J. Hum. Genet., 49, 7–16.
19. Heitz,D., Rousseau,F., Devys,D., Saccone,S., Abderrahim,H., Le Paslier,D., Cohen,D., Vincent,A., Toniolo,D., Della Valle,G., Johnson,S., Schlessinger,D., Oberlè,I., Mandel,J.L. (1991) Science, 251, 1236–1239.
20. Verkerk,A.J.M.H., Pieretti,M., Sutcliffe,J.S., Fu,Y.H., Kuhl,D.P.A., Pizzuti,A., Reiner,O., Richards,S., Victoria,M.F., Zhang,F., Eussen,B.E., van Ommen,G.J.B., Blonden,L.A.J., Riggins,G.J., Chastain,J.L., Kunst,C.B., Galjaard,H., Caskey,C.T., Nelson,D.L., Oostra,B.A., and Warren,S.T. (1991) Cell, 65, 905–914.
21. Patterson,M., Schwartz,C., Bell,M., Saver,S., Hofker,M., Trask,B., van den Engh,G., and Davies,K.E., (1987) Genomics, 1, 297–306.
22. Filippi,M., Tribioli,C., and Toniolo,D. (1990) Genomics, 7, 453–457.
23. Manoni,M., Tribioli,C., Lazzari,B., DeBellis,G., Patrosso,C., Pergolizzi,R., Pellegrini,M., Maestrini,E., Rivella,S., Vezzoni,P., and Toniolo,D. (1991) Genomics, 9, 551–554.
24. Kariya,Y., Kato,K., Hayashizaki,Y., Himeno,S., Tarui,S., and Matsubara,K. (1987) Gene, 53 1–10.
25. Ullu,E., and Weiner,A.M. (1984) EMBO J., 3, 3303–3310.
26. Harada,F., Tsukada,N., and Kato,N. (1987) Nucleic Acids Res., 15, 9153–9162.
27. Hohjoh, H., Minakami, R. and Sakaki, Y. (1990) Nucleic Acid Res. 18, 4099–4104
28. Schmid, C.W. (1991) Nucleic Acids Res. 19, 5613–5617
29. Gardiner,K., Horisberger,M., Kraus,J., Tantravahi,U., Korenberg,J., Rao,V., Reddy,S., and Patterson,D. (1990) EMBO J., 9, 25–34.
30. Davies,K. (1991) Nature, 349, 15.
31. Maestrini,E., Rivella,S., Tribioli,C., Rocchi,M., Camerino,G., Santachiara-Benerecetti,S., Parolini,O., Notarangelo.L.D.and Toniolo,D. (1992) Am.J.Hum.Genet. in press.