Identification of highly methylated human DNA sequences using a methyl-CpG binding domain column

Graham J. R. Brock

Thesis presented for the Degree of Doctor of Philosophy
at the University of Edinburgh
1997

Dedication

This thesis is dedicated to Julia and Kerry

Acknowledgements

Hopefully this thesis will soon be out of the way. What with the new job, it's time to move on. Time then to say thanks, naturally most go to Adrian for giving me this opportunity. It hasn't all been fun but overall it has definitely been worthwhile.

- Thanks must also go to all the members of the Bird Lab both past and present for all their help and advice.
- In particular Donald Macleod for reading various drafts and for advice and friendship during the past three years.
- Thanks as well to Julia for adding commas, asking questions and giving advice, the best is yet to come this year!
- Thanks to Wendy Bickmore and all in her lab for help with the *FISH* experiments.
- The Blood Pressure Unit at Glasgow University and Lisa Strain at the Human Genetics Unit of the WGH for providing DNA.
- Christine Struthers for organising the bird lab, Joan Davidson and Aileen Greig for providing all the solutions. Patrick , Simon and Dasha my fellow sufferers, (soon boys and girls soon).

- Thanks also to my Parents for being my parents, tough job but someone had to do it. Also John Brock, Art Burton and Peter Kao we'll never know if I should have stayed!!

- Finally its not all been work (maybe that shows) during the last three years so; for help in the hills, Drew, Ian, Colin, Bob and Rufus. When Saturday comes "way down in Gorgie, at the temple of the Maroon Magicians" Derek, Kevin, Brian and Ian again! Finally for pulling plastic when Swanny's was no more, Nigel, Anne and Simon.

Table of Contents.

	Page
Title	i
Declaration	ii
Dedication	iii
Acknowledgements	iv
Table of Contents	v
Abstract	x
Abbreviations	xi

Chapter 1 : General Introduction

		Page
1.1.	The methylation of cytosine in vertebrate genomes (m^5CpG)	1
1.2.	The distribution of m^5CpG in vertebrate genomes	9
1.3.	Isochores, chromosome bands and m^5CpG	11
1.4.	Unmethylated regions of the genome (CpG islands)	12
1.5.	The methylation of CpG islands	18
1.5.1.	The process of X-chromosome inactivation	18
1.5.2.	The process of parental imprinting	20
1.6.	Transcriptional repression through methylation of CpG	21
1.7.	Research objectives	22

Chapter 2 : Materials and methods

2.1.	Commonly used reagents and buffers	26
2.2.	Restriction enzyme analysis	27
2.3.	Agarose gel electrophoresis of nucleic acids	28
2.3.1.	Purification of Nucleic Acid fragments	28
2.4.	Purification of plasmid DNA	33
2.5.	Preparation and analysis of genomic DNA	37
2.6.	Polymerase chain reaction (PCR)	38
2.6.1.	Reaction conditions and primers for ZFX, ZFY and COMMON reactions	38
2.6.2.	Reaction conditions and primers for monitoring the fractionation of human DNA	39
2.6.3.	Reaction conditions for catch-linkers	40
2.6.4.	Reaction conditions for production of *FISH* probes	41

2.6.5.	Reaction conditions and protocol for cycle sequencing reactions	42
2.7.	Methylation of plasmid and genomic DNA	43
2.8.	Southern blot analysis of DNA	43
2.9.	Random prime labelling of DNA probes	45
2.9.1.	End-labelling of plasmid and DNA probes	48
2.10.	Preparation of competent cells	48
2.10.1.	Electroporation of plasmids into competent cells	49
2.11.	Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE)	50
2.12.	Attaching catch-linkers to the fractionated DNA	51
2.13.	Cloning of amplified DNA using the T-vector system	54
2.14.	Sequence analysis of cloned inserts	55
2.15.	Fluorescent in situ hybridisation	65
2.15.1.	*In situ* hybridisation of probes to chromosome spreads	66

Chapter 3 : MBD Column Preparation

3.1.	The methyl-CpG binding domain column	68
3.2.	Production and purification of the MBD protein	69
3.3.	Attaching the MBD protein to the nickel-agarose matrix.	71
3.4.	Testing the MBD column to determine methyl-CpG binding ability	76
3.4.1.	Binding of plasmids and plasmid fragments to the MBD column	83

3.5. Estimating the proportion of highly methylated DNA in the genome 83
3.6. Fractionating human DNA using the MBD column 86
3.7. Conclusion 92

Chapter 4 : Libraries of highly methylated MseI fragments from human blood DNA

4.1. Introduction 94
4.2. Analysis of the first library (MBDx3) 95
4.2.1. Conclusions concerning the MBDx3 library 95
4.3. Preliminary analysis of enriched sequences derived from male and female blood DNA (MBDx5) 101
4.3.1. Conclusions concerning preliminary analysis of enriched sequences (MBDx5) 104
4.4. Further analysis of enriched sequences derived from female blood DNA (MBDx5) 108
4.4.1. Conclusions regarding further analysis of enriched sequences from the MBDx5 library 117
4.4.2. Southern blot and sequence analysis of clones from the female MBDx5 library 120
4.5. Results 121
4.5.1. Conclusions 131

Chapter 5 : Fluorescent in situ hybridisation

5.1. Introduction 137
5.2. Results 146
5.2.1. Conclusions 148

Chapter 6 : The methylation status of the rDNA NTS

6.1. Introduction 150

6.2. Southern Blot analysis of the rDNA NTS 151

6.3. Conclusions 158

Chapter 7 : Discussion

7.1. Fractionation of the genome according to m^5CpG frequency 164

7.2. Characterisitics of the further purified DNA sequences 165

7.3. Distribution of cloned sequences on metaphase chromosomes 166

References 171

Appendix A 190

Appendix B 199

Appendix C 201

Abstract

A library of highly methylated sequences has been constructed from DNA derived from human blood. Genomic DNA was fractionated using a column which binds preferentially according to methyl-CpG frequency (Cross et al., 1994). The library contains known fragments of methylated CpG Islands (CGIs) and other novel regions with high numbers of m^5CpG. Analysis of cloned inserts demonstrated that most were more GC-rich and had a higher $CpG_{Obs/Exp}$ than bulk genomic DNA. Cloned inserts were used as probes against southern blots of DNA digested with methylation sensitive or insensitive isoschisomers. The hybridisation patterns demonstrated that majority of the cloned sequences tested were methylated in DNA when derived from blood. One of the most frequently occurring clones in the library matched to a fragment from the rDNA repeat non transcribed spacer (NTS). This entire region has subsequently been shown to be heavily methylated in DNA from somatic cells. The cloned fragments derived from both male and female DNA were used as probes in Fluorescent *in situ* Hybridisation (FISH) experiments. DNA derived from both male and female gave indistinguishable hybridisation patterns with many fragments apparently hybridising to subtelomeric regions. The other regions of significant hybridisation were to the short arms of the acrocentrics and the centromere of chromosome 9. The rDNA repeats are located on the short arms of the acrocentric chromosomes while the centromere of chromosome 9 has been reported to contain low copy number methylated repeats. These sequences, though methylated in DNA from somatic cells, are often unmethylated in sperm. This apparent lack of germ-line methylation may account for the high frequency of CpG in these regions.

Abbreviations

AdoMet	S-adenosyl-L-methionine
Ab	antibody
bp	base pair
CCD	charge coupled device
CGI(s)	CpG Island(s)
$CpG_{Obs/Exp}$	CpG Observed/Expected.
cpm	counts per minute
Dapi	4,6-Diamidino-2-phenylindole
ddH_20	double distilled water
dATP	deoxyadenosine triphosphate
dCTP	deoxycytidine triphosphate
dGTP	deoxyguanosine triphosphate
dTTP	deoxythymidine triphosphate
dNTPs	all four deoxy triphosphates
DMSO	dimethyl sulphoxide
DTT	dithiothreitol
DNA	deoxyribonucleic acid
EDTA	ethylenediaminetetraacetic acid
EtBr	ethidium bromide
EtOH	ethanol
FISH	fluorescent *in situ* hybridisation
Hepes	N-(Hydoxyethyl)piperazine-N'-[2-ethanosulphonic acid]
IPTG	isopropylthio-β-D-galactosidase
K_2HPO_4	potassium hydrogen phosphate
kb	kilo base
m^5C	5-methylcytosine
m^5CpG	methylated CpG
MBD	methyl binding domain
MeCPs	methyl-CpG binding proteins
$MgSO_4$	magnesium sulphate
min	minutes
ms	milliseconds
MTases	methyltransferases
Mwt	molecular weight

NaCl	sodium chloride
$(NH_4)_2SO_4$	ammonium sulphate
NTS	non-transcribed spacer
PCR	polymerase chain reaction
PMSF	phenylmethylsulfonyl flouride
SDS	sodium dodecyl sulphate
SDS-PAGE	SDS-polyacrylamide gel electrophoresis
TAE	tris acetate EDTA buffer
TE	tris-EDTA buffer
TEMED	N,N,N',N'-tetra-methylethylenediamine
(v/v)	volume : volume ratio
(w/v)	weight : volume ratio
X-Gal	5-bromo-4-chloro-3-indolyl-β-galactopyranoside

Chapter 1 : General Introduction

1.1.

The methylation of cytosine in vertebrate genomes (m^5CpG)

Deoxyribonucleic acid (DNA) consists of four bases; adenine, cytosine, guanine and thymine. The methylation of vertebrate DNA *in vivo* is a modification to cytosine catalysed by a cellular enzyme or enzymes. The methylation of cytosine involves the enzyme catalysed transfer of a methyl group from S-adenosyl-L-methionine (AdoMet) to the C-5 position on a cytosine ring (Figure 1.1.A). In vertebrates this occurs when the cytosine is followed by a guanine (5'-CG-3') in a CpG dinucleotide. Cytosine methylation produces a protrusion from the major groove of DNA which may have the affect of altering local protein recognition signals (Eden and Cedar, 1994). However, attachment of a methyl group does not alter inter-strand base pairing with guanine as there are still three hydrogen bonds. In vertebrates the modification of newly synthesised CpG dinucleotides takes place during the (S) phase of the cell cycle and may take several hours to complete (Adams and Burdon, 1985). Following replication, the unmethylated CpG in the daughter strand will be paired with the methylated CpG (m^5CpG) of the parental strand (Figure 1.1.B). These hemi-methylated DNA sequences are thought to be the preferred substrate for mammalian DNA methyltransferases (MTases). The mammalian MTases are related to bacterial methyltransferases which also modify cytosine bases but have a more complex sequence specificity (Bestor et al., 1988). In mammals, the best studied are the maintenance MTases which act upon hemi-methylated sequences (Gruenbaum et al., 1982). A pattern of DNA methylation once established will then be maintained through many

Figure 1.1.

The modification of cytosine by addition of a methyl-group

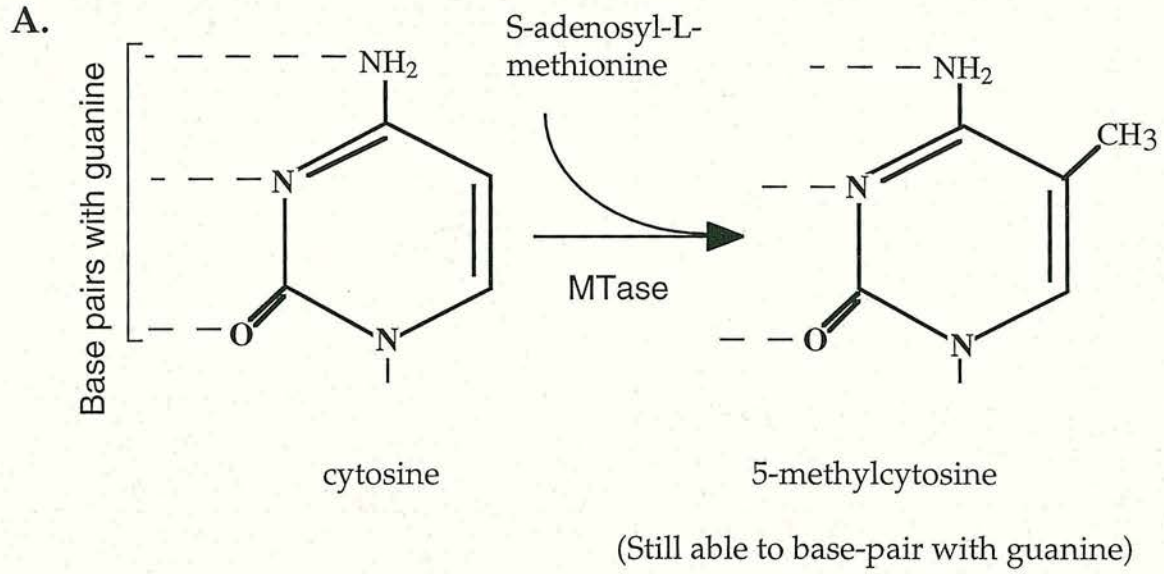
A.

The addition of a methyl group (CH_3) to the C-5 position of the DNA base cytosine to produce 5-methylcytosine. The reaction is enzymatic, with S-adenosyl-L-methionine (Ado Met) as the methyl group donor.

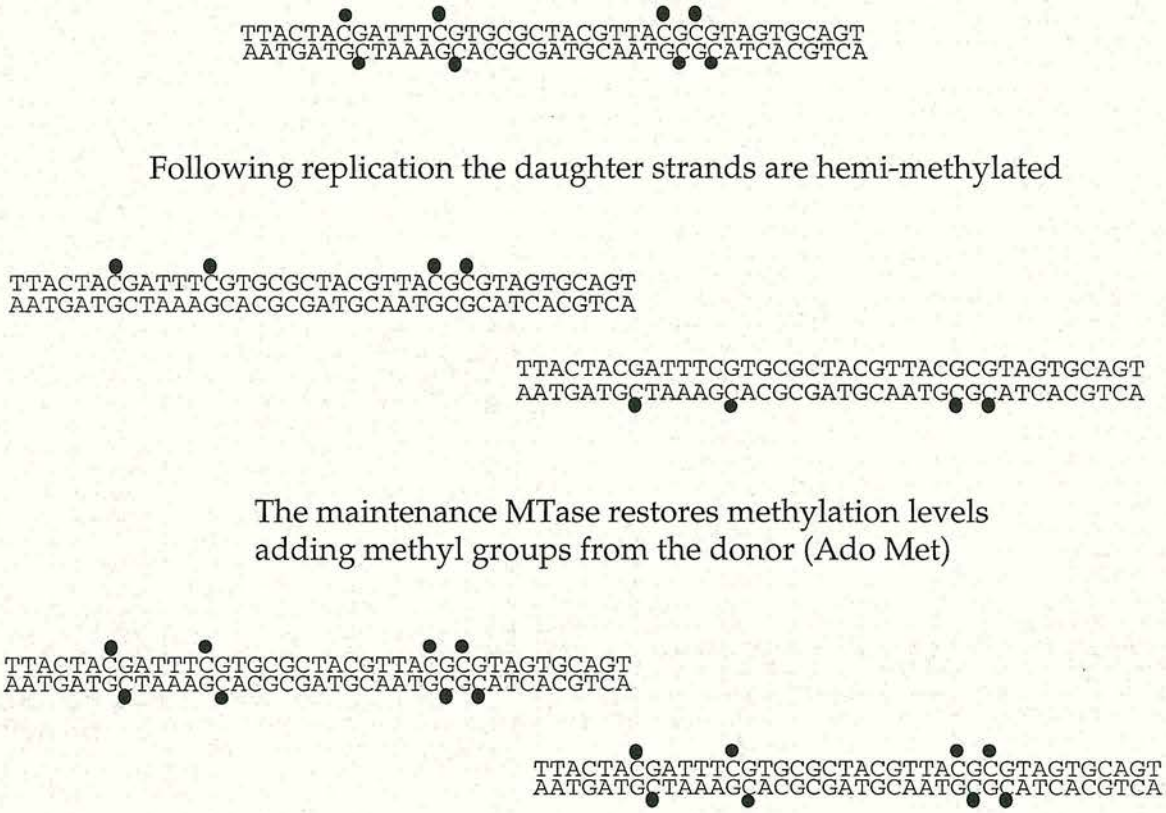
B.

Action of mammalian maintenance methyltransferases (MTase) on newly replicated DNA. In the original and daughter strands m5CpG is represented by a black solid circle. The newly synthesised strand has no methyl groups but is paired with a m5CpG from the original strand. This hemi-methylated DNA is thought to be the preferred substrate of mammalian maintenance MTases.

Figure 1.1.


A.
S-adenosyl-L-methionine
MTase
Base pairs with guanine
NH2
N
O
N
CH3
cytosine
5-methylcytosine
(Still able to base-pair with guanine)

B


TTACTACGATTTCGTGCGCTACGTTACGCGTAGTGCAGT
AATGATGCTAAAGCACGCGATGCAATGCGCATCACGTCA
Following replication the daughter strands are hemi-methylated
TTACTACGATTTCGTGCGCTACGTTACGCGTAGTGCAGT
AATGATGCTAAAGCACGCGATGCAATGCGCATCACGTCA
TTACTACGATTTCGTGCGCTACGTTACGCGTAGTGCAGT
AATGATGCTAAAGCACGCGATGCAATGCGCATCACGTCA
The maintenance MTase restores methylation levels
adding methyl groups from the donor (Ado Met)
TTACTACGATTTCGTGCGCTACGTTACGCGTAGTGCAGT
AATGATGCTAAAGCACGCGATGCAATGCGCATCACGTCA
TTACTACGATTTCGTGCGCTACGTTACGCGTAGTGCAGT
AATGATGCTAAAGCACGCGATGCAATGCGCATCACGTCA

generations with high fidelity (Figure 1.1.B). Establishing the pattern of methylation may require an MTase which can act *de novo* on DNA that is not hemi-methylated. This *de novo* DNA MTase would set the original methylation pattern which would then be maintained through subsequent generations by the maintenance MTase (Bestor and Verdine, 1994). Experiments have shown that the modification of cytosine by methylation is vital for the correct function of the vertebrate genome (Li et al., 1992). Transgenic mice homozygous for a mutated maintenance MTase had levels of m^5CpG approximately 30% of wild type. The resulting transgenics were stunted and failed to develop past mid-gestation (Li et al., 1992). It is probable that cytosine methylation is involved in various functions, including transcriptional control and the disruption to this regulatory process proves fatal (Section 1.6).

In vertebrates more than 80% of the CpG dinucleotides are modified through methylation. Why such modifications are so widespread is unclear, but, it has been suggested that as vertebrates evolved the proportion of their genome that was methylated has increased (Adams and Burdon, 1985; Bird et al., 1979). Although widespread and vital in vertebrates, some organisms, for example, sea-urchin only methylate 40% of their genome (Bird et al., 1979). No methylation has been detected in other organisms, for example *Drosophila melanogaster* (Urieli-Schoval et al., 1982) yeast and some non-vertebrates (Bird et al., 1979).

Although it is widespread in vertebrates, CpG is relatively rare, occurring with about one fifth of the frequency expected from its base composition (Swartz et al., 1962). One explanation for this relative rarity of CpG is as a result of its frequent methylation. Deamination of cytosine leads to the formation of uracil (Figure 1.2.B) but the resulting mismatch is recognised

and replaced by DNA repair mechanisms (Gates and Linn, 1977). Deamination of 5-methylcytosine leads to the formation of thymine (Coulondre et al., 1978) (See Figure 1.2.A) which as a 'normal' DNA base is not efficiently repaired (Bird and Taggart, 1980). There are examples of methylation and deamination leading to the depletion of CpG in the vertebrate genome. The sequences of the methylated α-globin pseudogene and the almost completely unmethylated and functional α-globin gene were compared (Bird et al., 1987). These two share almost 75% sequence homology overall but only 4 out of 70 CpGs in the gene are present in the pseudogene. Most have been replaced by TpG or CpA, the predicted outcome of m^5CpG deamination (Bird et al., 1987). In a second example, a comparison was made of CpG distribution. This was in the coding regions of 121 genes of six species, three with methylated genomes, three without (Schorderet and Gartler, 1992). The authors concluded that the overall base composition indicated that all species exhibit CpG suppression but that levels are higher in methylated genomes. They found that almost 90% of CpGs in species with non-methylated genomes had mutated in the homologous region in species with methylated genomes. In the majority of cases the mutation had changed CpG to TpG or CpA (Schorderet and Gartler, 1992). It is possible that deamination of m^5CpG could eventually result in the complete loss of CpG from vertebrate genomes. However the accelerated loss of m^5CpG through deamination is thought to be held in equilibrium by spontaneous point mutations that create new CpG dinucleotides (Sved and Bird, 1990).

.Calculations made to determine CpG depletion assumes that each base has an equal statistical chance of occurring in a given stretch of DNA. When the ratio of observed/expected (Obs/Exp) for the dinucleotide CpG

Figure 1.2.

Deamination reactions of 5-methyl cytosine and cytosine

A.

The deamination of 5-methylcytosine which gives rise to thymine. This is not efficiently recognised and rectified by the DNA repair mechanisms.

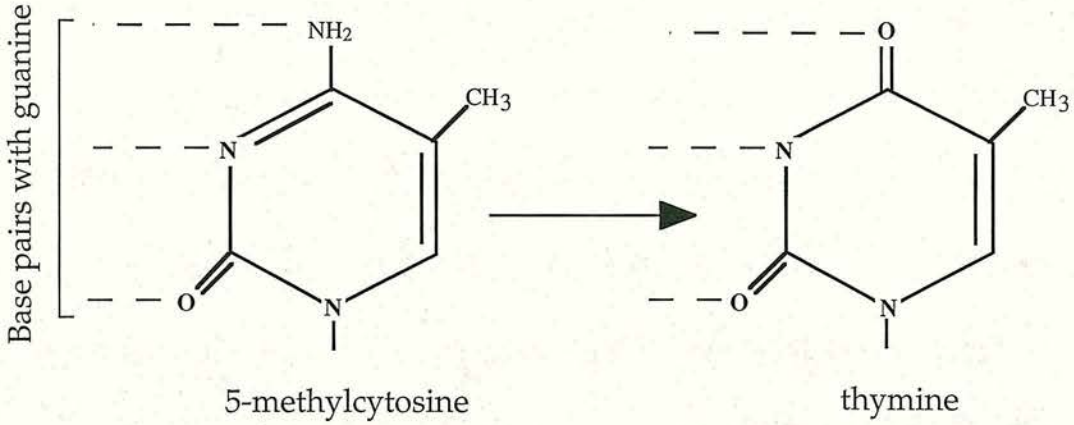
B.

The deamination of cytosine which gives rise to uracil. The resulting mismatch is recognised and replaced by DNA repair mechanisms.

Figure 1.2.

Deamination of 5-methylcytosine

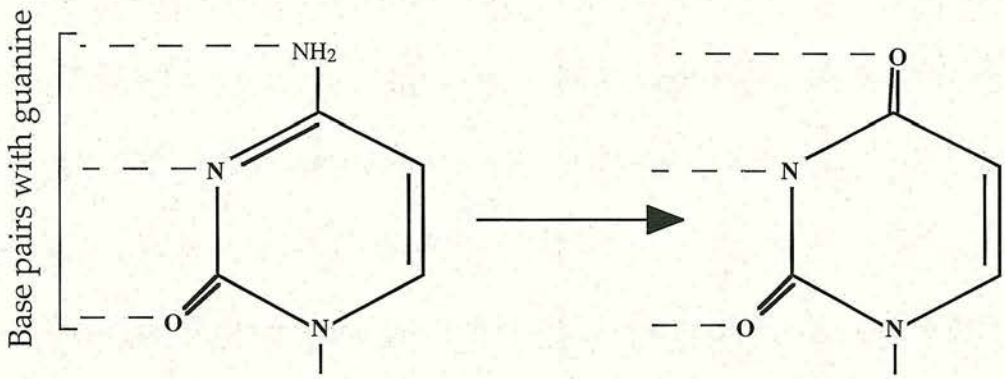
A



5-methylcytosine thymine

Deamination of cytosine

B



cytosine uracil

is calculated using the following formula it (CpG) appears to be depleted in the majority of the genome.

$$CpG_{Obs/Exp} = \frac{\text{Number of CpG}}{\text{Number of C x number of G}} \times N$$

Where N is the number of nucleotides in the sequence analysed (Gardiner-Gardner and Frommer, 1987).

It is estimated that one-third of all human diseases caused by point mutations are attributable to CpG - TpG transitions (Cooper and Youssoufian, 1988). One possible explanation for methylation of vertebrate genomes and the consequent increased mutational risk may be the benefit of better regulation. Vertebrate genomes contain numerous cryptic promoters that could cause widespread and spurious transcription (Bird, 1993). An example of low density methylation repressing transcription was demonstrated using a cloned γ-globin gene promoter. Upon transfection into HeLa cells the promoter was fully repressed by methylation of one CpG per 126 base pairs (Boyes and Bird, 1992). Another model system involved the artificial methylation of a transcription unit and its introduction into cells. Methylation was shown to dramatically reduce the transcription from the introduced fragment of DNA (Vardimon et al., 1982; Busslinger et al., 1983). The low level genome wide methylation of CpG may therefore be enough to prevent illegitimate transcription (see also Section 1.6).

1.2.

The distribution of m^5CpG in vertebrate genomes

The frequency of m^5CpG is on average one per 100 bp throughout 98% of the genome. However antibodies raised against 5-methylcytosine (m^5C) have demonstrated that its distribution is not even across all 46 human chromosomes. Regions of dense methylation occur on the secondary constrictions (near the centromere) of chromosomes 1, 9 and 16 and on the distal part of the long arm of Y (Miller et al., 1974). Experiments, employing an improved technique, also using a m^5C antibody, revealed additional locations of 5-methyl cytosine density (Barbin et al., 1994). The authors divided the distribution into four types. Type 1 agreed with the previous study by Miller et al. (1974) with additional sites at the juxtacentromeric regions of 2, 7, 10 and 17. These sites contain classical satellite DNA and repetitive sequences which are often methylated (Kokalj-Vokac et al., 1993). Type II binding was to the juxtatelomeric and intercalary bands that correspond to the T-bands of chromosomes (Barbin et al., 1994).The thermal denaturation resistant or T-bands are a subset of the reverse staining (R)-bands (Dutrillaux, 1977). Such regions are thought to be very GC-rich with a high gene concentration (Saccone et al., 1996) (see Section 1.3) and are also rich in CpGs. This is possibly due to a high concentration of unmethylated CpG islands in these regions (Ferraro et al., 1993)(see Section 1.4). The study by Barbin *et. al.* indicated that these areas additionally contain a high density of methylated CpGs (Barbin et al., 1994). Investigations of the telomeric regions of chromosomes (De Lange et al., 1990) may partially explain the Type II binding observed by Barbin *et al* (Barbin et al., 1994). A subtelomeric repeat was identified which was present at 10-25% of chromosome ends in the human genome (De Lange

et al., 1990). The minimal size of the repeat was 4 kb and it had a high GC content of 80%. However, unlike the majority of CpG islands, these repeats were extensively methylated in somatic cells (De Lange et al., 1990). Presumably this repeat would have been bound by the m^5C antibody. Other sequences, occurring near the telomeres, contained variable numbers of the 29 bp repeat as part of a larger GC-rich repeat (Brown et al., 1990; Cheng et al., 1990)(see also Section 1.3). Further sequences, showing enough similarity to the 29 bp repeat to suggest they were members of a family, have also been reported. These were also located in the subterminal regions of many human chromosomes (Cross et al., 1990). The Type III binding revealed a weak labelling of R-bands that was both low in intensity and uneven. It was thought to indicate short methylated CpG clusters amongst the sporadic GC-rich segments (Barbin et al., 1994). Finally Type IV binding was polymorphic and exhibited a wide range of fluorescent intensity on the short arms of the acrocentric chromosomes (Barbin et al., 1994). The authors noted that the acrocentric chromosomes contain various amounts of repeated DNA which might account for the staining pattern (Barbin et al., 1994). It has recently been shown that the rDNA repeat unit contains a large number of methylated CpGs (Brock and Bird 1997 and Chapter 6). The repeat unit is 43 kb in length and tandemly repeated at around 400 copies spread over the five acrocentric chromosomes. The transcribed sequence is comprised of 13 kb that is both GC and CpG rich and almost completely unmethylated. The remainder of the rDNA repeat is approximately 30 kb long and highly methylated with a slight CpG depletion (Brock and Bird, 1997). These highly methylated sequences clustered on the short arms of the acrocentric chromosomes may account for the hybridisation pattern seen in the Type IV binding (Barbin et al., 1994).

1.3.

Isochores, chromosome bands and m^5CpG

The Type II binding of the 5-methylcytosine antibody was localised in the juxtatelomeric and intercalary bands (Barbin et al., 1994). These corresponded to the T-bands or thermal denaturation resistant bands which occur at the telomeres on most chromosomes (reviewed by Holmquist, 1992). The T-bands had previously been reported to be both GC-rich and to have a high frequency of CpG islands (CGIs) (Section 1.4) (Holmquist, 1992; Saccone et al., 1992). As a result it was proposed that the "H3 isochore" would contain DNA which occurred in the T-bands of metaphase chromosomes (De Sario et al., 1991). Isochores are classified according to their GC content, with the GC-poorest isochores being LI and L2 which together comprise 60% of the genome. Isochores H1 and H2 form 10% and 20% of the genome respectively and are more GC-rich than L1 and L2. The GC-richest isochore is H3 which represents approximately 5% of the genome, with satellite sequences and rDNA making up the remaining 5% (Bernardi, 1993). In addition to being GC-rich, the H3 isochore has the highest concentration of genes and CGIs and the highest transcriptional and recombinational activity (Aissani and Bernardi, 1991a; Aissani and Bernardi, 1991b; Saccone et al., 1992). The apparent density of m^5CpG in these regions may arise as a consequence of the high GC content of T-bands. The apparent lack of CpG depletion may be attributable to the structure of chromosome bands or a selective pressure for GC-rich sequences. Alternatively, a lack of methylation in the germ-line may result in a higher frequency of CpG than that seen on the remainder of the chromosomes.

1.4.
Unmethylated regions of the genome (CpG islands)

There are distinct regions of DNA in which the dinucleotide CpG is not depleted and occurs at about its expected frequency from base composition. In the mouse these stretches of DNA were identified due to the presence of numerous *Hpa*II sites (CCGG) (Bird et al., 1985). As a result they were originally called *Hpa*II Tiny Fragment Islands (HTF Islands) (Cooper et al., 1983) and are now usually referred to as CpG islands (CGIs) (Bird, 1986; Gardiner-Gardner and Frommer, 1987). The number of these CGIs in the human genome is estimated at approximately 45,000 (Antequera and Bird, 1993). These amount to less than 2% of the total genome and usually cover the 5' end of genes but have been observed at both ends and even at the 3' end alone (Gardiner-Gardner and Frommer, 1987). They are on average 1 kb long and when located at the 5' end include the start of transcription and the first exon of approximately 60% of vertebrate genes. These include a number of the tissue specific and all constitutively expressed (housekeeping) genes (Gardiner-Gardner and Frommer, 1987; Larsen et al., 1992). In addition to the common characteristics of size, GC-richness and CpG frequency, they ordinarily lack methylation in all cells types (Bird et al., 1985). The lack of methylation in the germ-line is thought to account for the absence of CpG suppression. Spontaneous deamination from m^5CpG to TpG does not occur and the mutation will not be inherited.

The location of a number of CGIs has been reported in T-bands which are thought to be more GC-rich than other chromosome bands (Section 1.3). Experiments with antibodies for m^5CpG also indicates a high density of

methylated dinucleotides in these regions (Section 1.3). As one characteristic of CGIs is their lack of methylation, there appears to be a mechanism(s) which keeps them free of methylation in an otherwise heavily methylated genome. Transcription factor binding sites at the boundaries of a CGI may prevent access to the island and methylation by *de novo* MTases (Brandeis et al., 1994). Mutation of the binding site for the Sp1 transcription factor on the mouse *aprt* gene leads to the *in vivo* methylation of the 5' CpG island (Macleod et al., 1994). Alternatively, it has been suggested that CGIs are in some way actively demethylated after the genome wide wave of methylation that occurs early in development (Frank et al., 1991; Weiss et al., 1996). Whatever the process, it is clearly disrupted during X-inactivation when one of two identical, but randomly selected, chromosomes in the same cell has many of its CGIs modified through methylation (Section 1.5.1).

In the examples shown in Figures 1.3. and 1.4., sequences from the 5' end of four different genes are illustrated. The sequences shown include the start of transcription and first exon from genes both with and without CGIs. The depletion of CpG is noticeable in the 5' region of both the interferon alpha and the human G γ– and A γ-globin genes (Figure 1.3). These sequences are representative of the majority of the genome with a GC content of approximately 40% and a $CpG_{Obs/Exp}$ of 0.20 for both. In the second figure, the sequence of two genes with 5' CpG islands are shown, glucose-6-phosphate dehydrogenase (G6PD) and the human desmin gene. In these regions the frequency of the CpG dinucleotides is much closer to that of GpC and both also have a %GC of approximately 70 and an average Obs./Exp. around 0.85. The presence or absence of sites for two restriction enzymes is also plotted in the sequences shown.

Figure 1.3.

Examples of sequences which are CpG deficient and lack a CGI

A.

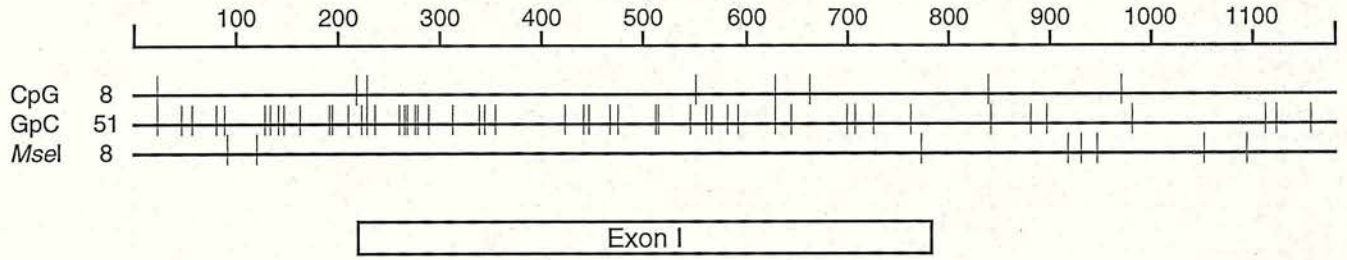
The sequence of the 5' end, start of transcription and entire first exon of the interferon alpha-d gene (IFNα) is shown (Accession No. J00210 and Mantei et al, 1980). The %GC of the entire 1179 bp shown is 42, the number of CpGs (8) and GpC (51) gives an Obs/Exp. of 0.15. There are eight sites for the enzyme *Mse*I (TTAA) but no sites for *Bst*UI (CGCG).

B.

The sequence for 2.4 kb of the duplicated human Gγ- and Aγ-globin genes is shown (Accession No. M91037 and Slightom et al, 1980). The region which covers the 5' end, start of transcription and first, second and third exons of both genes has a GC content of 41%. The dinucleotide CpG occurs 43 times with 144 GpCs giving the region an Obs/Exp of 0.19. The enzyme *Mse*I would cleave this sequence five times but the recognition sequence for *Bst*UI does not occur.

Figure 1.3.

5′ end of human IFNα-d Gene



Human Gγ-globin and Aγ-globin genes complete sequence

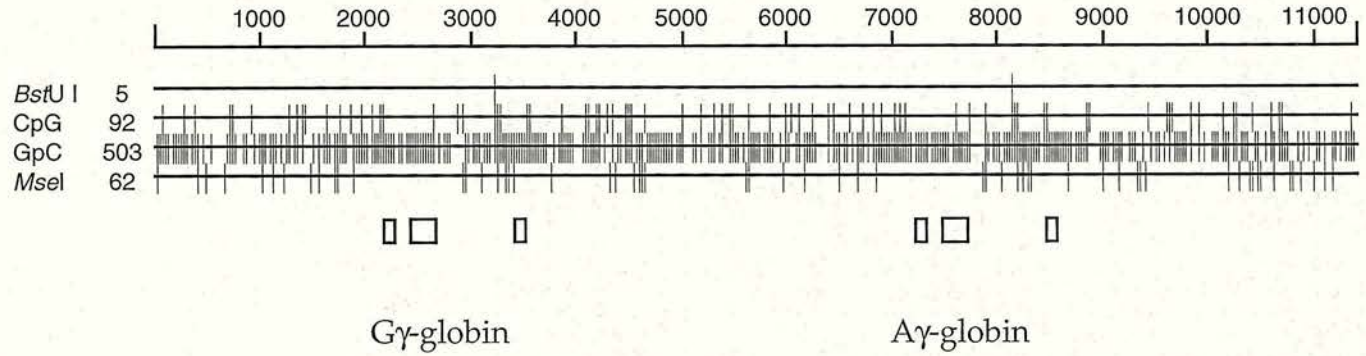


Figure 1.4.

Examples of two sequences both with CpG islands at their 5' end

A.

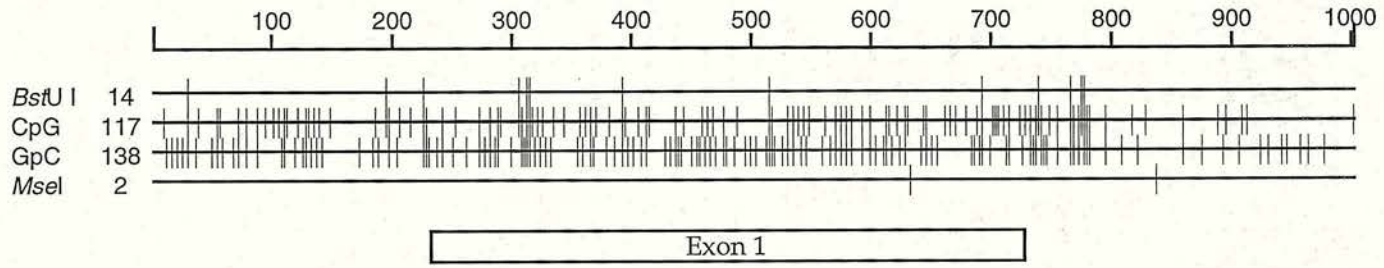
The first example shows 1.0 kb of sequence data from the 5' end of the human glucose-6-phosphate dehydrogenase (G6PD) gene (Accession No. X55448 and Chen et al, 1991). The region shown includes the start of transcription and the first exon with the relative frequencies of the two dinucleotides. The %GC of this region is 72 and 117 CpGs and 138 GpCs give an Obs/Exp of 0.90. The enzyme with the GC-rich recognition site (*Bst*UI) has a cluster of sites within the region covering the 5'end of the gene. The two *Mse*I sites both occur at the 3' end of the first intron.

B.

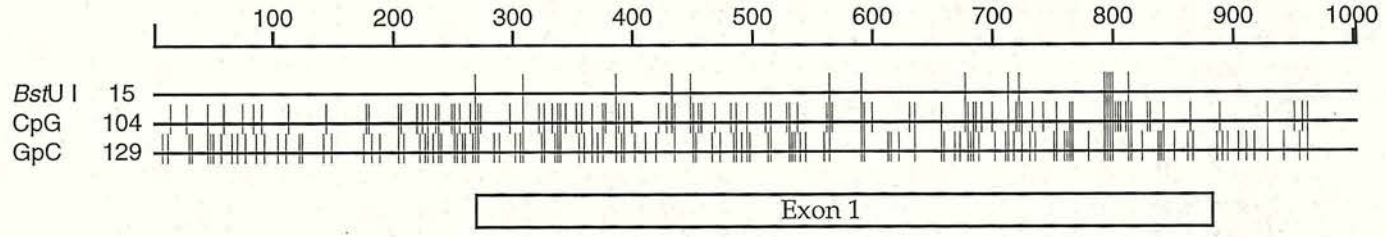
The first 1.0 kb of the 5' end of the human desmin gene showing the start of transcription and the first exon, (Accession No. M63391 and Li and Paulin, 1991). The %GC of this region is 71 with 104 CpGs and 129 GpCs giving an Obs/Exp of 0.83 . There is also a cluster of sites for *Bst*UI but none for the enzyme *Mse*I.

Figure 1.4.

Human G6PD gene 1.0 kb 5' end



Human desmin gene 1.0 kb 5' end



Although restriction enzyme sites are less specific, it demonstrates a difference between regions with the characteristics of CGIs and those of bulk genomic DNA. The use of the enzyme *Mse*I to fractionate the genome, leaving GC-rich sequences intact, is discussed in Chapter 3.

1.5.
The methylation of CpG islands

Although ordinarily free of methylation, CGIs can become modified and two important biological functions are associated with this transformation. The first is X-inactivation, a process of active methylation which acts on one X-chromosome modifying most (but not all) the CGIs. Methylation of CGIs appears to be involved in maintaining the inactive state of the selected chromosome (Riggs, 1975; Mohandas et al., 1981). The second case involves parental imprinting where CGIs from either the paternally or maternally derived allele are methylated in the offspring. As with X-inactivation, methylation of CGIs in itself may not be the primary imprint, but is thought to be an important part of the process (Li et al., 1993; Razin and Cedar, 1994).

1.5.1.
The process of X-chromosome inactivation

In 1962 the hypothesis was put forward that the normal method of dosage compensation in man was the inactivation in female somatic cells of either one of the two X-chromosomes (Lyon, 1962). Inactivation occurs early in development with the inactive X-chromosome (X_i) becoming visible at a cytological level due to the formation of the Barr body. Once inactivated the X-chromosome is heterochromatic and late replicating, with most, but not all, of its genes transcriptionally inactivated. Inactivation is random

with either one of the X-chromosomes (paternal or maternal) inactive in different cells of the same animal (Lyon, 1962). The mechanism involves counting and in rare cases of sex chromosome abnormalities (i.e. more than two X chromosomes in a cell) all but one are inactivated. The choice of which X-chromosome to inactivate is apparently random but once modified, the X_i is maintained in a silent state in all progeny cells. The overall methylation patterns of the active X (X_a) and the X_i are similar, the main difference being the methylation of CGIs associated with genes on the X_i (reviewed in Grant and Chapman, 1988). Examples of X-linked genes, with their 5' islands methylated on the X_i but unmethylated on the X_a, include glucose-6-phosphate dehydrogenase (G6PD) and hypoxanthine phosphoribosyl-transferase (HPRT) (Wolf et al., 1984; Toniolo et al., 1991). There are also examples of CGIs on the X_i which escape inactivation. These include those in the Pseudo Autosomal Region (PAR). The PAR lies on the short arm of the X-chromosome and shares sequence identity with the Y-chromosome. The CGIs in this region are not subject to inactivation, perhaps because dosage compensation is not required. However, other CGIs lying outside the PAR also escape inactivation and as more X-linked genes are studied, a fuller picture of the process including inactivation patterns may emerge (Goodfellow et al., 1988; Mondello et al., 1988; Ellison et al., 1992).

An examination of rearrangements involving X-chromosomes in human and mouse identified an X-inactivation centre (XIC) (Migeon, 1994). The XIC (Xic in mouse) is required for X-inactivation and its action in *cis* initiates the process. The gene XIST (Xist) maps to the XIC (Xic) and encodes a 17 kb untranslated RNA expressed exclusively from the X_i (Brown et al., 1992). The XIST gene is notably methylated on the X_a and

unmethylated and expressed from the X_i (Norris et al., 1994). Loss of methylation did cause expression of XIST from the X_a in somatic cells but not in embryonic cells suggesting another mechanism (Beard et al., 1995). Although the complete process and patterns of X-inactivation are not yet fully understood, the role of methylation appears to be in the maintenance of the inactive state in somatic cells. The picture may gain clarity with the identification and position of more methylated CGIs and the transcriptional activity of their associated genes.

1.5.2.
The process of parental imprinting

The process of parental imprinting occurs during development, when either the maternally or paternally derived allele is transcriptionally silenced (Surani, 1994). Three genes in mice, H19, insulin-like growth factor 2 (*Igf-2*), and *Igf-2* receptor (*Igf-2r*), are differentially methylated at CGIs (or sites) depending on their parent of origin (Surani, 1994). Levels of RNA from each of the genes listed was analysed in the MTase deficient mice. It was shown that a normal level of DNA methylation was required for controlling differential expression of the paternal and maternal alleles of imprinted genes (Li et al., 1993). Imprinted loci in humans, demonstrated to have differences in DNA methylation, include H19 (Zhang et al., 1993) and the small nuclear ribonucleoprotein-associated polypeptide N gene (SNRPN) (Glenn et al., 1993). As with X-inactivation, chromosome specific patterns of methylation may emerge when more imprinted genes are identified.

1.6.
Transcriptional repression through methylation of CpG

During X-inactivation, methylation of CGIs contributes to the transcriptional repression of the associated gene. The methylation of CpGs within a CGI may prevent transcription through a direct modification of the binding site (Boyes and Bird, 1991). Alternatively the mechanism may be indirect, with proteins bound to methylated CpGs, preventing access to the CGI by transcription factors (Boyes and Bird, 1992; Meehan et al., 1992). The indirect process could involve methyl-CpG binding proteins (MeCPs). Two such proteins have been characterised in mammalian nuclei. MeCP1 which binds a minimum of 12 symmetrically methylated CpGs (Meehan et al., 1989) and MeCP2 which binds specifically to a single methylated CpG pair (Lewis et al., 1992). MeCP2 is far more abundant than MeCP1 ($2.0x10^6$ versus 5.0×10^3 molecules in mammalian nuclei) (Meehan et al., 1992). These proteins are widely distributed in a variety of tissues. The involvement of MeCP1 in gene inactivation was demonstrated in cell lines (Boyes and Bird, 1991). For example the methylated promoter of the phosphoglycerate kinase gene (PGK1) was inactive but activity could be restored by competition with DNA specific for MeCP1.

Although precisely how MeCPs function is not clear, one hypothesis is that MeCP1 guides DNA into a heterochromatic structure involving stable association with MeCP2 (Boyes and Bird, 1991; Meehan et al., 1992).

1.7.
Research objectives

The purpose of this project was to fractionate the human genome using a methyl-CpG binding domain column (Cross et al., 1994). Using this column a minor fraction, with a high frequency of m^5CpG, can be separated from the remainder of the genome. The purified fraction can then be cloned and a representative library produced. Included in the library will be fragments containing novel methylated CGIs. The subsequent use of these fragments as probes, against cDNA libraries, would enable the isolation and characterisation of any associated genes. It was hoped that the identification of these genes would aid in a better understanding of the processes of X-inactivation and parental imprinting. As an example, of a CGI, the first 10.0 kb of the 5' end of the sequence of G6PD is shown including the downstream region and first two exons (Figure 1.5). The CGI at the 5' end of the G6PD gene can clearly be seen covering the first exon and start of transcription (Figure 1.5). In common with the majority of CGIs, the example shown contains a cluster of sites for *Bst*UI a restriction enzyme with the recognition sites CGCG. In contrast the restriction enzyme *Mse*I (TTAA) does not cut in the CGI shown, this is due to its recognition sequence. Genomic DNA was digested with *Mse*I, which should leave the majority of CGIs essentially intact. In the case of the G6PD gene two alleles occur in DNA derived from female blood. The CGI is unmethylated from the allele on the X_a and methylated on the X_i. The *Mse*I fragment from the later should contain sufficient m^5CpGs to bind to the column at high salt. This allows separation of the methylated CGI from its unmethylated counterpart and from the remaining sparsely methylated genome.

It was expected that in addition to fragments from methylated CGIs, the library might contain other highly methylated GC-rich regions of the genome. It was proposed to check this by characterisation of sequences bound to the column to determine whether they are both GC-rich and have a high frequency of CpG. The methylation status of sites within the cloned sequences would be determined using Southern blots. Cloned inserts would then be sequenced and used as queries in a search of the NCBI database. Finally the libraries of sequences prepared using the MBD domain column would be used as probes in Fluorescent *in situ* Hybridisation *(FISH)* experiments, enabling their location on the chromosomes to be identified. It may also be possible to indicate why there is a lack of CpG suppression in these regions, in a genome which is frequently methylated and depleted for CpG .

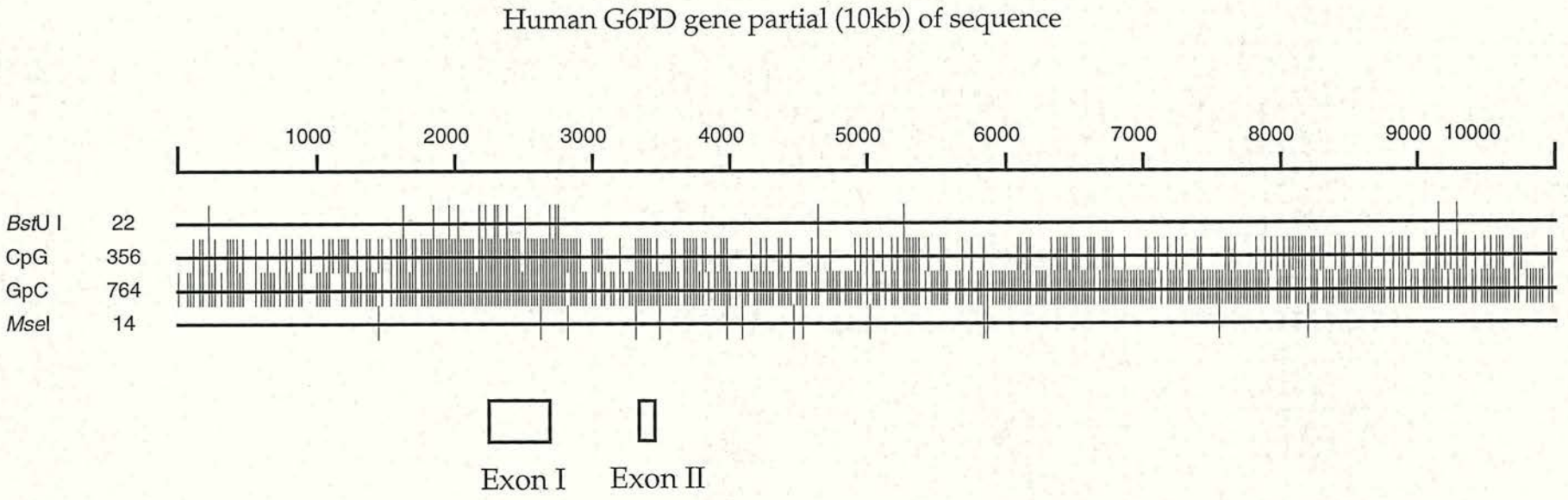
Figure 1.5.

Example of the 5' end of the G6PD gene showing the first two exons and the CGI

A.

Example showing the first 9984 bp of sequence from the glucose-6-phosphate dehydrogenase gene (Accession No. X55448 and Chen et al, 1991). The sequence includes the 5' end of the gene including the start of transcription and the first two exons. The CGI is clearly visible as a region were the frequency of CpG is roughly equal to that of GpC. The increase in GC-richness of this section is also indicated by the increase in restriction enzymes with GC-rich recognition sites and the absence of sites for *Mse*I.

Figure 1.5.


Human G6PD gene partial (10kb) of sequence
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
BstU I 22
CpG 356
GpC 764
MseI 14
Exon I
Exon II

2.1.

Commonly used reagents and buffers

Tris.HCl

Tris base (tris[hydoxymethyl] aminomethane) was dissolved in H_20 and adjusted to the correct pH by addition of HCl.

TE Buffer pH 7.4

10 mM Tris.Cl (pH7.4) 1 mM EDTA (pH8.0)

Use as a DNA solvent (Maniatis et al., 1982).

EDTA

EDTA (ethylenediaminetetramino acid di-sodium salt) was dissolved in H_20 and the pH adjusted to 8.0 using NaOH.

Loading Buffer 10x

10x loading buffer used in gel electrophoresis was prepared in a volume of 10 ml and stored at room temp.

Ficoll (Sigma) (11 gms) was added to EDTA pH8.0 (4 ml) and the volume made up to 10 ml with H_20. Add Orange G Dye (Sigma) until the loading buffer is a suitable colour.

Ethidium Bromide (EtBr).

Ethidium bromide (EtBr) was prepared in a stock solution of 10 µg/µl in H_20 and stored in the dark at 4°C.

TAE Buffer.

Used in the making and electrophoresis of agarose gels.

(Tris-acetate/EDTA) Concentrated stock solution (per litre)

50X : 242g Tris base, 57.1 ml glacial acetic acid, 100 ml 0.5M EDTA (pH8.0) (Maniatis et al., 1982).

Phenol / chloroform.

Phenol/chloroform solution was used to remove proteins and other impurities from nucleic acid preparations . The phenol was pre-equilibrated with 1M Tris.HCl (pH 7.5) and TE buffer before adding hydoxyquinoline 0.1% (v/v) and storing at 4°C in the dark.

2.2.
Restriction enzyme analysis

Digestion of genomic DNA, or of plasmids containing cloned inserts, was achieved using restriction enzymes. Unless otherwise stated all enzymes used were supplied by New England Biolabs with reaction conditions used according to the manufacturers protocol. A minimum of 1 unit of restriction enzyme was used per µg of genomic DNA with the volume of restriction enzyme not allowed to exceed 10% of the total digest. Digested DNA was resolved in agarose gels and visualised by staining with EtBr. Samples were carefully monitored to ensure that digestion of DNA derived from either lymphocytes or sperm had gone to completion. An aliquot of the digest was removed and added to 200 ng of a plasmid with appropriate restriction sites. Plasmid DNA control digests were then visualised on a 1.5% agarose gel (Figure 2.1.Ai & Aii).

2.3.

Agarose gel electrophoresis of nucleic acids

Agarose gels used to resolve genomic DNA digests were between 0.5% and 2.0% (w/v) of multi-purpose agarose (Boehringer Mannheim) in 1x TAE. The concentration used depended on the size of fragment(s) to be resolved. Small fragments were separated in high percentage gels (2% agarose for 0.1-2.5 kb); larger DNA fragments in 1% or less (Maniatis et al, 1982). The agarose gels used in southern blot analysis were ordinarily 1.5% and run at a low voltage for 14-16 h to achieve good resolution of bands. Gels used in the rDNA NTS experiments (Chapter 6) needed to resolve 19-5 kb fragments and were at a concentration of 0.5-0.4% agarose. These low percentage gels were overlaid on a higher percentage "base" making handling easier for photographs and Southern blots. Samples were mixed with loading buffer at 10-15% v/v before loading the gels. The molecular weight markers used were either Lambda DNA (Boehringer Mannheim) digested with *Sma*I, 1 kb molecular weight marker (Gibco BRL) or 100 bp marker (Gibco BRL). After staining with EtBr (0.5 μg/ml) the gels were visualised using a transilluminator (UVP) and photographed with a Video camera system (Mitsubishi) or an MP4 (Polaroid) camera.

2.3.1.

Purification of nucleic acid fragments

Following restriction enzyme digestion and other manipulations, DNA was purified by phenol/chloroform extraction followed by ethanol precipitation. Addition of an equal volume of 1:1 mixture of phenol/chloroform extracts proteins from solutions containing DNA. The solutions are mixed, vortexed and centrifuged for 5 min before removing the aqueous phase containing the

Figure 2.1.

Control digests of genomic DNA and an amplification reaction using PCR

Ai.

Example of DNA derived from either blood (B) or sperm (S) digested with *Mse*I (TTAA). As a control for complete digestion an aliquot of the reaction was added to a plasmid, pGem3Zf+ (Accession No. X65304) which was then electrophoresed through an agarose gel. The marker (M) is a 1 kb ladder (Gibco BRL). The size of the plasmid fragments generated indicates that the reaction was not inhibited and that the DNA was completely digested.

Aii.

Example of DNA derived from blood (B) digested with either *Msp*I or its methylation sensitive isoschisomer *Hpa*II (CCGG). Aliquots from digests were removed and added to a plasmid pGem3Zf+ then electrophoresed through an agarose gel. The marker (M) is a 1 kb ladder (Gibco BRL), undigested plasmid (U) is also shown. The size of the plasmid fragments indicates that neither reaction was inhibited and that the DNA was digested.

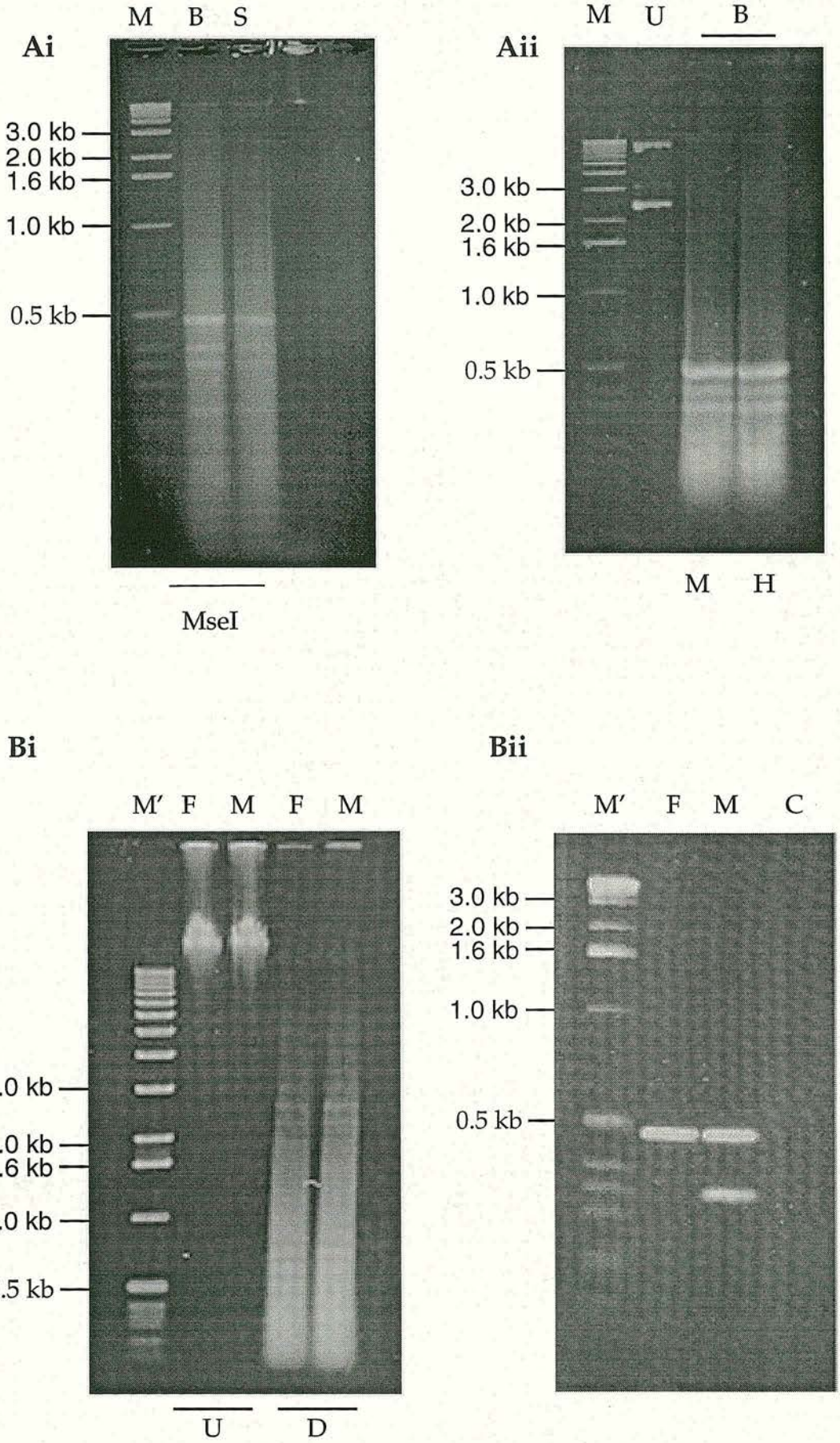
Bi.

Female (F) or male (M) genomic DNA was either undigested (U) or digested (D) with *Mse*I and electrophoresed through an agarose gel. The undigested DNA is of a high molecular weight and is undegraded. DNA digests were monitored as shown in Figure Ai, the marker (M') is a 1 kb ladder (Gibco BRL).

Bii.

The products generated using both ZFX, ZFY and the 'Common' primer in an amplification were electrophoresed through an agarose gel. The reaction enables accurate identification of male or female DNA from blood samples. When female DNA (F) is used as template the ZFX and the 'Common' primers amplify the same band from both the X-chromosomes. With male DNA (M) as template ZFX primers amplify the same band, in addition a smaller band is amplified with the ZFY and Common primers from the Y-chromosome. The marker (M') is a 1 kb ladder and the control reaction (C) contains no DNA, see Section 2.6.1. for reaction conditions.

Figure 2.1. .


Ai
M B S
3.0 kb
2.0 kb
1.6 kb
1.0 kb
0.5 kb
MseI
Aii
M U B
3.0 kb
2.0 kb
1.6 kb
1.0 kb
0.5 kb
M H
Bi
M′ F M F M
3.0 kb
2.0 kb
1.6 kb
1.0 kb
0.5 kb
U D
Bii
M′ F M C
3.0 kb
2.0 kb
1.6 kb
1.0 kb
0.5 kb

DNA. Ethanol precipitation of the aqueous phase involves the addition of a 1/10th volume of 3 M sodium acetate (pH 5.2) and 2 volumes of ice cold 100% ethanol. The solution was mixed, then either frozen on dry-ice for 15-20 min or overnight at -20°C, before centrifuging in a benchtop centrifuge (Eppendorf) at 15000 rpm for 20-30 min. The supernatant is removed and the resulting pellet washed in 70% ethanol, then dried down using a vacuum. The precipitation of small amounts of DNA can be improved by adding a carrier, glycogen (Boeringer Mannheim) before freezing. After precipitation and drying the DNA pellet was resuspended in TE buffer or distilled H_20. Resuspended pellets were usually stored at -20°C. The relative purity of the DNA could be assessed by measuring its absorbance at wavelengths of 260nm (A_{260}) and 280nm (A_{280}). At a wavelength of 260nm, 50 µg/ml of double stranded DNA gives a A_{260} reading of 1.0. The absorbance of the sample was then measured at a wavelength of 280nm (A_{280}) and the ratio between the two readings used to estimate the purity of nucleic acid. A ratio of 1.8 is ideal for purified DNA; variation from this figure indicates contamination or degradation.

Larger DNA fragments (250 bp-6 kb) could be isolated from agarose gels by spinning through glass wool. The glass wool was first cleaned as detailed in Maniatis (1982). Using a 25 gauge needle (Becton Dickenson) a hole was pierced in the bottom of a 0.5 ml eppendorf tube into which a plug of glass wool was placed. The DNA was cut from the gel using a clean scalpel and transferred to the 0.5 ml tube which was put in a 1.5 ml eppendorf tube. Both were then spun for 10 min at 5000 rpm in a benchtop centrifuge. After centrifuging the 1.5 ml eppendorf contains the solution of DNA which is then phenol/chloroform extracted and ethanol precipitated. This method

could also be used to separate small fragments of DNA, for example the products of a restriction enzyme digest.

2.4.
Purification of plasmid DNA

Plasmids were prepared using a modification of the alkaline lysis method of Birnboim and Doly (Birnboim and Doly, 1979). This revised method is detailed in Promega notes Edition 150 (Promega). This method removes chromosomal DNA by exploiting the difference in denaturation and renaturation characteristics between it and covalently closed circular plasmid DNA. Once precipitated the chromosomal DNA can be removed by centrifugation and the plasmid DNA further purified by selective absorption to a silica based resin, (Wizard, DNA purification resin, Promega). This revised method produces sufficient quantities of plasmids of a high enough quality for use as templates in cycle-sequencing reactions (Section 2.4.4). Bacterial colonies containing recombinant clones were picked and transferred to 10 ml of Luria-Bertani (LB Broth) with added ampicillin 50 μg/ml in a 50 ml screw cap tube (Falcon). LB broth was made by adding bacto-tryptone (10 g), bacto-yeast extract (5 g) and NaCl (10 g) to 950 ml of dd$H_2$0. This mixture was stirred until the solutes dissolve, then the pH adjusted to 7.0 with 5 N NaOH and the volume made up to 1 litre with ddH_2O. The LB broth was sterilised by autoclaving for 20 min at 15lb/sq. in. on liquid cycle (Maniatis et al, 1982) The Falcon tubes containing the LB Broth and the selected colony were then incubated for 14-16h in a 37°C shaking incubator at 250 rpm (New Brunswick Instruments). Following incubation, cultures were spun at 3000 rpm in a Beckman GP Centrifuge for 15 min and the resultant supernatant discarded. The bacterial pellets were gently resuspended in resuspension solution (300 μl). The resuspension

solution contained Tris-HCl (50 mM), EDTA (10 mM) and 10 μg/μl of RNase. The resuspended solution was then transferred to a 1.5 ml eppendorff. Lysis solution (300 μl) was added and the solutions mixed by inverting several times before adding neutralisation solution (300 μl). The lysis solution contained NaOH (0.2 M) and 1% SDS, the neutralisation solution was 1.32 M potassium acetate. The solutions were again mixed by inversion, then placed on ice for 2 min (to prevent plasmid degradation). The tubes were then spun at 15000 rpm for 10 min in a cold-room microcentrifuge. If clear the supernatant was transferred to a clean 2 ml eppendorff. If the supernatant was still cloudy tubes were respun in the microcentrifuge for a further 10 min until the solutions were clear. The Wizard miniprep resin (1 ml) was added and mixed by inverting several times over a 5 min period. The slurry was then added to a 2 ml syringe barrel attached to a Wizard minicolumn which in turn was inserted in a Vac-Man and attached to a vacuum. The resin was drawn through and contaminants removed by washing the mini-column twice. The column wash contained 55% EtOH, Tris-HCl pH 8.0 (8.0 mM) and EDTA (40 mM). A vacuum was applied to the column for 2-3 min followed by a 1 min spin at 14000 rpm to remove most of the EtOH. The mini-columns were then transferred to a 37^{o}C incubator and left for 10-15 min to remove any residual traces of EtOH. To elute the bound DNA ddH_20 heated to 80^{o}C was added. The columns were then allowed to stand for at least 1 min before spinning at 15000 rpm for 1 min. The resultant eluant was analysed on a 1.5% agarose gel for plasmid purity and concentration (Figure 2.2).

Figure 2.2.

Control digests to ensure complete methylation of plasmid and genomic DNA

A.

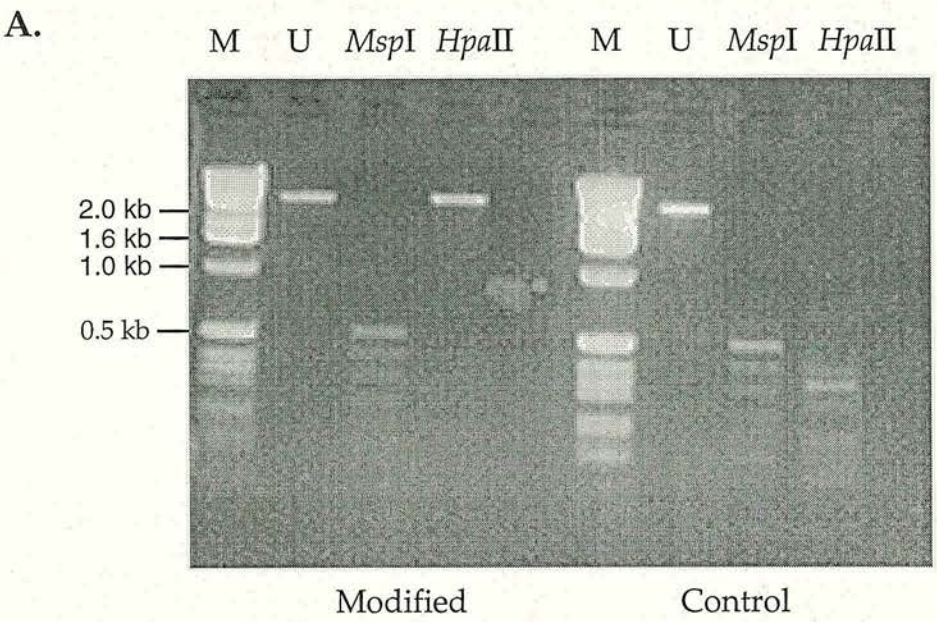
Methylation of plasmid and genomic DNA, see text for reaction conditions (Section 2.7.). Addition of an aliquot of the reaction, to 200 ng of pCG11, monitored the methylation reaction. *Msp*I, or its methylation sensitive isoschisomer *Hpa*II, were then used to digest the modified plasmid. In the example shown the pCG11 is not digested by *Hpa*II but is digested by *Msp*I. The undigested control is shown next to the size marker (M) a 1 kb ladder (Gibco BRL). The modified plasmid is protected from digestion indicating that the CpGs on pCG11 have been methylated. The unmodified control pCG11 is digested by both the *Hpa*II (H) and the *Msp*I (M).

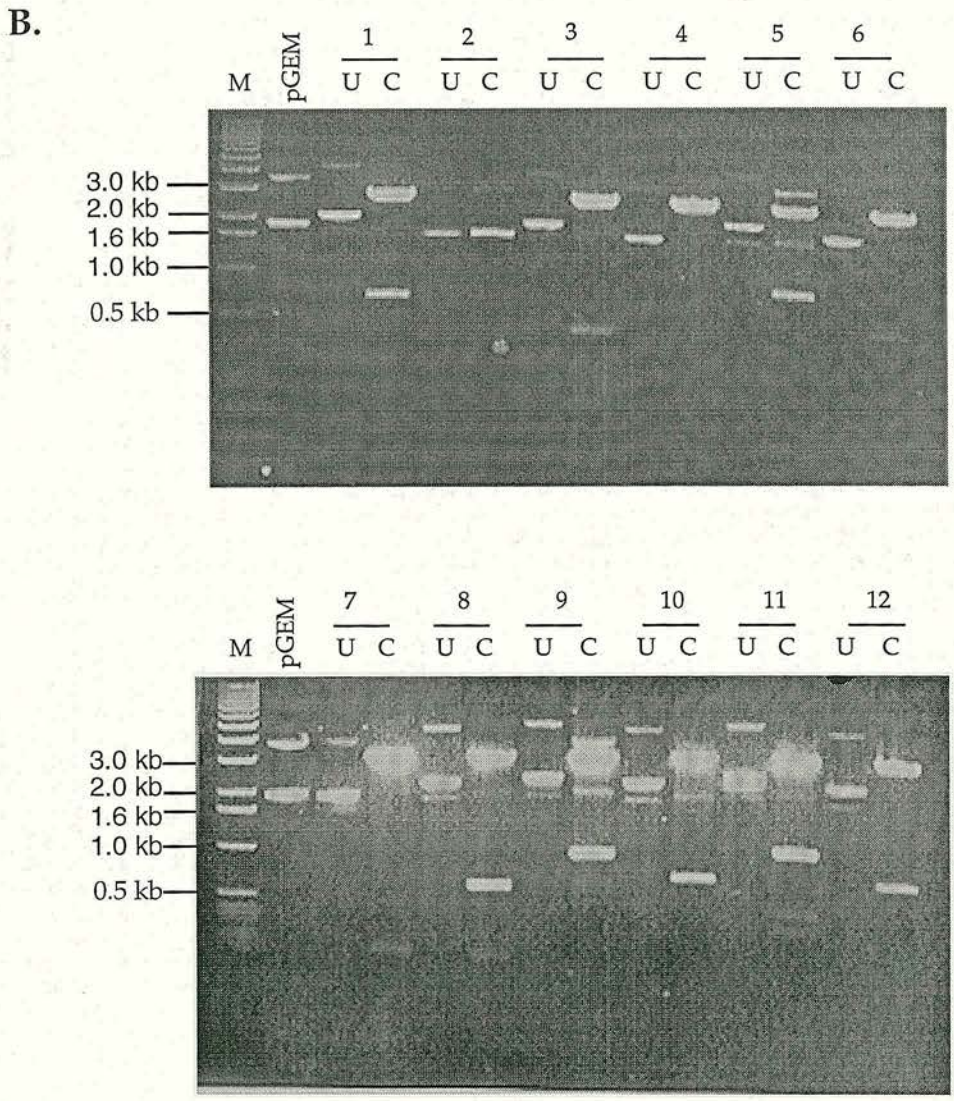
Digest to determine size of cloned inserts

B.

Preparation of plasmid DNA was by the modified alkaline lysis method (Promega) (Section 2.5.). In the example shown, the plasmids are digested with both *Nde*I and *Nco*I. These enzymes have sites in the plasmid flanking the cloning site and will to reveal if an insert has been cloned and its approximate size. Digested (C) and undigested (U) plasmids were electrophoresed through an agarose gel in adjacent lanes. Ten of the twelve plasmids have inserts, one plasmid, lane 2, does not and in addition is around 2 kb in size. Comparison with the undigested pGEM3zf+ control (500ng) provides an estimate of the concentration and purity of the plasmids. These plasmids were used as templates in dye-deoxyterminator reactions (Section 2.14.).

Figure 2.2.


A.
M
U
MspI
HpaII
M
U
MspI
HpaII
2.0 kb
1.6 kb
1.0 kb
0.5 kb
Modified
Control


B.
M
pGEM
1
2
3
4
5
6
U C U C U C U C U C U C
3.0 kb
2.0 kb
1.6 kb
1.0 kb
0.5 kb
M
pGEM
7
8
9
10
11
12
U C U C U C U C U C U C
3.0 kb
2.0 kb
1.6 kb
1.0 kb
0.5 kb

Plasmids were analysed using restriction digests to determine insert size (Section 2.2). Plasmids containing inserts were used in Dye deoxyterminator sequencing reactions (Section 2.6.5.).

2.5.
Preparation and analysis of genomic DNA

Human DNA used in the construction of the libraries was derived from blood (male or female) using standard methods (Maniatis et al., 1982). The amount and relative purity of the DNA was first assessed using a spectrophotometer (section 2.3.1). Both undigested DNA and DNA digested with *Mse*I were then visualised on an agarose gel. This was to ensure that the DNA was not degraded and that it would digest (Figure 2.1.B). The DNA was then analysed to ensure the samples had originated from either male or female blood. Fragments were amplified from a gene on the X-chromosome (ZFX) or on the Y-chromosome (ZFY) using two specific primers and one common primer (Strain et al., 1995). The fragment amplified from the ZFY gene is smaller than that amplified from the ZFX. The DNA derived from male blood will therefore test positive for both whereas DNA from female blood will give two fragments of the same size (Figure 2.1.Bii). For sequence of oligonucleotides and reaction conditions, see section 2.6.1. In addition to DNA derived from lymphocytes DNA from sperm was used in Southern blots (Section 2.7). DNA from sperm was extracted by addition of 50 mM EDTA pH8.0, 1% SDS and 500 μg/ml of Proteinase K with overnight incubation at room temperature. This was followed by addition of DTT to 50 mM and Proteinase K to 100 μg/ml and a further incubation overnight at 50°C (Cross, 1989). After phenol/chloroform extraction, the solution was ethanol-precipitated and resuspended in 10 mM TE.

2.6.

Polymerase Chain Reaction (PCR)

The polymerase chain reaction was used in order to amplify DNA fragments for use in further manipulations. These included, identification of the source of DNA fractionated with the MBD column, i.e. male or female blood (1). Monitoring the affinity for various fragments of DNA for the MBD column at different salt concentrations (2). Amplification of fractionated DNA following attachment of catch-linkers (3). Production of probes for use in *FISH* experiments (4). Cycle sequencing of inserts cloned into suitable plasmid vectors (5).

The reactions were carried out in either a Hybaid Omnigene Thermal Cycler or a PHC-2 (Techne), PCR machine. Reactions were ordinarily in a volume of 50µl unless otherwise stated. Reactions in the Techne PHC-2 were covered with mineral oil, the Hybaid thermocycler has a heated lid making this unnecessary. Mineral oil when present was removed by phenol chloroform extraction followed by EtOH precipitation before further analysis.

2.6.1.

Reaction conditions and primers for ZFX/ZFY and COMMON reactions

These primers, shown in table 1, were used in reactions to confirm that the original source of the DNA had been from either male or female blood (Section 2.4).

Table 1.

Primer	Sequence 5'-3'
ZFX	5'-AGACACACTACTGAGCAAAATGTATA-3'
ZFY	5'-CATCAGCTGAAGCTTGTAGACACACT-3'
COMMON	5'-ATTTGTTCTAAGTCGCCATATTCTCT-3'

Table 1 continued.

All three primers were used in each reaction using standard conditions, 10 mM Tris-HCl pH8.3, 1.5 mM $MgCl_2$, 200 mM each dNTP, 0.25 mM each primer and 250 ng template DNA.

Reaction conditions were 94°C for 5 min then 35 cycles of 94°C for 60 s, 65°C for 60 s and 72°C for 90 s followed by a 10 min extension at 72°C, (Strain et al., 1995).

2.6.2.

Reaction conditions and primers for monitoring fractionation of human DNA

These primers were used to monitor the purity of fractions eluting from the MBD column at different salt concentrations (Chapter 3).

Table 2.

Primer name	Sequence 5' - 3'
IFNαf	5'-GGATTGAAAACTGGTTCAACATGGC-3'
IFNαr	5'-TACTAGTGCCTGCACAGGTATACAC-3'
ApoA4f	5'-GGAGAAGTGAACACTTACGC -3'
ApoA4r	5'-TTTGAATTCGTCAGCGTAG -3'
G6PDf	5'-ATGGAACCCTGTCTTTGG -3'
G6PDr	5'-GGGGCTTGTGTTTTTACTTCCG -3'

Interferon Alpha (IFNα).

Apolipoprotein A (ApoA4).

Glucose-6-phosphate dehydrogenase (G6PD).

All three primers sets used standard reaction conditions, 10 mM Tris-HCl pH8.3, 1.5 mM $MgCl_2$, 200 mM each dNTP, 0.25 mM each primer and template DNA.

Reaction conditions for INFα primers were 95°C for 5 min followed by 30 cycles of 95°C for 60 s, 60°C for 90 s and 72°C for 3 min followed by an extension at 72°C for 10 min .

IFNα reaction conditions (Abbott and Povey, 1991).

Reaction conditions for ApoA4 and G6PD primers were 95°C for 5 min followed by 30 cycles of 95°C for 60 s, 55°C for 2 min and 72°C for 3 min followed by a 10 min extension at 72°C.

ApoA4 reaction conditions (Shemer et al., 1991) and G6PD (Zollo et al., 1994).

2.6.3.

Reaction conditions for Catch-linkers

Catch-linkers were used to amplify small quantities of DNA (Section 2.12.) They are designed to anneal giving an AT overhang compatible with the overhang generated by *Mse*I digestion (Section 2.12). Once ligated the catch-linkers could be used as primers in reactions to amplify trace amounts of DNA (Rothstein et al., 1979; Vooijs et al., 1993). Table 3 shows the sequence of the Catch-linkers used in construction of libraries

Table 3.

Primer	Sequence 5'-3'
Catch - 1	5'-TAAGTGCACGGTAGCGAATTCT-3'
Catch - 2	5'-GGAGAATTCGCTACCGTGCACT-3'
Catch-11	5'-TAAGACGATTTCCACTGAAGGCT-3'
Catch -12	5'-AGCCTTCAGTGGAAATCGTCT-3'
Catch-21	5'-TAGTTAACGCGCTGCATGAGTA-3'
Catch-22	5'-TACTCATGCAGCGCGTTAAC-3'

Equal volumes of each catch-linker were mixed (1 μg/μl) and heated to 80°C for 5 min then annealed by cooling to room temperature. Annealed catch-linkers were then attached to the *Mse*I digested DNA using T4 ligase . Either catch-linker could then be used as a primer in amplification reactions using PCR (see also Figure 2.4).

Reactions conditions for catch-linkers were as follows.

10 mM Tris-HCl pH8.3, 1.5 mM $MgCl_2$, 200 mM each dNTP, 1.0 mM primer and template DNA

Reaction conditions for both Catch 12 and 22 was 95°C for 5 min followed by 30 cycles of 95°C for 60 s, 60°C for 90 s and 72°C for 3 min followed by a 10 min extension at 72°C

2.6.4.

Reaction conditions for production of *FISH* probes

These reactions were similar to the amplification of trace amounts of DNA using catch-linkers as primers. The catch-linkers were again used as primers but Biotin-16-UTP was added to the reaction mixture and incorporated in the final product.

Table 4.

ddH_20	8.45µl
10x Buffer IV (Applied Biosystems)	5µl
25 mM $MgCl_2$ (Applied Biosystems)	5 µl
DMSO(Sigma)	5 µl
dATP (2 mM)(Applied Biosystems)	5 µl
dCTP (2 mM)(Applied Biosystems)	5 µl
dGTP (2 mM)(Applied Biosystems)	5 µl
Biotin-16-UTP (1 mM)	5 µl
dTTP (0.5 mM)(Applied Biosystems)	2.5 µl
Catch Linker see 2.6.3. (400ng)	4 µl
Taq Polymerase (Applied Biosystems)	0.05 µl
TOTAL	50 µl

Reaction conditions 95°C for 3 min followed by 30 cycles of 95°C for 1 min, 60°C for 30 s and 72°C for 3 min. The final extension time at 72°C was 10

min. The resulting labelled products were used as probes in the *FISH* experiments (Section 2.15. & Chapter 5).

2.6.5.
Reaction conditions and protocol for cycle sequencing reactions

Cycle sequencing, using dye-labelled terminators, is a method for performing enzymatic extension reactions for DNA sequencing. The DyeDeoxy Terminator Cycle Sequencing Kit uses a mutant Taq DNA polymerase which is less discriminating against the incorporation of dideoxynucleotides during the reaction (Perkin Elmer). This results in longer and better quality sequence from less starting template with fewer non-specific stop peaks. The template in all sequencing reactions was plasmid DNA (500µg) purified using the modified alkaline lysis method (Section 2.3.3). The plasmids included T7 and SP6 sequences which flanked the cloned insert. Each reaction was duplicated and carried out in a single 0.5 ml eppendorff using either T7 or SP6 primers.

Table 5

Plasmid DNA template	500µg
1µl T7 Primer	30ng/µl
1µl SP6 Primer	30ng/µl
8µl ABI premix	
ddH20	to total volume of 20µl

Fluorescent Dye-primer sequencing reactions were performed using a Perkin Elmer/Applied Biosystems ABI PRISM Dye Primer Cycle Sequencing Ready Reaction Kit with AmpliTaq FS DNA polymerase.

Reaction conditions for cycle sequencing reactions were 25 cycles of 98°C for 60 s, 55°C for 30 s and 60°C for 4 min.

2.7.
Methylation of plasmids and genomic DNA

In order to add methyl groups to fractionated DNA which had been amplified, or to plasmids, bacterial methylases *Sss*I (CpG) (Nur et al., 1985) or *Hha*I (NEB) were used according to the manufactures protocol. The *Sss*I (CpG) methylase recognises and methylates the dinucleotide CpG (Nur et al., 1985) *Hha*I methylase has the recognition site CCGG and methylates the internal CpG dinucleotide. Methylation reactions were carried out overnight at 37°C in a volume of 400 µl with S-adenosyl-L-methionine (Ado Met) as the methyl group donor. The concentration of Ado Met was either 80 µM when using *Hha*I or 160 µM when methylating with *Sss*I (CpG) methylase. To ensure the methylation of plasmids was complete, aliquots were removed and digested with either *Hpa*II or *Msp*I. Both have the recognition sequence CCGG but *Hpa*II will not cut if the internal CpG is methylated. When amplified DNA was treated with methylase an aliquot of the reaction was removed and added to 100ng of pCG11. Methylation of the amplified fraction could then be demonstrated by resistance of the plasmid to digestion by *Hpa*II (Figure 2. 2.A).

2.8.
Southern blot analysis of DNA

Southern blotting was used to transfer DNA from agarose gels to nylon membranes (Hybond N+ Amersham) by capillary action (Southern, 1975). The DNA from male and female blood or from sperm was first digested with *Mse*I (TTAA). This enzyme cuts frequently in bulk genomic DNA but due to its recognition sequence rarely in GC-rich DNA (Figure 1.5). This should leave CGIs and other regions of GC-rich DNA relatively intact. The digested

DNA was then redigested with either *Msp*I or its methylation sensitive isoschisomer *Hpa*II. Unmethylated GC-rich sequences would therefore be cut frequently but the same sites in methylated sequences would be resistant to digestion. All digests were monitored by removing aliquots and adding to 200ng of plasmid containing appropriate sites (Figure 2.1). The Southern blots used in this project usually contained 10 μg of DNA per lane of each digest. Sequences used as probes which occur at low or single copy in the genome will hybridise sufficiently to produce a signal with 10 μg of genomic DNA per lane. The DNA was electrophoresed in a 1.5% agarose gel at 50V for 14-16h to achieve good separation. Gels were then stained by shaking gently in EtBr (1μg/ml) for 10 min before destaining in 1xTAE for 30 min. The DNA was visualised using a transilluminator (UVP) and photographed with a MP4 camera (Polaroid). The transfer of larger DNA fragments was improved by depurinating gels in 0.3 M HCl for 10 min prior to transfer. Gels were then soaked in 0.4 M NaOH for 30 min with gentle shaking before setting up a Southern blot (Southern, 1975). Transfer of DNA was allowed to proceed overnight then the blot dismantled and the membrane marked to indicate the top of the gel. The membrane was gently washed in 2x SSC to remove fragments of agarose and transferred to prehybridisation solution (Section 2.9). Filters were prehybridised for 20-30 min at 68°C in a FHB11 hybridisation tube using a HB-1 rotating oven (Techne). These membranes could be probed with a cloned fragment revealing the methylation state of *Hpa*II sites within the sequence. Cloned inserts which were methylated in blood DNA would hybridise to high molecular weight bands in the lanes containing DNA digested with *Hpa*II. Inserts which are unmethylated hybridise to the same size bands in both the *Hpa*II and *Msp*I digested DNA.

2.9.

Random prime labelling of DNA probes

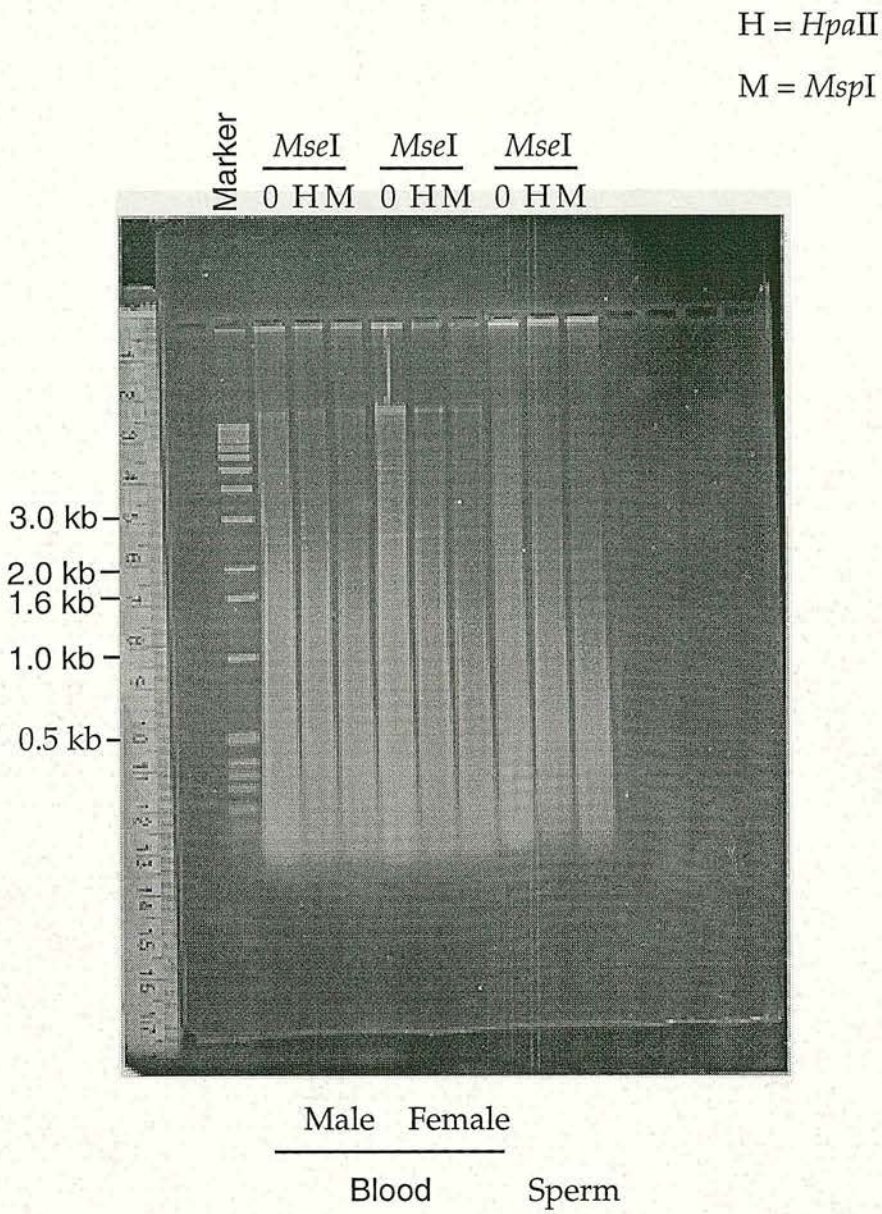
The cloned fragments used as probes against Southern blots were random prime labelled according to the method of Feinberg and Vogelstein (1983). The DNA fragment (50-100 ng) in a volume of 15-20 µl was heated to 100°C for 5 min to denature, 11 µl of denatured probe was then added to a 1.5 ml eppendorf tube on ice. The eppendorff contained 5 µl of the random hexanucleotide mixture (Boehringer Mannheim) with dTTP, dATP, dGTP (all at 2 mM) and 2 µl 10x buffer. The Klenow polymerase fragment 1 µl (Boehringer Mannheim 1 unit/1 µl) was added and the solution mixed well before adding 3 µl of α-32p dCTP (30 µCi) (Amersham) and incubating the reaction at 37°C for 30 min. To separate the unincorporated nucleotides from the labelled probe a G50 column (Sephadex) was prepared. A plug of glass wool was placed in a 1 ml syringe and the G50 slurry poured in and allowed to settle. After the slurry had settled the syringe/column was spun for 5 min at 1500 rpm in a S416 centrifuge (Eppendorf). The column was then transferred to a clean 15 ml tube (Falcon) and the labelled probe in 100 µl of TE added. The column was spun at 1500 rpm for 3 min and the activity in counts per minute (cpm) of 1-2 µl of the resulting flow through measured using a scintillation counter. The random prime labelled probe was then denatured by heating to 100°C for 5 min and transferred to the hybridisation tube (Techne). The tube contained the nitrocellulose membrane which had been washed in prehybridisation buffer (20 ml). The prehybridisation buffer contains 0.5 M Na_2HPO_4 pH7.2, 0.7 % SDS and 5% powdered milk. Hybridisation proceeded overnight at 68°C before the membranes were washed to remove unhybridised probe. Membranes were washed three

Figure 2.3.

Agarose gel electrophoresis of DNA extracted from blood and from sperm

DNA extracted from human blood, either male or female, or from sperm was digested with *Mse*I, which has the recognition sequence TTAA. The DNA was then redigested with *Msp*I (M) or its methylation sensitive isoschisomer *Hpa*II (H). Both enzymes have the recognition sequence CCGG, however *Hpa*II will not cut if the CpG is methylated. The *Mse*I digests and the *Mse*I /*Hpa*II or *Mse*I/*Msp*I double digests were loaded, 10 μg of each and electrophoresed through a 1.5% agarose gel for 14-16 h . Following depurination with HCl the gels were alkali blotted (NaOH) onto Hybond N+ filters (Amersham) following manufacturers protocol. The pattern of bands obtained during subsequent probing of the filters indicated the methylation status of the cloned inserts. Clones containing methylated CpGs will hybridise to smaller fragments in the *Mse*I/*Msp*I lanes than in the *MseI/Hpa*II lanes. The size of the bands can be confirmed by marking the position of the wells on the filter and measuring the relative distance travelled. For example a 500 bp fragment would have migrated 10.5 inches in the example shown.

Figure 2.3.


H = HpaII
M = MspI
Marker
MseI
MseI
MseI
0 H M 0 H M 0 H M
3.0 kb
2.0 kb
1.6 kb
1.0 kb
0.5 kb
Male Female
Blood Sperm

times for 20 min in 0.2 x SSC, 0.1% SDS at 68°C before transferring to a sealed bag and placing in a film cassette and exposing to film XAR (Kodak).

2.9.1.
End-labelling of plasmid and DNA probes

Fragments of DNA could also be labelled by using α–^{32}PdATP to fill in recessed 3' termini (Maniatis et al., 1982). End-labelling in this way uses the large fragment of DNA polymerase I (Klenow fragment). The probes labelled in this way did not have as high activity in cpm as those labelled using the random prime method. The DNA was first digested with an appropriate enzyme for example with *Mse*I (TTAA) or *Eco*RI (GAATTC). Digested DNA in a volume of 5 µl was added to dTTP 3µl (10 mM/µl), Klenow fragment 1µl (1 unit/µl), 10x buffer 4 µl and α–^{32}PdATP, 3 µl (10 µCi/µl) with the volume made up to 40 µl with double distilled H_2O (ddH_2O). The concentration of DNA depended on the probe being made, for example end-labelled plasmids, 100-200 ng, end-labelled DNA 100-200 µg. The reaction was allowed to proceed at 37°C for 30 min before removing the unincorporated α–^{32}PdATP by centrifugation through a G50 column (Sephadex). An aliquot (1-2 µl) of the flow through was removed and the cpm measured using a scintillation counter. Plasmids and DNA end-labelled in this way could be used to monitor the binding ability of the MBD column (Chapter 3).

2.10.
Preparation of competent cells

Epicurian Coli XLI-Blue MRF' Electroporation-Competent Cells (Stratagene) were used in the construction of libraries (Chapter 4). The cells are deficient in all known restriction systems and additionally are endonuclease and recombination deficient, making XL1-Blue MRF' cells ideal for construction

of genomic or methylated libraries (Stratagene Instruction Manual Cat No. 200158). A single colony was picked and used to seed 100 ml of LB broth which was transferred to a shaking incubator at 37°C overnight. The following day 15 ml of the cells were used to seed 500 ml of LB broth again incubated at 37°C. Aliquots of 1 ml were removed at regular time points and the OD_{600} measured with a U2000 spectrophotometer (Hitachi). When the density at OD_{600} had reached 0.6, the cells were transferred to 1 litre bottles (Beckman) on ice before placing in the J6B centrifuge and spinning at 4000 rpm for 15 min at 4°C. The supernatant was then poured off and the cells washed in ice cold ddH_20 and spun at 4000 rpm for 15 min at 4°C. This was repeated twice before the pellet was resuspended in 250 ml of ddH_20 with 10% glycerol and spun in a JA-14 rotor in the JS-21 centrifuge (Beckman). The resulting pellets were resuspended in 30 ml of 10% glycerol and transferred to 50 ml glass tubes (Corex) and spun at 3000 rpm in a JS-13.1 rotor in the J2-21 for 15 min. The final pellet was resuspended in an equal volume (200-300 µl) of 10% glycerol before aliquoting into 85 µl fractions and snap freezing on dry-ice before transfer to the -80°C freezer for storage.

2.10.1.
Electroporation of plasmids into competent cells

High voltage electroporation has been demonstrated to be efficient for introduction of genetic material into prokaryotic cells (Dower et al., 1988). After overnight ligation of the isolated DNA into a suitable vector the reactions were desalted by drop-dialysis through a 0.25µM membrane (Whatman). The removal of salt is necessary to reduce conductivity and prevent "arcing" (Bio-rad instruction manual). An aliquot of electrocompetent cells (85 µl) was thawed on ice before thoroughly mixing 40 µl with 1-2 µl of the de-salted ligation reaction. Electroporation cuvettes

with a 0.1 cm gap, were cooled on ice before addition of the cells and ligation reaction. Cells were then pulsed with the Gene-Pulser apparatus (Bio-Rad Laboratories Richmond CA) at 1.8 kV, 25 μFD and 200 ohms parallel resistance, time constants of greater than 4.5 ms were usually obtained. SOC medium 960μl was added immediately after electroporation and the suspension transferred to a 37°C shaking incubator for 45-60 min. SOC medium was made in batches of 1 litre by adding bacto-tryptone (20 g), bacto-yeast extract (5 g) and NaCl (0.5 g) to 950 ml of dd$H_2$0. This mixture was stirred until solutes dissolved then 10 ml of 250 mM KCL added and the volume to 1 litre with ddH_2O. The LB broth was sterilized by autoclaving for 20 min at 15 lb/sq. in. on liquid cycle. Once the solution had cooled 20 ml of a sterile solution of 1M glucose was added (Maniatis et al., 1982). The electroporation mixture (50-100 μl) was plated onto LB plates with ampicillin at 50 μg/ml for selection of competent cells with recombinant plasmids. Prior to plating out the electroporation mixture plates were also spread with 40 μl of X-gal (20 μg/ml in dimethylformamide) to allow for blue-white colony selection (Maniatis et al., 1982).

2.11.
Sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE)

Discontinuous SDS-PAGE was used to separate proteins according to their molecular weight (Laemmli, 1970). SDS is an anionic detergent, which, under the correct conditions, binds to proteins and confers a negative charge. The proteins are loaded onto a gel and a voltage applied causing them to move in an inverse relation to their size. The gels are poured in two parts with a separating layer below a stacking layer to give maximum resolution. The protein are first stacked in the upper layer before entering the separating

layer which, with its smaller pore size, resolves the proteins more effectively. The premix for both the stacking and separating gels is shown in Table 6. The SDS-PAGE gels were used to identify the recombinant MBD which moves an almost identical distance to cyctochrome C. Proteins were visualised after separation by staining with a dye, Coomasie blue . During purification of protein solutions, fractions were examined using SDS-PAGE. Those which contained proteins of the correct size could then be purified further (Chapter 3).

Table 6.

	Separating Gel	Stacking Gel
Acrylamide 29:1 30%	25 ml	5 ml
1.5M Tris HCl pH8.8	18.75 ml	N/A
0.5M Tris HCl pH6.8	N/A	7.5 ml
10% SDS	375 µl	150 µl
Temed	45 µl	20 µl
10% Ammonium Persulphate	375 µl	300 µl
Distilled H_20	31.25 ml	17.2 ml

2.12.

Attaching catch-linkers to the fractionated DNA

The profile of female DNA passed once over the MBD column indicates that a very small proportion of the genome is highly methylated enough to bind the MBD column at high salt (Figure 3.4). Monitoring with PCR also indicated that a further two passes were required to produce a pure sample (Figure 3.6). The quantity of DNA in the final bound fraction meant it was necessary to attach catch-linkers and amplify in order to efficiently clone the material (Rothstein et al., 1979; Vooijs et al., 1993). The catch-linkers were designed to give an AT overhang compatible with the TA overhang left after

Figure 2.4.

Attaching catch-linkers to *Mse*I digested DNA

A.

The catch-linkers are first mixed in equal amounts before heating to 80°C, cooling to room temperature allows them to anneal. This gives an AT overhang compatible with that generated by an *Mse*I digestion.

B.

Attaching the annealed catch-linkers to the *Mse*I digested DNA using T4 ligase.

C.

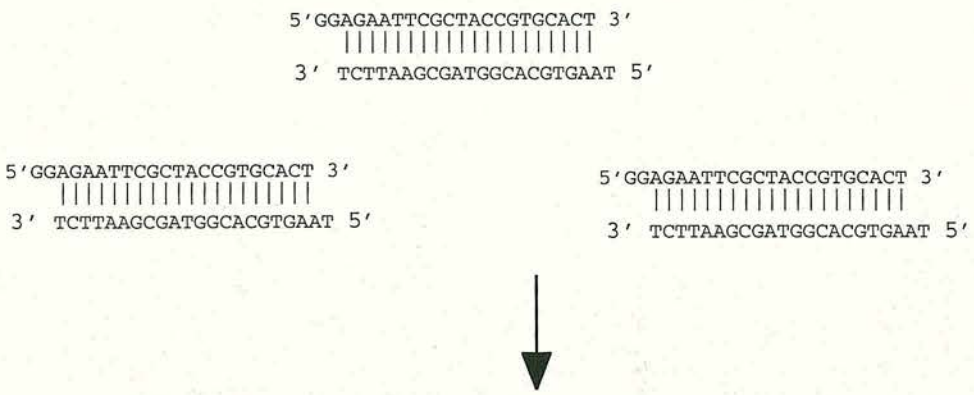
The amplification reaction showing the linkers used as primers (Section 2.6.3.).

D/E.

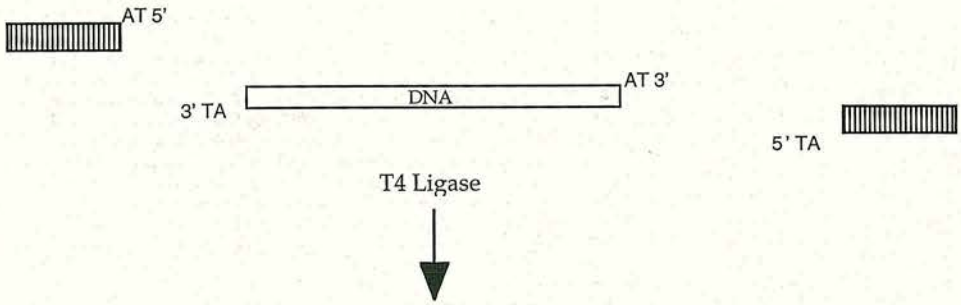
The amplified DNA has an added 3' adenosine, this improves the efficiency of cloning (Section 2.13.).

Figure 2.4.

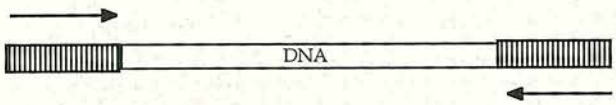
A. Catch-linkers are mixed together and heated to 80°C for two minutes. They are then annealed by cooling to room temperature.



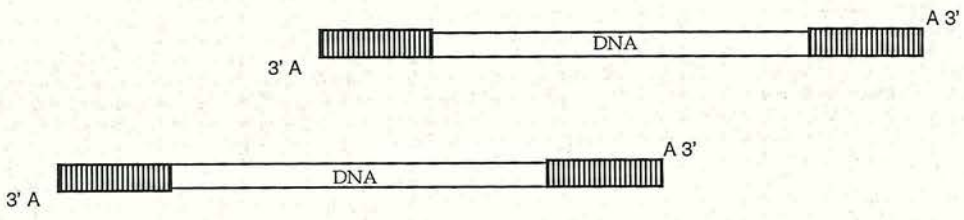
B Catch-linkers are then attached to *Mse*I digested DNA using T4 ligase.



C. Catch-linkers are used as primers to amplify the attached DNA



D. Polymerase has amplified the DNA and added a 3′ adenosine.



E. Clone into T-vector with improved efficiency due to 3′ adenosine.

*Mse*I digestion (Figure 2.4). The linkers (1 µg) were heated to 80°C for 5 min then allowed to anneal by cooling to room temp. DNA collected from the bound fraction (MBDx3) was mixed with the annealed catch-linkers (1 µl) 10x buffer 4 (2 µl NEB) T4 DNA ligase 1unit/µl (1 µl) and ddH_20 to a final volume of 20 µl. The reaction was then allowed to proceed overnight at 15°C. Once ligated to the DNA either one of the catch-linkers could be used as a primer in a reaction to amplify the attached material by PCR (Section 2.6.3).

2.13.
Cloning of amplified DNA using the T-vector system

The T-vector system (Promega) takes advantage of the fact that Taq polymerase adds a non-template dependant deoxyadenosine to the 3' hydroxyl terminus of blunt ended DNA (Clark, 1988). The T-vector is a modified pGEM-5Zf(+) which has been blunt-end digested with *Eco*RV (GAT | ATC) before adding a 3' terminal thymidine to both ends. These 3'-T overhangs at the insertion site greatly improve the efficiency of ligation compared to that of blunt-ended cloning (Promega Notes No.150). The digestion and removal of catch-linkers prior to cloning is also unnecessary when using this system.

The DNA amplified by PCR, with the catch-linker as primer, was phenol-chloroform extracted, ethanol precipitated and resuspended in ddH_20. Different ratios of vector and insert were then ligated until an optimum was determined. Having determined the optimum ratio, PCR amplified DNA, pGEM T-Vector (Promega), 10x buffer and T4 DNA ligase (NEB) were mixed and the ligation reaction proceeded overnight at 15°C in a LTD6 water bath (Grant).

The PCR amplified DNA ligated into pGEM T-Vectors was then introduced into electroporation competent cells (Section 2.10.1). The modified pGEM 5Zf(+) Vector contains T7 and SP6 RNA polymerase promoters flanking multiple cloning sites within the α-peptide coding region for β-galactosidase (Promega Catalogue). Cloning of DNA fragments causes insertional inactivation of the β-galactosidase enzyme and allows blue-white colony selection. Transformations were then plated onto LB plates with added ampicillin 50 μg/ml and β-gal 20 μg/ml for blue-white selection and incubated at 37°C overnight. White colonies containing inserts were picked the following day and transferred to 96-well plates containing 150 μl of YT media with 10% freezing solution. The YT media was made in 1 litre batches by adding bacto-tryptone (16 g), bacto-yeast extract (10 g) and NaCl (5 g) to 950 ml of ddH_20. This mixture was stirred until solutes dissolved then the pH was adjusted 7.0 with 5N NaOH and the volume adjusted to 1 litre with ddH_2O. Freezing solution contained K_2HPO_4 (360 mM), Na citrate (17 mM) $MgSO_4$ (4 mM), $(NH_4)_2SO_4$ (68 mM) and 44% glycerol. Selected clones were then picked from the 96-well plates, grown overnight in 10 ml of LB and the plasmids extracted using the modified alkaline lysis method detailed in section 2.5.

2.14.
Sequence analysis of cloned inserts

Fluorescent dye-primer sequencing reactions were performed using a Perkin Elmer/Applied Biosystems ABI PRISM Dye Primer Cycle Sequencing Ready Reaction Kit with AmpliTaq DNA polymerase (Section 2.6.5). The reactions were analysed using an automated 373 DNA sequencer (Applied Biosystems Perkin-Elmer). Unlike manual sequencing, automated sequencing does not use radioactive terminators followed by visualisation of the banding pattern

Figure 2.5.

Example of the data collected using a T7 primer and the ABI373 automated DNA sequencer

A.

The trace obtained from the ABI373 automated DNA sequencer. Plasmids were first purified using the Wizard modified alkaline lysis purification system (Promega). The plasmids were amplified and sequenced using fluorescent dye primer chemistry (Perkin Elmer/ABI PrismDye Primer Cycle Sequencing Ready Reaction Kit with AmpliTaq DNA polymerase). The entire insert is sequenced in two reactions, one starting at the T7 primer the other at the SP6.

B.

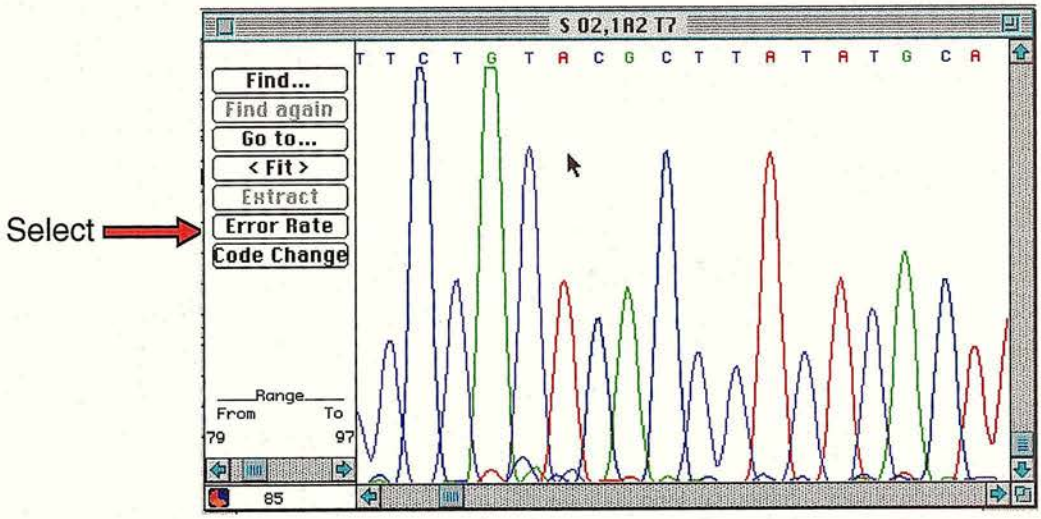
Using the Gene Jockey II program (Biosoft Cambridge) the collected data was analysed. The percentage of bases called in error was first shown, note that the sequence degenerates 5'-3' from the primer. The large peak at the start is the "dye-front", unincorporated dye-deoxyterminators moving rapidly through the gel. Extracted sequence with an error rate of less than 10% is shown in Figure 2.5.C.

C.

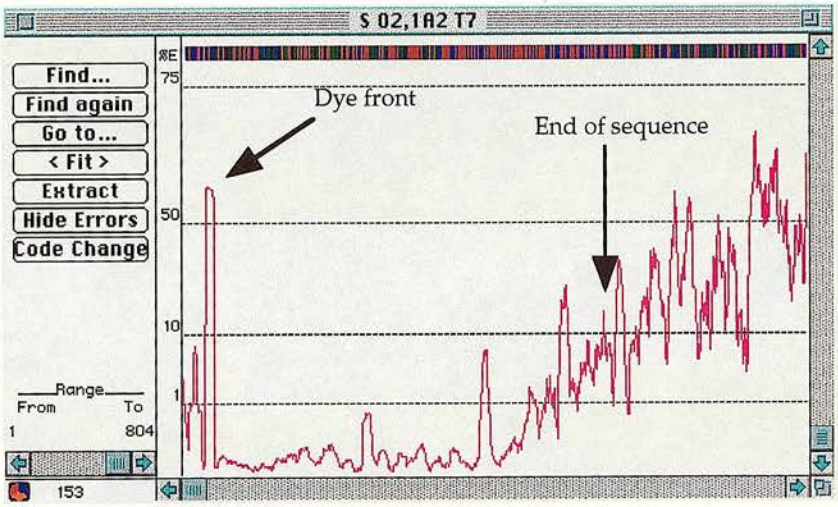
In the extracted sequence shown, the catch-linker (No. 12) is highlighted. After removing the sequence upstream (left) of the catch-linker, the remaining 'insert" sequence is saved.

Figure 2.5.

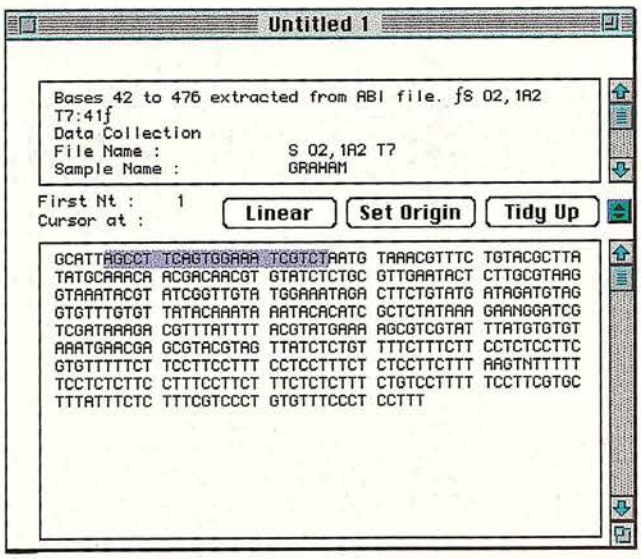
A.



B.



C.



through autoradiography. Instead, automated sequencers use labelled fluorescent dyes, one for each deoxynucleotide, visualised by a scanning laser (Ansorge et al., 1987). The cloned inserts were amplified from the plasmid template using the flanking primers, deoxynucleotides and the four dye-deoxynucleotides (Section 2.6.5). Once complete the entire sequencing reaction was added to a single lane on a the gel and the different bases visualised with the scanning laser. Template DNA for the cycle sequencing reaction is prepared following the miniprep protocol detailed (Section 2.4). The primary data from the ABI373 DNA sequencer was extracted using the Gene JockeyII program (Biosoft Cambridge). Inserts are sequenced in either orientation in two separate reactions using the flanking primers. This allows confirmation of the sequence obtained with one primer through alignment (where possible) with the sequence from the other (Figures 2.5. & 2.6). The data handling procedure used is shown with the insert sequence from plasmid 1A2 as an example. All data manipulation was performed using the GeneJockeyII program (Biosoft Cambridge). The data file was first opened (Figure 2.5.A) and the "error rate" button selected, the program reveals the percentage (%E) of ambiguous bases in the sequence (Figure 2.5.B). The large peak at the start of the sequence shown is the "dye-front" caused by unincorporated dye-deoxyterminators migrating rapidly through the gel. The remaining sequence shows an error rate below 10% for several hundred base pairs before rising to above 50% after almost 400 bp of sequence, "End of sequence". Sequence data with an error frequency of greater than 10% was only used if the ambiguous base could be identified from the raw data file or the reciprocal strand. The available sequence below 10% error was then extracted (Figure 2.6.B) and searched to determine the boundary between the vector/catch-linker sequence and the insert. In the example shown the sequence of the catch-linker is highlighted and the sequence of the insert is

Figure 2.6.

Example of the data collected using a SP6 primer and the ABI373 automated DNA sequencer

A.

Trace from the ABI373 DNA automated sequencer from a plasmid insert amplified using the SP6 primer. The plasmids were purified by the Wizard modified alkaline lysis method (Promega).

B.

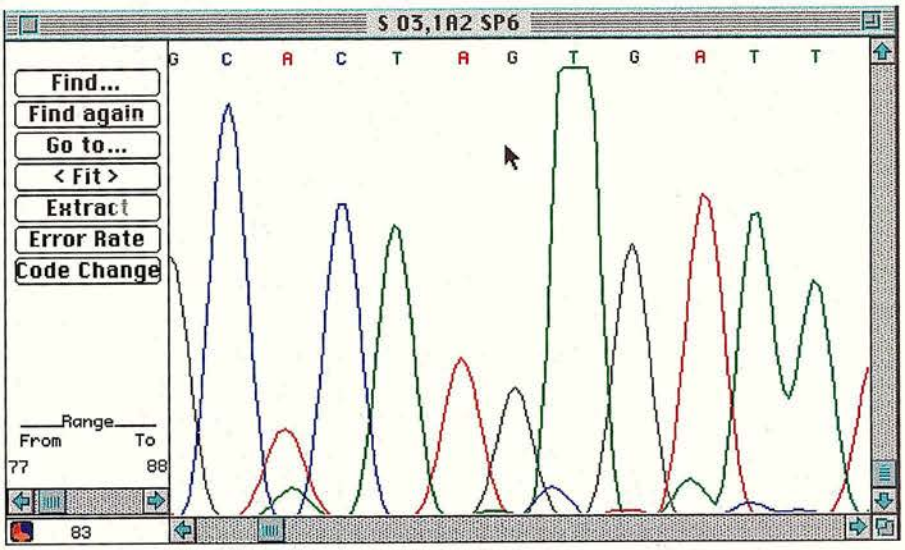
Using the Gene Jockey II program (Cambridge) the collected data was first inverted before analysis. The data from the SP6 primer is not as good quality as the data obtained from the T7 primer (Figure 2.5.) The error rate, though acceptable, rises above 10% a number of times.

C.

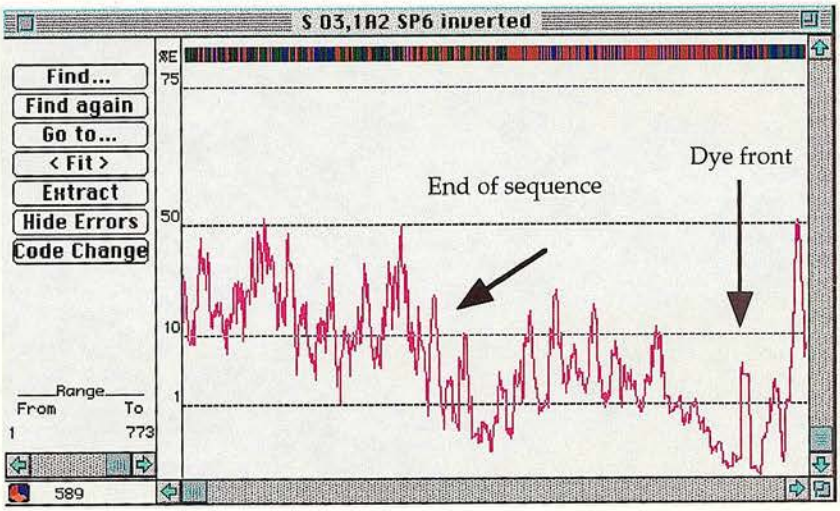
The extracted sequence is shown with the inverted sequence of the catch-linker (No. 12 inverted) highlighted. The sequence upstream (right) of the catch-linker sequence was removed and the remaining 'insert" sequence saved, the primer used to generate it is again indicated in the file name.

Figure 2.6.

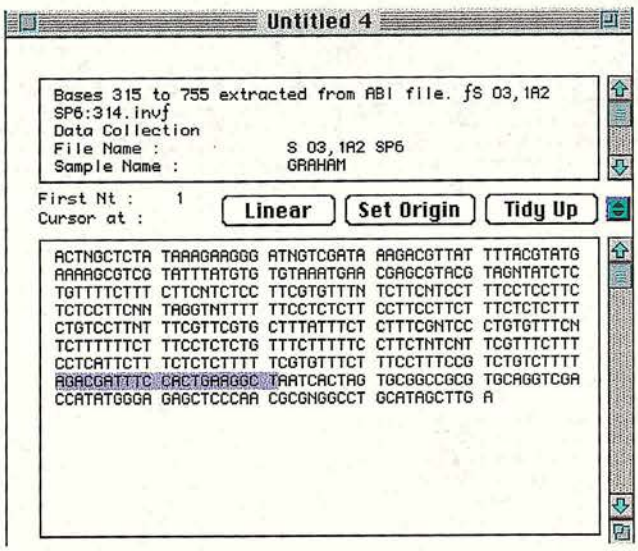
A.



B.



C.



downstream (Figure 2.5.C). Catch-linker and vector sequences are removed leaving only insert sequence which was saved. In the example shown the insert was saved as p1A2 T7, indicating the clone and the primer used to generate the sequence (Figure 2.5.C). The process was then repeated for the sequence obtained using the primer SP6. Before displaying the error rate the sequence is inverted giving the sequence of the reciprocal strand. Sequence with a %E of less than 10% is again selected, highlighted and then extracted as before (Figure 2.6.C). The sequence was again searched (using an inverted catch-linker sequence) and the boundary between catch-linker and vector and insert highlighted as shown (Figure 2.6.C). The sequence was then saved as p1A2 SP6, again indicating the clone and the primer used. Sequences were then aligned using the alignment function in GeneJockey II (Biosoft Cambridge) with a minimum base pair match of 6 (Figure 2.7). In the example shown there are several contentious bases between sequences. The original 'raw' data files were consulted and if the nucleotide is ambiguous in both sequences an "N" was used. The original raw data file was always saved in the unaltered form. In most cases the removal of the catch-linked sequences allowed confirmation of the TTAA junction and indicated that the clone had contained *Mse*I digested DNA with attached catch-linkers.

After aligning sequence from both forward T7 and reverse SP6 primers the resulting "contig" was analysed to determine base composition. This was done using either Gene Jockey II or MSP Crunch programs (Biosoft Cambridge). These calculate the %GC, the $CpG_{Obs/Exp}$ and the frequency of nucleotides, dinucleotides and trinucleotides in each sequence. The larger cloned inserts of 800-900 bp do not usually produce overlapping sequence with the T7 and SP6 primers as the sequence begins to degenerate around 350-400 bp. Sequence confirmation, or correction of ambiguous basepair in

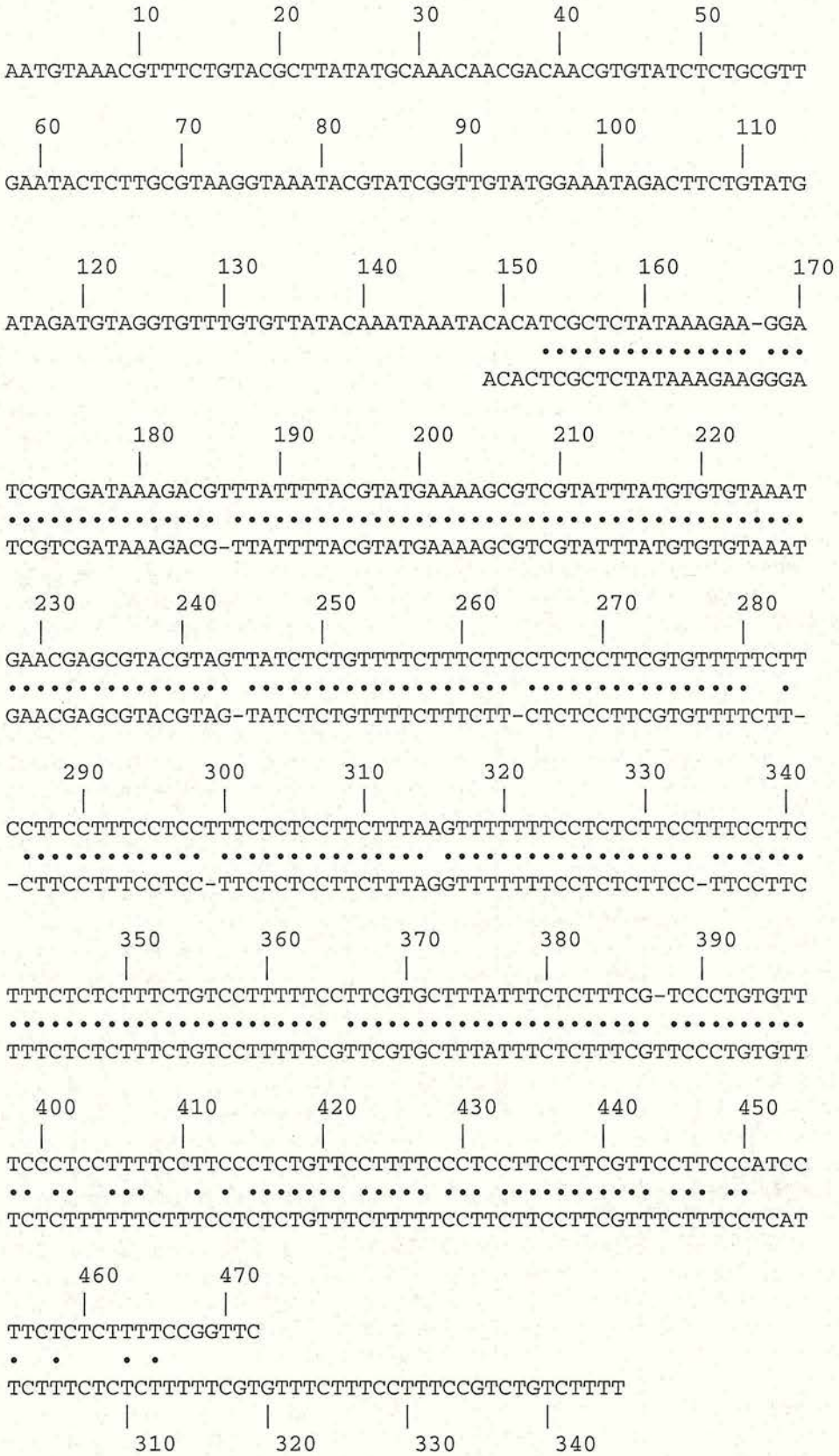
Figure 2.7.

Alignment of the two sequences generated using the T7 and SP6 primers

The sequences generated by the two primers using fluorescent dye primer reaction (Perkin Elmer/ABI PrismDye Primer Cycle Sequencing Ready Reaction Kit with AmpliTaq DNA polymerase) were then aligned. In the case highlighted, the primers have generated slightly different sequences with a number of dinucleotides either added or absent from each strand. The raw data files (2.5.A & 2.6.A.) were consulted, broad or tailing peaks are often interpreted as two dinucleotides. Generally the data file generated by primer T7 was of higher quality, however this sequences also degenerates towards the 3' end. This coincides with the start of the sequence generated by primer SP6 and in this case the quality of sequence was usually better. Using the original data file, nucleotides were added or subtracted from the sequence produced by each primer until the overlapping sequence aligned. If a nucleotide remained ambiguous even after consulting the data file "N" was used. The resulting contig was then analysed to determine the %GC and $CpG_{Obs/Exp}$. Finally the sequence was submitted to the NCBI data base as a query in a BLAST search (Altschul et al., 1990).

Figure 2.7.

Sequence 1A2 T7 (Top) 1A2SP6 (Bottom)



the form of a reciprocal strand, was therefore often unavailable for these larger clones. Having analysed and determined the %GC and $CpG_{Obs/Exp}$ of the available sequence it was submitted to the National Centre for Biotechnology Information (NCBI) database (see below and result section) and compared to the database using the BLAST program (Altschul et al., 1990). Files returned from NCBI were analysed using the MSP Crunch program which allows the filtering of database matches. The returned sequences were ranked according to the statistical probability that the query and subject (database) sequences were the same. If the query produced a high probability the database sequence was retrieved from the European Molecular Biology (EMBL) database. Both the database sequence and the cloned sequence were then compared using the GeneJockey II (Biosoft Cambridge) program. Examples of sequences which were similar to those from the database are shown in Appendix A.

2.15.

Fluorescent in situ Hybridisation (*FISH*)

Human male lymphoblast cell lines (a gift from Wendy Bickmore) were grown in RPM1 with 10% fetal calf serum. To obtain high mitotic indices, the cells were synchronised by addition of colcemid (0.1 μg/ml) for one hour. Approximately 10 ml of the cell suspension was dropped onto a clean slide and examined under a phase contrast microscope. The slides were allowed to air dry before placing in a desiccator for 2-3 days. Probes were prepared by amplifying MBDx3 or MBDx5 female DNA using PCR (Section 2.6.4). A nucleotide analogue, biotin-16-UTP, was added in order to label the probe, See section 2.4.5. for reaction mixture and conditions. Unincorporated nucleosides were removed by spinning through a G50 column for 3 min at 15000 rpm and retaining the flow-through. To ensure optimum *FISH* conditions the concentration of probe should be greater than 2 ng/ml. To estimate the concentration of reaction products they were diluted in water to 10^{-2} and 10^{-3} and spotted onto pretreated nitrocellulose filters (Schleicher and Scheul). Biotinylated lambda DNA standards at 1, 2, 10 and 20 pg (Gibco BRL) were also spotted onto the filters. After fixing the reaction products and standards to the filters by UV cross-linking, they were washed in buffer 1 for 5 min, followed by 6 min in buffer 2 at 60°C. Filters were immersed for 30 min in 10 ml of buffer 1, Tris pH 7.5 (0.1M) and NaCl (0.15M) containing 10 μl of streptavidin-alkaline phosphatase, then washed three times in buffer 1 (15 min) and once in buffer 3 (5 min) Tris pH 9.5 (0.1M). The filters were sealed in polythene bags with 5 ml of buffer 3 containing alkaline phosphatase substrate kit IV BCIP/NBT (Vector) and placed in the dark for 2-4h. The amount of probe solution required was estimated by a visual comparison with the lambda DNA standards.

2.15.1.

In situ hybridisation of probes to chromosome spreads

The desiccated slides prepared earlier were incubated in RNase (100 µg/ml in 2x SSC) for 1 hour at 37°C. After washing quickly in 2x SSC, the slides were dehydrated by placing sequentially in 70%, 90% and 100% EtOH for 2 min each then drying in a speedvac. The biotinylated probe was prepared by mixing with Cot-1 DNA (Gibco BRL), sonicated salmon sperm (Sigma) and either an rDNA probe or an X-linked cosmid probe, M3 Dig (both gifts from W. Bickmore). Two volumes of ethanol were then added and the probe solution dried down in a speedvac. The hybridisation solution 10 µl was added and the pellet left to dissolve for one hour. Hybridisation solution contained 50% formamide, 2x SSC, 1% Tween 20 and 10% dextran sulphate. Before use, the slides were warmed at 70°C for 5 min, then the chromosome spreads were denatured in prewarmed 70% formamide, 2x SSC for 3 min at 70°C. Slides were then transferred to ice-cold 70% ethanol for 2 min, then 2 min each in 90% and 100% EtOH and dried as before. The mixture of probe and Cot-1 DNA was denatured at 75°C for 5 min then transferred to a 37°C water bath and pre-annealed for 15 min. The slides were then warmed to 37°C and the coverslips thoroughly cleaned. The pre-annealed mixture of probe and competitor (10 µl) was pipetted onto the coverslip and picked up with the slide. Any trapped bubbles were gently removed with light pressure before sealing the coverslip to prevent drying out. The slides were then placed in a box lined with wet paper towels (to keep the atmosphere moist) and incubated overnight at 37°C. Slides were washed four times for 3 min in 50% formamide, 2x SSC at 45°C, followed by four washes for 3 min in 2x SSC at 45°C and four times for 3 min in 1x SSC at 60°C. The slides were placed in 4x SSC 0.1% Tween 20 before incubating each slide in 40 µl of

blocking buffer (4x SSC, 5% powdered milk) to reduce the background signal. The diluted antibodies and conjugates had earlier been prepared in blocking buffer. Biotin labelled MBDx5 and MBDx3 probes were detected by first using avidin-texas red then biotynalyted anti-avidin and finally avidin texas red, chromosomes were stained with 2 μg/ml DaPi. The rDNA and M3DIG probes, when included, were detected using anti DIG FITC and avidin texas red followed by FITC anti-sheep, anti avidin biotin and then avidin texas red. Chromosomes were again stained with 2 μg/ml DaPi. Images were captured using a Photometrics cooled CCD camera and a Zeiss Axioplan fluorescence microscope. Data was analysed using Digital Scientific (Cambridge) or IPC Photolab.

Chapter 3 : MBD column preparation

3.1.

The methyl-CpG binding domain column

The methyl-CpG binding domain (MBD) affinity matrix column binds DNA according to its degree of methylation (Cross et al., 1994). Highly methylated DNA fragments bind to the column at a high salt concentrations and unmethylated or sparsely methylated fragments are eluted. If DNA is first digested with the restriction enzyme *Mse*I (TTAA), the bulk of the genome will be reduced to small fragments (Chapter 1). DNA with a high GC content is left relatively intact as the recognition sequence does not occur as frequently in these regions (Figure 1.5). If, in addition, the GC-rich sequences are heavily methylated, for example if they are derived from a methylated CGI, they will bind tightly to the column even at high salt. These fragments can then be separated from the remainder of the sparsely methylated genome and the unmethylated CGIs.

The MBD column is comprised of the 85 amino acid long MBD region of the rat MeCP2 protein (Nan et al., 1993). The MeCP2 protein has a strong affinity for methylated CpG residues though it does bind some DNA non-specifically . The non-specific binding can be significantly reduced by using only the MBD portion of the protein (Nan et al., 1993). The MBD portion was produced as a recombinant protein fused to a Histidine Tag which allowed it to be attached to a nickel-agarose matrix. Each recombinant MBD protein can bind to an individual m^5CpG as it passes through the column matrix. The result is that the longer and more highly methylated the fragment is, the greater its interaction with the matrix. This allows DNA to be fractionation with its the main criteria for binding being the number of methyl-CpGs. The

interaction between m^5CpG and MBD is reversible at high salt concentrations allowing the elution of the bound fragments.

3.2.
Production and purification of the MBD protein

Production of the recombinant protein MBD involves the plasmid (pET6H) which is based on the pET system (Novagen). The plasmid contains the sequence of MBD cloned adjacent to six histidines and an initiation methionine. The resulting plasmid with its cloned insert was called pET6HMBD (Cross et al., 1994) and the cloned sequence can be expressed by virtue of a T7 promoter system (Studier et al., 1993). The expressed protein also includes six histidines which allow its purification using metal chelation chromatography (Hochuli et al., 1987). In order to produce enough MBD protein, a 100 ml culture of pET6HMBD was grown overnight in *E. coli* strain *BL21 (DE3) pLyss* with ampicillin at 50 μg/ml for resistance selection (Studier et al., 1993). This *E. coli* strain is a lysogen of bacteriophage lDE3 which contains the T7 RNA polymerase gene under the control of an inducible lacUV5 promoter. The RNA polymerase of bacteriophage T7 is very selective for specific promoters that are rarely encountered in DNA unrelated to T7 DNA (Studier et al., 1993). Strain *BL21 (DE3) pLyss* also lacks the ompT outer membrane protease that can degrade proteins during purification. Additionally the strain carries the plasmid *pLyss* which provides a small amount of T7 lysozyme. This has a dual function, it can act as a natural inhibitor of T7 RNA polymerase thereby decreasing basal activity but not preventing induction of high levels of a target protein. The T7 lysozyme also cleaves a bond in the peptidoglycan layer of the *E.coli* cell wall and allows cells to be lysed under mild conditions such as the addition of 0.1% Triton X-100 (Studier et al., 1993). Aliquots (15 ml) from the 100 ml overnight

culture were used to seed 500 ml of LB broth with ampicillin (50 μg/ml). Chloramphenicol was also added (10 μg/ml) to increase plasmid copy number and IPTG (isopropylthio-β-galactoside) added (0.4 mM) for induction. When the OD_{600} reached 1.0, the cultures were spun for 10 min at 3600 rpm in 1 litre bottles in a J-6B rotor (Beckman) at 4°C. The resulting pellets were resuspended in 30 ml of bacterial extract solution. The extract solution contained, urea (5 M), NaCl (50 mM), Hepes (20 mM) EDTA (1 mM) pH 8.0, PMSF (0.5 mM) and 10% glycerol. Protease inhibitors, pepstatin A, leupeptin, antitrypsin and apoprotinin were added each at 5 μg/ml.

The recombinant protein could now be extracted. Following addition of 0.1% TritonX-100, resuspended pellets were sonicated (power 6 cycle 40) on ice. The suspension was then spun at 15000 rpm at 4°C in a JA-20 rotor (Beckman) to pellet the bacterial debris before removing the supernatant and transferring to a dialysis tube. Dialysis was in 5 steps of (1L) with the amount of urea being gradually reduced (2.5 M, 1.25 M, 0.6 M, 0.3 M and finally no urea) in the dialysis buffer allowing refolding of the protein. The dialysis buffer also contained NaCl (50 mM), sodium phosphate pH7.0 (50 mM), TritonX-100 (0.1%), PMSF (0.5 mM), β-mercapthoethanol (10 mM) and 10% glycerol. The extract was respun at 16000 rpm for 30 min at 4°C and loaded onto a Fractogel EMD SO_3^{e-}, 650M column (Merck) for preliminary clean up. The Fractogel column had first been equilibrated by washing with low salt buffers (0.1 M NaCl) followed by high salt (1 M NaCl) then the low salt again. In addition to the variable concentrations of NaCl, buffers contained Hepes pH7.9 (20 mM) and 10% glycerol. The buffers were also filtered through a 0.2 μM filter then degassed using a vacuum before adding PMSF (0.5 mM) and 1% TritonX-100. These buffers were also used when running the MBD column (Section 3.4.). The proteins (in low salt buffer)

were loaded onto the column, before reloading the flow-through. The proteins were then fractionated using low salt (50 mM NaCl) then high salt buffer (1 M NaCl). In order to determine which fractions contained the MBD an aliquot (7.5 µl) of each was analysed using a 15% SDS-PAGE gel (Figure 3.1.A). The recombinant protein (MBD) was known to co-migrate with cytochrome C in polyacrylamide gels. The fractions containing proteins of the correct size were pooled (Figure 3.1.A). The pooled solution was dialysed against low salt buffer and loaded onto a HR 10/10 column (Pharmacia) containing Fractogel EMD SO_3^{e-}, 650M (Merck) for further clean-up of the protein. The column was equilibrated by washing with low salt (50 mM NaCl), then high salt (0.4 M NaCl). The protein solution was loaded in low salt and eluted from the HR 10/10 column using a 60 ml linear salt gradient. Samples of 2 ml were collected using an automatic sample collector (Pharmacia) and aliquots of the first 30 fractions run on polyacrylamide gels (Figure 3.1.B). The bulk of the MBD protein elutes between 0.6 M and 0.85 M NaCl, that is fractions 17 through 23 (Figure 3.1.B).

3.3.
Attaching the MBD protein to the nickel-agarose matrix

The MBD protein was further purified using a nickel-agarose matrix. Proteins with a histidine tag have a strong affinity for this matrix. The buffer used to wash the column contained imidazole (a histidine analogue) at a concentration of 8 mM. The buffer also contains NaCl (50 mM), TritonX-100 (0.1%), β-mercapthoethanol (10 mM) PMSF (0.5 mM) and 10% glycerol. After equilibrating the column the protein was loaded in imidazole free buffer. The use of 8 mM imidazole buffer will remove any impurities in the final protein solution before it gradually removes the bound MBD.

The nickel-agarose slurry (2 ml) was poured into a small column and equilibrated with the imidazole free buffer before loading the protein solution. The flow through was reloaded twice to ensure maximum binding of the MBD protein (Figure 3.1.C. FTI &FTII). Next the slurry was washed with imidazole free buffer (2 ml A1-A3) followed by buffer with imidazole at 8 mM (2 ml B1-B3). Fractions were collected from each wash and an aliquot run on a 15% SDS-PAGE gel (Figure 3.1.C). A protein of approximately the right size is visible in the load but this is significantly reduced in the first and second flow through. The other proteins present do not appear to bind significantly and further more are eluted after 6 ml of the buffer with imidazole along with a small amount of the MBD (Figure 3.1.C. Wash B3). An aliquot (100 µl) of the column slurry was then completely stripped using 80 mM imidazole, this allowed the concentration of protein bound to the matrix to be calculated (Figure 3.1.D). The amount of protein eluting was measured using a Bradford assay and calculated to be approximately 16 mg per ml of column slurry (Figure 3.1.D). The remaining slurry with the bound MBD was then loaded onto a HR 10/10 column (Pharmacia) according to the manufacturers protocol. This column has been equilibrated by washing with 1 M NaCl followed by washing with 0.1 M NaCl.

The protein sample stripped from the nickel-agarose matrix is almost pure, but several proteins of a different size are visible (Figure 3.1 D washes B1 &B2). The concentration of protein was also slightly higher than the figure given as all the MBD has not been stripped from the 100 µl aliquot, (Figure 3.1.D. wash B5). The MBD protein was now tested for its ability to separate fragments of DNA according to the number of m^5CpG dinucleotides they contained.

Figure 3.1.

Extracting, isolating and purifying the recombinant protein MBD

A.

Following extraction and dialysis of the proteins, a Fractogel EMD SO_3^e-650M column (Merck) was used for preliminary clean up. After equilibration of the column, the protein solution was loaded in low salt (L). The flow through was then reloaded twice (FTI & FTII). The column was then washed twice with 7.5 ml of 50 mM NaCl solution (LSI & LSII) followed by four washes with 7.5 ml of 1 M NaCl (HSI, II, III & IV). The MBD, co-migrates with cytochrome C (C) and elutes in the 1M NaCl washes. The fractions containing the MBD (HSI, II, III & IV) were dialysed against a 50 mM NaCl buffer.

B.

After dialysis the proteins were loaded onto a HR 10/10 column (Pharmacia) which also contained Fractogel EMD SO_3^e-, 650M (Merck) for further clean-up. The recombinant protein was then eluted using a linear salt gradient. Aliquots of 7.5 µl from the collected fractions were run on SDS-PAGE gels. Fractions 17-23 between, 0.50 and 0.85 M NaCl, contained the bulk of the MBD (Figure 3.1.B. arrow). These fractions, containing the MBD, were collected and pooled.

C.

The recombinant MBD was further purified by attaching it to a column through an interaction between its histidine tag and a nickel-agarose matrix. After equilibration of 2 ml of nickel-agarose slurry using the imadazole free buffer, the proteins, also in imadazole free buffer, were loaded and the flow through was collected (L). The flow through was then added and collected

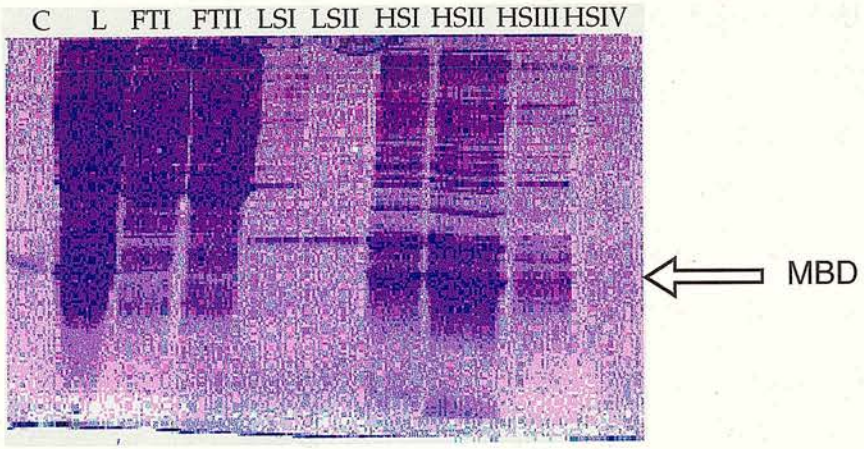
twice more to ensure maximum binding of any protein (FTI & FTII). The column was then washed three times with 2 ml each of imadazole free buffer (AI, AII, AIII) followed by three washes with buffer containing 8 mM imadazole (BI, BII, BIII). After 6 ml of the buffer containing 8 mM imadazole, a small amount of protein of the size of the putative MBD can be seen eluting. The majority of the other proteins present have less affinity for the matrix and eluted in the flow through.

D.

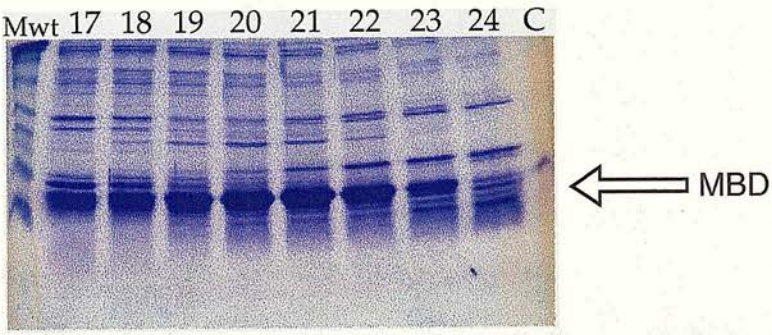
The amount of the MBD bound to the nickel-agarose was calculated by stripping a 100 µl aliquot from the column using 80 mM imidazole, a histidine analogue. The 100 µl aliquot was first washed with 800 µl of the imadazole free buffer (AI). This was followed by five 800 µl washes with the buffer containing 80 mM imidazole (BI through B5). The amount of protein bound per ml of matrix was then be calculated using a Bradford assay.

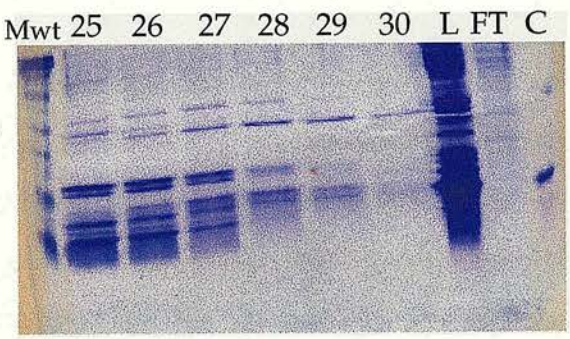
Figure 3.1.

A.

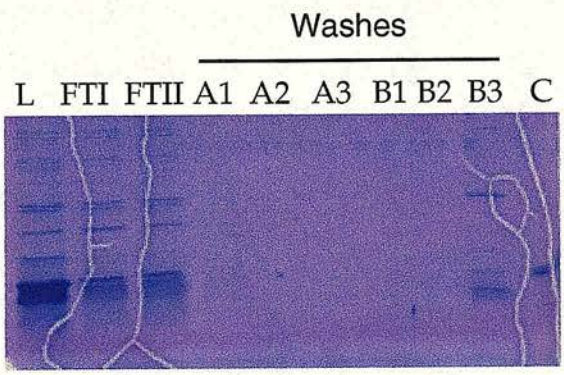


B.

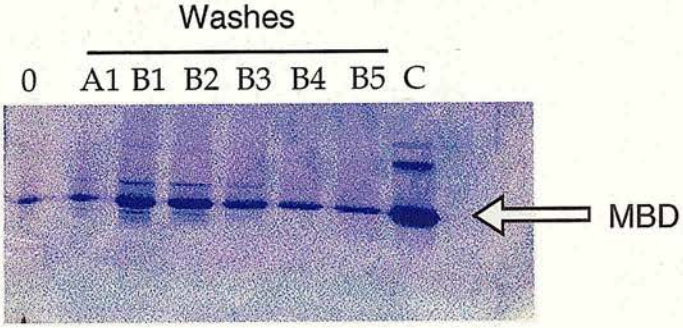




C.



D.



3.4.
Testing the MBD column to determine methyl-CpG binding ability

To demonstrate that the MBD column could separate DNA according to its level of methylation, three plasmids were used. Each has a different level of methylation, starting with pCG11 which had no methyl groups. Modification of pCG11 with *Hha*I methylase adds methyl groups to 37 CpGs (Figure 3.2.A). Plasmids modified with *Sss*I (CpG) methylase (Nur et al., 1985) have methyl groups added to 200 CpGs (Figure 3.2.A). The plasmids were linearised by digestion with *Eco*RI (NEB) then end-labelled (Section 2.9.1). After end-labelling, plasmids were centrifuged through G50 columns (Sephadex) to remove the unincorporated nucleotides. An aliquot from each plasmid end-labelling reaction was removed from the G50 flow through and the activity in counts per minute (cpm) measured using a scintillation counter (Beckman). Having measured the cpm of each plasmid, equal activities were then mixed and loaded onto the MBD column. A linear salt gradient was then run through the MBD column using Flow Pressure Liquid Chromatography (FPLC) L100 (Pharmacia). Fractions were collected at 2 ml intervals and the activity in cpm of each measured using the scintillation counter. These measurements were then plotted against molarity of NaCl as shown (Figure 3.2.B). The fractions showing the highest activity (Peaks A, B & C. figure 3.2.B) were ethanol precipitated and resuspended in 50 µl of ddH_2O. These samples were then divided into three and either digested with *Hpa*II (NEB) *Cfo*I (Boehringer Mannheim) or left undigested. Unmodified plasmid pCG11 (200 ng) was added to each mixture before digestion and after 1h the digests were electrophoresed on an agarose gel. The gels were stained with EtBr, visualised using a UV light box and photographed with

the Hitachi video camera (see Figure 3.2.C). The gels were placed on a dryer (Savant) and dried down onto DE81 Paper (Whatman) before placing in a PhosphoImager cassette (Molecular Dynamics) (Figure 3.2.D).

The result of digesting the plasmids, contained in the peaks giving the highest activity measurements, is shown in Figure 3.2.C & D. The plasmid eluting at low salt (peak A) has been digested by both enzymes and therefore contained the unmodified plasmid pCG11. The plasmid which eluted at 0.5-0.6 M NaCl (peak B) is protected from digestion by *Cfo*I but not by *Hpa*II showing this peak contained pCGII treated with *Hha*I methylase. The plasmid eluting in the final peak (peak C) is protected from digestion by both *Hpa*II and *Cfo*I, showing it contained pCG11 treated with *Sss*I (CpG) methylase. The unmodified plasmid added to the digest and visualised using EtBr demonstrated that both enzymes were digesting unmodified plasmid. This demonstrated that the MBD column was capable of separating these three plasmids according to their degree of methylation. Unmodified pCG11 elutes from the column at a NaCl concentration of 0.5 M and the *Hha*I methylase treated pCG11 elutes at approximately 0.65 M. However the fully modified *Sss*I (CpG) methylase treated pCG11 continues to bind to the column until the salt concentration reaches 0.80 M.

The column was, therefore, capable of separating three plasmids each 2.8 kb in length. To test the MBD column's ability to separate smaller fragments the *Sss*I (CpG) methylase treated pCG11 was digested with *Mse*I and the resulting fragments end-labelled. The fragment sizes and the number of m^5CpGs contained in each region are shown in the legend to Figure 3.3.A. The fragments were loaded onto the MBD column and eluted using the salt gradient as before. Aliquots of the collected fractions were EtOH precipitated, and separated on an agarose gel which was dried down onto

Figure 3.2.

Separation of modified plasmids using the MBD column

A.

Linear schematic of plasmid pCG11, showing the sites modified by addition of methyl groups. The *Sss*I (CpG) methylase modifies 200 CpGs and the *Hha*I methylase 37.

B.

The activity in count per minute (cpm) of the fractions collected from the column. End-labelled plasmids were added and a linear salt gradient passed over the MBD column. The activity of the fractions was measured in cpm and graphed against the molarity of NaCl as shown. Three peaks can be seen, peak A at 0.5 M NaCl, peak B at 0.65 M NaCl and peak C at 0.85 M NaCl.

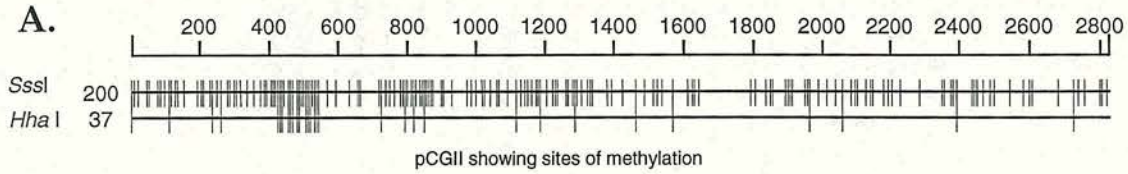
C.

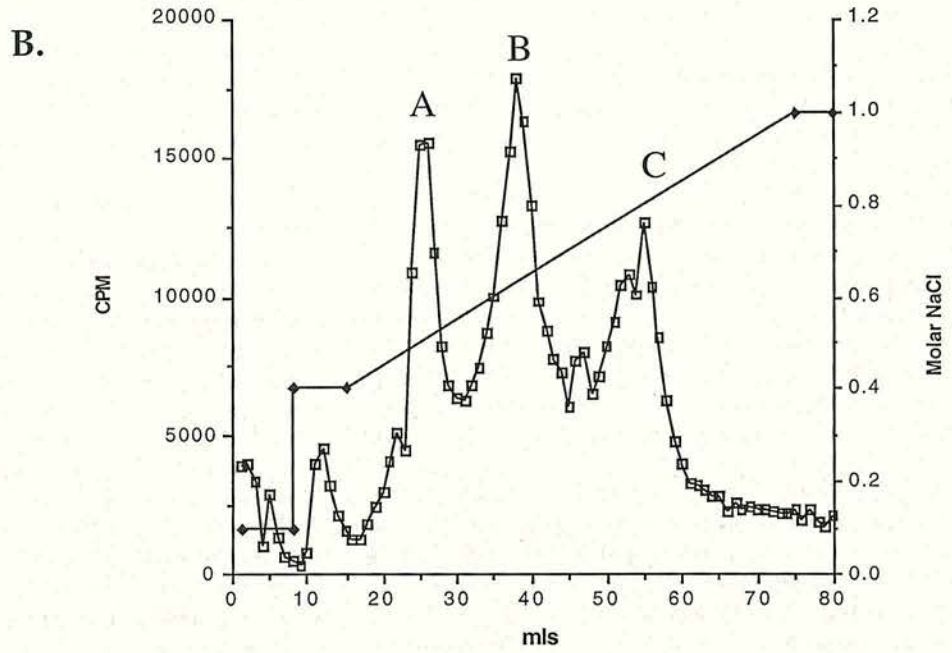
Two fractions corresponding to the three peaks with the highest activity were precipitated. These were then digested with the methylation sensitive enzymes *Hpa*II (CCGG) or *Cfo*I (GCGC). As a control unmodified plasmid pCG11 (200ng) was also added to each digest. The digestion reactions were then run out on agarose gels and visualised with EtBr.

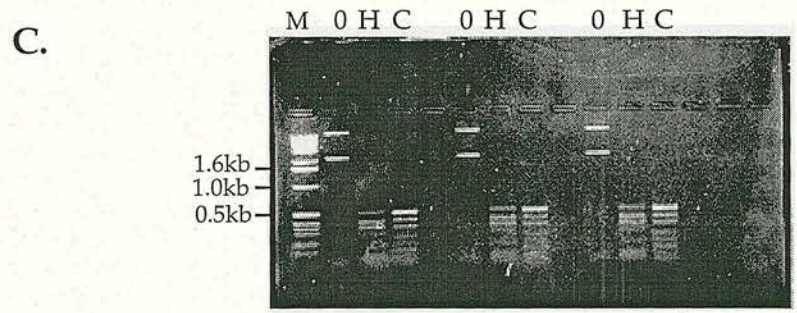
D.

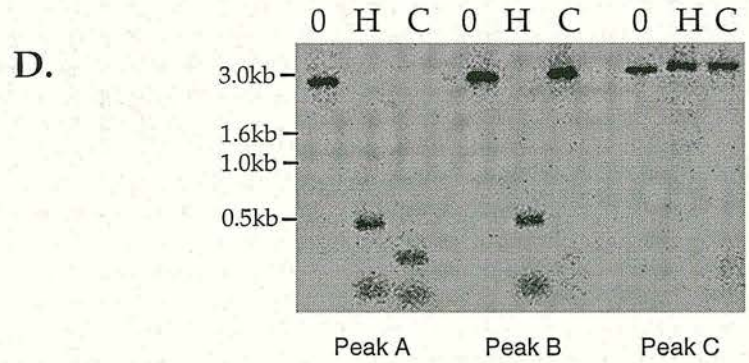
The gel shown in 3.2.C. was dried down onto DE81 paper (Whatman) and placed against a Phopho-imager screen overnight (Molecular dynamic). The screen was scanned the following day and the image generated shown (Figure 3.2.D).

Figure 3.2.


A.
200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400 2600 2800
SssI 200
Hha I 37
pCGII showing sites of methylation


B.
A
B
C
CPM
Molar NaCl
mls
20000
15000
10000
5000
0
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0 10 20 30 40 50 60 70 80


C.
M 0 H C 0 H C 0 H C
1.6kb
1.0kb
0.5kb


D.
0 H C 0 H C 0 H C
3.0kb
1.6kb
1.0kb
0.5kb
Peak A
Peak B
Peak C

DE81 paper. The DE81 paper was then placed against film overnight before developing (Figure 3.3.B). Unidentified material, possibly DNA contamination or a product of degradation is seen eluting at low salt concentration.

The MBD column appears to have much less affinity for the smaller fragments of pCG11. Most elute in fractions 12 through 17 between 0.5 and 0.6 M NaCl on the gradient. There are two larger fragments which have a strong affinity even at 1.0 M though most elute around 0.8 M. One of these is the 876 bp fragment with 66 m^5CpGs, the other is a mixture of the three fragments of between 340 and 376 bp. It is not possible to accurately differentiate between the three but there appears to be two peaks of this ~360 bp band. The first elutes at the lower salt concentration but a second peak indicates a tightly bound fragment which also elutes at 0.80 M and above. Of the three fragments of approximately 360 bp two have 21 and 15 m^5CpGs respectively while the third has 40.

In order to demonstrate that fragments would bind to the column again after dialysis. Those fractions eluting at 0.85 M NaCl and above were dialysed against the 0.1 M NaCl buffer and loaded onto the column as before. Fractions were collected, precipitated and electrophoresed as before. The dried gel was then exposed to film (Figure 3.3.C). The two large fragments again show strong affinity for the MBD matrix and are still bound at high salt. The unidentified material which does not match the size of any of the pCG11 fragments again elutes in the low salt fractions 14-16.

Figure 3.3.

Separation of fragments of the modified plasmid pCG11 using the MBD column.

A.

Linear schematic of plasmid pCG11 shows the relative positions of CpG, the sites of an *Mse*I digest and the fragments generated. The size of the larger fragments in base-pairs with the %GC and number of m5CpG per fragment is shown in the table below

Size	%GC	No. CpG	Obs/Exp.	Size	%GC	No. CpG	Obs/Exp.
181 bp	60	18	1.14	253 bp	54	14	0.85
197 bp	55	24	1.70	340 bp	49	21	1.0
373 bp	63	40	1.10	372 bp	41	15	0.97
876 bp	55	66	1.0				

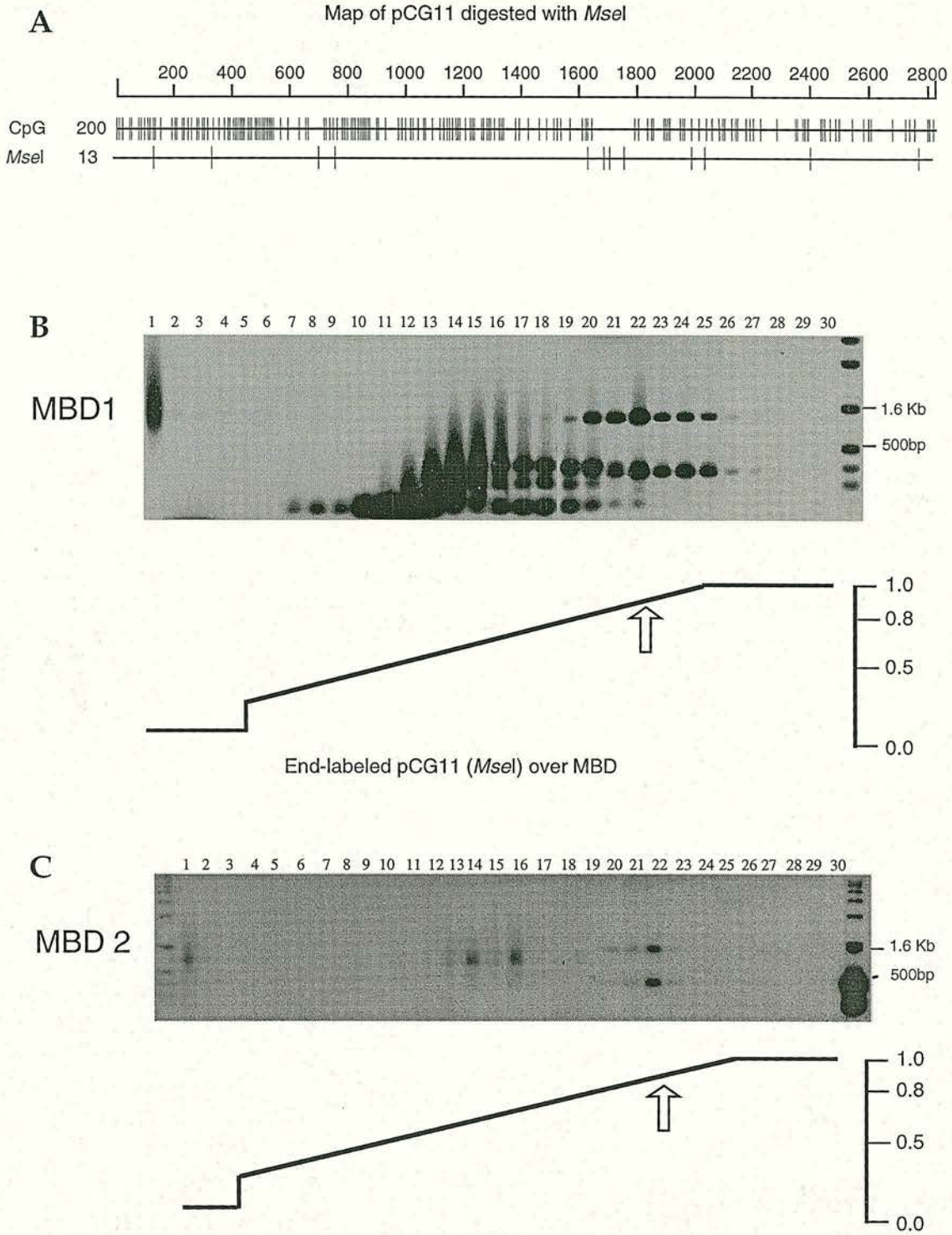
B.

The fully methylated *Mse*I digested pCG11 was loaded onto the MBD column and eluted using a 60 ml linear salt gradient, as shown. Fractions of 2 ml were collected and an aliquot precipitated before drying down and exposing overnight to Kodak XAR film. The approximate salt concentration at which the various fragments elute is shown below. The large open arrow indicates the point at which pCG11 methylated at 200 CpGs elutes from the MBD column (0.85M NaCl) see Figure 3.2.

C.

Fragments which bound to the column above 0.85M NaCl were precipitated, dialysed and resuspended in 0.1M NaCl buffer. The fragments in 0.1M NaCl were then reloaded and eluting using a linear salt gradient. The precipitated aliquots were run on a 1.5% agarose gel, dried down and exposed to film. Figure 3.3.C.

Figure 3.3.


A
Map of pCG11 digested with MseI
200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400 2600 2800
CpG 200
MseI 13
B
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
MBD1
1.6 Kb
500bp
1.0
0.8
0.5
0.0
End-labeled pCG11 (MseI) over MBD
C
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
MBD 2
1.6 Kb
500bp
1.0
0.8
0.5
0.0

3.4.1. Binding of plasmids and plasmid fragments to the MBD column

The column was first shown to be capable of separating 2.8 kb plasmids according to their degree of methylation (Figure 3.2). The affinity of the MBD matrix for smaller fragments of DNA was then investigated. Small fragments of less than 200 bp do not show strong affinity for the column at high salt even when they contain 24 m^5CpGs. At concentrations of 0.80 M NaCl and above, only two large fragments from the digested pCG11 can be detected binding to the column. One of these is the 876 bp fragment, the other is one of or a mixture of the three ~370 bp fragments. This shows that in addition to larger fragments of 2.8 kb the MBD column can separate smaller fragments of different sizes and varying degrees of methylation (Figure 3.3).

3.5. Estimating the proportion of highly methylated DNA in the genome

In the previous section it was demonstrated that the MBD column could separate both 2.8 kb plasmids and smaller fragments according to their degree of methylation. Most of the vertebrate genome is known to be sparsely methylated but there are regions of dense methylation. Digestion of DNA derived from human blood with *Mse*I will produce fragments of different sizes and methyl-CpG densities (Section 1.7). The proportion of *Mse*I fragments of the human genome which were highly methylated enough to bind to the MBD column at high salt was then estimated. The fragments were produced by digesting DNA derived from female blood with *Mse*I. The digested DNA was then end labelled with α–^{32}PdATP and loaded onto the MBD column before eluting using the linear salt gradient.

Figure 3.4.

Separation of human DNA using the MBD column

A.

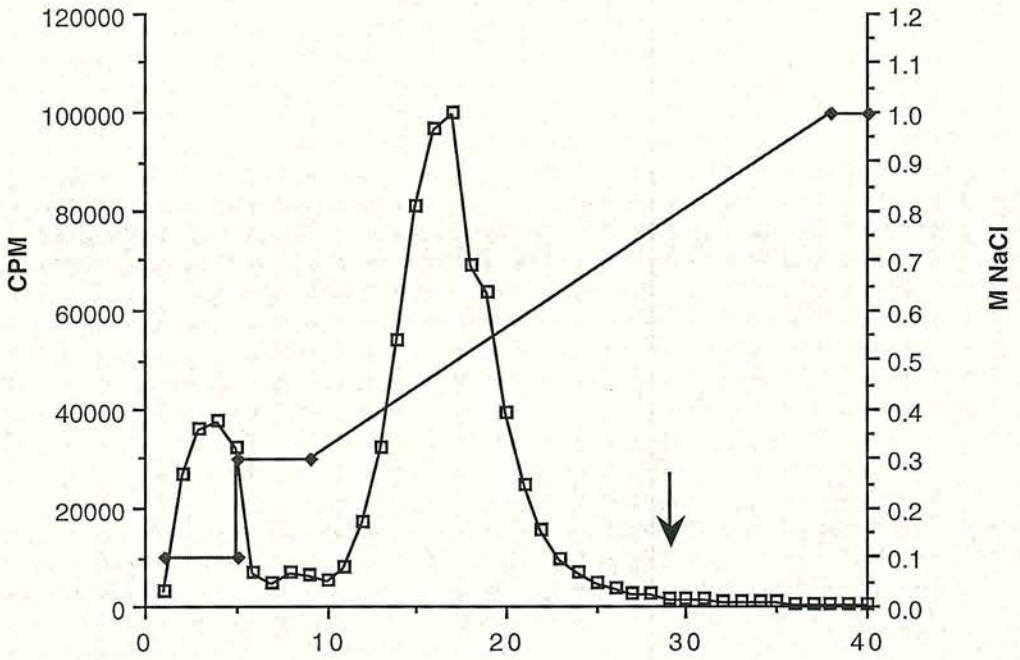
Human DNA derived from female blood was digested with *Mse*I and end-labelled before loading onto the MBD column. The DNA was then eluted using a linear salt gradient and the fractions collected and their activity in cpm measured. The activity of each fraction in cpm was then plotted against NaCl molarity as shown. The arrow indicates the point at which the plasmid pCG11, modified by *Sss*I methylase, elutes from the column.

B.

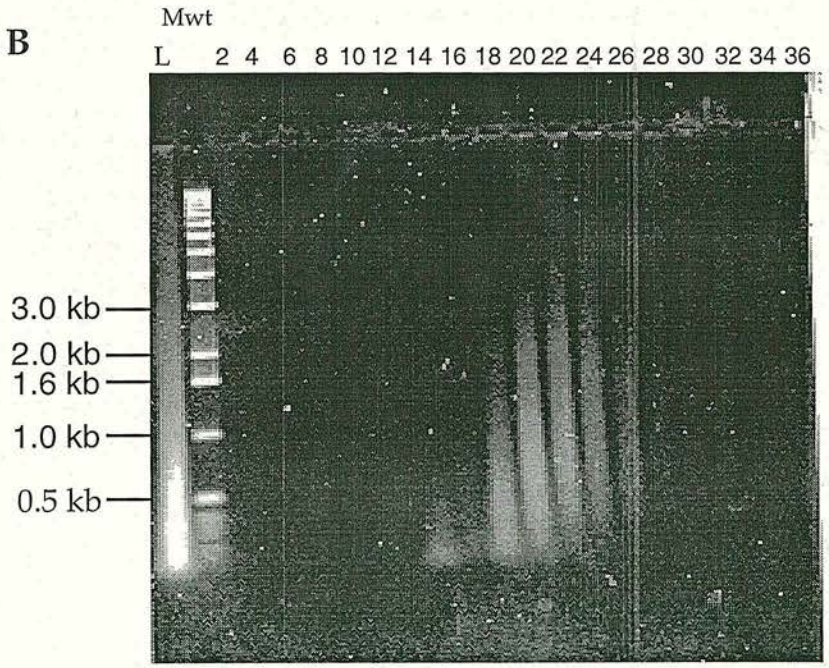
Fractions giving the highest activity (Fractions 1-36) were also ethanol precipitated and run on a 1.5% agarose gel. The size of fragments appears to increase at the higher salt concentrations (see text).

Figure 3.4.

A


CPM
M NaCl
120000
100000
80000
60000
40000
20000
0
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40

B


Mwt
L 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36
3.0 kb
2.0 kb
1.6 kb
1.0 kb
0.5 kb

The activity of the fractions which were collected was measured in cpm and plotted against NaCl (Figure 3.4.A). An aliquot from each fraction was also ethanol precipitated and run on a 1.5% agarose gel before visualising with EtBr. (Figure 3.4.B). The small arrow (Figure 3.4.A) indicates the point at which pCG11, modified by *Sss*I (CpG) methylase elutes (approximately 0.85 M NaCl). This plasmid consists of 200 m^5CpGs or one every 14 bp. The proportion of human DNA which is methylated at a sufficient density to bind the MBD column at high salt is therefore very low. The bulk of the digested DNA elutes between 0.5 and 0.65 M NaCl (Figure 3.4.A). This is also demonstrated by the DNA visible on the agarose gel with the bulk eluting from the column at 0.5-0.6 M. However the trace of radioactivity and the DNA visible in the agarose gel do not appear to be in complete agreement. The cpm suggests a peak of DNA in fraction 18 falling sharply as the salt gradient increases. The size of the fragments visible on the gel increases across the gradient, with the fragments bound at high salt apparently larger. The DNA had first been digested with *Mse*I which produced large numbers of small fragments each of which was end-labelled. The larger number of 'ends' produces a higher activity for example five 100 bp fragments are expected to label with a higher activity than one 500 bp fragment. Thus a larger proportion of the sample of genomic DNA binds to the column than is suggested by the graph (Figure 3.4.A). However the amount of DNA bound to the column at the salt concentration at which pCG11, modified with *Sss*I (CpG) methylase, elutes is still very low.

3.6.
Fractionating human DNA using the MBD column.

The end-labelling of *Mse*I digested genomic DNA in the previous section demonstrates that only a very small proportion shows affinity for the MBD

Figure 3.5.

Diagrams of the sequences used to monitor the separation of human genomic DNA by the MBD column

A.

Diagram of the interferon alpha (IFNα) gene which is depleted for CpG and representative of bulk DNA. The location of the PCR primers on the *Mse*I fragment and the size of product amplified and the number of CpGs are all indicated.

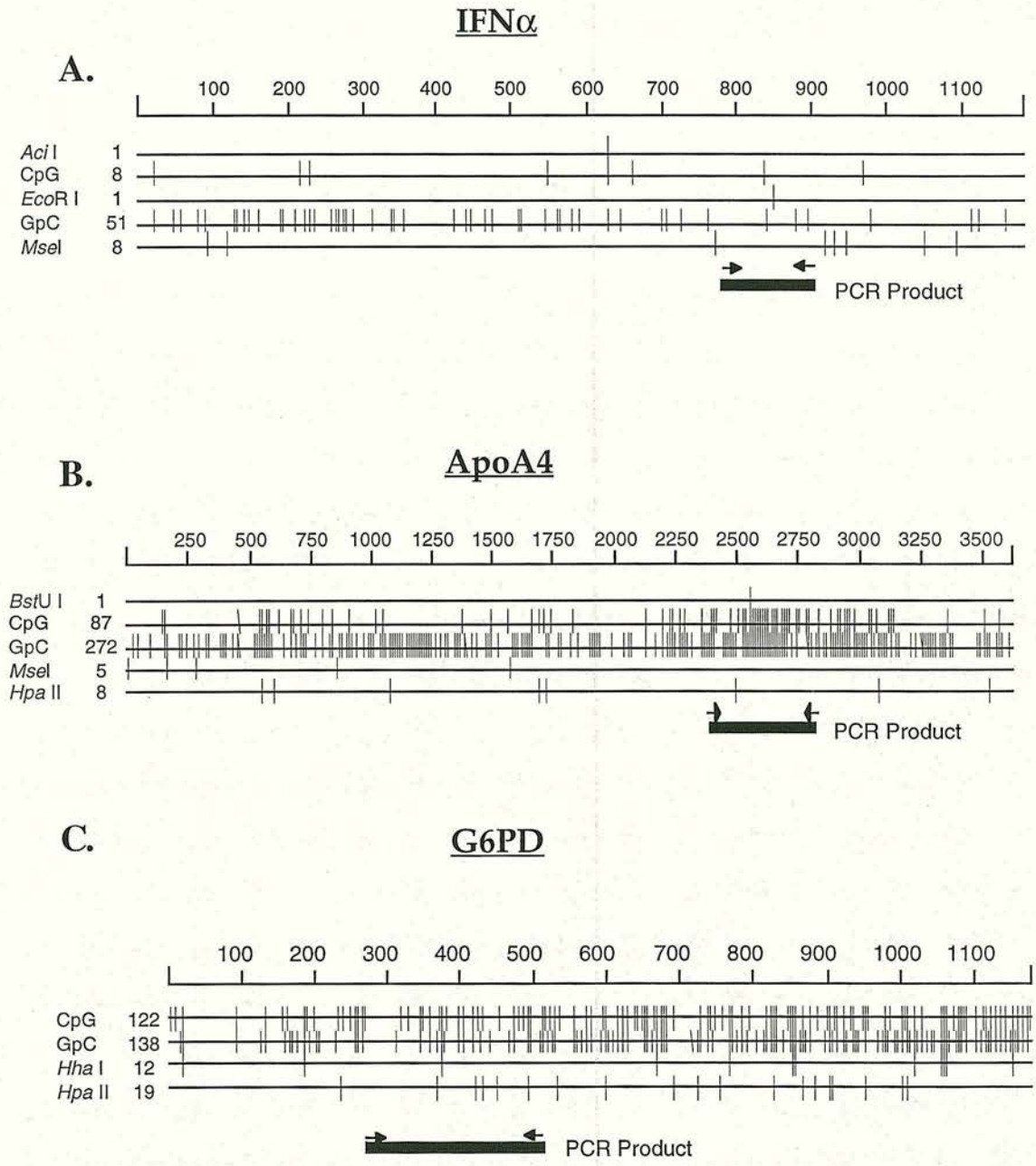
B.

Diagram of ApoA4 showing the 3' CpG island which is heavily methylated in both male and female DNA , Shemer et al (1991). The location of the PCR primers on the *Mse*I fragment and the size of product amplified and the number of CpGs are all indicated.

C.

Diagram of G6PD showing the 5' CpG island covering the first exon. Two copies of this X-linked island will be found in DNA from female blood one methylated and one unmethylated, male blood will contain a single unmethylated copy. The location of the PCR primers on the *Mse*I fragment and the size of product amplified and the number of CpGs are all indicated.

Figure 3.5.


IFNα
A.
100 200 300 400 500 600 700 800 900 1000 1100
AciI 1
CpG 8
EcoR I 1
GpC 51
MseI 8
PCR Product
B.
ApoA4
250 500 750 1000 1250 1500 1750 2000 2250 2500 2750 3000 3250 3500
BstU I 1
CpG 87
GpC 272
MseI 5
Hpa II 8
PCR Product
C.
G6PD
100 200 300 400 500 600 700 800 900 1000 1100
CpG 122
GpC 138
Hha I 12
Hpa II 19
PCR Product

column at high salt. However the previous experiment did not determine if the bound fraction contained only highly methylated fragments. The methylation density of bound fragments was then assessed using primers and amplification. The primers used were from *Mse*I fragments from human blood DNA for which the numbers of CpG were known (Chapter 2).

In order to investigate DNA (500µg) derived from male blood was digested with *Mse*I and loaded onto the MBD column. The samples were then collected using a linear salt gradient as before (Figures 3.2 & 3.4). An aliquot of 500 µl from each 2 ml fraction was EtOH precipitated and resuspended in 10 µl of ddH_20. The column was then cleaned by flushing with 1 M NaCl buffer (40 ml) to ensure no residual DNA remained bound. It was then re-equilibrated with 20 ml of 0.1 M NaCl buffer. Fractions eluting at 0.85 M NaCl and above were pooled and dialysed overnight against 0.1 M NaCl buffer. The dialysed material was then reloaded onto the column at 0.1 M NaCl and the gradient run and fractions collected as before. After the third pass over the MBD column the final bound fractions were pooled and precipitated. The column was cleaned and equilibrated before loading DNA derived from female blood and treating in exactly the same manner as the DNA from the male.

The primers used in the amplification reactions were from specific *Mse*I fragments containing different numbers of m^5CpGs. By monitoring the fractions which eluted from the column, the salt concentration at which these fragments eluted could be determined. This allowed the column's ability to fractionate specific fragments of DNA derived from blood to be assessed. The reactions used 0.5-1.0 µl of the resuspended fractions as template with the primers detailed below. Aliquots were tested from both the unbound,

Figure 3.6.

Amplification of DNA which had been fractionated by the MBD column

MBDx1 indicates the first pass over the column with unbound material eluting at 0.5 M and bound eluting above 0.85 M NaCl.
MBDx2 and MBDx3 are the second and third passes respectively.

A.

Reactions with primers for IFNα and either male or female DNA as template showing product generated using PCR. The *Mse*I fragment from IFNα can no longer be detected in either sample after three passes. The marker lane (M) is a 123 bp ladder (Gibco BRL) . The control (C) contains genomic DNA as template the blank (0) contains the reaction premix with no template DNA.

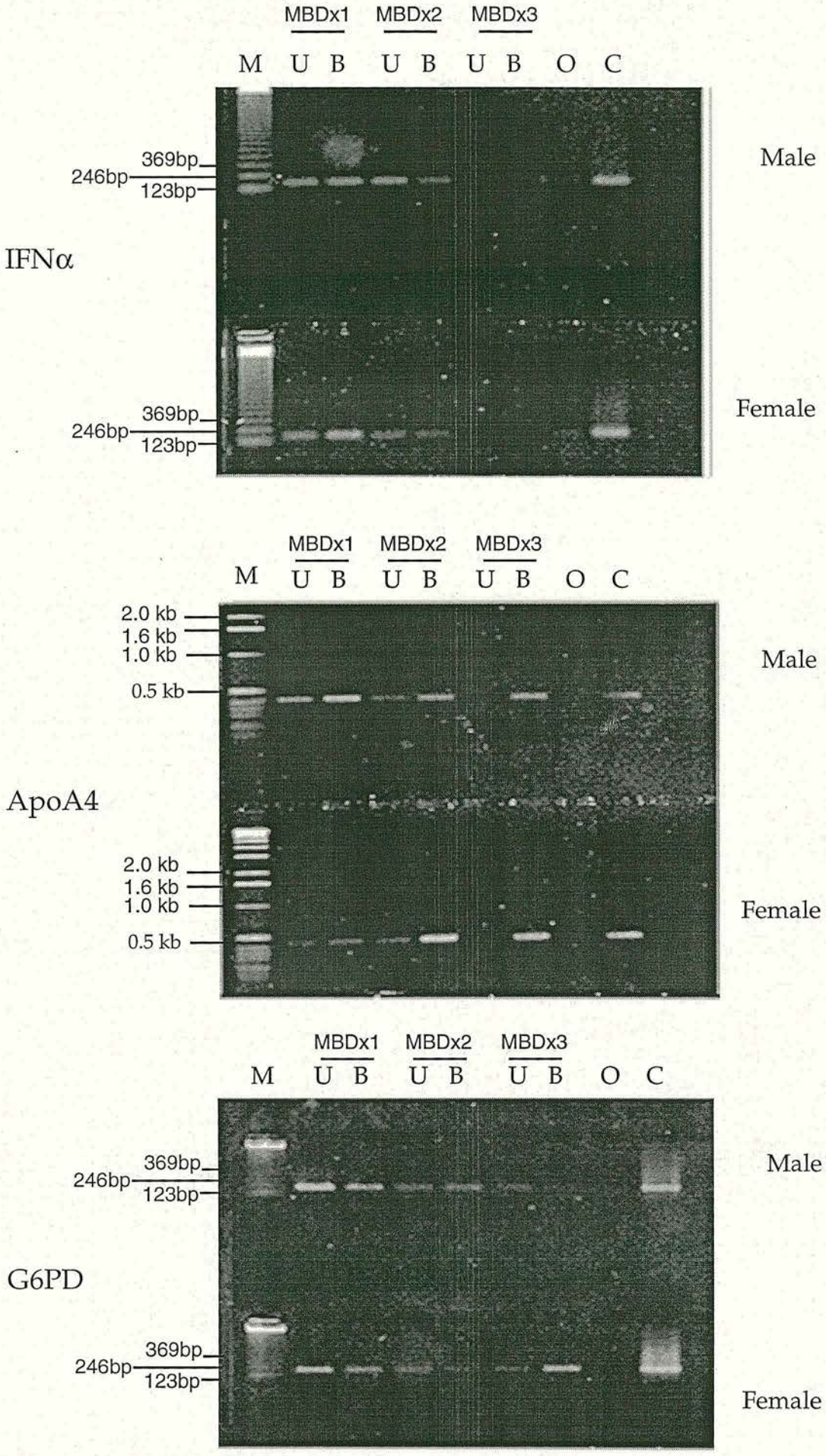
B.

Reactions with primers for ApoA4 and male and female DNA as template showing product generated using PCR. After three passes the *Mse*I fragment can still be detected in both male and female, bound to the column. The marker lane (M) is a 1 kb ladder (Gibco BRL) . The control (C) contains genomic DNA as template the blank (0) contains the reaction premix with no template DNA.

C.

Reactions with primers for G6PD and male and female DNA as template. After three passes only the female derived *Mse*I fragment can still be detected bound to the column. The marker lane (M) is a 123 bp ladder (Gibco BRL). The control (C) contains genomic DNA as template the blank (0) contains the reaction premix with no template DNA.

Figure 3.6.


MBDx1
MBDx2
MBDx3
M U B U B U B O C
369bp
246bp
123bp
Male
IFNα
369bp
246bp
123bp
Female
MBDx1
MBDx2
MBDx3
M U B U B U B O C
2.0 kb
1.6 kb
1.0 kb
0.5 kb
Male
ApoA4
2.0 kb
1.6 kb
1.0 kb
0.5 kb
Female
MBDx1
MBDx2
MBDx3
M U B U B U B O C
369bp
246bp
123bp
Male
G6PD
369bp
246bp
123bp
Female

fractions eluting at 0.5 M NaCl and the bound, which eluted at 0.85 M NaCl and above.

Interferon alpha (IFNα). The primers for IFNα amplify a 178 bp product from a 200 bp *Mse*I fragment from the 3' untranslated end of the gene (Abbott and Povey, 1991) (Figure 3.5.A). The amplified fragment contains a single m^5CpG and was considered to be representative of bulk genomic DNA which is sparsely methylated. The fragment could not be detected in either male or female DNA bound to the column after the third pass (Figure 3.6.A).

Apolipoprotein A (ApoA4). These primers amplify a 429 bp product from a 3 kb *Mse*I fragment. The *Mse*I fragment is from the 3' CpG island which is methylated in both male and female somatic cells (Shemer et al., 1991). (Figure 3.5.B). The primers are positive in the bound fraction after the third pass for both male and female DNA showing that the highly methylated fragment is still bound (Figure 3.6.B).

Glucose-6-phosphate Dehydrogenase (G6PD) These primers amplify a 202 bp product from a 1.1 kb *Mse*I fragment from the 5' X-linked CpG island (Figure 3.5.C) (Zollo et al., 1994). DNA derived from male blood will contain a single unmethylated copy but when derived from female blood the DNA will contain one unmethylated and one methylated copy. After three passes of female DNA the *Mse*I fragment from G6PD can still be detected but none can be detected in the male (Figure 3.6.C).

3.7. Conclusions

The MBD column separated plasmids and fragments of plasmids according to the number of m5CpG. The separation of human blood derived DNA suggested that very few fragments generated by an *Mse*I digestion contain

enough methyl groups to bind to the column at high salt. Additionally the results of the PCR showed that it was necessary to pass DNA over the column at least three times in order to purify these methylated fragments. During the first two passes fragments of DNA which are methylated can be detected in the unbound and vice-versa. However after the third pass, only fragments known to be highly methylated in blood DNA could be detected in the retained fractions (ApoA4). The column also separated an X-linked fragment from the female DNA (G6PD). As it was from a CGI, this fragment would be present in both a methylated and unmethylated form in female blood DNA. The same fragment was unmethylated in DNA derived from male blood and could not be detected in the final bound fraction. It was therefore considered that the fragment detected in the bound fraction of female DNA was from the methylated copy. At least three passes were required but the column had apparently separated DNA fragments according to the number of m5CpGs present. The separated fragments could now be cloned and further investigated.

Chapter 4 : Libraries of highly methylated *Mse*I fragments from human blood DNA

4.1.
Introduction.

In the previous chapter, plasmid DNA and digested genomic DNA was fractionated according to the frequency of m5CpG. Using an identical procedure, DNA derived from human blood was also fractionated, according to m^5CpG numbers, using the MBD column. The results obtained in Chapter 3 indicated that three rounds of purification were required to effectively separate highly methylated fragments. Following this procedure a library of fragments with strong affinity for the column matrix at high salt was generated. Only a very small fraction of the starting amount of DNA would consistently bind to the column after three rounds of purification (MBDx3). In order to efficiently clone this material catch-linkers were attached and the DNA was amplified. It was anticipated that these cloned sequences would have a significantly higher %GC and $CpG_{Obs/Exp}$ than the majority of the genome. The sparsely methylated remainder of the genome being defined as having 40% GC and a $CpG_{Obs/Exp}$ of 0.20. (Gardiner-Gardner and Frommer, 1987). In addition to these sequence characteristics, the cloned fragments should, unlike most CGIs, be methylated in DNA derived from blood. Methylated CGIs from the inactive X and from imprinted genes will also be represented. The methylation status of sites within the cloned sequences could be confirmed using southern-blots.

4.2.
Analysis of the first library (MBDx3)

A random selection of clones were sequenced to determine the %GC and $CpG_{Obs/Exp}$. The results are shown in Table 4.1. and Figure 4.1. The preliminary sequence data indicated that this first library contained few inserts with the characteristics of a CGI. Figure 4.1. shows the %GC plotted against the $CpG_{Obs/Exp}$ of sequences from the database with known 5' CGIs (black solid squares) or of sequences known to be depleted for CpG (solid triangles) (Gardiner-Gardner and Frommer, 1987). The %GC and the $CpG_{Obs/Exp}$ of the cloned sequences from the female blood DNA (MBDx3) library are shown as open triangles. With the exception of clones p4C and p6D, the sequences are not significantly more GC-rich and only have a marginally higher $CpG_{Obs/Exp}$ than those which represent CpG depleted DNA.

4.2.1.
Conclusion concerning the MBDx3 library

Analysis of the MBDx3 sequences suggested that the library was not, as expected, comprised exclusively of highly methylated sequences. The criteria for highly methylated sequences, representative of a CGI, was a %GC of greater than 50 and a $CpG_{Obs/Exp}$ of greater than 0.60 (Gardiner-Gardner and Frommer, 1987). Examples of sequences with these characteristics are represented by black squares in the graph shown (Figure 4.1). Sequences which were representative of CpG depleted fragments had a %GC of less than 45 and a $CpG_{Obs/Exp}$ of less than 0.4 (solid triangles, Figure 4.1). The cloned inserts examined from the library (MBDx3) cover a wide range and had an average %GC of 50 with an average $CpG_{Obs/Exp}$ of 0.45. With the

exception of p4Cf and p6Df, the sequences cloned were only slightly more GC-rich than those which represent DNA depleted for CpG. It was unclear why sequences with CpG frequencies of one every 50 or 100 bp consistently bound to the MBD column at a salt concentration above 0.6 M. (Table 4.1.) When testing the column, using both plasmid and human blood derived DNA, sequences with few m^5CpGs were monitored eluting at low salt after three round of purification (Chapter 3). Several of the cloned inserts had the %GC and a $CpG_{Obs/Exp}$ of depleted sequences, examples include p1Af (38% and 0.11) p2Cf (47% and 0.17) p3Dr (50% and 0.21) and p2Fr (42% and 0.32) (Table 4.1). It was thought possible that during the amplification step, prior to cloning, the polymerase may have preferentially amplified non GC-rich sequences. This would have had the effect of increasing low levels of 'contamination' by any GC-poor sequences still present in the fractions eluting at high salt. In an attempt to avoid potential polymerase bias, DNA bound at high salt was cloned directly. The number of colonies obtained was not however sufficiently large enough to be representative. As a result a second library was made using the same column and methods as before. However an additional step was added in order to further purify fragments after the amplification step. This involved the methylation of the amplified DNA followed by two further rounds of selection using the MBD column (Section 4.3.). After methylation fragments containing large numbers of CpGs, now m^5CpGs, would be again be retained. These would then be purified from the fragments containing few methylatable CpGs.

Table 4.1.

Sequences of a random selection of clones from the MBDx3 library

The identification number of each plasmid containing a cloned insert is shown in the first column. Inserts were sequenced using either a forward (f) or reverse (r) primer, the inserts cloned in p2D and p4E were sequenced completely (see section 2.14.) . In the remaining columns, reading from left to right, the number of base pairs , the %GC, the number of CpGs, number of GpCs, number of CpGs divided by GpCs and the $CpG_{Obs/Exp}$ of the available sequence is shown.

Table 4.1.

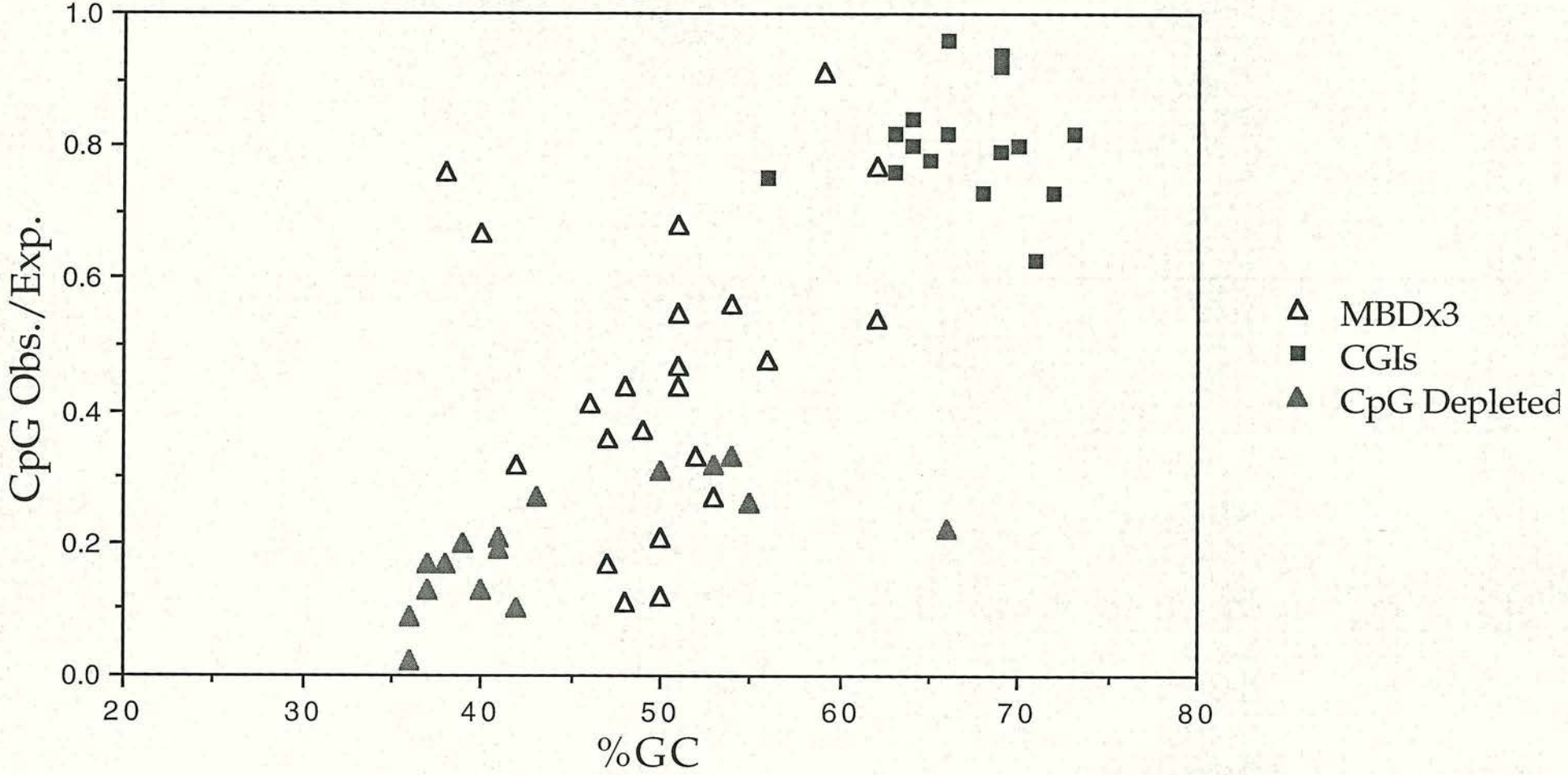
Clone ID	Seq'd	%GC	No. CpG	No. GpC	CpG/ GpC	Obs./ Exp.
p1Af	327	48	2	18	0.11	0.11
p2Ar	429	38	9	6	1.50	0.76
p2Bf	222	51	6	7	0.85	0.44
p2Br	399	52	9	12	0.75	0.33
p2Cf	441	47	4	11	0.36	0.17
p2Cr	271	47	5	10	0.50	0.36
p3Cf	268	48	6	9	0.66	0.44
p4Cf	372	62	27	29	0.93	0.77
p4Cr	387	62	20	40	0.50	0.54
p2D	473	40	13	22	0.59	0.67
p3Df	378	46	8	15	0.53	0.41
p3Dr	329	50	4	8	0.50	0.21
p4Df	454	49	10	29	0.34	0.37
p4Dr	435	50	3	22	0.14	0.12
p6Df	385	59	30	28	1.1	0.91
p6Dr	268	54	11	17	0.64	0.56
p2Ef	159	53	3	9	0.33	0.27
p3Er	248	51	7	13	0.54	0.47
p4E	370	51	16	27	0.59	0.68
p6Ef	215	56	8	15	0.53	0.48
p6Er	309	51	11	20	0.55	0.55
p2Fr	225	42	3	40	0.75	0.32

Figure 4.1.

Graph of %GC versus $CpG_{Obs/Exp}$ for the sequences from Table 4.1.

The graph shows the of %GC and $CpG_{Obs/Exp}$ of the cloned sequences from the MBDx3 library (white triangles) shown in Table 4.1. The sequence data from fragments of known CGIs (black squares) and the sequence data from fragments depleted for CpG (black triangles) are graphed for comparison. The sequence data for CGIs and depleted sequences was taken from Table 1 of the study by Gardiner-Garden (1987).

Figure 4.1.


1.0
0.8
0.6
0.4
0.2
0.0
CpG Obs./Exp.
20
30
40
50
60
70
80
%GC
MBDx3
CGIs
CpG Depleted

4.3.

Preliminary analysis of enriched sequences derived from male and female blood DNA (MBDx5)

The addition of a further step was intended to counter the possibility that the amplification step had been biased towards non GC-rich sequences. The same procedure used in chapter 3 was again used to generate a fresh batch of both male and female DNA (MBDx3). The DNA (MBDx3) was amplified and treated with *Sss*I (CpG) methylase in order to methylate all available CpGs (Section 2.5). An aliquot of the methylated DNA was then removed and end-labelled (Section 2.9.1). This allowed the affinity of the MBD column for the artificially methylated DNA to be monitored at increasing salt concentration. The graphs indicate that the MBD column had little affinity at a high salt concentration for the majority of the amplified and methylated male and female DNA (MBDx4) (Figure 4.2). The small fraction of DNA which showed affinity (black arrow) was then dialysed. When this was reloaded the majority of the DNA (both the male and female) now showed affinity for the MBD column above 0.8 M NaCl (Figure 4.2).

In order to investigate the bound sequences further, the remaining *Sss*I (CpG) methylase treated male (MBDx3) DNA was passed over the column. A linear salt gradient was used and the fractions binding above 0.85 M retained and dialysed before reloading. After a second pass, the fractions eluting above 0.85 M were retained, precipitated and resuspended in ddH_2O (MBDx5). The MBD column was then cleaned and re-equilibrated before the *Sss*I (CpG) methylase modified female DNA (MBDx3) was treated in exactly the same manner. Prior to cloning these sequences, they were again amplified. When these reactions were visualised on a gel, a smear of DNA fragments from around 2-3 kb down to 200 bp could be seen (Figure 4.3.A).

Figure 4.2.

The amount of purified DNA showing affinity for the MBD column following amplification and artificial methylation

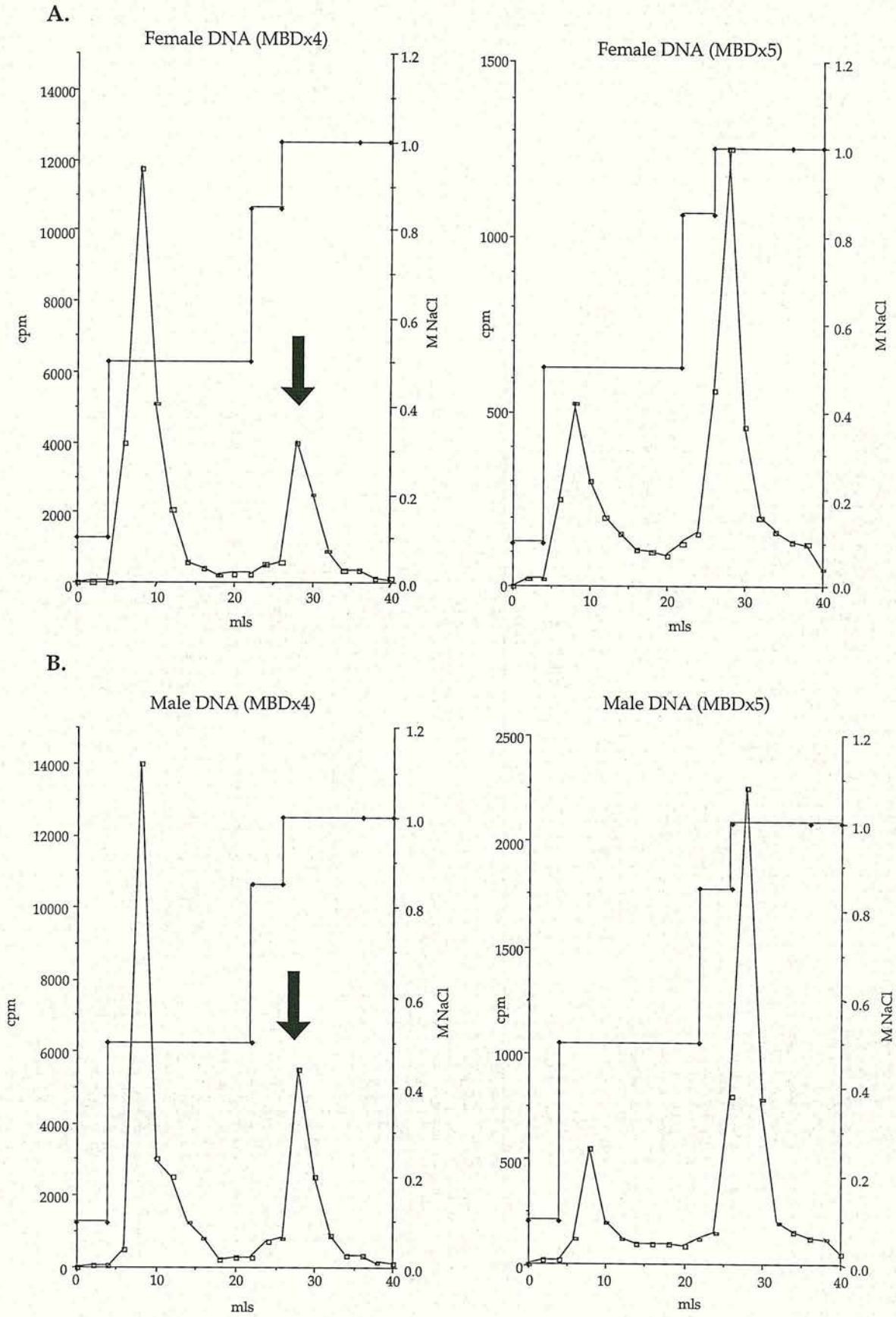
A.

An aliquot of the female MBDx3 DNA treated with *Sss*I (CpG) methylase was end-labelled. The DNA was then passed rapidly over the column using a step gradient. At a salt concentration of 0.5 M the bulk of the modified DNA does not bind to the MBD matrix (Female MBDx4). The small proportion that did bind (black arrow) was dialysed, reloaded and monitored as before. Some of the monitored DNA (Female MBDx5) does not bind and elutes from the column at low salt concentration. However the bulk of the dialysed and reloaded DNA does not elute until the salt concentration reaches 0.85 M. All the bound material has eluted at a concentration of 1.0 M (MBDx5).

B.

An aliquot of the amplified and artificially methylated DNA from male blood was end-labelled and passed over the column. As with the DNA derived from female blood most of the material does not bind to the column (Male MBDx4). The small amount that does bind (black arrow) continues to show affinity when dialysed and reloaded onto the MBD column (Male MBDx5).

Figure 4.2.


A.
Female DNA (MBDx4)
Female DNA (MBDx5)
B.
Male DNA (MBDx4)
Male DNA (MBDx5)
cpm
M NaCl
mls

In addition to the smear of heterogeneous fragments three bands could be seen. Although much fainter, in relation to the smear, the same three bands could also be seen when DNA bound to the column after three rounds (MBDx3) was amplified. The smallest band in the MBDx5 amplified DNA (Figure 4.3.A. arrow 1) was around 550 bp in size, the other two were approximately 800 and 900 bp respectively (Figure 4.3.A. arrows 2 & 3). Amplification of unfractionated DNA digested with *Mse*I and with attached catch-linkers failed to produce any visible bands (Figure 4.3. see control). Notably, both male and female DNA (MBDx5) produced the same three major bands when amplified. The MBDx5 DNA derived from both male and female blood was then analysed further to detect fragments containing known numbers of m^5CpG (Chapter 3). The primers for both G6PD and IFNα were used to amplify from the male and female DNA (MBDx5). The results are shown in Figure 4.3.B. The primers for IFNα give a negative result with both male and female DNA (MBDx5) as expected as they are representative of bulk DNA. However the primers for G6PD, an X-linked fragment, gave a positive result for female but not for male (Chapter 3 section 3.6). The primers for monoamine oxidase B (MOAB) another X-linked fragment also gave a positive result for female but were negative for male (Figure 4.3.B.).

4.3.1. Conclusions regarding the preliminary analysis of enriched sequences (MBDx5)

When amplified by PCR and artificially methylated, the majority of the DNA (MBDx3) derived from both male and female blood showed little affinity for the column even at low salt (MBDx4, Figure 4.2). However the small amount which did show affinity for the column continued to bind when it was dialysed and reloaded (MBDx5). It is possible that the PCR reaction

Figure 4.3.

The amplification of specific fragments from MBDx5 DNA

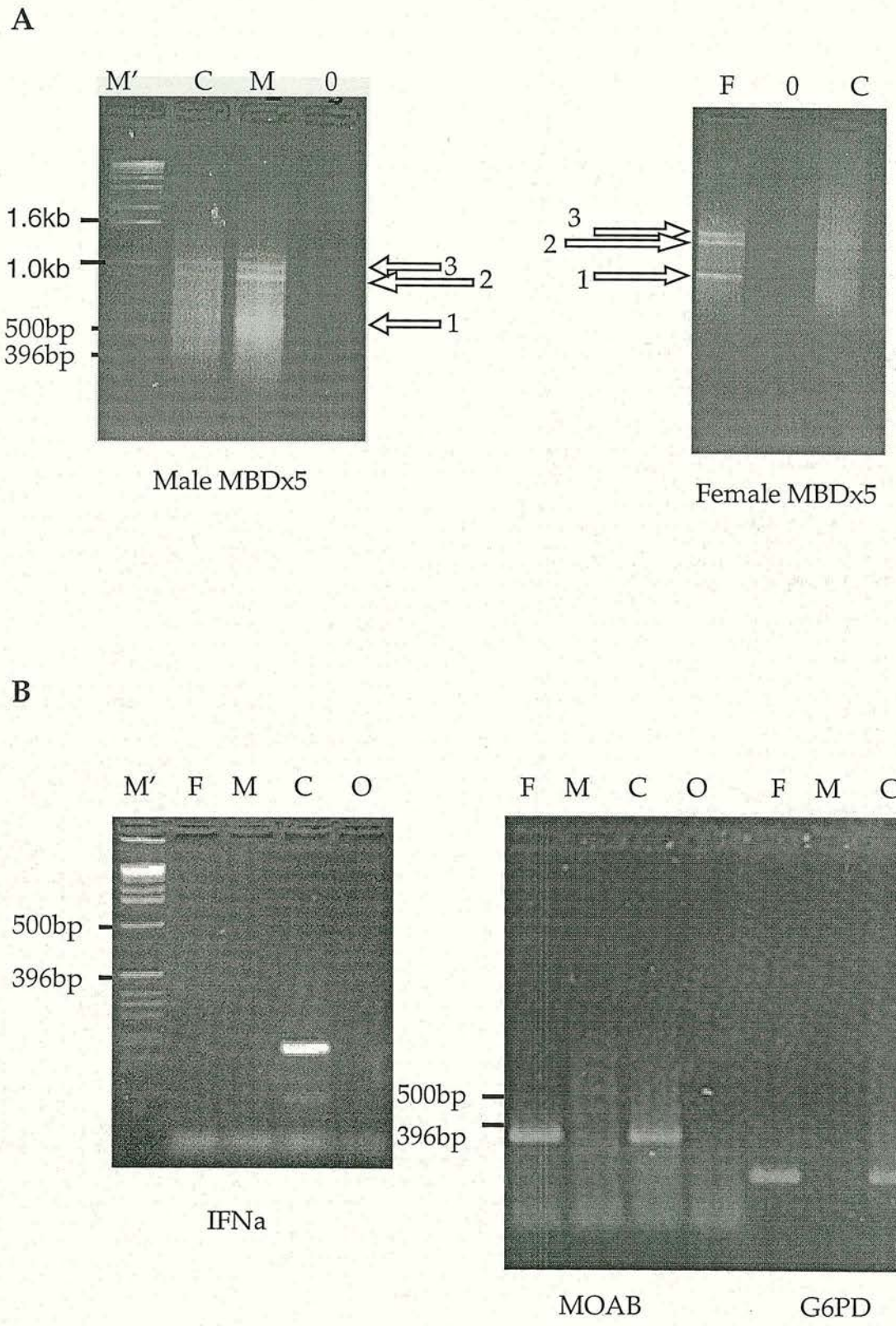
A.

Male and female DNA, treated with *Sss*I methylase was passed over the MBD column as detailed in the text. The fractions still bound to the column at 0.85 M NaCl and above were pooled and precipitated. An aliquot of the MBDx5 DNA was amplified using the attached catch-linkers as primers. The product of this reaction was electrophoresed on an agarose gel as shown. The marker (M') lane contains a 1 kb ladder (Gibco BRL). The products of the reaction vary from 2-3 kb in size down to 200 bp with three clearly visible bands (arrows 1, 2 & 3). Amplification of both male (M) and female (F) DNA (MBDx5) produced the same size of bands. The control reactions (C) contained human DNA digested with *Mse*I, with attached catch-linkers, the blank reactions (O) contained no DNA.

B.

After five passes over the column (MBDx5) Male (M) and female (F) DNA was further analysed by amplifying with primers for IFNα and G6PD (Section 3.6.). Primers for a region on a second X-linked 5'CGI, Monoamine oxidase B (MOAB) were also used. The female MBDx5 DNA tests positive for both the 5' X-linked fragments but not for the IFNα representing bulk DNA. The male MBDx5 DNA sample tests negative for all three. The marker lane (M') contains a 1 kb ladder (Gibco BRL). The reaction control (C) contains unfractionated human DNA, the reaction blank (O) contains none.

Figure 4.3.


A
M′ C M 0
1.6kb
1.0kb
500bp
396bp
3
2
1
Male MBDx5
F 0 C
3
2
1
Female MBDx5
B
M′ F M C O
500bp
396bp
IFNa
F M C O F M C O
500bp
396bp
MOAB
G6PD

preferentially amplified shorter or less GC-rich fragments which when artificially methylated, would not have a strong affinity for the MBD column. The amplification of such sequences would however lead to contamination of the final fraction. This may explain the GC-poor and CpG depleted characteristics of some of the inserts analysed from the MBDx3 library (Figure 4.1).

When the DNA, derived from either male or female blood, which did show affinity for MBD was amplified, the results were very similar (Figure 4.2.A). In addition to the heterogeneous smear of DNA, three distinct bands were visible at approximately 550, 800 & 900 bp in size in both the male and female MBDx5. This material was further analysed using primers for fragments which contain either a few (IFNα) or a large number of CpGs (G6PD & MOAB). The fragment representative of bulk DNA (IFNα) could again not be detected in either (Figure 4.2.B). The fragments from two X-linked islands could be detected in female but not in the male MBDx5. Such fragments are expected to remain bound to the column at high salt and to be purified. As they contain large numbers of CpGs they would again be heavily methylated after the treatment with *Sss*I (CpG) methylase. Two representatives of methylated CGIs were therefore still present in the female MBDx5 fraction. In addition to testing positive for these fragments, when labelled and monitored, the MBDx5 also shows strong affinity for the column at high salt concentration. It was therefore cloned and the library then analysed to see if the sequences did possess the correct characteristics.

4.4. Further analysis of enriched sequences derived from female blood DNA (MBDx5)

The DNA (MBDx5) originally derived from female blood was cloned for further analysis, using the T-Vector system (Promega). Individual clones were picked and transferred to 96-well plates which allowed duplicate colony blots to be produced. During the preliminary analysis, amplification of the cloned DNA (MBDx5) had revealed three distinct bands (Figure 4.3.A) which were present in both male and female fractions. There are a number of highly repetitive sequences present in the human genome. Many of these sequences are known to be both GC-rich, and, in some cases, methylated. It was therefore necessary to first determine if these sequences were over-represented in the library. Such sequences would include short interspersed elements, SINES and long interspersed repeated elements, LINES. Members of the SINE family include *Alu* elements and examples of LINES include the LI (*Kpn*I) sequence families (Labuda et al., 1995). The human *Alu* repeats are reported to comprise almost 5% of the genome and are often highly methylated (Kochanek et al., 1993). The *Alu* elements belong to subfamilies with the 'young' subfamily being nine-fold enriched in CpG compared to total human DNA (9% vs 1%) (Schmid and Maraia, 1992). It has been calculated that *Alu* repeats account for about one-third of the potential methylation sites in human DNA (Rubin et al., 1994). It was considered that the MBD column would have a strong affinity for methylated *Alu* sequences and as a result they could be represented in the library. In order to screen for repeated sequences, the 96-well plates were blotted onto Hybond C extra membrane (Amersham) and probed with DNA digested with *Mse*I (total genomic probe) (Figure 4.4.Ai). This probe will hybridise strongly to sequences which are abundantly represented in the genome. An identical

colony blot was then probed with an insert from a plasmid containing a consensus *Alu* sequence (pBL8 probe, a gift from A. Bird) (Figure 4.4.Aii). The sequences of an insert which hybridised strongly to both these probes is shown (Figure 4.5.A). The cloned insert shows sequence similarity to a consensus *Alu* sequence from the Sc-subfamily (Jurka and Milosavljevic, 1991; Claverie and Makalowski, 1994). A second insert which hybridised to the 'total human' probe but not to pBL8 showed good sequence similarity to the 5' end of the human transponson L1.2. (Crowther et al., 1991) (Figure 4.5.B). The L1 or LINE-1 family of long dispersed repetitive elements are related evolutionary to known retrotransposons. In mammals the 5' region of these sequences has been reported to be GC-rich with a high content of CpG and is methylated in a number of cell types (Crowther et al., 1991; Schmid and Maraia, 1992). The total genomic and pBL8 probes indicate that some highly repeated sequences in the genome are represented in the final library. Frequently occurring sequences in the genome did not appear to be over-represented in the library. However, in order to avoid repetitive sequencing, the 96-well plates were colony blotted and pre-screened using the cloned material (MBDx5) as a probe (Figure 4.4.Bi). Any cloned inserts represented frequently in the library would hybridise strongly to this probe and could be investigated further.

A number of positive clones were identified in this way and one insert which did occur frequently in the library was sequenced. Five different clones were analysed to give a consensus sequence for this insert. The cloned fragment was 534 bp in size with a %GC of 59 and a $CpG_{Obs/Exp}$ of 0.67. This sequence was used as a query in a search of the NCBI data base using the

Figure 4.4.

Probing colony blots of 96-well plates to estimate the percentage of repeated clones in the library

Ai. Colony blot from a 96-well plate probed with human DNA that had been digested with *Mse*I and end-labelled. The number of clones testing positive suggested that the library did contain significant numbers of fragments from sequences that occur frequently in the genome, i.e. SINES. The average figure over the five plates examined was 17% of clones testing strongly positive for the "total human DNA" probe.

Aii. Colony blot of the same plate probed in Ai, probed with insert from the plasmid pBL8 which contains an *Alu* consensus sequence. Fourteen clones gave a positive signal when probed with both the *Alu* consensus sequence and the total human DNA probe (see text for details). Three examples of clones which hybridise pBL8 and the total human probe are indicated with white arrows (see also Ai).

Bi. Example of a colony blot from a 96-well plate probed with the amplified DNA (MBDx5) used to make the library. Clones that hybridised strongly were thought to contain inserts that occur frequently in the library, an average of 40% of the clones hybridise strongly. One of these frequently occurring clones was determined by database searches to be an *Mse*I fragment of the rDNA NTS

Bii. Example of a colony blot from a 96-well plate probed with the insert from the clone p1B12. The insert from this clone was an *Mse*I fragment from the rDNA NTS and was in an averages 25% of the clones tested (see text for details).

igure 4.4.

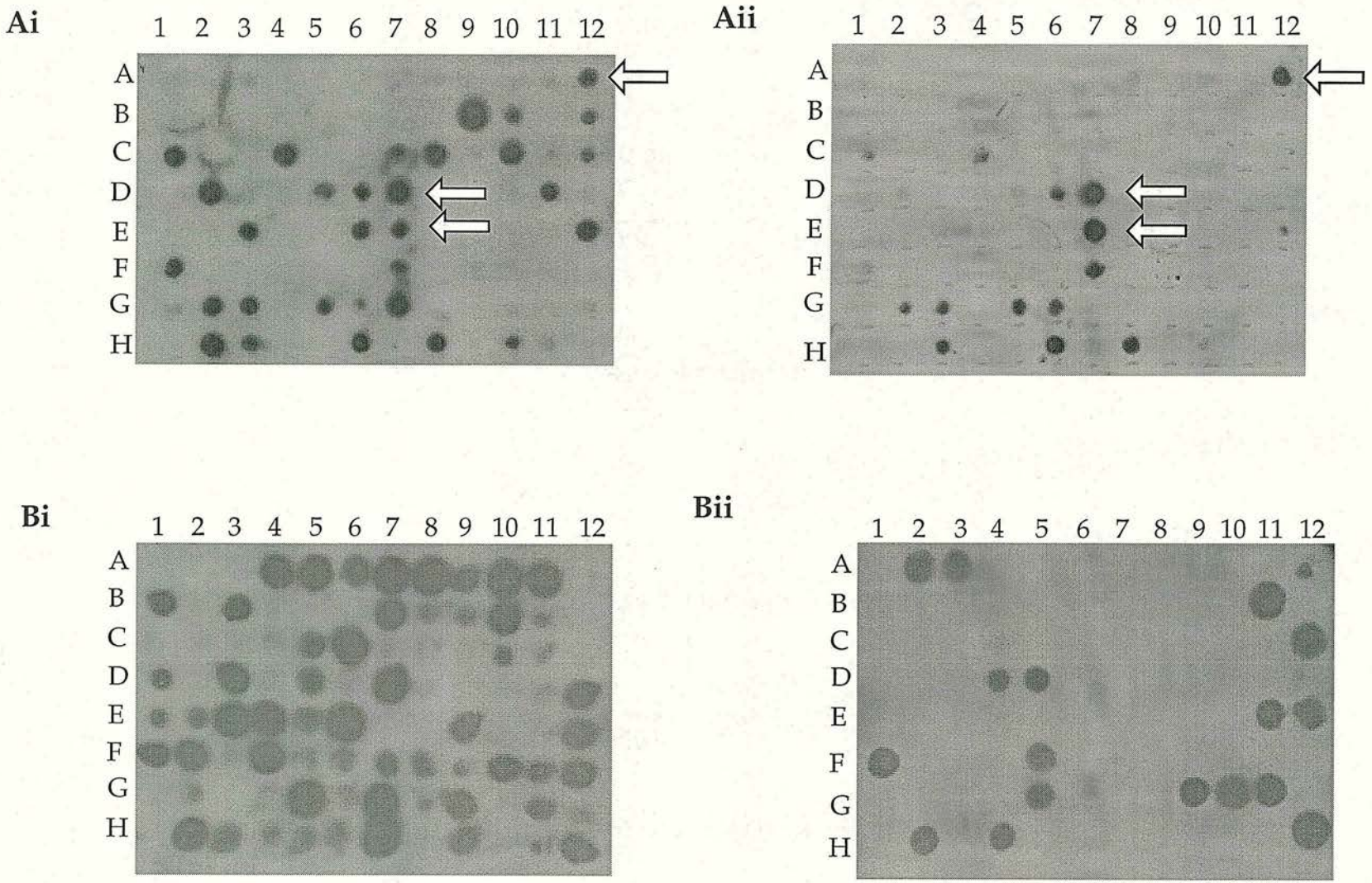

Ai
1 2 3 4 5 6 7 8 9 10 11 12
A
B
C
D
E
F
G
H
Aii
1 2 3 4 5 6 7 8 9 10 11 12
A
B
C
D
E
F
G
H
Bi
1 2 3 4 5 6 7 8 9 10 11 12
A
B
C
D
E
F
G
H
Bii
1 2 3 4 5 6 7 8 9 10 11 12
A
B
C
D
E
F
G
H

Figure 4.5.

Alignment of cloned sequences that hybridised strongly to pre-screening probes and sequences from the data-base

A.

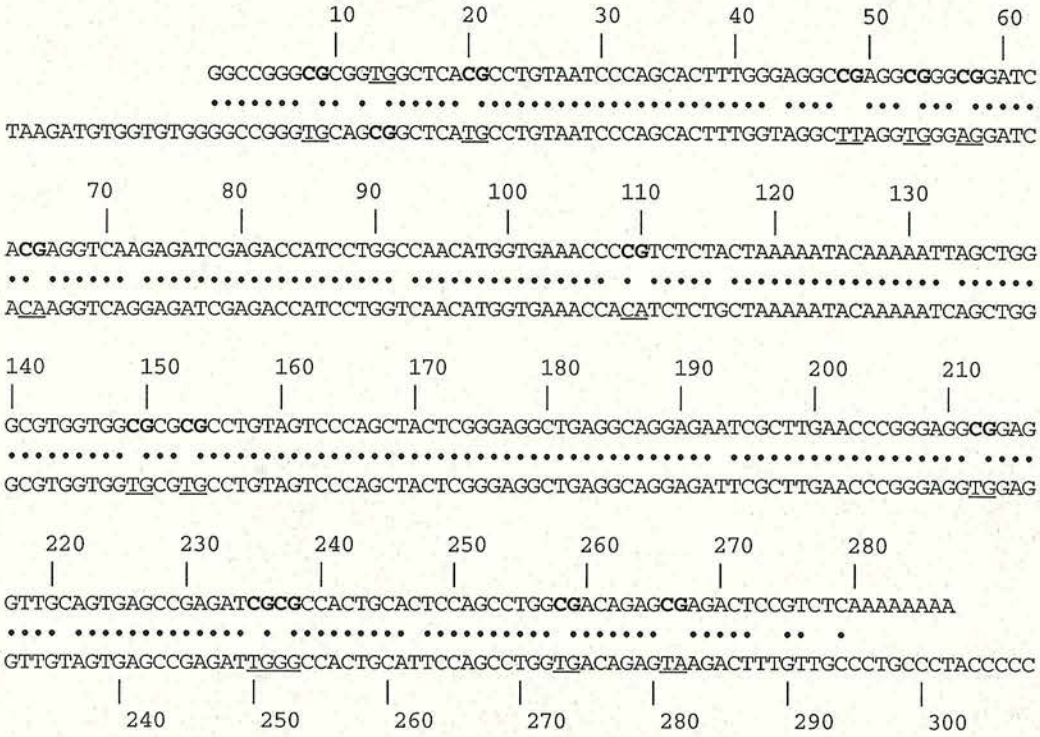
Comparison of the consensus sequence from the inserts of clones p1B6 and p1A8 and an *Alu* Sc. subfamily consensus sequence (Accession No.U14571). The consensus sequence of clones p1B6 and p1A8 is 593 bp in size of which 275 bp match to the *Alu* consensus as shown. Fifteen CpGs in the *Alu* Sc consensus sequence are shown in bold, the CpGs at the same position in the cloned sequence have often mutated to TpG or CpA. A single CpG in the cloned sequence occurs as a TpG in the *Alu* Sc. sequence, the remaining CpGs are common to both. The remaining 318 bp of insert p1B6/p1A8 contains a further 11 CpGs, giving the entire insert a CpG Obs./Exp of 0.48 and a %GC of 52. The consensus sequence from the two clones also gives a very high sequence similarity to the *Alu*-Sb subfamily consensus sequence . The *Alu*-Sb subfamily sequence matches to 138 bp out of 593 bp (not shown).

B.

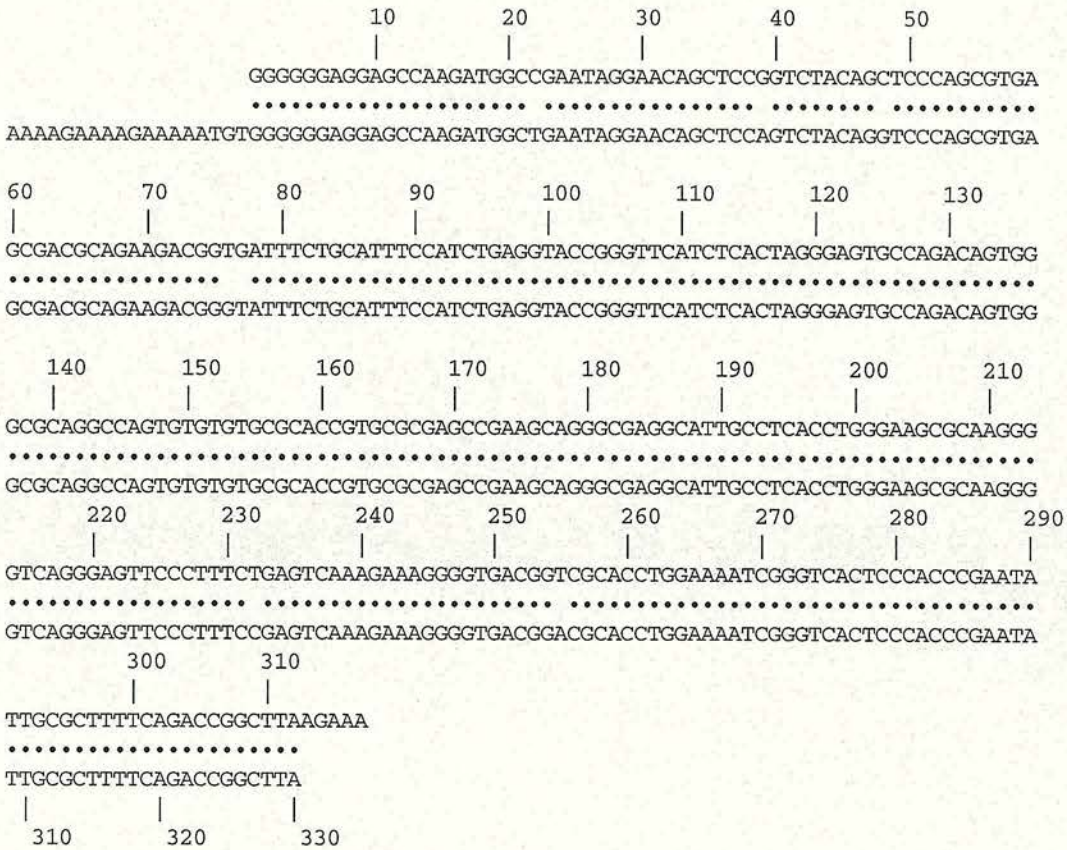
Comparison of insert from clone p4H8 and the database sequence for the 5' end of a human transponson L1.2. (Accession No. X58075). Almost the entire 330 bp of the cloned sequence matches the sequence from the database including 20 out of 22 CpGs. Note also that the cloned sequence ends in an *Mse*I site (TTAA).

Figure 4.5..

A. Alu Sc. (Top) subfamily consensus sequence p1B6/p1A8 (Bottom)



B. Human transposon L1.2. (Top) vs p4H8 (Bottom)



BLAST program (Altschul et al., 1990). The result showed that there was sequence similarity to a 536 bp *Mse*I fragment from the non-transcribed spacer (NTS) of the human ribosomal DNA (rDNA) complete repeating unit (Figure 4.6) (Accession Number U13369). The plasmid from one of the recombinant clones (p1B12) was purified and the insert was used as a probe in a Southern blot to determine its methylation status in the genome. The result of probing DNA (5 μg per lane) derived from male or female blood or from sperm then digested with *Mse*I followed by either *Hpa*II or *Msp*I is shown (Figure 4.7.A). The hybridisation pattern of insert from p1B12 to this blot shows that the four *Hpa*II sites in the sequence are methylated in blood but hypomethylated in sperm. Further investigation of the methylation status of the entire rDNA NTS was subsequently performed (see Chapter 6 and Brock and Bird, 1997). As the insert was of approximately the same size as the smallest band in the amplified material it was also used as a probe (p1B12) against the colony blot filters (Figure 4.4.Bii). In total seven filters were screened and this sequence was found to comprise almost 25% of the cloned inserts in the library.

During the pre-screening a second frequently occurring clone had been identified. A consensus sequence was obtained from two such clones giving an insert size of 857 bp with a %GC of 58.6 and a $CpG_{Obs/Exp}$ of 0.69. When this was used as a query against sequences in the NCBI data-base using the BLAST program, no sequence similarity was found. The insert from one of the clones (p1A12) was used as a probe in a southern-blot (Figure 4.7.A). As with insert of p1B12, genomic DNA corresponding to the insert of p1A12 was methylated at *Hpa*II sites (seven) tested in DNA derived from blood. Unlike p1B12 when the DNA was derived from sperm was probed, with the insert of p1A12, the seven sites were completely unmethylated. It was noted

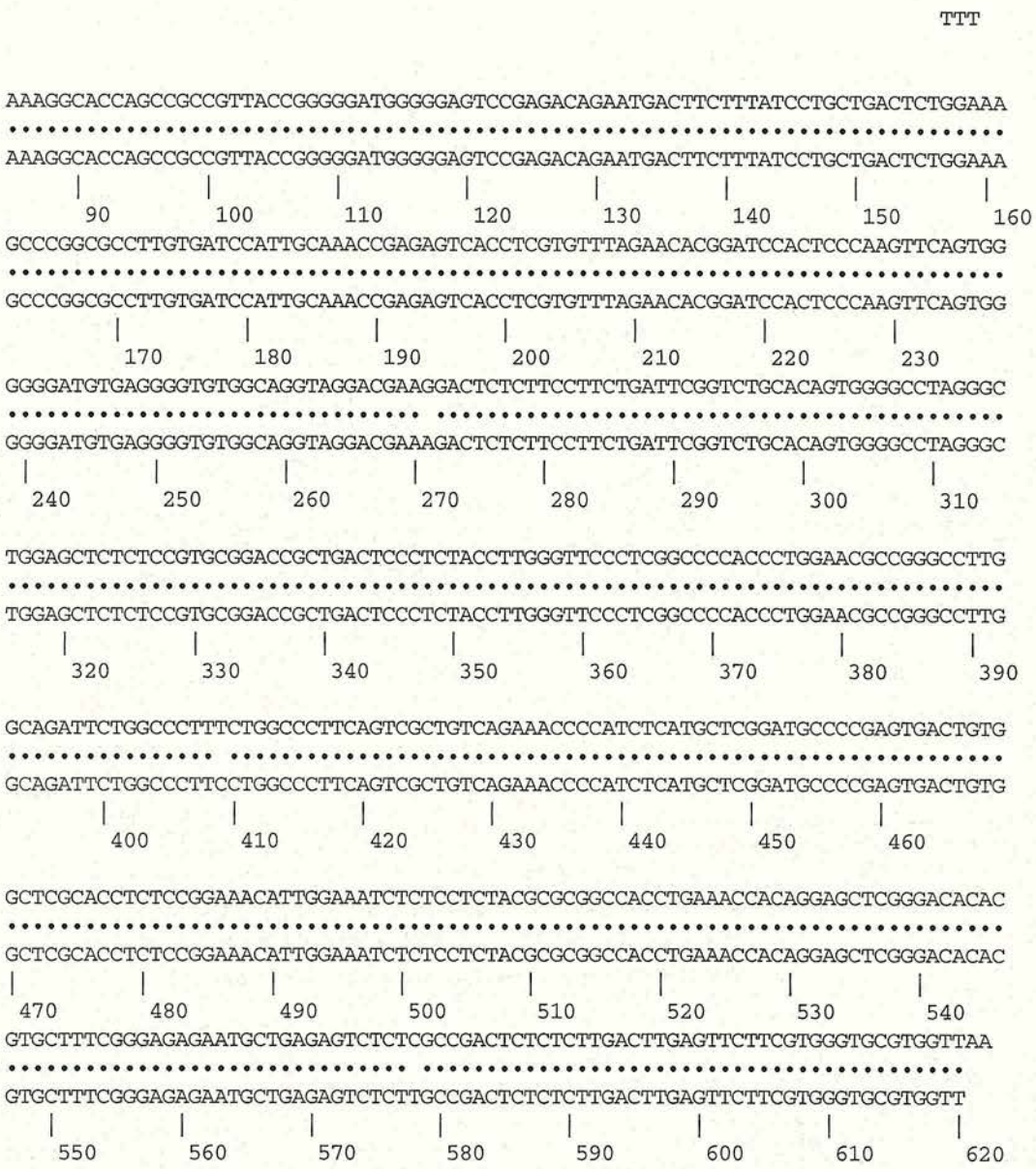
Figure 4.6.

Alignment of the sequence of the most frequently occurring clone in the MBDx5 library and an *Mse*I fragment from the rDNA NTS

Consensus sequence of the frequently occurring cloned insert ("p1B12") compared to a region at the 3' end of the rDNA NTS (Accession No. U13369). In addition to the sequences being almost completely identical the rDNA sequence ends in *Mse*I sites.

Figure 4.6.

Alignment of rDNA NTS (Top)and the Consensus sequence (p1B12) (Bottom)



that both p1B12 (534 bp) and p1A12 (857 bp) contain inserts of approximately the same size as two of the bands visible when the DNA (MBDx5) is amplified. An aliquot of the amplified DNA was electrophoresed through an agarose gel and blotted. This membrane was then probed with either p1B12 or p1A12, both hybridised very strongly, suggesting that they may be major components of two of the bands (Figure 4.7.B).

4.4.1.
Conclusions regarding further analysis of enriched sequences from the MBDx5 library

Both SINES and LINES which are GC-rich, often methylated and occur frequently in the genome, were represented in the library. In addition to these inserts, two clones, occurring frequently in the library but not the genome, were also identified in the previous section (Section 4.3). The most frequently occurring insert was an *Mse*I fragment from the rDNA repeat NTS, which is present in the genome at around 400 copies. This insert (p1B12) appears, using colony hybridisation blots, to be represented in 25% of the clones analysed (Figure 4.4). Amplification of the MBDx5 DNA had indicated that a number of clones would be derived from three distinct bands (Figure 4.3.A). The most frequently occurring insert (p1B12) was of approximately the same size as the smallest and strongest band. This may partially explain its high frequency in the library. A second insert (p1A12) which occurred less frequently, showed no sequence similarity to the database. The insert from p1A12 was the same size (857 bp) as the middle band produced during amplification of MBDx5 DNA (Figure 4.3.A, Arrow 2). A Southern blot of the MBDx5 DNA probed with these two inserts indicated that the major components of two of the three bands had been identified (Figure 4.7.B). Both the sequences, when examined further, were

Figure 4.7.

Methylation status of two sequences frequently cloned in the MBDx5 library

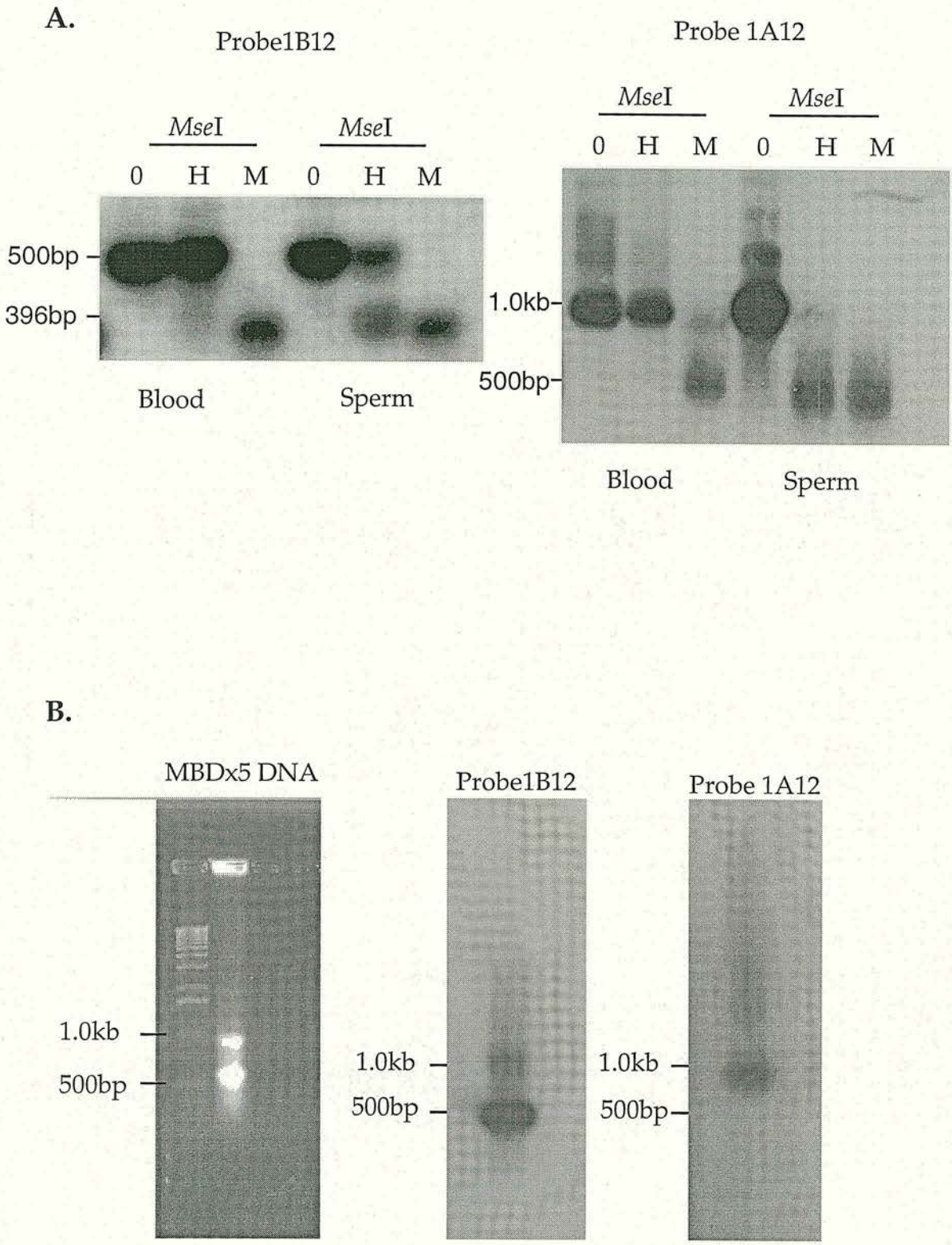
A.

Southern blots of DNA derived from blood and sperm. The DNA was digested with *Mse*I followed by either *Msp*I (M) or *Hpa*II (H). After alkali transfer to a Hybond N^+ membrane, the DNA was probed with the insert from either p1B12 or p1A12 as shown. The inserts from both clones are methylated in somatic cell DNA. The insert from p1B12 hybridises to a digested and therefore unmethylated fragment in the DNA from sperm. It also hybridises to an undigested band suggesting that these sites are occasionally methylated in DNA derived from sperm. The second probe p1A12 hybridises only to digested fragments indicating that these seven *Hpa*II sites are unmethylated in DNA from sperm.

B.

The first panel shows a photograph of the product of amplification of DNA (MBDx5) electrophoresed through a 1.5% agarose gel. The amplified MBDx5 was then blotted onto a Hybond N+ membrane. Duplicate membranes were then probed with the insert from either p1B12 or p1A12, as shown in the second and third panels. The inserts from these two clones appear to hybridise to the major components of the bands visible in the PCR reaction.

Figure 4.7.


A.
Probe1B12
Probe 1A12
MseI
MseI
0 H M 0 H M
500bp
396bp
Blood
Sperm
MseI
MseI
0 H M 0 H M
1.0kb
500bp
Blood
Sperm
B.
MBDx5 DNA
Probe1B12
Probe 1A12
1.0kb
500bp
1.0kb
500bp
1.0kb
500bp

found to be GC-rich with a high frequency of CpGs. In addition Southern blot analysis demonstrated that both sequences were methylated at sites tested in DNA derived from blood. This preliminary analysis had shown that after enrichment, two of the most frequently occurring clones were both GC-rich and methylated in blood DNA. As a result it was decided to investigate the less frequently occurring clones, i.e. those which did not hybridise to the probes (Figure 4.4). No insert had been identified which hybridised the MBDx5 probe and was the same size as the upper band (Figure 4.3.A, Arrow 3). The major component(s) of this band therefore remain unidentified.

4.4.2.

Southern blot and sequence analysis of clones from the female MBDx5 library

The preliminary analysis had shown that the library contained a number of repeated sequences, eg. *Alu* sequences. Two other frequently occurring sequences (p1A12 and p1B12) were examined and both were found to be methylated in blood, GC-rich and with a high $CpG_{Obs/Exp}$. It was decided to examine a selection of those clones which had not hybridised to any of the pre-screening probes. A selection of inserts from these clones were also used as probes against Southern blots, this determined if they were methylated in blood DNA (Figure 4.8.A & B) or in DNA derived from sperm (Figure 4.8.C). The Southern blots indicated that all of the clones examined were methylated in blood DNA. The majority of the inserts hybridised to uncut *Mse*I fragments when DNA was cut with *Mse*I and *Hpa*II. The same probes hybridised to low molecular weight bands when DNA was cut with *Mse*I and *Msp*I. When sperm DNA was probed the majority of clones tested were unmethylated (Figure 4.8.C). As the majority of CpGs in DNA derived from

blood will be methylated it was also important to show that the cloned sequences were GC-rich and had a high frequency of CpG. A selection of inserts which had not hybridised to the pre-screening probes were therefore sequenced (Chapter 2 section 2.13). The %GC, number of GpCs and CpGs, number of CpG divided by GpC and the $CpG_{Obs/Exp}$ for all clones of which there is complete sequence are shown in Tables 4.2. The %GC of the sequences is plotted against their $CpG_{Obs/Exp}$ with the data from known CGI sequences and CpG depleted sequences plotted for comparison (Figure 4.9). Not all the recombinant clones were fully sequenced and the details of those which were only partially sequenced are shown in Table 4.3. In the final graph the %GC and $CpG_{Obs/Exp}$ of sequenced inserts from the MBDx3 library is compared to those from the MBDx5 library (Figure 4.10).

4.5.
Results

The use of the various cloned inserts from the MBDx5 library demonstrated that most , including the frequently occurring clones were methylated at sites tested in DNA derived from blood (Figure 4.7 & 4.8). After the additional purification step the clones from MBDx5, had an average %GC of 55 and an average $CpG_{Obs/Exp}$ of 0.65. compared to 50 and 0.45 respectively for the MBDx3 sequences (see also Figure 4.10). Several inserts cloned from the MBDx5 library had sequence characteristics close to the anticipated %GC and $CpG_{Obs/Exp}$. Examples from Table 4.2 include pCPA1 (58%, 0.70), pCPD3 (56%, 0.63), pCPD8 (56%, 0.66), pCPF2 (54%, 0.61), pCPG2 (60%,0.98) p4B7 (60%, 0.66) and p4G2 (55%, 0.80). With the exception of four clones (pCPB8, pCPE8, pCPH9 and p5A11) all the clones that were fully sequenced had a %GC of greater than 50. Three of the four inserts with a %GC of less than 50 had a $CpG_{Obs/Exp}$ of greater than 0.70 (the fourth is 0.61).

Figure 4.8.

Southern-blot analysis to determine methylation status of sites in sequences cloned in the MBDx5 library

A. Examples of Southern blots of DNA derived from blood and probed with the inserts from the three clones shown. All three are methylated at the sites tested in DNA from blood. The insert from clone p4C3 was 595 bp in size with a %GC of 67 and a $CpG_{Obs/Exp}$ of 0.70. The insert was sequenced and found to have four *Hpa*II sites. No sequence data was available for the inserts from clones pA3 and pCPC8

B. Examples of Southern blots of DNA derived from male and female blood and probed with the two inserts shown. Both are methylated in male and female blood at the sites tested. The insert from clone p4F9 was 460 bp in size with a %GC of 54 and a $CpG_{Obs/Exp}$ of 0.70. The insert was sequenced and found to have a single *Hpa*II sites (see also Appendix A). No sequence data was available for the inserts from clone p6R.

C. Examples of Southern blots with DNA from male and female blood and DNA from sperm, probed with the inserts shown. All inserts are methylated in DNA derived from blood. The insert from clone pCPC7 was partially sequenced and had a %GC of 56 and a $CpG_{Obs/Exp}$ of 0.54. The insert from clone pCPE3 was also partially sequenced and had a %GC of 55 and a $CpG_{Obs/Exp}$ of 0.88. The insert from clone p4A12 is 465 bp in size and had a %GC of 61 and a $CpG_{Obs/Exp}$ of 0.42. The insert was sequenced and found to have five *Hpa*II sites. The insert from clone p5H12 is 795 bp in size and had a %GC of 60 and a $CpG_{Obs/Exp}$ of 0.38 The insert was sequenced and also found to have five *Hpa*II sites. The insert from p5H12 is methylated at the sites tested in sperm the remainder are unmethylated.

Figure 4.8.

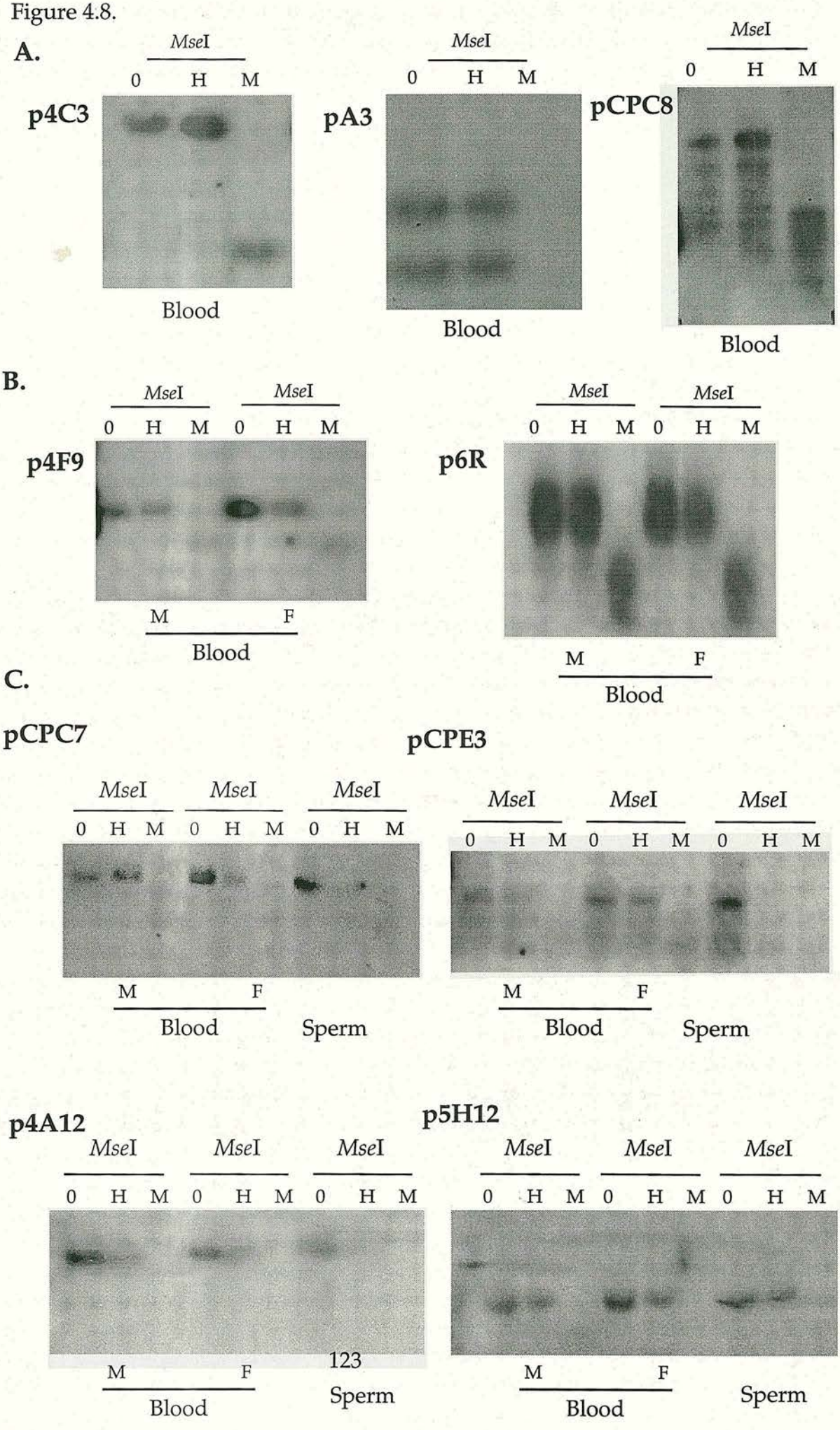

A.
MseI
0 H M
p4C3
Blood
pA3
Blood
pCPC8
Blood
B.
p4F9
M F
Blood
p6R
M F
Blood
C.
pCPC7
MseI MseI MseI
M F
Blood
Sperm
pCPE3
Blood
Sperm
p4A12
Blood
Sperm
p5H12
Blood
Sperm

Table 4.2

Entire sequence data of a random selection of clones from the MBDx5 library

The identification number of each plasmid containing a cloned insert is shown in the first column. In the remaining columns, reading from left to right, the number of base pairs , the %GC , the number of CpGs, number of GpCs, number of CpGs divided by GpCs and the $CpG_{Obs/Exp}$ of each sequenced insert is shown.

Table 4.2.

Clone ID	Seq'd	%GC	No. CpG	No. GpC	CpG/GpC	Obs./Exp.
pCPA1	857	58	51	60	0.85	0.70
pCPA3	707	58	27	61	0.44	0.45
pCPA7	472	56	19	34	0.56	0.52
pCPA12	421	52	16	23	0.69	0.57
pCPB1	677	59	32	50	0.64	0.54
pCPB2	488	60	36	49	0.73	0.83
pCPB6	399	50	11	26	0.42	0.44
pCPB8	560	48	23	41	0.56	0.70
pCPC4	507	55	22	40	0.55	0.57
pCPC5	815	54	60	76	0.79	1.01
pCPD3	451	56	21	32	0.65	0.63
pCPD4	666	56	30	50	0.60	0.57
pCPD5	437	51	27	39	0.69	0.95
pCPD8	545	56	28	44	0.64	0.66
pCPE2	418	57	18	33	0.55	0.53
pCPE8	576	42	14	12	1.20	0.74
pCPE10	315	50	20	15	1.30	1.02
pCPF2	662	54	29	55	0.53	0.61
pCPF4	573	59	29	51	0.57	0.59
pCPF6	579	58	25	41	0.61	0.52
pCPF7	628	62	42	61	0.69	0.69
pCPF9	462	58	21	34	1.70	0.54
pCPF12	398	68	48	27	0.73	1.09

Table 4.2 Continued.

Clone ID	Seq'd	%GC	No. CpG	No. GpC	CpG/GpC	Obs./Exp.
pCPG3/G9	611	46	19	26	0.73	0.65
pCPG11	654	60	56	45	1.20	0.98
pCPG12	510	53	20	34	0.59	0.55
pCPH8	330	57	20	26	0.77	0.76
pCPH9	743	45	20	44	0.45	0.61
p1B12	534	59	31	34	0.91	0.67
p1A12	857	58	51	60	0.85	0.69
p4A12	465	61	18	43	0.42	0.42
p4B7	536	60	31	34	0.92	0.66
p4B8	458	54	23	26	0.88	0.74
p4C1	559	54	15	38	0.39	0.38
p4C2	740	57	34	59	0.57	0.57
p4C3	595	67	46	57	0.81	0.70
p4C12	498	52	20	39	0.51	0.59
p4E5	558	40	26	14	1.88	1.24
p4F9	460	54	22	25	0.88	0.70
p4G2	436	55	26	29	0.89	0.80
p4G3	540	57	25	43	0.58	0.58
p4G6	533	59	22	40	0.55	0.49
p4G12	450	52	12	23	0.52	0.40
p4H5	646	56	24	62	0.39	0.47
p5A6	569	58	32	21	1.5	0.89

Table 4.2 Continued.

Clone ID	Seq'd	%GC	No. CpG	No. GpC	CpG/GpC	Obs./Exp.
p5A7	654	57	21	63	0.33	0.40
p5A9	792	55	56	75	0.75	0.93
p5A11	569	49	25	17	1.50	0.78
p5B6	684	65	45	75	0.60	0.63
p5C10	602	62	31	60	0.52	0.54
p5D3	661	52	22	43	0.51	0.49
p5E3	520	53	50	30	1.70	1.36
p5G5	677	55	20	35	0.57	0.40
p5H11	597	56	20	49	0.41	0.43
p5H12	795	60	27	64	0.42	0.38

Table 4.3.

Partial sequence data of a random selection of clones from the MBDx5 library

The identification number of each plasmid containing a cloned insert is shown in the first column. In the remaining columns reading from left to right, the number of base pairs sequenced , the %GC , the number of CpGs, number of GpCs, number of CpGs divided by GpCs and the $CpG_{Obs/Exp}$ of each sequenced insert is shown.

Table 4.3.

Clone ID	Seq'd	%GC	No. CpG	No. GpC	CpG/GpC	Obs./Exp.
pCPA1T7	364	60	25	29	0.86	0.75
pCPA2T7	335	58	13	29	0.45	0.46
pCPA2SP6	372	58	14	32	0.44	0.45
pCPA4SP6	234	45	9	8	1.10	0.75
pCPA5T7	341	56	12	14	0.86	0.46
pCPA5SP6	251	60	16	14	1.10	0.71
pCPB12T7	377	59	15	36	0.42	0.46
pCPC1T7	116	62	8	12	0.67	0.72
pCPC7T7	336	52	11	20	0.55	0.50
pCPC7SP6	291	60	15	25	0.60	0.57
pCPC9T7	378	65	11	38	0.29	0.28
pCPC9SP6	419	54	15	34	0.44	0.49
pCPC10T7	170	54	8	11	0.73	0.64
pCPC10SP6	178	51	11	13	0.85	0.95
pCPC11T7	333	64	16	35	0.46	0.47
pCPC11SP6	330	57	19	30	0.63	0.72
pCPC12T7	260	59	14	25	0.56	0.65
pCPC12SP6	366	51	22	26	0.85	0.91
pCPD6T7	335	54	20	18	1.10	0.84
pCPD10T7	366	63	29	40	0.73	0.82
pCPD10SP6	233	59	12	22	0.55	0.57
pCPE1T7	483	65	35	56	0.63	0.69
pCPE1SP6	366	65	24	36	0.67	0.63

Table 4.3. Continued

Clone ID	Seq'd	%GC	No. CpG	No. GpC	CpG/GpC	Obs./Exp.
pCPE3T7	269	56	14	15	0.93	0.69
pCPE3SP6	278	55	22	18	1.20	1.08
pCPE4T7	271	57	16	21	0.76	0.75
pCPE6SP6	338	54	14	23	0.61	0.56
pCPH2SP6	328	58	15	23	0.65	0.54
pCPH3SP6	433	62	30	38	0.79	0.72
pCPH5T7	363	61	24	41	0.58	0.72
pCPH5SP6	480	60	31	39	0.79	0.73
pCPH7T7	413	56	13	34	0.38	0.40
pCPH7SP6	413	54	11	30	0.37	0.36
pCPH10SP6	467	62	19	36	0.53	0.43

When the cloned sequences in Table 4.2 are plotted (Figure 4.9, white squares) it can be seen that they are more GC-rich than depleted sequences (Figure 4.9, black triangles). However the majority are still only matching the borderline characteristics of representative fragments of CGIs (Figure 4.9, black squares). The data from sequences in Table 4.3 were not plotted as the sequences were incomplete. Two sequences (p4E5 and p5E3) which have a $CpG_{Obs/Exp}$ greater than 1.1 are also not plotted. All available sequence was used as queries in NCBI database searches. The majority of the sequences cloned showed little or no similarity to sequences in the database. Those sequences which did have good similarity are shown in Appendix A and Figures 4.5 and 4.6. The insert from clone pCPD8 showed similarity to an *Mse*I fragment from the α-adducin gene. The insert was 545 bp long and contained 26 CpGs, the methylation status of this clone was not investigated. The clone p4F9 contained a 476 bp insert with 22 CpGs, this fragment was almost identical to an MseI fragment from the VNTR locus DXZ4 (Giacalone et al., 1992). The single HpaII site was shown to be methylated in DNA derived from blood (Figure 4.8.B). The remaining cloned inserts are similar but only partially match sequences from the database (Appendix A).

4.5.1. Conclusions

Plotting the data from Tables 4.2 against the data from Table 4.1 indicates that enrichment has been achieved (Figure 4.10). Although some overlap between MBDx3 and MBDx5 is evident, overall, the later sequences were more GC-rich and had a higher $CpG_{Obs/Exp}$. The extra purification step had therefore increased the %GC and $CpG_{Obs/Exp}$ of the MBDx5 DNA. Additionally fragments from CGIs are present in MBDx5 and can be

detected using specific primers (Figure 4.3.B). The library does not however consist mainly of fragments from CGIs. Although detectable, these sequences are obviously still in the minority, even after the extra purification step. The majority of sequence cloned have a higher %GC and a higher $CpG_{Obs/Exp}$. than the bulk of the genome. Those which were tested were also shown to be methylated in DNA derived from blood. In addition to methylation in blood, in common with some CGIs, the majority of those sequences examined lacked methylation in germ-line DNA (Figures 4.7 and 4.8) The higher levels of CpG in the cloned sequences may be maintained by a lack of methylation in sperm.

Figure 4.9.

Graph of %GC versus $CpG_{Obs/Exp}$ for the sequences from Table 4.2.

The sequences shown were cloned after five passes through the MBD column (MBDx5) white squares. The sequence data from known CpG islands (blue squares) and the sequence data from sequences depleted for CpG (red triangles) are graphed for comparison. The data was taken from Table 1 of the study by Gardiner-Garden (1987). The sequences from the MBDx5 library lie in the upper region of the graph close to the GCI sequences. However they are still not as GC-rich nor do they have the same frequency of $CpG_{Obs.Exp}$

Figure 4.9.

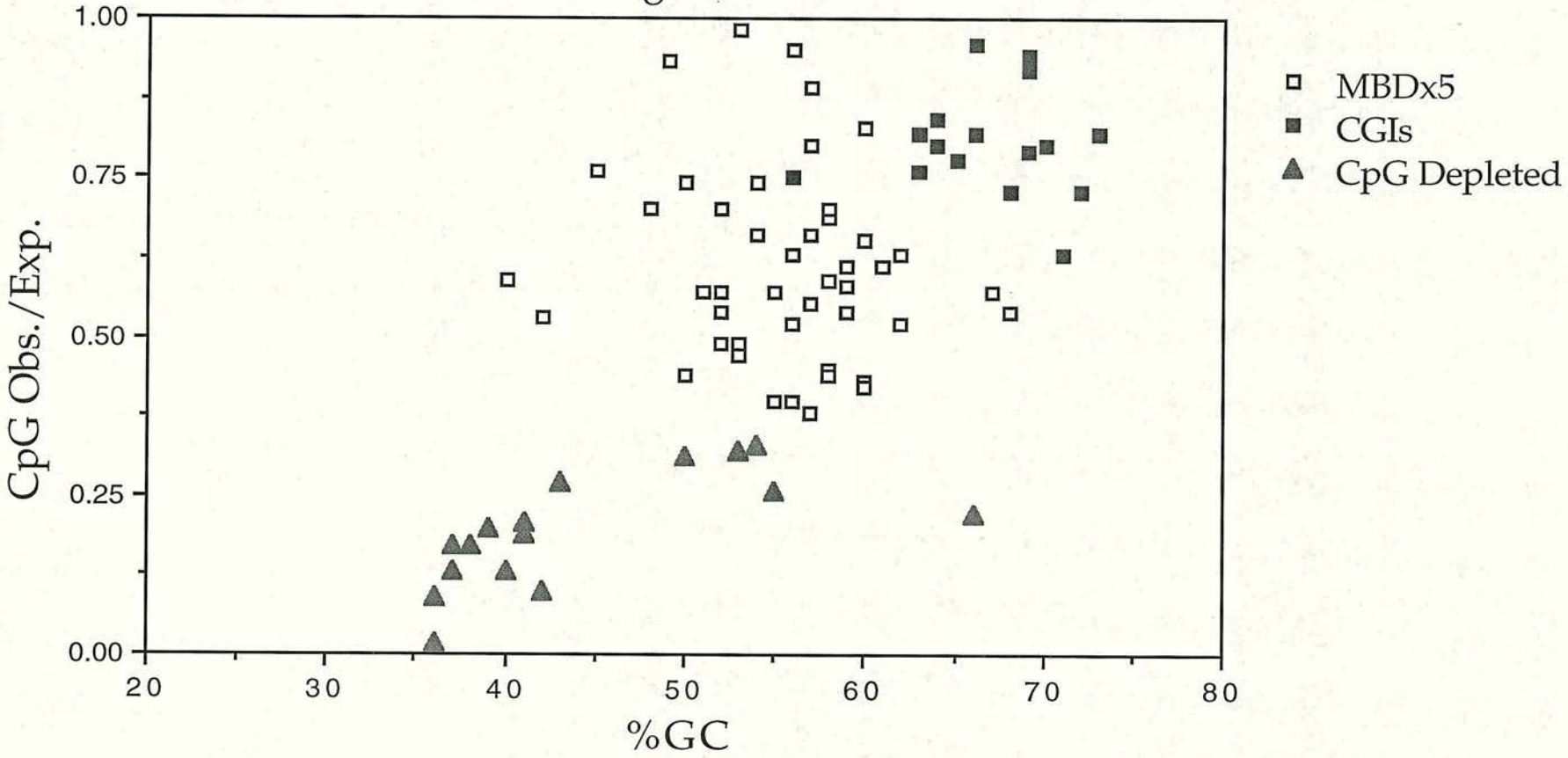
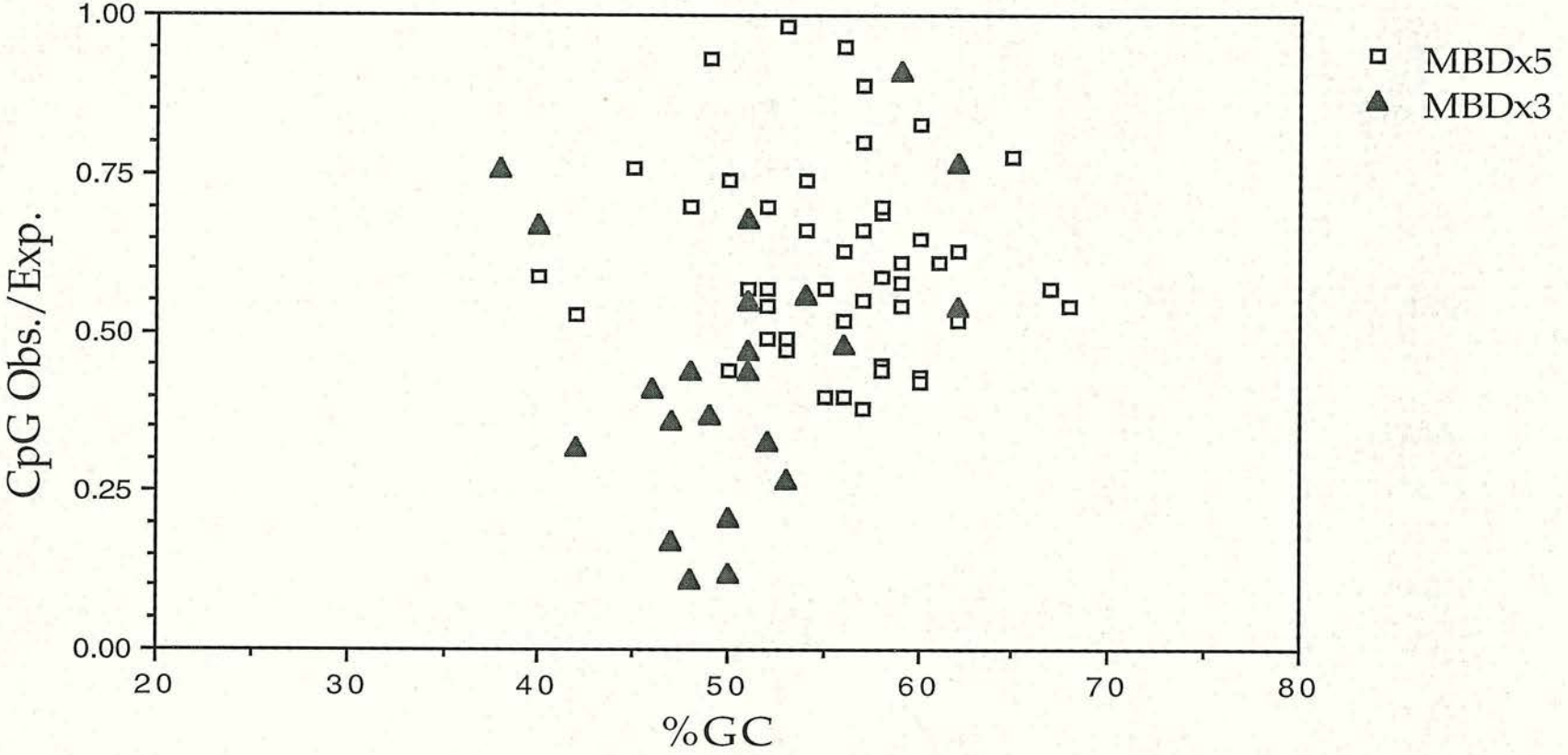

1.00
0.75
0.50
0.25
0.00
CpG Obs./Exp.
20
30
40
50
60
70
80
%GC
MBDx5
CGIs
CpG Depleted

Figure 4.10.

Graph of %GC versus $CpG_{Obs/Exp}$ for the sequences the MBDx3 and MBDx5 libraries.

The sequences shown were from clones from the MBDx3 (black triangles) or MBDx5 (white squares) libraries. The sequences from the MBDx5 library are consistently more GC-rich and have a higher $CpG_{Obs.Exp}$ than those from MBDx3.

Figure 4.10.


1.00
0.75
0.50
0.25
0.00
CpG Obs./Exp.
20
30
40
50
60
70
80
%GC
MBDx5
MBDx3

Chapter 5 : Fluorescent *in situ* Hybridisation

5.1.
Introduction

In the previous chapter the further analysis of the MBDx5 library clones showed that most were more GC-rich with a higher frequency of CpG than the majority of the genome. A number of the inserts were also shown to be methylated in DNA from blood but not in DNA from sperm. It was decided to examine the location of the cloned sequences on chromosomes by using the DNA from the MBDx3 and MBDx5 libraries as probes in Fluorescent *in situ* Hybridisation (*FISH*) experiments. In addition the insert from clone p1A12 was also used (Figure 4.7). This would determine the chromosomal location of both libraries and the location of one of the three major bands in the cloned MBDx5 DNA. The location of the rDNA repeat, and therefore a major component of the smallest band (p1B12) was already known to be on the short arms of the acrocentric chromosomes.

The application of *FISH* to study the location on chromosomes of cloned sequences was developed in the early 1970s (Rudkin and Stoller, 1977). *FISH* involves denaturing the probe and the target (in this study the DNA in a male metaphase spread) then adding together and allowing any complementary strands to re-anneal. The labelled probes can then be visualised by incubating with fluorescent reagents with an affinity for the probe label (Korenberg, 1992). The probes in these experiments were labelled with biotin which was then visualised using antibodies and conjugates. The antibodies and conjugates gave the probes a red fluorescent label clearly visible against the chromosomes which were stained blue with the DNA stain DaPi.

Figure 5.1.

Male metaphase spread probed with MBDx3 DNA derived from female blood

Two examples are shown of chromosomes from a male metaphase spread (stained blue with DaPi) probed with MBDx3 DNA from female blood (stained with Texas Red).

A.

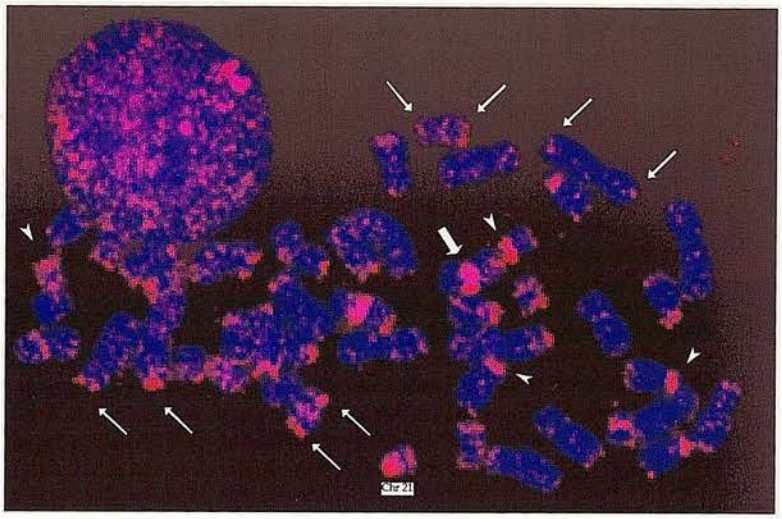
The first example shows the pattern of hybridisation when MBDx3 was used as a probe in the presence of competitor DNA. Although there is some overall background staining the strongest hybridisation is to the subtelomeric regions of the majority of the chromosomes, a few examples are indicated by thin white arrows. One chromosome, thought to be 21, stains almost entirely (indicated by Chr 21). In common with all the FISH hybridisation results shown in this chapter the most intense staining is seen at the centromere of Chromosome 9 (indicated by short arrow). Intense staining is also seen on the short arms of the acrocentric chromosomes, a few examples are indicated with large white arrowheads.

B.

Enlarged section of a different chromosome spread again probed with the MBDx3 library, in the presence of competitor DNA. In addition to the short arms of the acrocentric chromosomes (examples indicated by arrowheads) strong staining is again evident on the telomeric regions (examples indicated by long white arrows).

Figure 5.1.

A.



B.

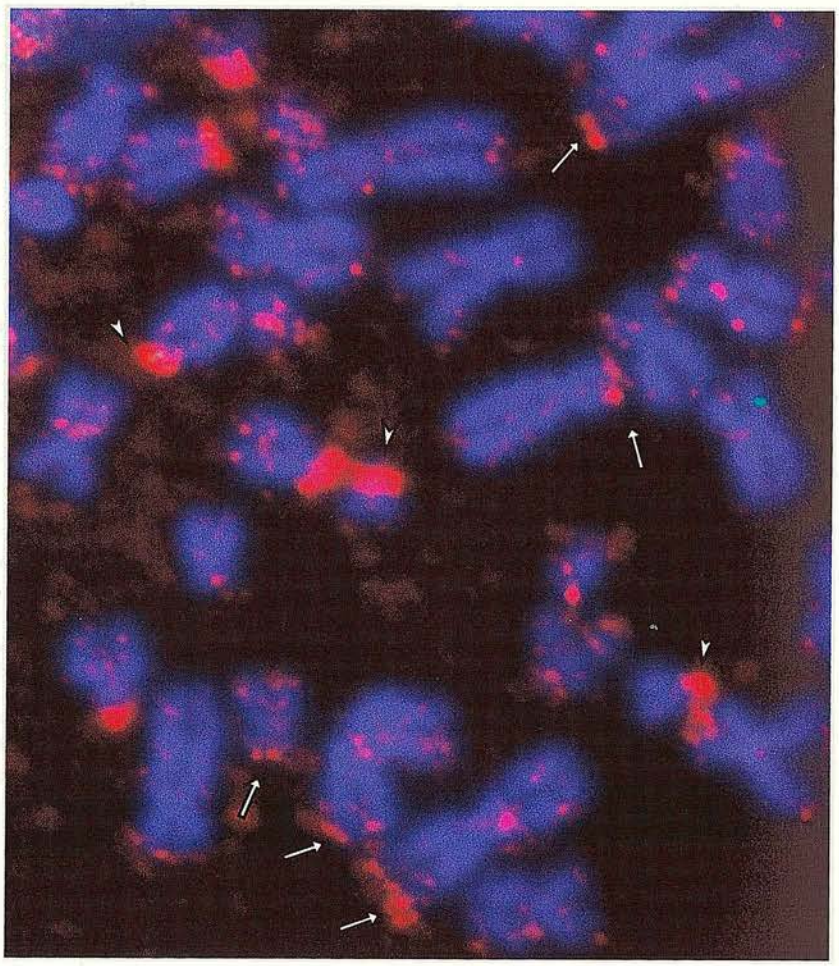


Figure 5.2.

Male metaphase spread probed with MBDx5 DNA derived from female blood

A.

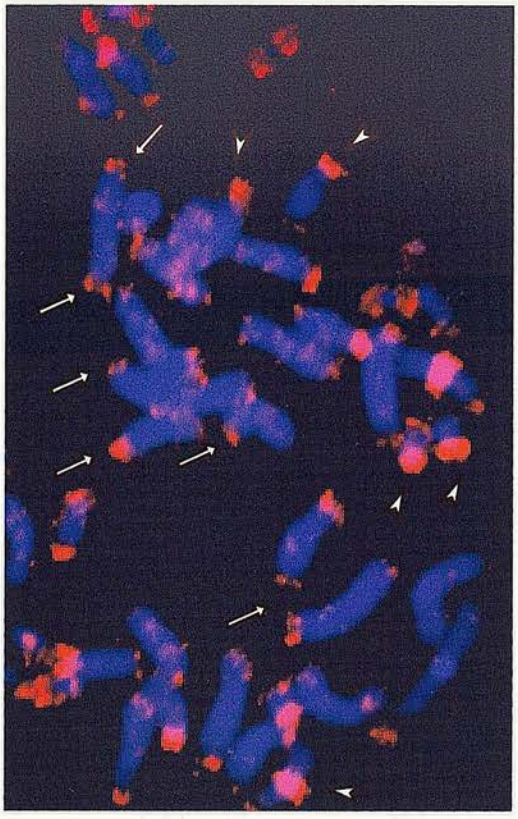
An example of a male metaphase spread probed with MBDx5, derived from female blood, in the presence of competitor. Subtelomeric regions on many chromosomes are again hybridised to strongly, examples are indicated by long white arrows. The staining pattern is very similar to that seen when MBDx3 was used as a probe.

B.

An example of a male metaphase spread probed with MBDx5, derived from female blood, in the presence of competitor. The staining pattern is similar to that seen in Figure 5.1., examples of subtelomeric staining are indicated with long white arrows. The centromere of chromosome 9 and examples of staining on the short arms of the acrocentric chromosomes are indicated with white arrowheads.

Figure 5.2.

A.



B.

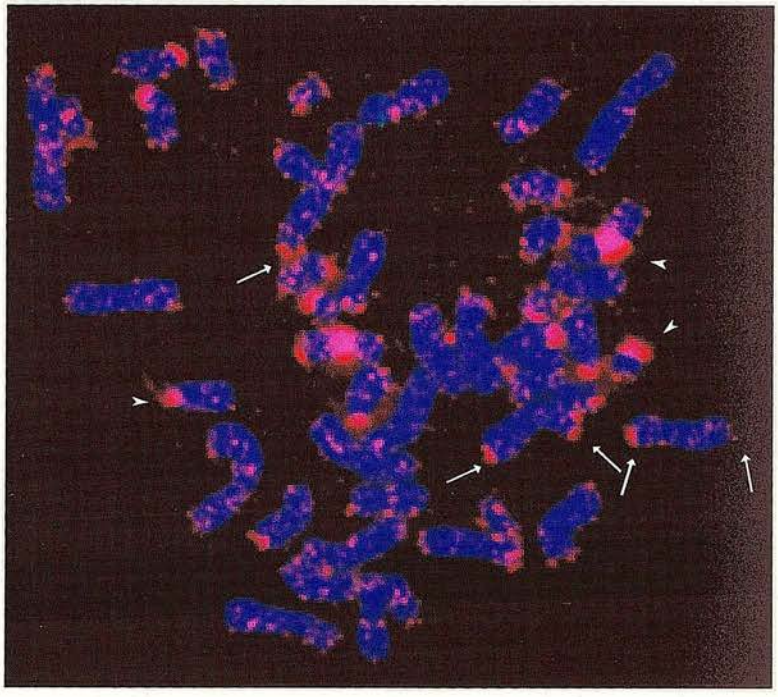


Figure 5.3.

Karyotyped male metaphase spread probed with MBDx5 DNA derived from male blood

A.

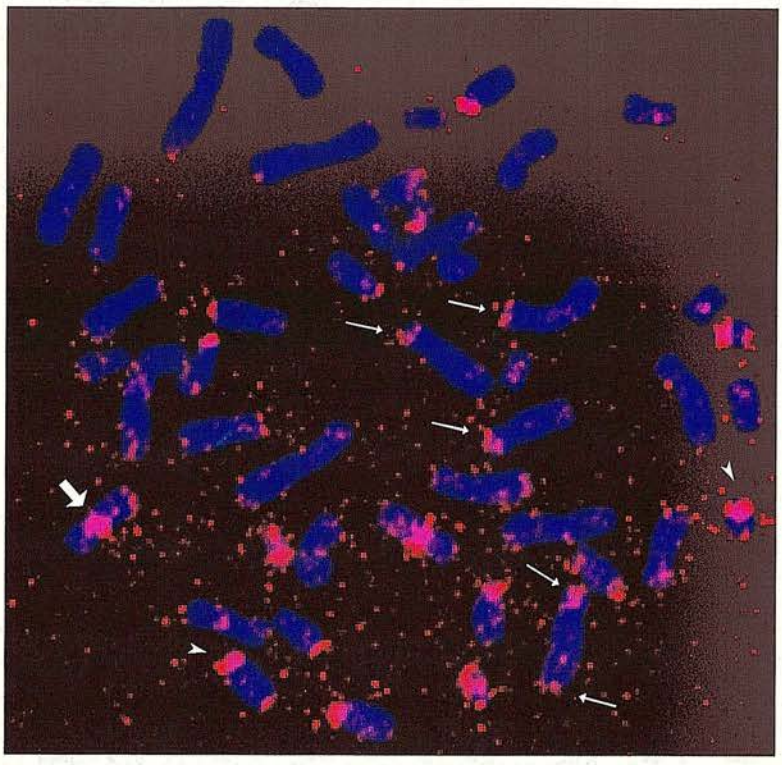
Metaphase spread probed with male MBDx5 DNA in the presence of competitor. The staining pattern is similar to that observed in Figures 5.1 and 5. 2. examples of subtelomeric staining are indicated with long white arrows.

B.

Chromosomes from 5.3.A. were organised in pairs after being karyotyped (W. Bickmore). Only one example of chromosome 19 could be positively identified. The slight difference in staining seen between pairs was due to technical problems with the light source. This has resulted in more light at the bottom of the image than at the top.

Figure 5.5:

A.



B.

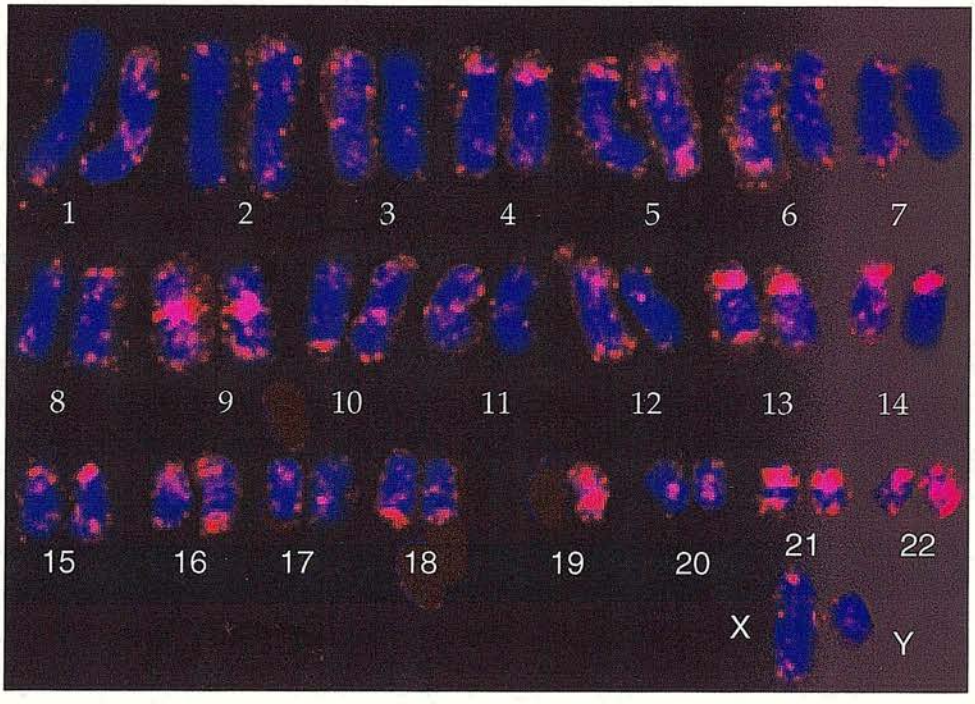


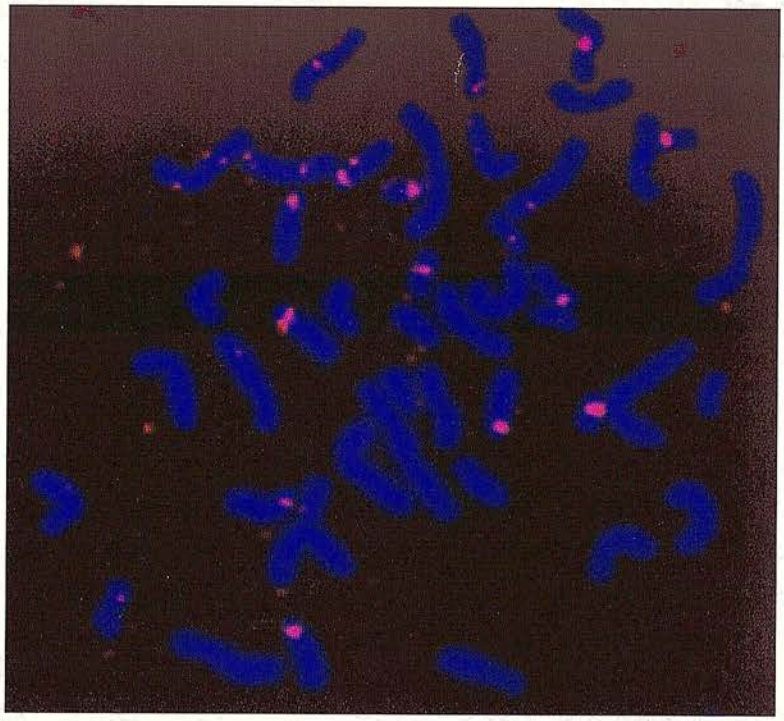
Figure 5.4.

Male metaphase spread probed with the insert from the p1A12

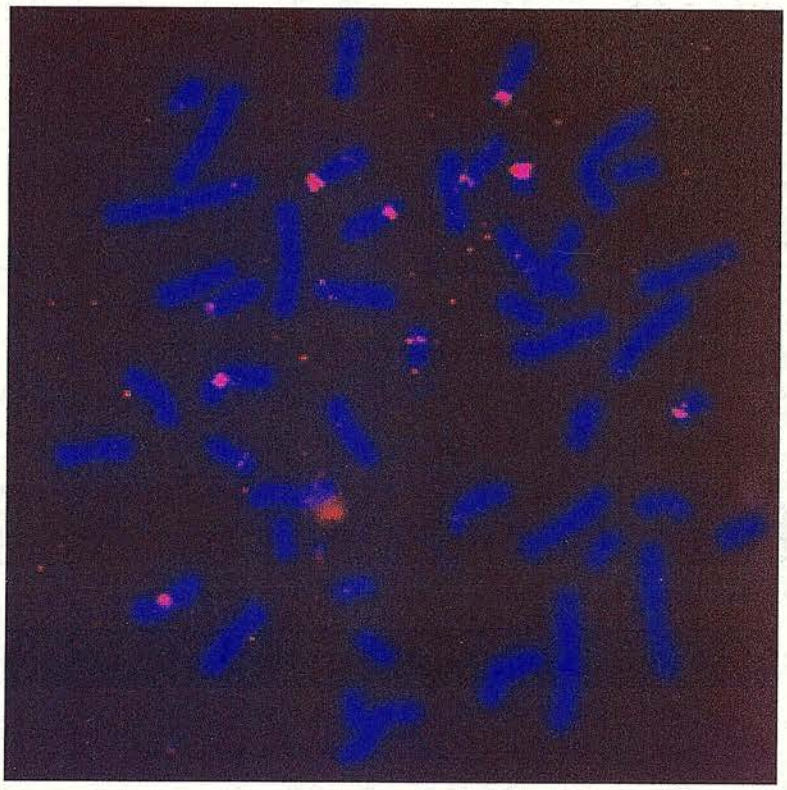
The insert from the probe p1A12 had been sequenced and shown to have a high %GC and $CpG_{Obs/Exp}$. Use of this insert as a probe in a southern-blot had also shown that this sequence was methylated in DNA derived from blood (Figure 4.7). The probe appears to hybridise to two acrocentric chromosomes (Chr 13 & Chr 14) and to a sequence some distance from the telomeres on several chromosomes. There also appears to be relatively intense staining on chromosomes 21 or 22. Absolute identification of the chromosomes was not possible due to problems with karyotyping these slides.

Figure 5.4.

A.



B.



5.2.
Results

It was necessary to add Cot-1 DNA to compete out any repetitive sequences present in both the MBDx3 and MBDx5 probes. Removal of Cot-1 DNA from the prehybridisation mixture during *FISH* experiments had the effect of increasing the overall signal and in particular the signal at the centromeric regions of chromosomes 1 and 16 (not show). The increase in overall signal is probably caused by SINES such as *Alus* which are known to be present in the MBDx5 library. The centromeres of chromosomes 1, 9 and 16 and the short arms of the acrocentric chromosomes are the location of the C-bands. The C-bands are the regions of chromosomes which contain a large amount of constitutive heterochromatin. A close relationship between the heterochromatin of chromosomes 1 and 16 had previously been shown. The study also indicated a distinction between the heterochromatin of these two and that of chromosome 9 (Schwarzacher-Robinson et al 1988). With both the MBDx3 and MBDx5 probes a strong hybridisation signal was observed at the centromere of chromosome 9. This hybridisation was apparently unaffected by the addition of competitor DNA (Cot-1). The secondary constriction of chromosome 9 has been reported to contain unique methylated repetitive sequences (Mitchell et al., 1986). Two repeats, a 68 bp *Sau* 3A (54% GC) β-satellite sequence and a 545 bp *Sau* 3A (68% GC) sequence, have also been reported (Meneveri et al., 1993). These occur on the pericentromeric region of chromosome 9 and are thought to be organised as part of a 2.7 kb higher order repeat unit (Meneveri et al., 1993). It appears that these sequences have the correct characteristics to show strong affinity for the MBD column. As they are methylated they will therefore be represented in the library which would account for the signal seen at the centromere of chromosome 9.

In the presence of Cot-1 DNA the hybridisation patterns obtained using the MBDx3 and MBDx5 probes appears only slightly different (Figure 5.1 and 5.2.). The MBDx3 probe can be seen to hybridise strongly to many subtelomeric regions and it also gives an overall stain on the chromosomes (Figure 5.1.A). In the enlarged figure the most intense staining, with the exception of the acrocentric chromosomes, occurs at the telomeric regions of many chromosomes (Figure 5.1.B). The overall stain on the chromosomes appears to be reduced in some of the images generated with the MBDx5 library probe when compared to MBDx3 (Figure 5.2.A). This may be a result of enrichment of the sample, indicated by the difference in sequence characteristics between the two libraries (Chapter 4). Both libraries hybridise strongly to the short arms of the acrocentric chromosomes (chromosomes 13, 14, 15, 21, & 22). This is the region in which the rDNA repeats are located. The probes also hybridised to subtelomeric regions on a number of other chromosomes, in particular both stain chromosomes 21 and 22 strongly. A karyotype of the image generated by the MBDx5 probe allows identification of specific chromosome staining (Figure 5.3.B). The chromosomes 4, 5 10, 12, 16, 18 and 19 all show strong staining at both telomeres. Chromosomes 1, 2, 3 and 17 show weak staining at the telomeres and chromosomes 7, 8, 11, 20 and the X and Y all show intermittent staining.

The second *FISH* experiment used p1A12 alone to determined the location of this insert (Figure 5.4). Although it did not clone as frequently as p1B12, it was thought to be a major component of the middle band of the PCR product (Figure 4.7). The probe p1A12 appears to hybridise to two of the acrocentric chromosomes and a number of other subtelomeric regions. There is also hybridisation to some lower bands on other chromosomes, mainly on the short arms.

5.2.1.
Conclusions

The most striking result of the FISH analysis is that, once the repetitive fraction of the libraries is competed out, both hybridised preferentially to the subtelomeric regions of the majority of human chromosomes. The location of the most frequently occurring clone (p1B12) was already known as it was derived from a rDNA repeat. There are approximately 400 rDNA repeats located on the short arms of the acrocentric chromosomes. In addition to this clone, the MBDx5 library is thought to contain other fragments of the rDNA NTS (see also Chapter 6). It is not known however if fragments of the rDNA repeat are present in the MBDx3 library at the same level. Both the MBDx3 and MBDx5 probes hybridise strongly to the short arms of the acrocentric chromosomes (Figures 5.1, 5.2 and 5.3) suggesting that fragments from the rDNA NTS are present in both libraries. The other significant site of hybridisation from both libraries is to the centromere of Chromosome 9. This is thought to be due to the presence in the libraries of low copy number methylated repeats as described by Meneveri et al, (1985). Examination of the remaining signal generated by the MBDx5 library probe shows that the strongest signal is frequently found in the subtelomeric regions of most chromosomes (Figure 5.2.A). Hybridisation to the subtelomeric regions of chromosomes 4, 5 10, 12, 16, 18 and 19 is particularly strong as is the signal on the long arm of chromosomes 21 and 22. The clone p1A12, which may account for a large number of sequences in the MBDx5 probe, also hybridises to a number of subtelomeric regions (Figure 5.4). The *FISH* hybridisation pattern appears to indicate that a large number of the sequences in the both libraries are derived from subtelomeric regions of a number of chromosomes. A random selection of inserts from the MBDx5 library have

been shown to be both GC-rich and have a higher $CpG_{Obs/Exp}$ than the bulk of the genome (Chapter 4). In addition to being GC-rich, all sequences tested were methylated in blood and the majority were unmethylated in sperm. In Chapter 4 it was suggested that this lack of methylation in germ-line was a possible explanation for the lack of CpG suppression. The *FISH* hybridisation pattern indicates that such sequences are also found in specific regions of the genome. These coincide with regions which have previously been reported as being the most GC-rich of the genome (Saccone et al., 1992). Lack of methylation in germ-line may therefore contribute to the high GC content in these areas.

Chapter 6 : The methylation status of the rDNA NTS

6.1.

Introduction

During the analysis of the library derived from human female blood DNA (MBDx5) an insert which occurred in 25% of the clones was identified. The insert was found to be an *Mse*I fragment from the rDNA NTS (Chapter 4). Previous studies of methylation levels in the transcribed region of the rDNA repeat of various species had demonstrated that these regions of the repeat unit exist in a predominantly unmethylated form in DNA from a variety of mammals, including man (Bird and Taggart, 1980). The frequent occurrence of rDNA fragments in a the library was therefore unexpected as the entire rDNA repeat was thought to be unmethylated. In addition to the frequently occurring insert (p1B12) other less frequently cloned fragments of rDNA (though always from the NTS) were also found in the library. A Southern blot using the frequently occurring clone demonstrated that this particular *Mse*I fragment from the NTS was indeed methylated in DNA from blood (Figure 4.6.A). An investigation of the methylation status of the entire rDNA repeat was then undertaken using two fragments from the NTS cloned in the library. The location of both p1B12 and p1A7 had been determined as a result of a NCBI database search. The probes were used in conjunction with restriction enzymes and their methylation sensitive isoschizomers to determine the methylation status of the entire rDNA repeat.

The human 18S and 28S ribosomal RNA (rRNA) genes are present at about 400 copies per human haploid genome, clustered on the short arms of the five acrocentric chromosomes (Worton et al., 1988). Each gene is part of a 43 kb repeat unit that can be divided into two regions: a 13.3 kb transcribed

region which contains the highly conserved genes for 18S, 5.8S and 28S rRNA subunits of the ribosome, and the 30 kb NTS (Gonzalez et al., 1992). Repeat unit clusters consist of head-to-tail arrays of about eighty repeats (Sakai et al., 1995).

6.2.
Southern blot analysis of the entire rDNA NTS.

In order to reduce its molecular weight, the DNA from human blood and sperm was digested first with *Eco*RV, which has no recognition site within the rDNA repeat. The DNA was then redigested with *Xma*I or its methylation-sensitive isoschizomer *Sma*I (CCCGGG) (McClelland et al., 1994). Where DNA had been digested first with *Eco*RI or *Hind*III followed by digestion with methylation-sensitive restriction enzymes, *Eco*RV was not used. After probe p1A7 had been denatured, Cot-1 DNA (30μg) was added. Competition was necessary due to the presence of a 100 bp *Alu* consensus sequence within p1A7 (Sealey et al., 1985). The base pair co-ordinates of all the probes correspond to those published in the GenBank database for the human ribosomal DNA complete repeating unit (Accession Number U13369). The coordinates of restriction sites shown in the figures was determined using either the GeneJockey II (Biosoft Cambridge) or DNAStar (DNAStar Incoporated) programs. These agree with previously published studies (Maden et al., 1987; Gonzalez and Sylvester, 1995). None of the cloned inserts examined in the library showed sequence similarity to the transcribed region of the rDNA repeat. In order to investigate methylation in this region two probes, pHsrDNA5.1 and pHsrDNA7.9, which contain *Eco*RI fragments (coordinates 1 to 5900 bp and from 5900 to 13000 bp respectively) were used (a gift from Rakesh Anand). The other two probes p1A7 and

p1B12 contain *Mse*I fragments from the NTS (coordinates 17062 - 17369 bp and 35695 - 36231 bp).

Blots were prepared using genomic DNA digested with *Eco*RV, then redigested with the methylation-sensitive enzyme *Sma*I or its methylation-insensitive isoschizomer *Xma*I. Probes p1A7 and p1B12 both hybridised to very high molecular weight *Sma*I fragments, around 28 kb, whereas *Xma*I generated smaller fragments as predicted by the sequence of the spacer (Figure 6.1.A). Weaker *Sma*I bands below 27 kb indicate that sites in this part of the repeat unit are occasionally non-methylated. The large size of the major band in the *Sma*I lanes shows that p1A7 and p1B12 are part of long tracts of DNA in which multiple (at least 14) *Sma*I/*Xma*I sites are usually in a methylated state. The equivalent experiment with sperm DNA again showed that although methylated the level is significantly lower than in blood DNA. Most of the rDNA repeats in sperm DNA are digested by *Sma*I to give fragments less than 20 kb, and some fragments are of the same size as in the *Xma*I digested lane. Hybridisation of a similar blot with the probes from the transcribed region (pHsrDNA5.1/7.9) gave a contrasting result. The *Sma*I and *Xma*I patterns of both blood and sperm DNA were predominantly the same, indicating that the majority of sites in this part of the repeat unit are not methylated. A notable feature of blots that were probed with the transcribed regions of the repeat unit is the small but significant fraction of hybridisation in the unresolved high molecular weight regions of the gel (Figure 6.1.A. arrow). This suggests that, although most repeats are non-methylated in this region, a minor proportion of repeat units are heavily methylated in blood cells. However no fragments from this region have yet been identified in the library though such sequences would be expected to be in a minority if present at all. The methylated fraction is

Figure 6.1.

Investigation of the methylation pattern of human rDNA repeat units using southern blot analysis of DNA from lymphocytes or sperm

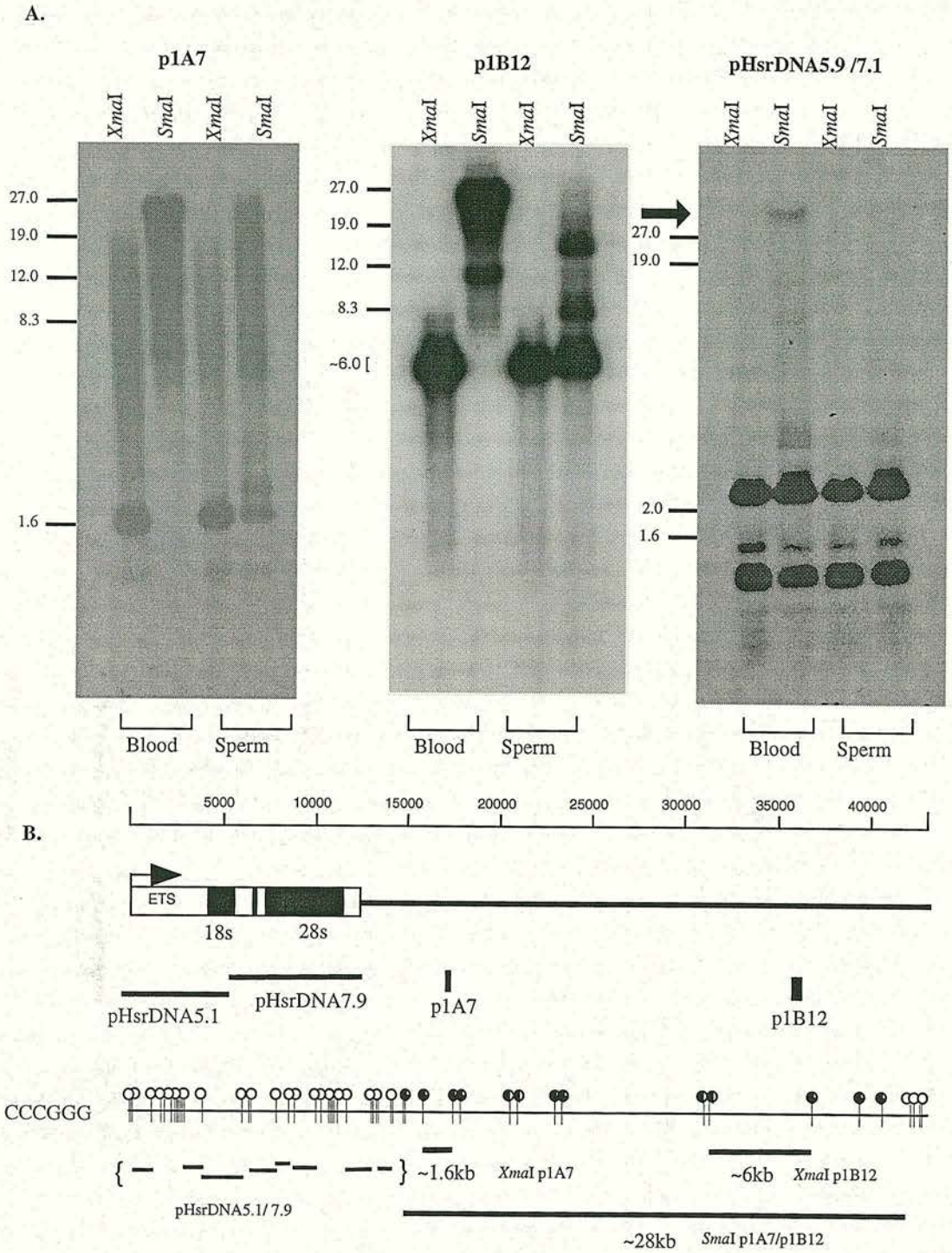
A.

Hybridisation is shown for probes corresponding to the transcribed region of the repeat unit (pHsrDNA5.9 and pHsrDNA7.1) and the NTS probes p1A7 and p1B12. The arrow highlights very large fragments in the *Sma*I lane of blood DNA after probing with the transcribed region.

B.

Map of a single rDNA repeat unit showing the positions of probes, sites for *Xma*I/*Sma*I and their methylation status. Open circles denote non-methylated CpGs; partially filled circles depict sites which are methylated with a high frequency. Prominent bands on the blot probed with pHsrDNA5.9 and 7.1 correspond in size with the fragments shown below the map and it was therefore concluded that sites flanking these fragments are non-methylated. The diagram of the transcription unit shows 18S, 5.8S and 28S rRNA in black, and the ETS, ITS1 and ITS2 regions of the rRNA precursor unshaded. The NTS region is represented by a black line.

Figure 6.1.


A.
p1A7
p1B12
pHsrDNA5.9 /7.1
XmaI
SmaI
27.0
19.0
12.0
8.3
~6.0 [
1.6
2.0
Blood
Sperm
B.
5000
10000
15000
20000
25000
30000
35000
40000
ETS
18s
28s
pHsrDNA5.1
pHsrDNA7.9
p1A7
p1B12
CCCGGG
~1.6kb
XmaI p1A7
~6kb
XmaI p1B12
pHsrDNA5.1/ 7.9
~28kb
SmaI p1A7/p1B12

not apparent in sperm DNA and may be absent. Alternatively, the fraction may be present, but the reduced level of methylation seen in the spacer regions of sperm DNA (Figure 6.1.A) may give a dispersed and therefore indistinct pattern of large fragments.

In order to map the boundary between methylated and non-methylated regions of the repeat, genomic DNA was digested with *Eco*RI or *Hind*III together with one of a selection of restriction endonucleases which contain CpG in their recognition sequence and are sensitive to methylation. The CpG enzymes were *Aci*I (CCGC), *Bst*UI (CGCG), *Hha*I (GCGC) and *Hpa*II (CCGG). *Msp*I, a methylation-insensitive isoschizomer of *Hpa*II, was used as a control. To map the 5' boundary of the methylated domain, blots were probed with p1A7 (Figure 6.2.A). The 15 kb *Hind*III fragment that hybridised to the probe spans the 3' half of the transcription unit and 6.9 kb of downstream spacer. All four methylation-sensitive enzymes in combination with *Hind*III produced bands in the region of 5.5-6.0 kb. This locates the furthest downstream site for each enzyme that is consistently non-methylated at 1.0-1.5 kb downstream of the 3' end of the transcription unit (Figure 6.2.B. , broken vertical line close to nucleotide 14,500). Digestion with *Eco*RI plus methylation-sensitive enzymes confirmed the location of the boundary between methylated and non-methylated domains of the repeat unit. Digestion with *Eco*RI alone gave the expected band of 18.1 kb when probed with p1A7. Further digestion with methylation-sensitive enzymes gave bands clustered at 16 kb due to cleavage at position 14,500. These enzymes also generated bands at approximately 6 kb, indicating a discrete hypomethylated region about 200 bp downstream of the *Hind*III site (Figure 6.2.B., open arrow at nucleotide 20,500). The hybridisation signal is distributed roughly evenly between the 16 kb and 6 kb bands in these lanes.

Figure 6.2.

Investigation of the methylation status of sites downstream of the termination of transcription at the 5' end of the NTS

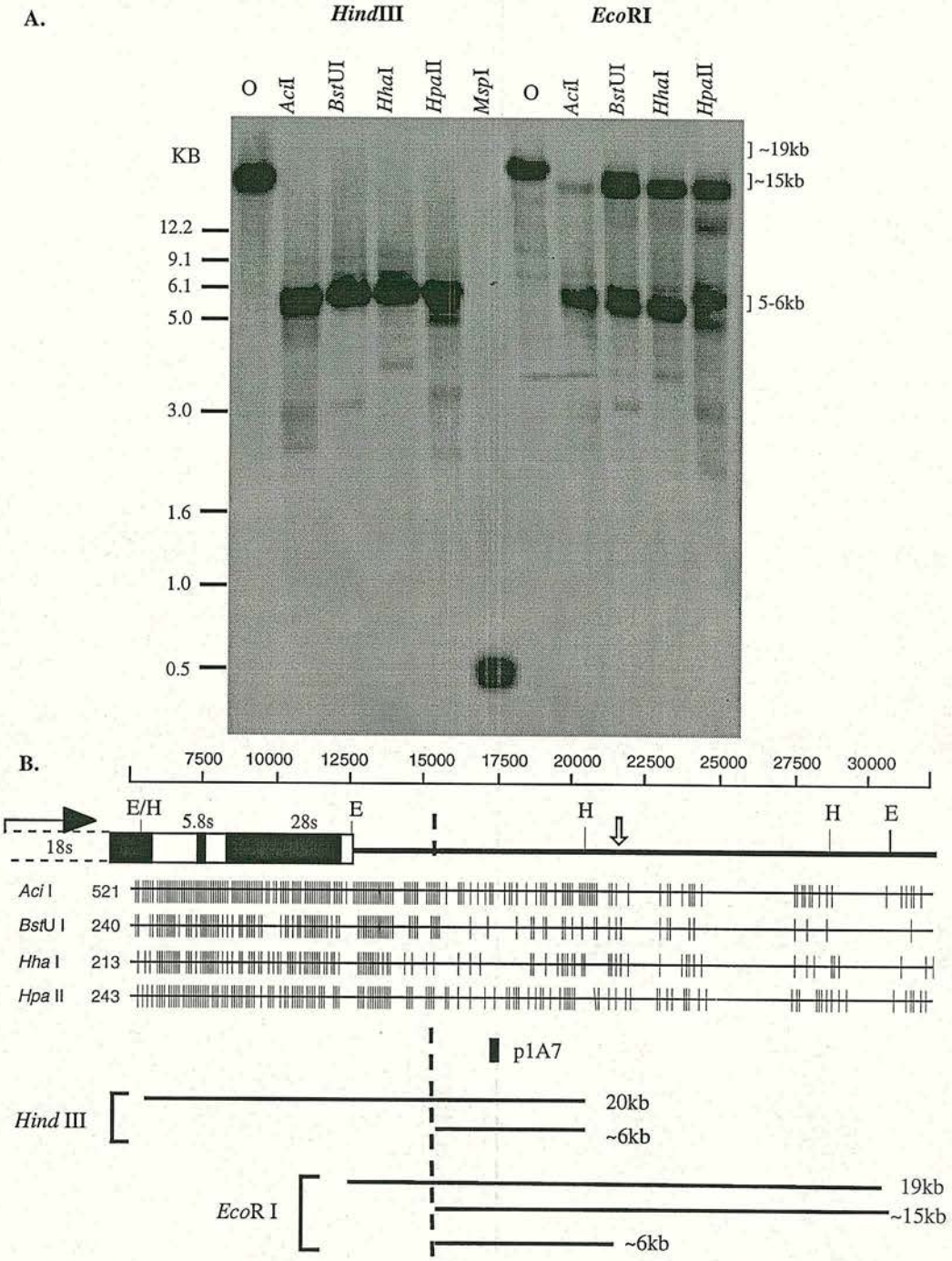
A.

Southern blot analysis of normal male lymphocyte DNA digested first with *Hind*III or *Eco*RI followed by the methylation-sensitive enzymes shown and then hybridised with p1A7.

B.

Restriction map of part of the rDNA repeat unit showing the sites for *Eco*RI (E) and *Hind*III (H) and site maps for each of the methylation-sensitive restriction enzymes that were used. The boundary between methylated DNA (to the right) and non-methylated DNA (to the left) is indicated by a dotted vertical line. A hypomethylated region is marked by an open arrow. The position of probe p1A7 is shown. The origin of prominent bands seen on the autoradiographs is diagrammed below.

Figure 6.2.


A.
HindIII
EcoRI
O
AciI
BstUI
HhaI
HpaII
MspI
O
AciI
BstUI
HhaI
HpaII
KB
12.2
9.1
6.1
5.0
3.0
1.6
1.0
0.5
] ~19kb
]~15kb
] 5-6kb
B.
7500
10000
12500
15000
17500
20000
22500
25000
27500
30000
E/H
5.8s
28s
E
H
H
E
18s
Aci I 521
BstU I 240
Hha I 213
Hpa II 243
p1A7
Hind III
20kb
~6kb
EcoR I
19kb
~15kb
~6kb

Survival of a high proportion of the 16 kb fragments, in spite of the presence of multiple sites for the methylation-sensitive enzymes, indicates that the 186 CpGs in the NTS that were tested in these experiments are methylated in a high proportion of repeat units.

Similar experiments were used to map the methylation boundary at the 5' end of the spacer. Figure 6.3.A. shows that the 13 kb *Hind*III band is reduced to about 7 kb by all four methylation-sensitive enzymes. This places the boundary between methylated spacer and the non-methylated transcription unit at about 1 kb upstream of the transcription start site. In addition there is a collection of weaker bands just below 3 kb, suggesting that this discrete region of the repeat unit is somewhat undermethylated (Figure 6.3.B , see open arrow at position 38,000).

6.3. Conclusions

The Southern blot hybridisation patterns obtained using the mixture of probes and methylation sensitive restriction enzymes indicates methylation at ~300 CpG sites tested in the NTS. If the entire NTS is methylated, this may explain why such a large number of clones in the library are derived from this region. Overall when compared to the transcribed regions of the rDNA repeat the NTS is depleted for CpG. In certain regions, it is not as depleted as the bulk of the genome. It has a %GC which often rises above 50 and a $CpG_{Obs/Exp}$ which is close to 0.75 for large tracts of DNA (Figure 6.4.B). This is particularly true for the 3' and 5' ends which border the unmethylated and transcribed regions. The fragments from the NTS which have been identified in the library are from these areas. A third clone, not used as a probe, was from an *Mse*I fragment close to p1B12. Those sequences cloned and analysed

Figure 6.3.

Investigation of the methylation status of sites upstream of the start of transcription at the 3' end of the NTS

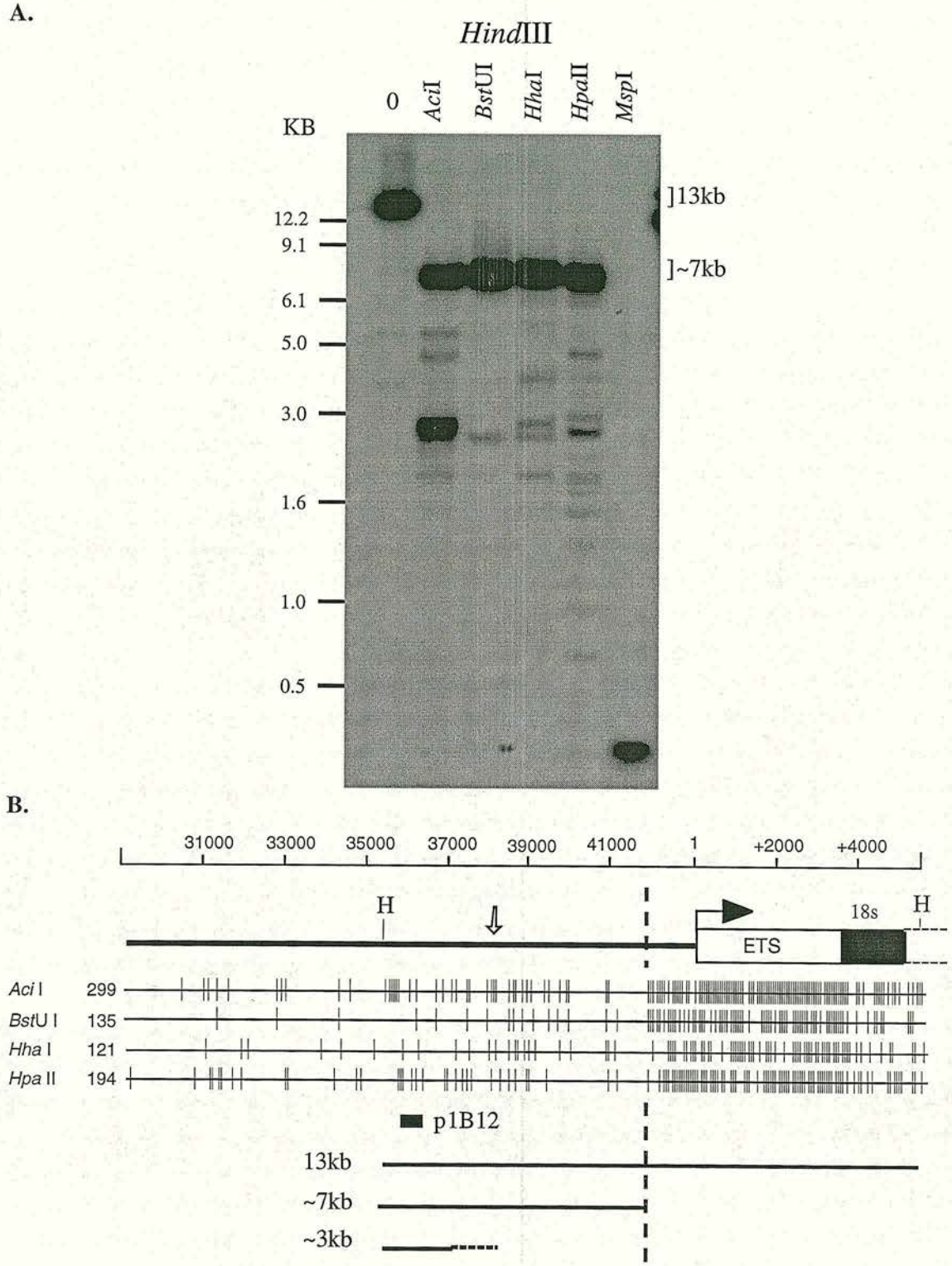
A.

Southern blot analysis of normal male lymphocyte DNA digested with *Hind*III and the methylation-sensitive enzymes shown and then hybridised with p1B12.

B.

Restriction map of part of the rDNA repeat unit showing sites for *Hind*III (arrows marked H) and the methylation-sensitive enzymes that were used. The origin of prominent bands seen on the autoradiograph is diagrammed below. The boundary between methylated DNA (to the left) and non-methylated DNA (to the right) is indicated by a dotted vertical line. The open arrow marks a hypomethylated region.

Figure 6.3.


A.
HindIII
0
AciI
BstUI
HhaI
HpaII
MspI
KB
12.2
9.1
6.1
5.0
3.0
1.6
1.0
0.5
]13kb
]~7kb
B.
31000
33000
35000
37000
39000
41000
1
+2000
+4000
H
18s
H
ETS
Aci I 299
BstU I 135
Hha I 121
Hpa II 194
p1B12
13kb
~7kb
~3kb

are of a high enough m^5CpG frequency to show affinity for the MBD column at high salt. The frequency in the library of such clones may be due the large number of rDNA repeats in the genome (~400) and the methylation level of the NTS. When digested with *Mse*I, a considerable number of the fragments produced from the NTS should bind to the MBD column. During the preliminary analysis only three fragments from the NTS were positively identified in the MBDx5 library. However when the DNA cloned to produce the libraries (either MBDx3 or x5) was used as a probe in a FISH experiment a strong signal was always seen on the short arms of the acrocentric chromosomes (Chapter 5). This is in agreement with the finding that other fragments from the NTS were cloned in MBDx5. The short arms of the acrocentric chromosomes also gave strong signal when antibodies raised against m^5C were used (Barbin et al., 1994). The authors of that study noted that the acrocentric chromosomes contained various amounts of repeated DNA which are often methylated and might account for the staining pattern (Barbin et al., 1994). It is possible that large number of rDNA repeats and the frequency of m^5CpGs in the NTS produced the signal seen with the antibody.

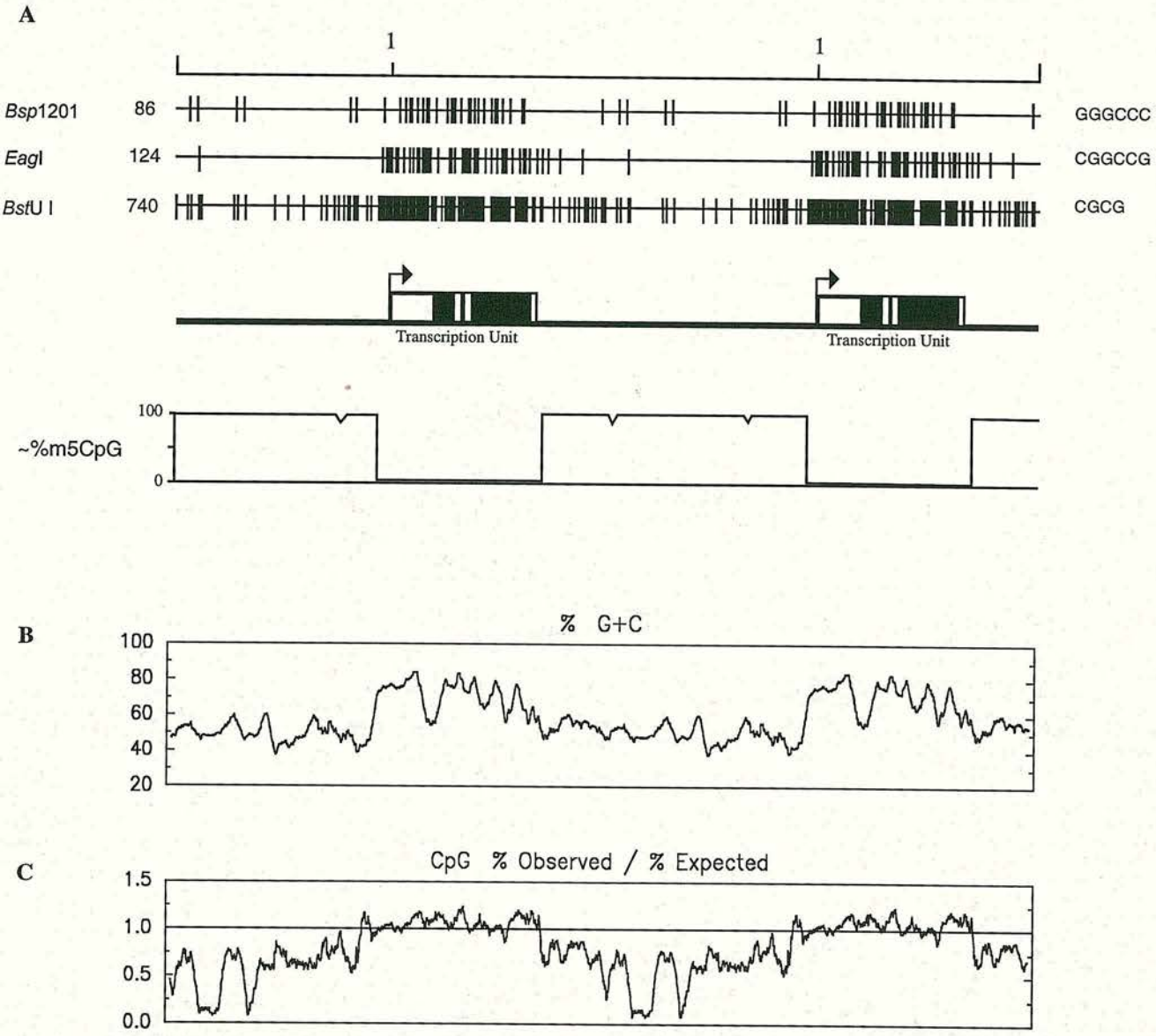
Figure 6.4.

Schematic representation of two rDNA repeat units

A.

Schematic representation of two rDNA repeat units showing the frequency of GC-rich sites for three restriction enzymes *BSp*1201, *Eag*I and *Bst*UI. The transcription unit is represented by a box above the line with an arrow indicating the start of transcription. The 18S, 5.8S and 28S genes are shown in black. Approximate methylation levels are plotted below. (**B**) The plot of GC content across the repeat unit based on a window size of 1000 base pairs and a step of 30 base pairs. (**C**) The plot of the frequency of CpG across the repeat based on the observed frequency in a 1000 base pair window divided by the expected frequency, the step is 30 base pairs.

Figure 6.4


A
1
1
Bsp1201
86
GGGCCC
EagI
124
CGGCCG
BstU I
740
CGCG
Transcription Unit
Transcription Unit
~%m5CpG
100
0
B
% G+C
100
80
60
40
20
C
CpG % Observed / % Expected
1.5
1.0
0.5
0.0

7.1.

Fractionation of the genome according to m^5CpG density

The distribution of m^5CpG is non-uniform throughout the vertebrate genome with distinct regions containing this modified dinucleotide at a higher frequency. The best characterised of these regions are the methylated CGIs which are involved in parental imprinting and X-inactivation. To further investigate such regions a library of highly methylated sequences was constructed. The MBD column allowed methylated CGIs and other regions of dense methylation, to be fractionated from the remainder of the genome. This was due to the strong affinity for highly methylated sequences, shown by the MBD column at high salt concentration (Chapter 3). It was anticipated that in common with known CGIs, sequences from the library would have a %GC of greater than 50 and a $CpG_{Obs/Exp}$ higher than 0.69 (Gardiner-Gardner and Frommer, 1987). The sequenced libraries were all derived from female blood DNA as this source would contain methylated X_i island fragments in addition to those associated with imprinted genes. After three rounds of purification many of the fragments still had sequence characteristics representative of the majority of the genome (Gardiner-Gardner and Frommer, 1987). In order to obtain a sample of the minor highly methylated fraction of the genome, further purification was required. In an attempt to achieve this, the DNA (MBDx3) was first amplified then methylated using *Sss*I (CpG) methylase. The modified DNA was then purified with further passes over the MBD column (Chapter 4). When this final bound fraction of DNA was amplified a heterogeneous smear was produced which was visible from 2 kb down to 200 bp. Three distinct bands could also clearly be seen in both male and female DNA (Figure 4.3.A).

These same three bands could also be seen, though faintly, in the DNA prior to amplification and methylation (MBDx3). Before cloning the purified DNA (MBDx5) was again screened to detect fragments of X-linked CGIs or a fragment which contained a single CpG. As anticipated, the purified DNA contained the two X-linked CGI fragments (Figure 4.3.B). Amplification by PCR followed by treatment with *Sss*I (CpG) methylase, results in these and other CpG-rich fragments becoming extensively methylated and binding at high salt.

7.2.
Characteristics of the further purified DNA sequences.

In addition to these fragments, monitored using PCR, there were also two frequently cloned inserts. These were subsequently shown to be both methylated and to have a high frequency of CpGs. A number of other cloned sequences were examined further and shown to be more GC-rich and have a higher frequency of CpG than the bulk of the genome. The use of cloned sequences as probes against southern-blots indicated that all were methylated at sites tested in DNA derived from blood. In the case of the fragment from the rDNA repeat this was unexpected. The methylation status of the entire rDNA repeat region was then examined using methylation sensitive restriction enzymes and the cloned sequences as probes for southern blots (Chapter 6). This revealed that restriction enzyme sites throughout the NTS were methylated in DNA derived from blood. The purified library sequences (MBDx5) which were analysed were slightly more GC-rich and CpG-rich than the majority of the genome and also methylated in DNA derived from blood.

Although present in the library (Figure 4.3.B.), methylated CGI fragments are still in the minority. It had been anticipated that, as most of the genome is depleted for m5CpG, most fragments generated by *Mse*I digestion would show little affinity for the MBD column at high salt. This would have allowed the exclusive purification of methylated CGIs. Analysis of the cloned inserts indicated that this was not the case and considerably more fragments than anticipated bound to the MBD column at high salt concentration (Figure 4.8). After further purification and pre-screening with various probes, analysis of the remaining sequences revealed a higher frequency of methylated CpGs than the bulk of the genome (Figure 4.9). The high frequency of CpG in methylated CGIs is thought to be due to a the lack of methylation in the germline (Tykocinski and Max, 1984). An alternative explanation is that such regions are in some way protected from deamination, thus preventing the depletion of CpG(Adams and Eason, 1984). Most of the sequences in the final library (MBDx5) are not derived from island regions. However in common with the majority of CGIs, many were not methylated in DNA derived from sperm. This may explain the lack of CpG depletion in these sequences.

7.3.
Distribution of cloned sequences on metaphase chromosomes.

Many of the cloned sequences in the MBDx5 library had been shown to be GC-rich and methylated in blood, but are not from methylated CGIs. The MBDx5 DNA was used as a *FISH* probe to determine the physical position of the constituents of the library on the chromosomes. The MBDx3 library was also used for comparison, sequencing having indicated that the MBDx5 library inserts were enriched for %GC and CpG (Figure 4.10). With both probes a strong hybridisation signal was observed at the centromere of

chromosome 9. The remaining signal when using either the MBDx3 or MBDx5 probes, is predominately seen in the subtelomeric regions of many chromosomes (Chapter 5). Several of the sequences are similar to database sequences for which the relative position on the chromosome is known and these often map to subtelomeric regions (Appendix A). Other sequences such as clone p4F9 map to the DXZ4 locus, which is located on the long arm of the X-chromosome, and reported to be methylated. The sequence that was cloned contains a single *Hpa*II site which has been shown to be methylated in DNA derived from blood (Figure 4.8). Its methylation status in sperm DNA is not known. There are two inserts which are very similar to sequences in the database, clones p4C12 and pCPC3. The former shows sequence similarity to the human ERV9LTR3 gene and to a human telomere associated repeat. The latter is similar to the tandem repeat region of FSHD which contains a satellite 4 sequence (Appendix A). Finally, a number of sequences contained internal repeats which were often very GC-rich and have a large number of CpGs but do not match known sequences in the database. There are several examples in the database of GC-rich repeated sequence, reported to be located in subtelomeric regions (Cheng et al, 1990; Cross et al, 1990; De Lange et al., 1990; Brown et al, 1990).

Antibodies raised against m^5C had previously revealed a localisation to the constitutive heterochromatin of chromosomes 1, 9, and 16 (Miller et al., 1974). These areas have been shown to contain large numbers of alphoid and classic satellite DNA sequences which frequently contain m^5CpG (Meneveri et al., 1985; Almeida et al., 1993; Kokalj-Vokac et al., 1993). More recently, a study used antibodies raised against m^5C, but with a modified method of chromosome preparation (Barbin et al., 1994). With this new method antibody signal was again observed on the secondary constrictions of

chromosomes 1, 9, 16, and also the juxtacentromeric regions of 2, 7, 10 and 17 (Barbin et al., 1994). These additional sites also contain some satellite DNA and other repetitive sequences containing m^5CpGs (Jackson et al., 1993; Kokalj-Vokac et al., 1993; Barbin et al., 1994). The *FISH* hybridisation pattern seen with the MBDx5 is in the presence of competitor. Hybridisation to these regions (with the exception of chromosome 9) is therefore not evident using the MBDx5 probe (Figure 5.1, 5.2 and 5.3).

Signal from the m^5C antibody was also observed on the thermal denaturation resistant or T-bands (Barbin et al., 1994) which are a subset of the reverse staining (R)-bands (Dutrillaux, 1977) (Section 1.2). Of the two main chromosome bands, R-bands and Giemsa staining (G)-bands, the former are more GC-rich. An additional staining procedure can subdivide the R-bands further (Ambros and Sumner, 1987). These GC-richest R-bands correspond almost exactly to the T-bands. These occur on the majority of human chromosomes adjacent to the telomeres and on a minority of intercalary bands (Holmquist, 1992). Another characteristic of T-bands was noted when human DNA extracted from placenta was fractionated into isochores (Saccone et al., 1992). After fractionation, the DNA which had a G+C level of greater than 54.9% and corresponded to the H3 isochore was isolated and used as a probe in a FISH experiment. This method of separation did not produce an entirely pure probe and it contained trace amounts of the H2 isochore and some rDNA sequences (Saccone et al., 1992). The resulting hybridisation pattern was to the telomeric T-bands on seventeen chromosomes. Signal was also observed at intercalary T-bands located on a further six (Saccone et al., 1992). The hybridisation pattern roughly agrees with a report by Holmquist which defined the chromosomal location of 'R' and 'G' bands (Holmquist, 1992). T-bands, a subset of 'R'

bands, represent 15% of all chromosome bands and contain 65% of all mapped genes. With the exception of the T-bands located on 16p, 17p, 19q and the intercalary bands on chromosome 6 and 7, the pattern of hybridisation by H3 is very similar (Holmquist, 1992). Recently an improved protocol for in situ hybridisation and for the isolation of H3 isochore DNA was developed (Saccone et al., 1996). This showed that the previous hybridisation pattern could be better defined with respect to the location of the H3 isochore. The strongest signal was located at 28 R(everse) bands which the authors called $H3^{+}$. An additional 31 'R' bands called H3* also contained H3 isochore material but at a lower concentration. The remaining R bands called $H3^{-}$ did not contain any H3 isochore material (Saccone et al., 1996) . The hybridisation resolved the banding pattern and was in broad agreement with the earlier study and closer to that reported by Holmquist (Holmquist, 1992; Saccone et al., 1992). The most GC-rich regions of many chromosomes appear to be located near the telomeres. Many chromosomes are relatively GC poor with the exception of their terminal R-band (chromosomes 4, 5, 8, 9, 10 and 16). Others, such as chromosomes 1, 17 and 19 have a high GC level on both telomeric and intercalary T-bands. Although neither of these studies examined the sex chromosomes in detail, the short arm and telomere of the X has been studied. Analysis of 36 YACs for almost the entire chromosome band Xq28 an 'R' band identified a 2.2-Mb region formed by H2 and H3 isochores (47.2%GC). This was adjacent to a proximal (telomeric) region of 3.5Mb of essentially L and H1 isochores (43% GC). On the distal (centromeric) side was a 1.3 Mb region again formed by L-isochores (39.5% GC) DeSario et al, (1990). Use of the antibody has showed that these GC-rich T-bands have the additional characteristic of containing a high density of m^5CpGs in somatic cells (Barbin et al., 1994). The m^5C antibody shows regions of methyl-CpG density and the H3 isochore probe

GC-richness. Sequences which are both GC-rich and have a high density of CpG are represented in the MBDx5 library. In addition the FISH experiment using MBDx5 does show some overlap with the pattern seen in T-banding (Figure 5.3). It appears, that highly methylated sequences are located subtelomerically on a number of chromosomes. The function of these GC-rich regions and why many are located close to the ends of some but not all chromosomes remains to be determined.

The main conclusion of the thesis, based on analysis of the cloned inserts, was that the library predominantly contained CpG-rich sequences which were heavily methylated in DNA derived from blood. A large number of these were frequently occurring or repeated sequences. Subtraction of repeated sequences and analysis of the remainder indicated that the were not derived exclusively from methylated CGIs, nor were they from the bulk of the genome. In addition the *FISH* experiments show that many of these sequences are located subtelomerically on some but not all human chromosomes.

References.

Abbott, C. and Povey, S. (1991). Development of Human Chromosome-Specific PCR Primers for Characterization of Somatic Cell Hybrids. *Genomics* **9**: 73-77.

Adams, L. P. R. and Burdon, H. R. (1985). Molecular Biology of DNA Methylation. New York, Springer-Verlag.

Adams, R. L. and Eason, R. (1984). Increased G+C content of DNA stabilizes methyl CpG dinucleotides. *Nucleic Acid Research* **12**: 5869-5877.

Aissani, B. and Bernardi, G. (1991a). CpG islands: features and distribution in the genomes of vertebrates. *Gene* **106**: 173-183.

Aissani, B. and Bernardi, G. (1991b). CpG islands,genes and isochores in the genomes of vertebrates. *Gene* **106**: 185-195.

Almeida, A., Kokalj-Vokac, N., Lefrancois, D., Viegas-Pequignot, E., Jeanpierre, M., Dutrillaux, B. and Malfoy, B. (1993). Hypomethylation of classical satellite DNA and chromosome instability in lymphoblastoid cell lines. *Human Genetics* **91**: 538-546.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403-410.

Ambros, P. F. and Sumner, A. T. (1987). Correlation of pachytene chromomeres and metaphase bands of human chromosomes, and

distinctive properties of telomeric regions. *Cytogenetics and Cell Genetics* **44**: 223-228.

Ansorge, W., Sproat, B., Stegemann, J., Schwager, C. and Zenke, M. (1987). Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. *Nucleic Acids Research* **15**: 4593-4602.

Antequera, F. and Bird, A. (1993). Number of CpG islands and genes in human and mouse. *Proceedings of the National Academy Sciences* **90**: 11995-11999.

Antequera, F., Boyes, J. and Bird, A. (1990). High levels of de novo methylation and altered chromatin structure at CpG islands in cell lines. *Cell* **62**: 503-514.

Barbin, A., Montpellier, C., Kokalj-Vokac, N., Gibaud, A., Niveleau, A., Malfoy, B., Dutrillaux, B. and C., B. A. (1994). New sites of methylcytosine-rich DNA detected on metaphase chromsomes. *Human Genetics* **94**: 684-692.

Beard, C., Li, E. and Jaenisch, R. (1995). Loss of methylation activates Xist in somatic but not in embryonic cells. *Genes and Development* **9**: 2325-2334.

Bernardi, G. (1993). The isochore organization of the human genome and its evolutionary history a review. *Gene* **135**: 57-66.

Bernardi, G., Olofson, B., Kilipski, J., Zerial, M., Salinas, J., Curry, G., Meunier-Rotival, M. and Rodier, F. (1985). The mosaic genome of warm-blooded vertebrates. *Science* **228**: 953 - 958.

Bestor, T., Laudano, A., Mattaliano, R. and Ingram, V. (1988). Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. *Journal of Molecular Biology* **203**: 971 - 983.

Bestor, T. H. and Verdine, G. L. (1994). DNA methyltransferases. *Current Opinion in Cell Biology* **6**: 380-389.

Bird, A., Taggart, M., Frommer, M., Miller, O. J. and Macleod, D. (1985). A fraction of the mouse genome that is derived from islands of non-methylated, CpG-rich DNA. *Cell* **40**: 91 - 99.

Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature* **321**: 209 - 213.

Bird, A. P. (1993). Functions for DNA methylation in vertebrates. *Cold Spring Harbor Symp of Quant Biol* **LVIII**: 281-285.

Bird, A. P. and Taggart, M. H. (1980). Variable patterns of total DNA and rDNA methylation in animals. *Nucleic Acid Research* **8**: 1485 - 1497.

Bird, A. P., Taggart, M. H., Nicholls, R. D. and Higgs, D. R. (1987). Non-methylated CpG-rich islands at the human alpha-globin locus: implications for evolution of the alpha-globin pseudogene. *EMBO Journal* **6**: 999 -1004.

Bird, A. P., Taggart, M. H. and Smith, B. A. (1979). Methylated and unmethylated DNA compartments in the Sea Urchin genome. *Cell* **17**: 889 - 901.

Birnboim, H. C. and Doly, J. (1979). A rapid alkaline extraction method for screening recombinant plasmid DNA. *Nucleic Acid Research* **7**: 1513-1523.

Boyes, J. and Bird, A. (1991). DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. *Cell* **64**: 1123-1134.

Boyes, J. and Bird, A. (1992). Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. *EMBO J ournal***11**: 327-333.

Brandeis, M., Frank, D., Keshet, I., Siegried, Z., Mendelsohn, M., Nemes, A., Temper, V., Razin, A. and Cedar, H. (1994). Sp1 elements protect a CpG island from de novo methylation. *Nature* **371**: 435-438.

Brock, G. J. R. and Bird, A. P. (1997). Mosaic methylation of the repeat unit of the human ribosomal RNA genes. *Human Molecular Genetics* **6**(3): 451-456.

Brown, A. R. W., MacKinnon, J. P., Villasante, A., Spurr, N., Buckie, J. V. and Dobson, J. M. (1990). Structure and Polymorphism of Huma Telomere-Associated DNA. *Cell* **63**: 119-132.

Brown, C. J., Hendrich, B. D., Rupert, J. L., Lafreniere, R. G., Xing, Y., Lawrence, J. and Willard, H. F. (1992). The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell* **71**: 527-542.

Busslinger, M., Hurst, J. and Flavell, R. A. (1983). DNA methylation and the regulation of globin gene expression. *Cell* **34**: 107 - 206.

Chen, E. Y., Cheng, A., Lee, A., Kuang, W. J., Hillier, L., Green, P., Schlessinger, D., Ciccodicola, A. and D'Urso, M. (1991). Sequence of human glucose-6-phosphate dehydrogenase cloned in plasmids and a yeast artificial chromosome. *Genomics* **10**(3): 792-800.

Cheng, F.-J., Smith, L. C. and Cantor, R. C. (1990). Structural and transcriptional analysis of a human subtelomeric repeat. *Nucleic Acid Research* **19**(1): 149-154.

Clark, M. J. (1988). Novel non-templated nucleotide addition reactions catalyzed by procaryotic and eukaryotic DNA polymerases. *Nucleic Acid Research* **16**(20): 9677-9686.

Claverie, J. M. and Makalowski, W. (1994). Alu alert. *Nature* **371**: 752-752.

Cooper, D. N., Taggart, M. H. and Bird, A. P. (1983). Unmethylated domains in vertebrate DNA. *Nucleic Acids Research* **11**: 647 - 658.

Cooper, D. N. and Youssoufian, H. (1988). the CpG dinucleotide and human genetic disease. *Human Genetics* **78**: 151-155.

Coulondre, C., Miller, J. H., Farabaugh, P. J. and Gilbert, W. (1978). Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* **274**: 775-778.

Cross, H. S., Lindsey, J., Fantes, J., McKay, S., McGill, N. and Cooke, H. (1990). The structure of a subterminal repeated sequence present on many human chromosomes. *Nucleic Acid Research* **18**(22): 6649-6657.

Cross, S. H. (1989). Isolation and Characterisation of Human Telomeres. Edinburgh University.

Cross, S. H., Charlton, J. A., Nan, X. and Bird, A. P. (1994). Purification of CpG islands using a methylated DNA binding column. *Nature Genetics* **6**: 236-244.

Crowther, P. J., Doherty, J. P., Lisenmeyer, M. E., Williamson, M. R. and Woodcock, D. M. (1991). Revised genomic consensus for the hypermethylated CpG island region of the human L1 transponson and integration sites of full length L1 elements from recombinant clones made using methylation-tolerant host strains. *Nucleic Acid Research* **19**(9): 2395-2401.

De Lange, T., Shiue, L., Myres, M. R., Cox, R. D., Naylor, L. S., Killery, M. A. and Varmus, E. H. (1990). Structure and variability of human chromosome ends. *Molecular and Cellular Biology* **10**(2): 518-527.

Dower, W. J., Miller, J. F. and Ragsdale, C. W. (1988). High efficiency transformation of *E.coli.* by high voltage electroporation. *Nucleic Acid Research* **16**: 6127-6145.

Driscoll, D. J. and Migeon, B. R. (1990). Sex difference in methylation of single-copy genes in human meiotic germ cells: Implications for X chromosome inactivation, parental imprinting, and origin of CpG mutations. *Somatic Cell & Mol.ecular Genet.ics***16**: 267-282.

Dutrilluax, B. (1977). New chromosome techniques. Molecular structure of human chromosomes . New York, Acedemic Press. 233-265.

Eden, S. and Cedar, H. (1994). Role of DNA methylation in the regulation of transcription. *Current Opinions in Genetics and Development* **4**: 255-259.

Ellison, J., Passage, M., Yu, L., Yen, P., Mohandas, T. K. and Shapiro, L. (1992). Directed isolation of human genes that escape X-inactivation. *Somatic Cell and Molecular Genetics* **18**: 259-268.

Feinberg, A. P. and Vogelstein, B. (1983). A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. *Analytical Biochemistry* **132**: 6-13.

Ferraro, M., Predazzi, V. and Prantera, G. (1993). In human chromosomes telomeric regions are enrcihed in CpGs relative to R-bands. *Chromosoma* **102**: 712-717.

Frank, D., Keshet, I., Shani, M., Levine, A., Razin, A. and Cedar, H. (1991). Demethylation of CpG islands in embryonic cells. *Nature* **351**: 239-241.

Gardiner-Gardner, M. and Frommer, M. (1987). CpG islands in vertebrate genomes. *J.ournal of Mol.ecular Biol.ogy***196**: 261 - 282.

Gates, F. T. and Linn, S. (1977). *Journal of Biological Chemistry* **252**: 1647-1653.

Giacalone, J., Friedes, J. and Francke, U. (1992). A novel GC-rich human macrosatellite VNTR in Xq24 is differentially methylated on active and inactive X chromosomes. *Nature Genetics* **1**: 137-143.

Glenn, C. C., Porter, K. A., Jong, M. T. C., Nicholls, R. D. and Dirscoll, D. J. (1993). Functional imprinting and epigenetic modification of the human SNRPN gene. *Human Molecular Genetics* **2**: 2001-2005.

Goldberg, P. Y., Rommens, M. J., Andrew, E. S., Hutchinson, B. G., Lin, B., Theilmann, J., Graham, R., Glaves, L. M., Starr, E., McDonald, H., Nasir, J., Schappert, K., Kalchman, A. M., Clarke, A. L. and Hayden, R. M. (1993). Identification of an Alu retrotransposition event in close proximity to a strong candidate gene for Huntington's disease. *Nature* **362**: 370-373.

Gonzalez, I. L. and Sylvester, J. E. (1995). Complete sequence of the 43-kb human ribosomal DNA repeat: analysis of the intergenic spacer. *Genomics* **27**(2): 320-328.

Gonzalez, L. I., Wu, S., Li, W., Kuo, A. B. and Svlvester, E. J. (1992). Human ribosomal RNA intergenic spacer sequence. *Nucleic Acid Research* **20**(21): 5846-5847.

Goodfellow, P. J., Mondello, C., Darling, S. M., Pym, B., Little, P. and Goodfellow, P. N. (1988). Absence of methylation of a CpG-rich region at the 5' end of the MIC2 gene on the active X, the inactive X, and the Y chromosome. *Proceedings of the National Academy Sciences* **85**: 5605 - 5609.

Grant, S. G. and Chapman, V. M. (1988). Mechanisms of X-chromosome regulation. *Annual Review of Genetics* **22**: 199-233.

Gruenbaum, Y., Cedar, H. and Razin, A. (1982). Substrate and sequence specificity of a eukaryotic DNA methylase. *Nature* **295**: 620 - 622.

Hewitt, E. J., Lyle, R., Clark, N. L., Valleley, M. E., Wright, J. T., Wijmenga, C., Deutekom van, T. C. J., Francis, F., Sharpe, T. P., Hofker, M., Frants, R. R. and Williamson, R. (1994). Analysis of the tandem repeat locus D4Z4 associated with facioscapulohumeral muscular dystrophy. *Human Molecular Genetics* **3**(8): 1287-1295.

Hochuli, E., Dobeli, H. and Schacher, A. (1987). New metal chelate adsorbents selective for proteins and peptides containing neighbouring histidine residues. *Journal of Chromatography* **411**: 177-184.

Holmquist, P. G. (1992). Chromosome Bands their chromatin flavors and their functional features. *American Journal of Human Genetics* **51**: 17-37.

Jackson, M. S., Slijepcevic, P. and Ponder, B. A. (1993). The organisation of repetitive sequences in the pericentromeric region of human chromosome 10. *Nucleic Acids Research* **21**: 5865-5874.

Jurka, J. and Milosavljevic, A. (1991). Reconstruction and analysis of human *Alu* genes. *Journal of Molecular Evolution.* **32**: 105-121.

Kochanek, S., Renz, D. and Doerfler, W. (1993). DNA methylation in the *Alu* sequences of diploid and haploid primary human cells. *EMBO Journal* **12**: 1141-1151.

Kokalj-Vokac, N., Almeida, A., Viegas-Pequignot, E., Jeanpierre, M., Malfoy, B. and Dutrillaux, B. (1993). Specific induction of uncoiling and recombination by azacytidine in classical satellite-containing constitutive heterochromatin. *Cytogenetics and Cell Genetics* **63**: 11-15.

Korenberg, J. R., Yang-Feng, T., Schreck, R. and Chen, X. N. (1992). Using fluorescence in situ hybridisation (FISH) in genome mapping. *Trends in Biotechnology* **10**: 27-32.

Labuda, D., Zietkiewicz, E. and Mitchell, A. G. (1995). Alu elements as a source of genetic variation: Deleterious effects and evolutionary novelties. The impact of short interspersed elements (SINEs) on the host genome . Austin, R. G. Landes. 1-24.

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**: 680-685.

Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992). CpG islands as gene markers in the human genome. *Genomics* **13**: 1-14.

Lewis, J. D., Meehan, R. R., Henzel, W. J., Maurer-Fogy, I., Jeppesen, P., Klein, F. and Bird, A. (1992). Purification, sequence and cellular localisation of a novel chromosomal protein that binds to methylated DNA. *Cell* **69**: 905-914.

Li, E., Beard, C. and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. *Nature* **366**: 362-365.

Li, E., Bestor, T. H. and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**: 915-926.

Lin, B., Nasir, H., McDonald, H., Graham, R., Rommens, J. M., Goldberg, P. Y. and Hayden, M. R. (1995). Genomic organisation of the human alpha-adducin gene and its alternately spliced isoforms. *Genomics* **25**: 93-99.

Lyon, F. M. (1962). Sex Chromatin and Gene Action in the Mammalian X-Chromosome. *American Journal of Human Genetics* **14**: 135-148.

MacLeod, D., Charlton, J., Mullins, J. and Bird, A. P. (1994). Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. *Genes and Development* **8**: 2282-2292.

Maden, E. B., Dent, L. C., Farrell, E. T., Garde, J., McCallum, F. S. and Wakeman, A. J. (1987). Clones of human ribosomal DNA cotaining the

complete 18S-rRNA and 28S-rRNA genes. Characterization, a detailed map of the human ribosomal transcription unit and diversity among clones. *Biochemistry Journal* **246**: 519-527.

Maniatis, T., Fritsch, E. F. and Sambrook, J. (1982). Molecular cloning: A laboratory manual (Cold Spring Harbor, New York).

Mantei, N., Schwarzstein, M., Streuli, M., Panem, S., Nagata, S. and Weissmann, C. (1980). The nucleotide sequence of a cloned human leukocyte interferon gene. *Gene* **10**(1): 1-10.

McClelland, M., Nelson, M. and Raschke, E. (1994). Effect of site-specific modification on restriction endonucleases and DNA modification methyltransferases. *Nucleic Acid Research* **22**: 3640-3659.

Meehan, R., Lewis, J., Cross, S., Nan, X., Jeppesen, P. and Bird, A. (1992). Transcriptional repression by methylation of CpG. *Journal of Cell Science* **16**: 9-14.

Meehan, R. R., Lewis, J. D., McKay, S., Kleiner, E. L. and Bird, A. P. (1989). Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. *Cell* **58**: 499-507.

Meneveri, R., Agresti, A., Della Valle, G., Talarico, D., Siccardi, A. and Ginelli, E. (1985). Identification of a human clustered G+C-rich DNA family of repeats (*Sau* 3A family). *Journal of molecular Biology* **186**: 483-490.

Meneveri, R., Agresti, A., Marozzi, A., Saccone, S., Rocchi, M., Archidiacono, N., Corneo, G., Valle, D. G. and Ginelli, E. (1993). Molecular organisation and chromosomal localisation of human GC-rich heterochromatin blocks. *Gene* **123**: 227-234.

Migeon, B. R. (1994). X-chromosome inactivation: molecular mechanisms and genetic consequences. *Trends in Genetics* **10**: 230-235.

Miller, O. J., Schendl, W., Allen, J. and Erlanger, B. F. (1974). 5-Methylcytosine localised in mammalian costitutive heterochromatin. *Nature* **251**: 636-637.

Mitchell, A. R., Ambros, P., McBeath, S. and Chandley, A. C. (1986). Molecular hybridization to meiotic chromosomes in man reveals sequence arrangement on the No.9 chromosome and provides clues to the nature of "parameres". *Cytogenetics Cell Genetics* **41**: 89-95.

Mohandas, T., Sparkes, R. S. and Shapiro, L. J. (1981). Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. *Science* **211**: 393-396.

Mondello, C., Goodfellow, P. J. and Goodfellow, P. N. (1988). Analysis of methylation of a human X located gene which escapes X inactivation. *Nucleic Acids Research* **16**: 6813 - 6824.

Morris, J. D. and Reis, A. (1994). A YAC contig spanning the nevoid basal cell carcinoma syndrom, Fanconi anaemia group C and xeroderma pigmentosum group A loci on chromosome 9q. *Genomics* **23**: 23-29.

Nan, X., Meehan, R. R. and Bird, A. (1993). Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. *Nucleic Acids Research* **21**: 4886-4892.

Neufeld, E. J., Skalnik, D. G., Lievens, P. M. J. and Orkin, S. H. (1992). Human CCAAT displacement protein is homologous to the Drosophila homeoprotein, cut. *Nature Genetics* **1**: 50-55.

Nickoloff, J. A. (1992). Transcription enhances intrachromosomal homologous recombination in mammalian cells. *Molecular Cell Biology* **12**: 5311-5318.

Norris, D. P., Patel, D., Kay, G. F., Penny, G. D., Brockdorff, N., Sheardown, S. A. and Rastan, S. (1994). Evidence that random and imprinted Xist expression is controlled by preemptive methylation. *Cell* **77**: 41-51.

Nur, I., Szyf, M., Razin, A., Glaser, G., Rottem, S. and Razin, S. (1985). Procaryotic and eucaryotic traits of DNA methylation in spiroplasmas (mycoplasmas). *Journal of Bacteriology* **164**: 19-24.

Razin, A. and Cedar, H. (1994). DNA methylation and genomic imprinting. *Cell* **77**: 473-476.

Riggs, A. D. (1975). X-inactivation, differentiation and DNA methylation. *Cytogenet. Cell. Genet.* **14**: 9 - 25.

Rothstein, R. J., Lau, F. L., Bahl, P. C., Narang, S. A. and Wu, R. (1979). Synthetic Adaptors for Cloning DNA. *Methods in Enzymology* **68**: 98-108.

Rubin, C. M., VandeVoort, C. A., Teplitz, R. L. and Schmid, C. W. (1994). Alu repeated DNAs are differentially methylated in primate germ cells. *Nucleic Acids Research* **22**: 5121-5127.

Rudkin, G. T. and Stollar, B. D. (1977). High resolution detection of DNA-RNA hybrids in situ by indirect immunofluorescence. *Nature* **265**: 472-473.

Saccone, S., Caccio, S., Kusuda, J., Andreozzi, L. and Bernardi, G. (1996). Identification of the gene-richest bands in human chromosomes. *Gene* **174**: 85-94.

Saccone, S., de Sario, A., Valle, G. D. and Bernardi, G. (1992). The highest gene concentrations in the human genome are in telomeric bands of metaphase chromosomes. *Proceedings of the National Academy Sciences* **89**: 4913-4917.

Sakai, K., Ohta, T., Minoshima, S., Kudoh, J., Wang, Y., De Jong, J. P. and Shimizu, N. (1995). Human Ribosomal RNA Gene Cluster: Identification of the Proximal End Containing a Novel Tandem Repeat Sequence. *Genomics* **26**: 521-526.

Schmid, C. and Maraia, R. (1992). Transcriptional regulation and transpositional selection of active SINE sequences. *Current Opinion in Genetics & Development* **2**: 874-882.

Schorderet, F. D. and Gartler, M. S. (1992). Analysis of CpG suppression in methylated and nonmethylated species. *Proceedings of the National Academy of Sciences* **89**: 957-961.

Schwarzacher-Robinson, T., Cram, L. S., Meyne, J. and Moyzis, R. K. (1988). Characterisation of human heterochromatin by in situ hybridisation with satellite DNA clones. *Cytogenetics and Cell Genetics* **47**: 192-196.

Sealey, P. G., Whittaker, P. A. and Southern, E. M. (1985). Removal of repeated sequences from hybridisation probes. *Nucleic Acid Research* **13**: 1905 - 1921.

Shemer, R., Eisenberg, S., Breslow, J. L. and Razin, A. (1991). Methylation patterns of the human ApoA-I/C-III/A-IV gene cluster in adult and embryonic tissues suggest dynamic changes in methylation during development. *Journal of Biological Chemistry* **266**: 23676-23681.

Slightom, J. L., Blechl, A. E. and Smithies, O. (1980). Human fetal G gamma- and A gamma-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. *Cell* **21**: 627-638.

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology* **98**: 503 - 517.

Strain, L., Warner, P. J., Johnston, T. and Bonthron, T. D. (1995). A human parthenogenetic chimaera. *Nature Genetics* **11**: 164-169.

Studier, W. F., Rosenberg, H. A., Dunn, J. J. and Dubendorff, W. J. (1993). Use of T7 RNA Polymerase to direct expression of cloned genes. *Methods in Enzymology* **185**: 60-89.

Surani, M. Z. (1994). Genomic imprinting: control of gene expression by epigenetic inheritance. *Current Biology* **6**: 390-395.

Sved, J. and Bird, A. (1990). The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proceedings of the National Academy Sciences* **87**: 4692-4696.

Swartz, M. N., Trautner, T. A. and Kornberg, A. (1962). *Journal of Biological Chemistry* **237**: 1961-1967.

Taylor, S. A. M., Snell, R. G., Buckler, A., Ambrose, C., Duyao, M., Church, D., Lin, C. S., Altherr, M., Bates, G. P., Groot, N., Barnes, G., Shaw, D. J., Lehrach, H., Wasmuth, J. J., Harper, P. S., Housman, D. E., MacDonald, M. E. and Gusella, J. F. (1992). Cloning of the α-adducin gene from the Huntington's disease candidate region of chromosome 4 by exon amplification. *Nature Genetics* **2**: 223-227.

Toniolo, D., Filippi, M., Dono, R., Lettieri, T. and Martini, G. (1991). The CpG island in the 5' region of the G6PD gene of man and mouse. *Gene* **102**: 197-203.

Tykocinski, M. L. and Max, E. C. (1984). CG clusters in MHC genes and 5' demethylated genes. *Nucleic Acids Research* **12**: 4385 - 4396.

Urieli-Schoval, S., Gruenbaum, Y., Sedat, J. and Razin, A. (1982). The absence of detectable methylated bases in Drosophila melanogaster DNA. *FEBS Letters* **146**: 148 - 151.

Vardimon, L., Kressmann, A., Cedar, H., Maechler, M. and Doerfler, W. (1982). Expression of a cloned adenovirus gene is inhibited by in vitro methylation. *Proceedings of the National Academy Sciences.* **79**: 1073 - 1077.

Vooijs, M., Yu, L. C., Tkachuk, D., Pinkel, D., Johnson, D. and Gray, W. J. (1993). Libraries for each human chromosome, constructed from sorter-enriched chromosomes by using linker-adaptor PCR. *American Journal of Human Genetics* **52**(586-597):

Weiss, A., Keshet, I., Razin, A. and Cedar, H. (1996). DNA Demethylation In Vitro: Involvement of RNA. *Cell* **86**: 709-718.

Wilkie, O. M. A., Higgs, R. D., Rack, A. K., Buckle, J. V., Spurr, K. N., Fischel-Ghodsian, N., Ceccherini, I., Brown, R. A. W. and Harris, C. P. (1991). Stable length polymorphism of up to 260 kb at the tip of the short arm of human chromosome 16. *Cell* **64**: 595-606.

Wolf, S. F., Jolly, D. J., Lunnen, K. D., Friedman, T. and Migeon, B. R. (1984). Methylation of the hypoxanthine phosphoribosyltransferase locus on the human X-chromosome: Implications for X-chromosome inactivation. *Proceedings of the National Academy Sciences* **81**: 2806 - 2810.

Worton, R. G., Sutherland, J., Sylvester, J. E., Willard, F. H., Bodrug, S., Dube, I., Duff, C., Kean, V., Ray, N. P. and Schmickel, R. D. (1988). Human Ribosomal RNA Genes: Orientation of the Tandem Array and conservation of the 5'end. *Science* **239**: 64-68.

Zhang, Y., Shields, T., Crenshaw, T., Hao. Y, Moulton, T. and Tycko, B. (1993). Imprinting of human *H19*: Allele-specific CpG methylation, loss of the active allele in Wilms Tumor, and potential for somatic allele switching. *American Journal of Human Genetics* **53**: 113-124.

Zollo, M., Mazzarella, R., Bione, S., Toniolo, D., Schlessinger, D., D'Urso, M. and Chen, E. Y. (1994). Sequence and gene content of the RCP/GCP-g6PD region in human Xq28: the first 52 kb. *Unpublished* :

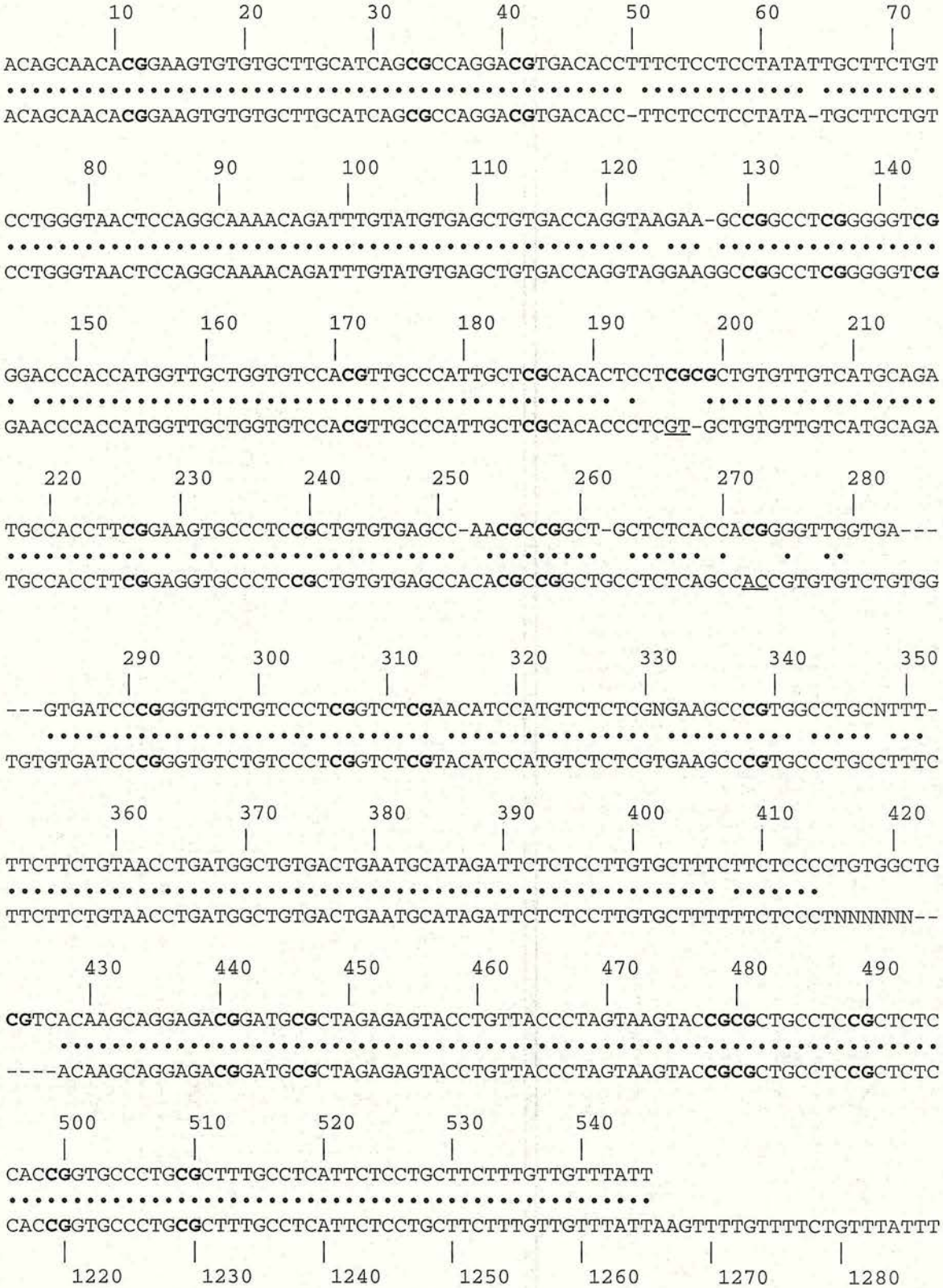
Appendix A

Inserts with sequence matches in the database.

Clone pCPD8 (p2B7)

Clone pCPD8 originally p2B7 had a 545 bp insert with a GC of 56% and a 26 CpGs giving an Obs./Exp. of 0.66. Using the insert as a query in a search of the NCBI database revealed significant similarity to an *Mse*I fragment from the α-adducin gene. The human α-adducin gene spans approximately 85 kb and comprises 16 exons ranging in size from 34 to 1892 bp (Lin et al., 1995). The cDNA has been cloned and different spliced forms have been identified one truncated at exon 10 and another at exon 15 (Lin et al., 1995). The α-adducin gene is located at the telomere of chromosome four in band 4p16.3 (Taylor et al., 1992; Lin et al., 1995). The region contains an Alu retrotransposition event which has occurred within an intron at the 5' end of the α-adducin gene (Goldberg et al., 1993).

Clone pCPD8 (Top) versus α-adducin (Bottom)

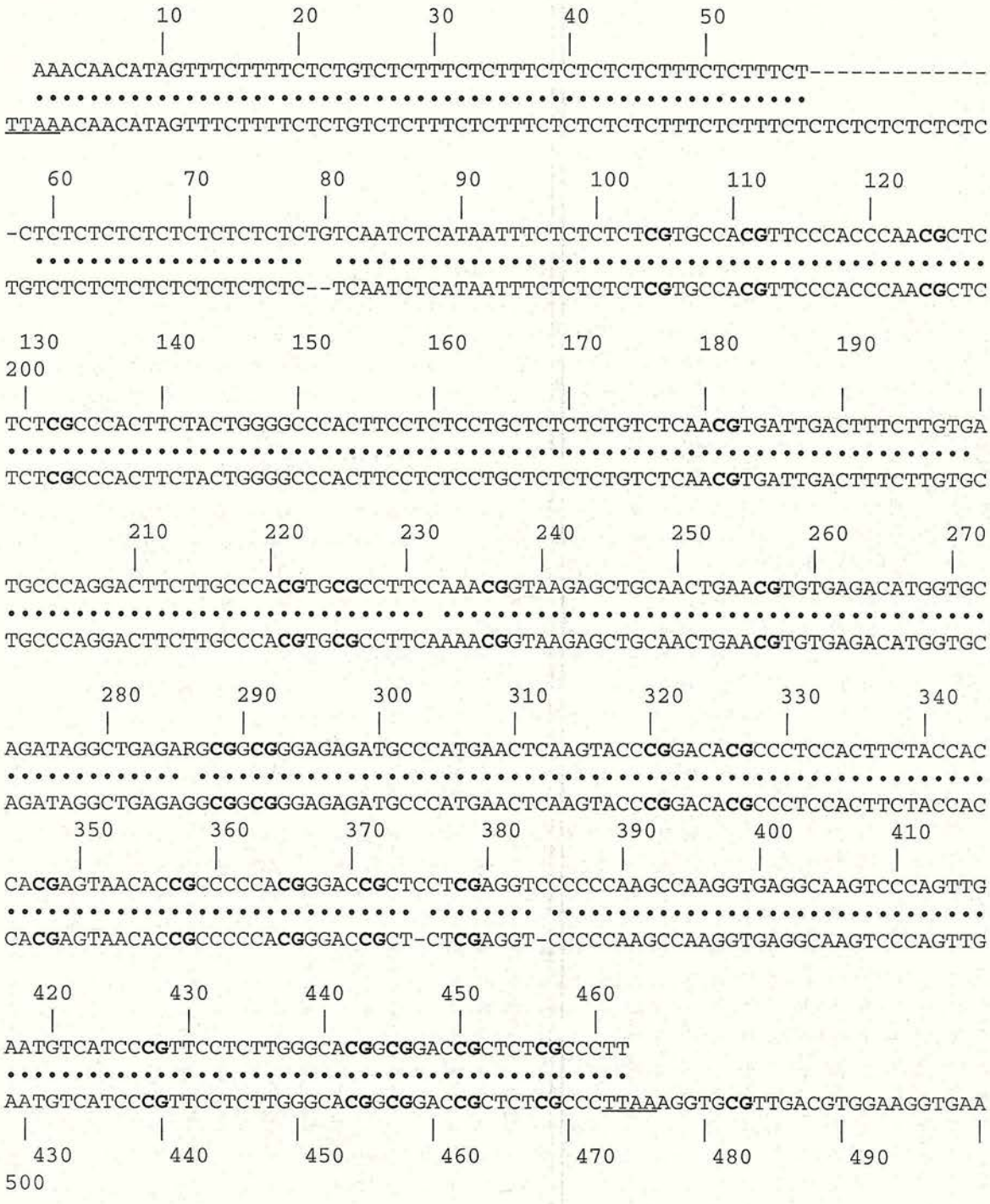


Clone p4F9

Three clones in the library contained an insert with this sequence, the consensus called p4F9 had a 476 bp insert with a %GC of 54 and 22 CpGs giving an Obs./Exp. of 0.70. The insert sequence was used as a query in a database search which revealed similarity to an *Mse*I fragment from the VNTR locus DXZ4 (Giacalone et al., 1992) (Appendix A). The locus DXZ4 consists of a major cluster of 50-100 copies of a unique 3 kb sequence that maps to chromosome band Xq24 (Giacalone et al., 1992).The 3 kb sequence is tandemly repeated and highly polymorphic. The authors looked at the methylation levels of the entire repeat and found that most of the 28 *Hpa*II sites were methylated. The study also claimed to show that the repeat cluster is highly methylated on the active X-chromosome and hypomethylated on the inactive-X (Giacalone et al., 1992).

Alignment of the two sequences reveals a slight gap after nucleotide 57, this may have arisen as the result of either polymorphism or a sequencing error. The relative positions of CpG are shown in bold with the *Mse*I sites underlined.

Sequence of MseI fragment from VNTR DNA (Bottom)
compared to sequence of p4F9



Clone p4C12

Clone p4C12 has been completely sequenced and contains an insert of 498 bp with a %GC of 51 and 20 CpGs giving an Obs./Exp of 0.59. Use of the insert sequence as a query against the NCBI database resulted in several sequences with good similarity. These included the sequence of the human (ERV9LTR3) gene and a human telomere associated repeat sequence, both a aligned on the following pages.

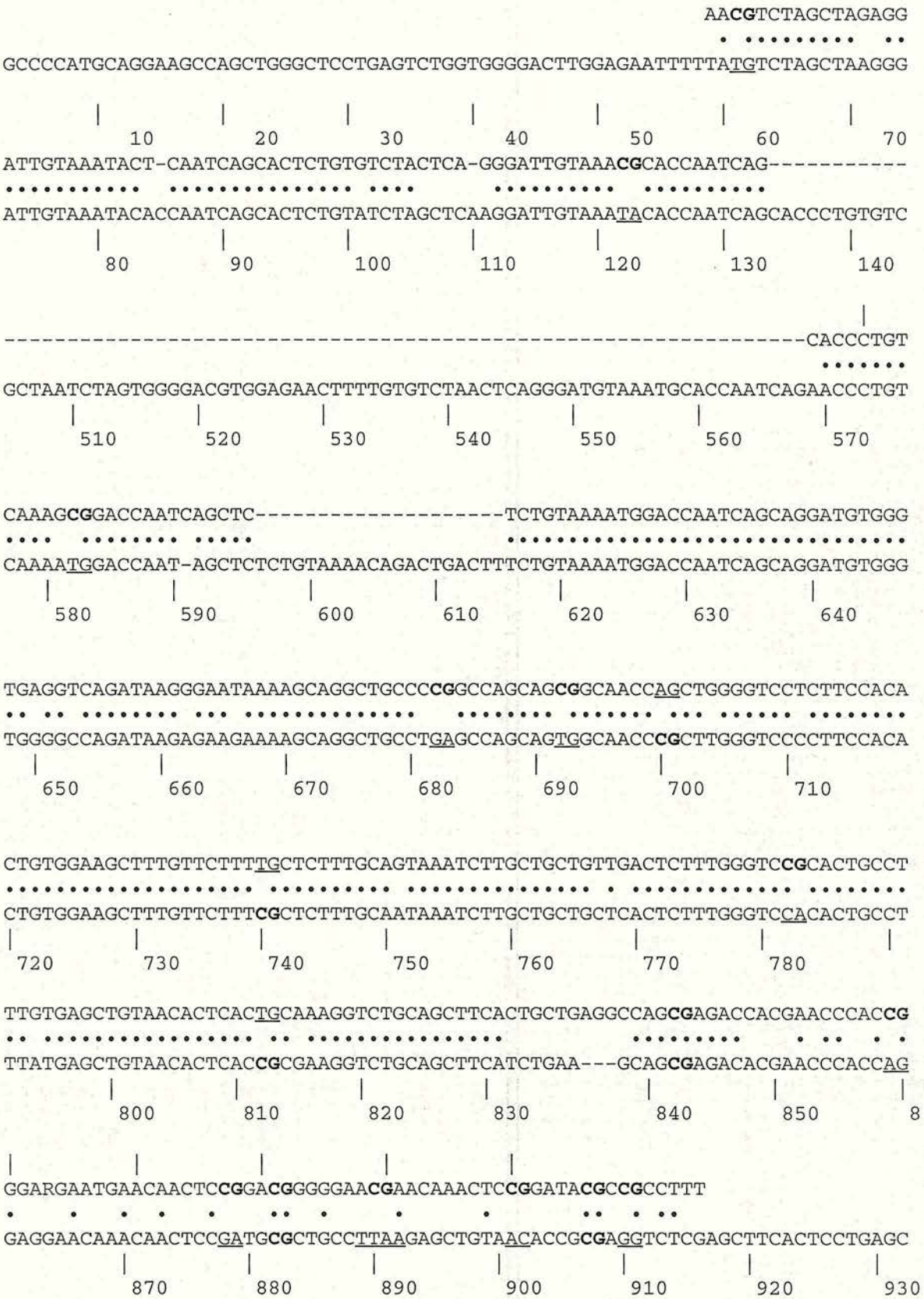
1. Alignment of p4C12 with the ERV9LTR3 gene.

Of the 430 bp of p4C12 almost 220 are almost identical to the a section of the ERV9LTR3 gene sequence. The differences between CpG dinucleotides in the region between the two sequences which gives the best match is shown. The CpGs are in bold and the dinucleotides on the other strand are underlined. There are a number of differences between the two sequences and in addition the ERV9LTR3 gene sequence contains an *Mse*I site at position 890.

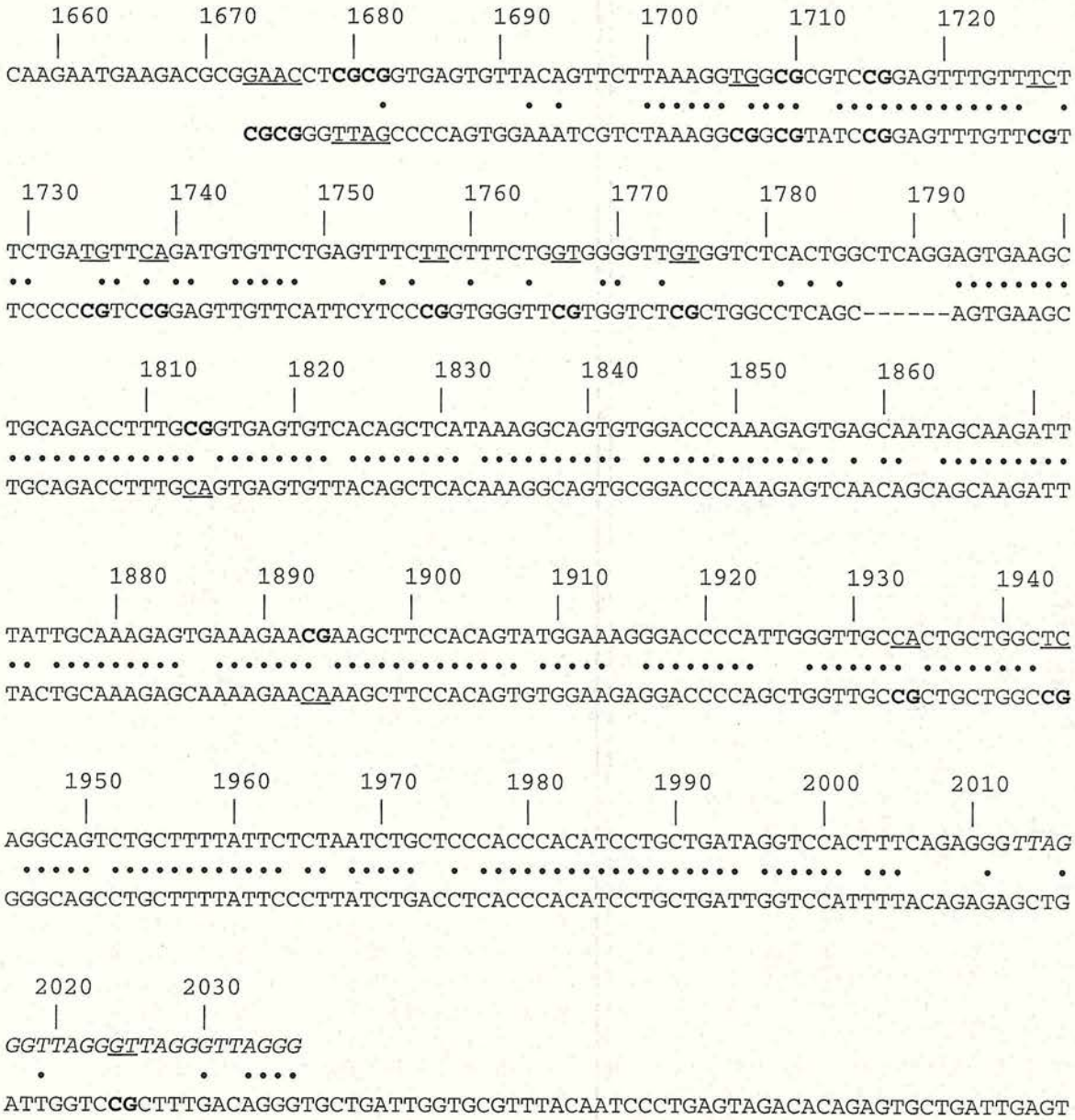
Alignment of p4C12 (inverted) with the human telomere associated repeat sequence.

Approximately 270 bp of the 430 bp sequence of p4C12 are very similar to a portion of the human telomere associated repeat sequence. The CpGs are in bold and the dinucleotides on the other strand are underlined. As with the ERV9LTR3 gene sequence there are a number of differences between the two sequences. The portion of the human telomere associated repeat sequence does not contain any *Mse*I sites and the 3' end includes the T_2AG_3 repeat .

Human ERV9LTR3 gene (Bottom) versus clone p4C12 (Top)



Human telomere associated repeat sequence (Top) clone p4C12 (Bottom)

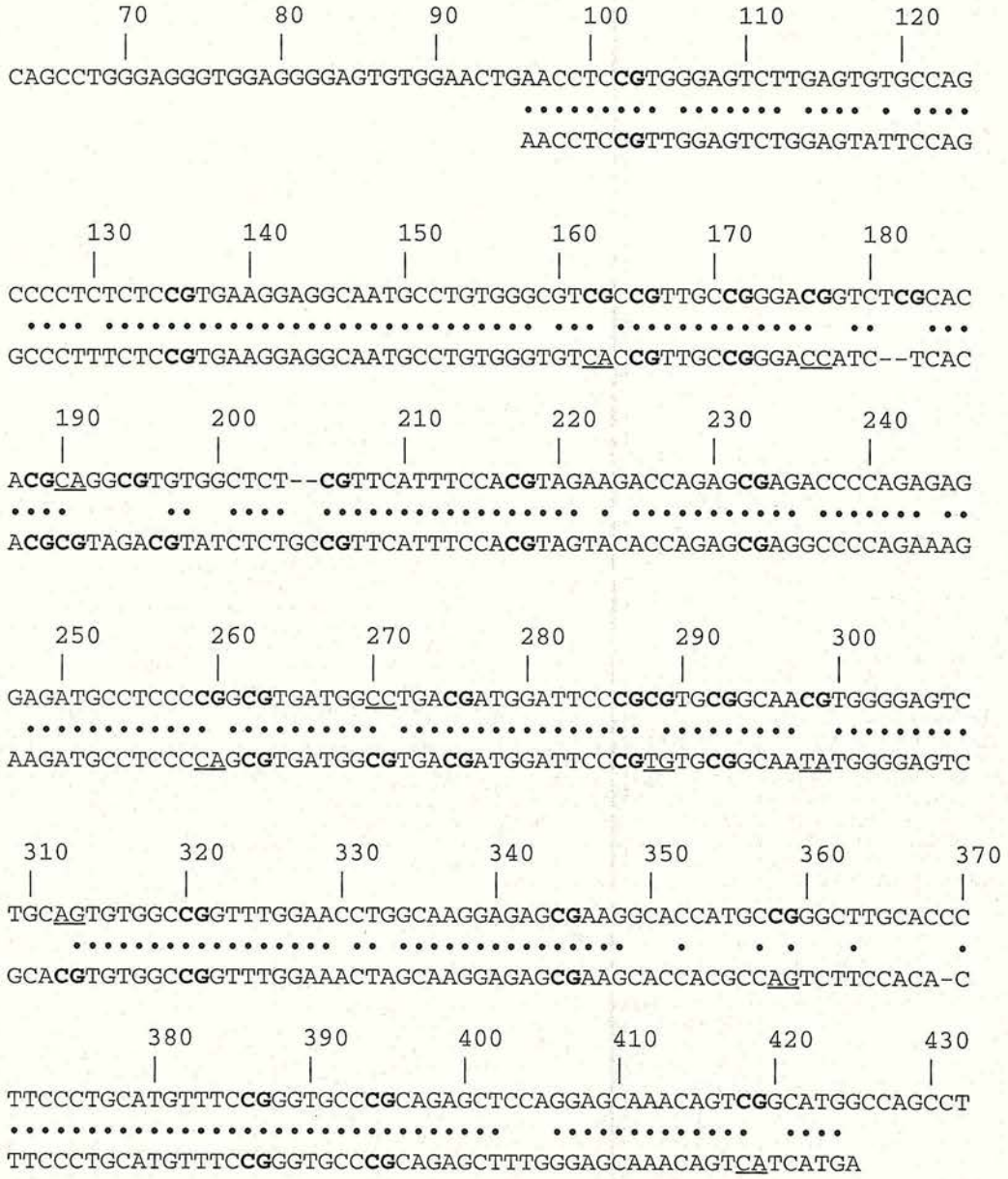


Clone pCPC3

Clone pCPC3 contains an insert of approximately 950 bp which has not been fully sequenced. The T7 primer produced 329 bp of sequence with a %GC of 57 and 22 CpGs giving an Obs./Exp. of 0.81. The SP6 primer produced 462 bp of sequence with a %GC of 48 and 4 CpGs giving an Obs./Exp. of 0.15.

When the available sequences were used as a query in a search of the NCBI database sequence similarity was found to the human tandem repeat region from facioscalpulohumeral muscular dystrophy (FSHD) (Hewitt et al., 1994). The tandem repeat sequence (D4Z4) is associated with the FSHD 3.3 kb repeat and contains two homeoboxes and two previously described repetitive sequences, LSau and a GC-rich low copy repeat hhspm3 (Hewitt et al., 1994). The authors show that D4Z4 is located on chromosome 4 in band q35 but that it also occurs on other chromosomes.

Human chromosome 4 satellite repeat DNA sequence (Top) pCPC3 (Bottom)

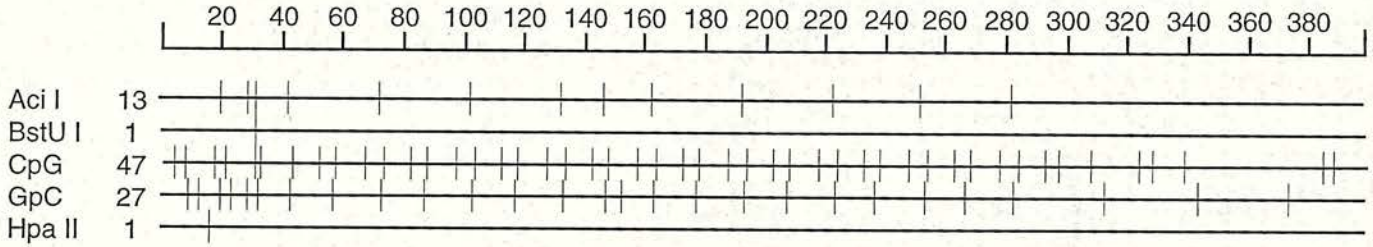


Appendix B

Clones which contain inserts with internal repeats.

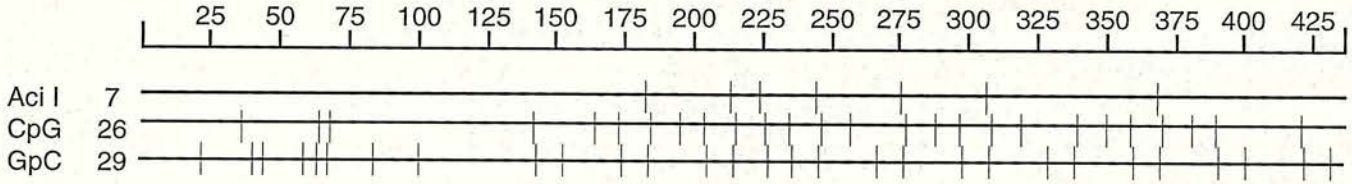
A selection of inserts contained internally repeated sequences but show no sequence similarity to internal repeats in the database. Examples of three of the internal repeats cloned are shown with the internally repeated sequences highlighted (Figure 7.10) .

Contig CP F12



CGACGTCGCATGCTCCCGGCCGCCATGGCCGCGGGATTTGTGCGGTGAGGACGGAGCGTG
TCCCTCCGTGTGCGGTGAGGACGGAGCGTGTCCCTCCGTGTGCGGTGAGGACGGAGCGTG
TCCCTCCGTGTGCGGTGAGGACGGAGCGGGTGCCTCCGTGTGCGGTGAGGACGGAGCGTG
TCCCTCCGTGTGCGGTGAGGACGGAGCGTGTCCCTCCGTGTGCGGTGAGGACGGAGCGTG
TCCCTCCGTGTGCGGTGAGGACGGAGCGTGTCCCTCCGTGTGCGGTGAGGACGGARCGTG
TCCCTCCGTGTGCTGATGAGGACGGAACGTGTCTCACCGTGTGCNGTGAGGACAGANCCT
GTCCCTCCNTGTGCNGTGAGGACCGAACGTGTCCCTCC

Probe 4G2



AAATGATACAACTGTATANGGCACTTACCATGAACGGGGCTTGCAGGAYTA
GAAGTTGCCTGCGTGCGTCAGGGAGTGAGTGGCAAGTGAATGTGAAGGCCT
GGGACATTACTGTACACTTTGGGGGACTTTATAACAGACGCTGTACACTGCTGGGGACTCCG
TAACAGACGCTGTACACTGCGGGGGACTCCGTAACAGACGCTGTACACTGCGGGGGACTCCG
CAACAGACGCTGTACACTGCGGGGGACTCCGTAACAGAAGCTGTACACTGCGGGGGACTCCG
TAACAGACGCTGTACACTGCGGGGGACTCCGTAACAGAAGCTGTACACTGCGNNAGANTCCG
TAACAGACGCTGTACACTGCGGGGGACTCCGTAACAGACGCTGTACACTGCTGGGGACTCTG
TAACAGACGCTGTACACTGCTGGG

Appendix C

Publications arising

Human Molecular Genetics, 1997, Vol. 6, No. 3 451–456

Mosaic methylation of the repeat unit of the human ribosomal RNA genes

Graham J. R. Brock and Adrian Bird*

Institute of Cell and Molecular Biology, University of Edinburgh, Kings Buildings, Edinburgh EH9 3JR, UK

Received October 22, 1996; Revised and Accepted December 11, 1996

The pattern of methylation in human genes for 18S and 28S ribosomal RNA has been investigated using methylation-sensitive restriction enzymes. We find that the transcribed region of the repeat unit is predominantly unmethylated, in agreement with previous studies. In contrast the non-transcribed spacer, which makes up the majority of the 43 kb repeat unit, is highly methylated in blood cell DNA. The boundaries between methylated and non-methylated domains appear to be relatively sharp, and occur ~1.5 kb upstream of the 5′ edge of the proximal promoter and ~1.0 kb downstream of the 3′ end of the transcribed region. A small proportion of all repeat units are methylated throughout the transcribed region, and may represent silent genes. The coincidence between the methylation pattern, the transcription pattern and other features of the repeat unit has implications for our understanding of the mechanism by which patterns of DNA methylation are generated.

INTRODUCTION

The human 18S and 28S ribosomal RNA (rRNA) genes are present at ~400 copies per human haploid genome, clustered on the short arms of the five acrocentric chromosomes (1). Each gene is part of a 43 kb repeat unit that can be divided into two regions: a 13.3 kb transcribed region which contains the highly conserved genes for 18S, 5.8S and 28S rRNA subunits of the ribosome, and a 30 kb non-transcribed spacer (NTS) (2). Repeat unit clusters consist of head-to-tail arrays of ~80 repeats (3).

Vertebrate DNA is frequently modified at the dinucleotide CpG by addition of a methyl group to give m^5CpG (4). The modification can alter the ability of various restriction enzymes to cleave DNA, and this facilitates investigation of methylation at these sites (5,6). A previous study of methylation levels in the transcribed region of the rDNA repeat of various species demonstrated that the conserved region of the repeat unit exists in a predominantly unmethylated form in DNA from a variety of mammals, including man (7). In the mouse, a small fraction of all rDNA repeats are methylated throughout the transcribed region, and indirect evidence has suggested that these correspond to transcriptionally inactive repeats (8). The findings in mammals contrast with those in amphibia and fish, where chromosomal rDNA is heavily methylated in both transcribed and non-transcribed regions of the repeat unit (9,10).

This study began with isolation of fragments of the rDNA spacer from a library of densely methylated DNA fragments. The library was constructed by fractionating human blood DNA over a column that specifically binds to densely methylated DNA (11). The frequent occurrence of rDNA fragments in the library was unexpected given previous evidence that the transcribed region of mammalian rDNA is predominantly unmethylated. Further experiments showed that, in fact, human rDNA is mosaic with respect to CpG methylation. The transcribed region is indeed largely methylation-free at testable sites, but the spacer is densely methylated.

RESULTS

A library of densely methylated sequences was constructed after fractionating human blood DNA on a methyl-CpG binding column (G.J.R.Brock, unpublished). Fragments that repeatedly bound to the column were cloned in plasmids. The library was tested by examination of the DNA sequence and methylation status of random inserts. Clones containing p1A7 and p1B12 showed DNA sequence identity with the NTS of human rDNA. Each clone was represented many times in the library. Methylation was analysed by using inserts from the library as probes against Southern blots of genomic DNA that had been digested with the methylation-sensitive enzyme *Sma*I (CCCGGG) or its methylation-insensitive isoschizomer *Xma*I. Probes p1A7 and p1B12 both hybridised to very high molecular weight *Sma*I fragments, around 28 kb, whereas *Xma*I generated smaller fragments as predicted by the sequence of the spacer (Fig. 1). Weaker *Sma*I bands <27 kb indicated that sites in this part of the repeat unit are occasionally non-methylated. The large size of the major band in the *Sma*I lanes showed that p1A7 and p1B12 are part of long tracts of DNA in which multiple (at least 14) *Sma*I/*Xma*I sites are usually in a methylated state. The equivalent experiment with sperm DNA again showed that the sequence environment of p1A7 and p1B12 is methylated, but the level of methylation is significantly lower than in blood DNA. Most of the rDNA repeats in sperm DNA were digested by *Sma*I to give fragments <20 kb, and some fragments were of the same size as in the *Xma*I digested lane.

Hybridisation of a similar blot with probes from the transcribed region (pHsrDNA5.1/7.9) gave a contrasting result. *Sma*I and *Xma*I patterns of both blood and sperm DNA were predominantly the same, indicating that the majority of sites in this part of the

*To whom correspondence should be addressed

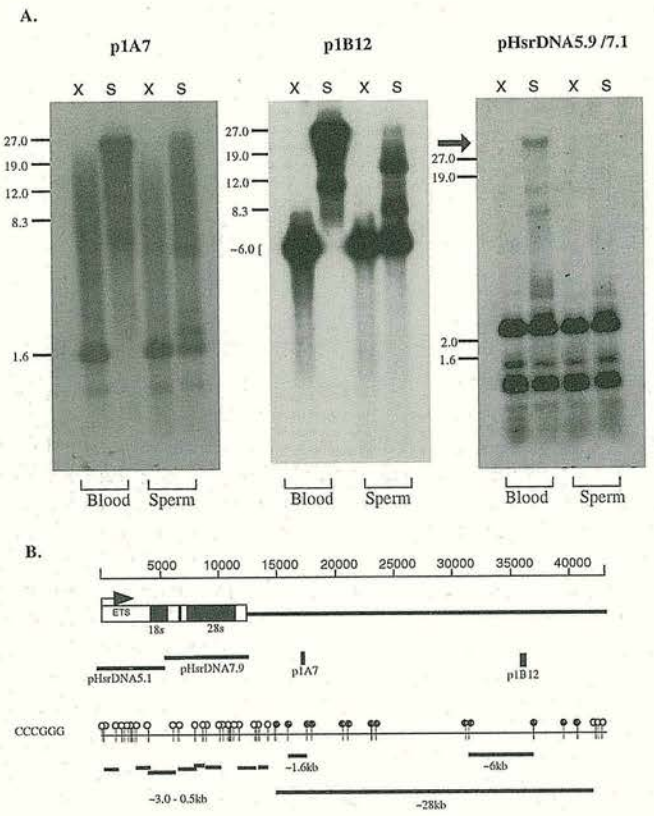


Figure 1. Investigation of the methylation pattern of human rDNA repeat units using Southern blot analysis of DNA from lymphocytes or sperm. DNA was digested with *Eco*RV to reduce the molecular weight of non-ribosomal sequences and then further digested with *Xma*I or its methylation-sensitive isoschizomer *Sma*I. (**A**) Hybridisation is shown for probes corresponding to the transcribed region of the repeat unit (pHsrDNA5.9 and pHsrDNA7.1) and the NTS probes p1A7 and p1B12. The arrow highlights very large fragments in the *Sma*I lane of blood DNA after probing with the transcribed region. (**B**) Map of a single rDNA repeat unit showing the positions of probes, sites for *Xma*I/*Sma*I and their methylation status. Open circles denote non-methylated CpGs; partially filled circles depict sites which are methylated with a high frequency. Prominent bands on the blot probed with pHsrDNA5.9 and 7.1 correspond in size with the fragments shown below the map and it was therefore concluded that sites flanking these fragments are non-methylated. Identity of *Msp*I and *Hpa*II digests probed with pHsrDNA5.9 and 7.1 also indicated that this region lacks detectable methylation (data not shown). The large *Sma*I-resistant band on blots probed with NTS fragments is interpreted to correspond to the complete NTS sequence. This deduction depends on experiments shown in Figures 2 and 3, and on unpublished data. The diagram of the transcription unit shows 18S, 5.8S and 28S rRNA in black, and the ETS, ITS1 and ITS2 regions of the rRNA precursor unshaded. The NTS region is represented by a black line.

repeat unit are not methylated. Similar results were obtained when *Hpa*II and *Msp*I were used in place of *Xma*I and *Sma*I (data not shown), in agreement with an earlier study (7). A notable feature of blots that were probed with the transcribed regions of the repeat unit is the small but significant fraction of hybridisation in the unresolved high molecular weight regions of the gel (Fig. 1A, arrow). This suggests that, although most repeats are non-methylated in this region, a minor proportion of repeat units are heavily methylated in blood cells. The methylated fraction is not apparent in sperm DNA and may be absent. Alternatively, the fraction may be present, but the reduced level of methylation seen in the spacer regions of sperm DNA (Fig. 1) may give a dispersed and therefore indistinct pattern of large fragments.

In order to map the boundary between methylated and non-methylated regions of the repeat, genomic DNA was digested with *Eco*RI or *Hin*dIII together with a selection of restriction endonucleases which contain CpG in their recognition sequence and are sensitive to methylation. The CpG enzymes were *Aci*I (CCGC), *Bst*UI (CGCG), *Hha*I (GCGC) and *Hpa*II (CCGG). *Msp*I, a methylation-insensitive isoschizomer of *Hpa*II, was used as a control. To map the 5′ boundary of the methylated domain, blots were probed with p1A7 (Fig. 2). The 15 kb *Hin*dIII fragment that hybridised to the probe spans the 3′ half of the transcription unit and 6.9 kb of downstream spacer. All four methylation-sensitive enzymes in combination with *Hin*dIII produced bands in the region of 5.5–6.0 kb. This locates the furthest downstream site for each enzyme that is consistently non-methylated at 1.0–1.5 kb downstream of the 3′ end of the transcription unit (Fig. 2, broken vertical line close to nucleotide 14 500). Digestion with *Eco*RI plus methylation-sensitive enzymes confirmed this location for the boundary between methylated and non-methylated domains of the repeat unit. *Eco*RI alone gave the expected band of 18.1 kb when probed with p1A7. Further digestion with methylation-sensitive enzymes gave bands clustered at 16 kb due to cleavage at position 14 500. These enzymes also generated bands at ~6 kb, indicating a discrete hypomethylated region ~200 bp downstream of the *Hin*dIII site (Fig. 2B, open arrow at nucleotide 20 500). The hybridisation signal is distributed roughly evenly between the 16 and 6 kb bands in these lanes. Survival of a high proportion of the 16 kb fragments, in spite of the presence of multiple sites for the methylation-sensitive enzymes, indicates that the 186 CpGs in the NTS that were tested in these experiments are methylated in a high proportion of repeat units.

Similar experiments were used to map the methylation boundary at the 5′ end of the spacer. Figure 3 shows that the 13 kb *Hin*dIII band is reduced to ~7 kb by all four methylation-sensitive enzymes. This places the boundary between methylated spacer and the non-methylated transcription unit at ~1 kb upstream of the transcription start site. In addition there is a collection of weaker bands just below 3 kb, suggesting that this discrete region of the repeat unit is somewhat undermethylated (Fig. 3B, see open arrow at position 38 000).

DISCUSSION

We have demonstrated a mosaic pattern of CpG methylation in the majority of human rDNA repeat units. The 13.3 kb transcribed region is apparently free of methylation, whereas the NTS which lies between consecutive transcribed regions is highly methylated at ~300 tested CpG sites (Fig. 4A). The transition between methylated and non-methylated domains appears to be sharp and occurs near the boundaries of the transcribed region. Previous studies have only determined the methylation status of the transcribed region of the repeat unit. In this respect there is a great contrast within the vertebrates. Fish and amphibia have heavily methylated transcribed regions, whereas the mammals (and perhaps birds and reptiles) have transcribed regions that are predominantly non-methylated (7). The present results reduce the contrast somewhat, as it now seems likely that a major part of the chromosomal rDNA repeat unit may be methylated in most or all vertebrates.

The sequence of the human rDNA repeat unit shows that the dramatic variation in methylation is matched by less dramatic, but

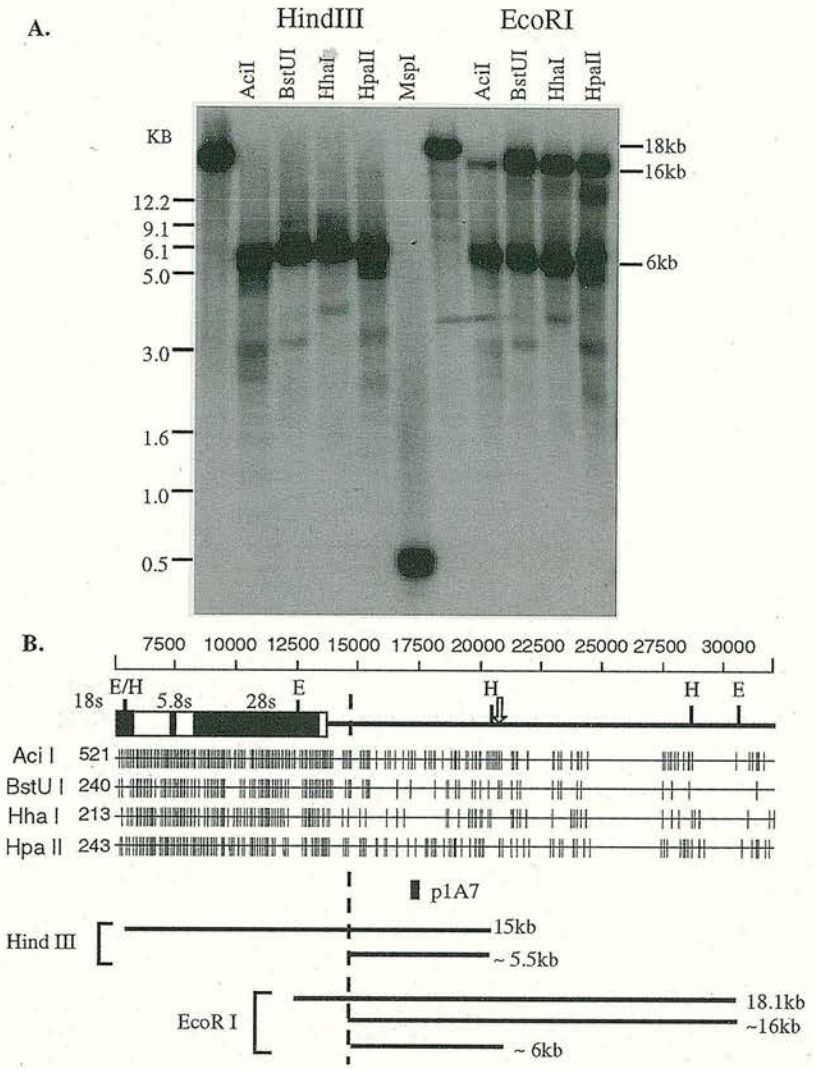


Figure 2. Investigation of the methylation status of sites downstream of the termination of transcription at the 5′ end of the NTS. (**A**) Southern blot analysis of normal male lymphocyte DNA digested first with *Hin*dIII or *Eco*RI followed by the methylation-sensitive enzymes shown and then hybridised with p1A7. (**B**) Restriction map of part of the rDNA repeat unit showing the sites for *Eco*RI (E) and *Hin*dIII (H) and site maps for each of the methylation-sensitive restriction enzymes that were used. The boundary between methylated DNA (to the right) and non-methylated DNA (to the left) is indicated by a dotted vertical line. A hypomethylated region is marked by an open arrow. The position of probe p1A7 is shown. The origin of prominent bands seen on the autoradiographs is diagrammed below.

easily detectable, differences in sequence composition. Firstly, base composition fluctuates through the repeat unit from a high average GC content over the transcribed region, to a lower GC content over the NTS. This is apparent from the uneven frequency of GC-rich restriction enzyme sites (Fig. 4A), and also from a plot of average GC-content across the repeat unit (Fig. 4B). In some respects the structure of the repeat resembles the large scale isochore structure of the genome in which regions of differing average base composition are juxtaposed (20). One possible reason for the GC-richness of the transcribed region is selective pressure to conserve the rRNA precursor sequence. This seems an inadequate explanation, however, as the 5′ transcribed spacer evolves quite rapidly (there is little overall sequence similarity between human and rat or mouse in this region) and is therefore not apparently under strong sequence constraint (21–25). In spite of the sequence variation, the GC-rich character of the external transcribed spacer is maintained in mammals (23,24). Indeed, the human 18S sequence is less GC-rich than the transcribed spacers that flank it (see Fig. 4B). It is also apparent that the GC-rich domain, like the non-methylated domain, extends outside the region that is transcribed, to include the proximal promoter sequences (see also ref. 26) and regions downstream of the transcription termination site. This would not be expected if selection only for GC-rich RNA was operating. It seems more likely that the phenomena that relate lack of methylation to transcription also lead to GC-richness in this domain.

Another prominent feature of the repeat unit sequence is the fluctuation in CpG frequency (observed/expected). In the transcribed region, CpG occurs at near the expected level, but in the NTS there is a significant CpG deficiency (Fig. 4C). This finding fits well with the observation that methylation occurs in the NTS, as the primary cause of CpG deficiency is the mutability of m^5C (26). It is possible to see analogies between the non-methylated GC-rich domain of rDNA and CpG islands, which have similar properties (26). Both are also associated with transcription, although the details differ. CpG islands usually

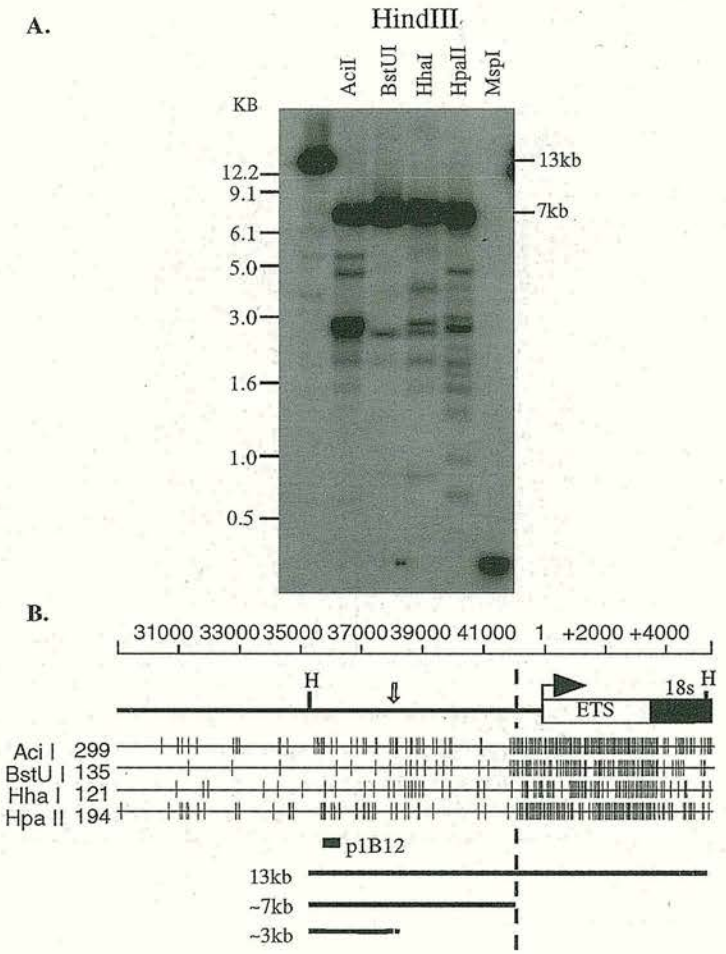


Figure 3. Investigation of the methylation status of sites upstream of the start of transcription at the 3′ end of the NTS. (**A**) Southern blot analysis of normal male lymphocyte DNA digested with *Hin*dIII and the methylation-sensitive enzymes shown and then hybridised with p1B12. (**B**) Restriction map of part of the rDNA repeat unit showing sites for *Hin*dIII (arrows marked H) and the methylation-sensitive enzymes that were used. The origin of prominent bands seen on the autoradiograph is diagrammed below. The boundary between methylated DNA (to the left) and non-methylated DNA (to the right) is indicated by a dotted vertical line. The open arrow marks a hypomethylated region.

extend for ~1 kb downstream from the promoter of a gene and rarely encompass the entire transcription unit. Only in cases where the gene is unusually small does it lie entirely within the island (for example, α globin; see ref. 27). In rDNA the non-methylated domain is much larger than a typical CpG island (13 kb) and includes all of the transcription unit. Although the mosaic character of the rDNA repeat unit cannot yet be satisfactorily explained, it is notable that the region that is methylation-free and GC-rich corresponds with parts of the repeat unit that are probably associated with proteins. These are bound to the promoter and termination sites, and pass repeatedly through the transcribed region during transcription. It is conceivable that protein–DNA complexes block access to the DNA methyltransferase. A more radical possibility is that the act of transcription actively demethylates this part of the repeat unit. Preliminary evidence for RNA-mediated demethylation has been reported (28). Whatever the mechanism, it is important to keep in mind that the transcription unit of amphibian rDNA is maintained in a heavily methylated state in spite of protein associations and transcription (10).

Analysis of amplified ribosomal RNA genes in a human lymphoblastoid cell line has suggested that the methylated genes are transcribed at low levels (29). Active and relatively inactive populations of rDNA were also detected in HeLa cells (30). In the mouse, it has been shown that a minority of rDNA repeats are methylated in the transcribed region and are probably transcriptionally inactive (8). The present work shows a parallel phenomenon in normal human cells, as a small proportion of rDNA from blood is resistant to methylation-sensitive restriction enzymes in the transcribed region of the repeat unit. By analogy with the mouse, the methylated repeats may represent inactive human rRNA genes. There is evidence that lack of methylation at CpG islands depends on the binding of factors that are also required for transcription to occur (31,32). A similar requirement may apply to rDNA. Inactivity of a gene may invite *de novo* methylation. Once methylated, the silent repeat unit may be unable to return to transcriptional activity due to the repressive effect of methylation (33). In this way CpG methylation may stabilise the repressed state.

MATERIALS AND METHODS

DNA

DNA derived from blood was extracted using standard protocols (12), a gift from L. Strain (MRC Human Genetics Unit,

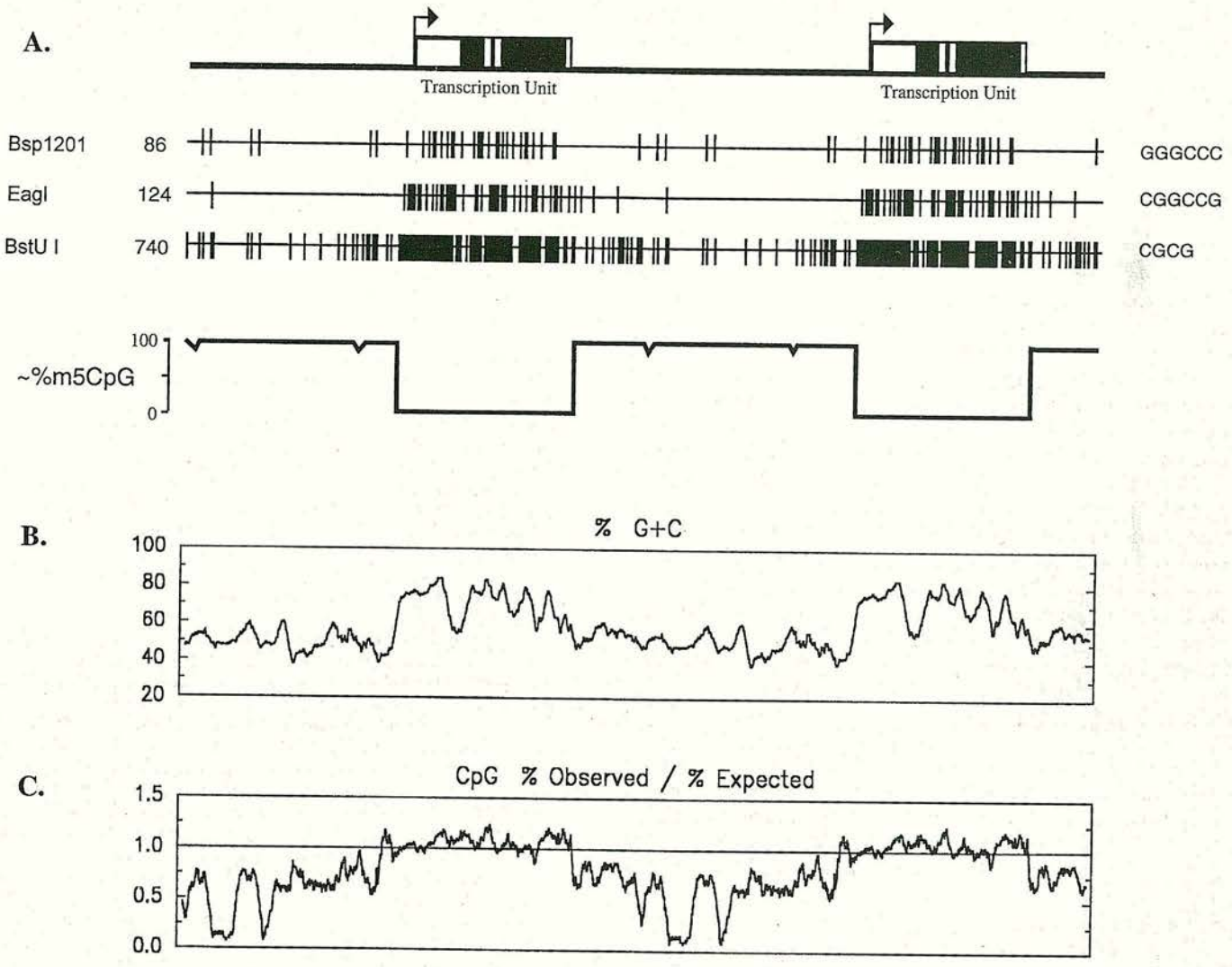


Figure 4. (**A**) Schematic representation of two consecutive rDNA repeat units showing the frequency of GC-rich sites for three restriction enzymes *Bsp*1201, *Eag*I and *Bst*UI. The transcription unit is represented by a box with an arrow indicating the start and direction of transcription. The 18S, 5.8S and 28S genes are shown in black. Approximate methylation levels across the repeat are plotted below. (**B**) A plot of %GC content across the repeat unit based on a window size of 1000 bp and a step of 30 bp. (**C**) A plot of the frequency of CpG across the repeat based on the observed frequency in a 1000 bp window divided by the expected frequency. The step is 30 bp.

Edinburgh). Restriction enzyme digests were performed according to the manufacturer's instructions (New England Biolabs). As a control for complete digestion, an aliquot was removed and incubated with plasmid DNA (200 ng) that contained sites for the relevant enzyme. The plasmid and an undigested control were then visualised on an agarose gel stained with ethidium bromide to check for complete digestion (data not shown). DNA derived from sperm was extracted by addition of 50 mM EDTA pH 8.0, 1% SDS and 500 µg/ml Proteinase K with overnight incubation at room temperature. This was followed by addition of DTT to 50 mM and proteinase K to 100 µg/ml and further incubation overnight at 50°C (13). After phenol/chloroform extraction, the solution was ethanol-precipitated and resuspended in 10 mM Tris pH 7.5, 1 mM EDTA.

Southern blot analysis

In some cases DNA from human blood and sperm was first digested with *Eco*RV, which has no recognition site within the rDNA repeat, in order to reduce the molecular weight of bulk DNA. The DNA was then redigested with *Xma*I or its methylation-sensitive isoschizomer *Sma*I (14). Where DNA was first digested with *Eco*RI or *Hin*dIII followed by digestion with methylation-sensitive restriction enzymes, *Eco*RV was not used. Digested DNA (5 µg/lane) was separated by electrophoresis in 0.8% or 0.5% agarose gels and blotted onto Hybond N+ filters (Amersham) following the manufacturer's protocol. Probes were labelled using the random priming method to incorporate [α-^{32}P]dCTP (15). Hybridisation was performed overnight at

68°C in 0.5 M Na_2HPO_4 (pH 7.2), 7% SDS with 5% powdered milk added to reduce background signal. After probe p1A7 had been denatured, 30 µg Cot-1 DNA (Gibco BRL) was added, competition was necessary due to the presence of a 100 bp Alu consensus sequence within p1A7 (16). Filters were washed three times for 20 min in 0.1× SSC, 0.2% SDS at 68°C (12).

Probes

Base pair co-ordinates used in this study correspond to those published in the GenBank database for the human ribosomal DNA complete repeating unit (accession number U13369). The coordinates of restriction sites shown in the figures were determined using either the GeneJockey III (Biosoft Cambridge) or DNAStar (DNAStar Incoporated) programs. Locations of restriction sites agree with previously published studies (17,18). Probes pHsrDNA5.9 and pHsrDNA7.1 comprise *Eco*RI fragments from coordinates 1 to 5900 bp and from 5900 to 13 000 bp, respectively, cloned into pUC9 (a gift from Rakesh Anand). Probes p1A7 and p1B12 contain *Mse*I fragments from coordinates 17 062 to 17 369 bp and from 35 695 to 36 231 bp cloned into pGem T-vectors (Promega). Probes p1B12 and p1A7 were both obtained during the construction of a library of densely methylated sequences from human blood DNA (G.J.R.Brock, unpublished). Both sequences were used as queries in BLAST (19) searches of the databases maintained by the NCBI Bethesda.

ACKNOWLEDGEMENTS

We thank Vicky Clark for DNA sequencing, Martin Simmen for computing assistance, Joan Davidson and Aileen Greig for technical help and Christine Struthers for secretarial support. We also thank Donald MacLeod, Susan Tweedie, Xinsheng Nan and Sally Cross for comments on the manuscript. GJRB was supported by a studentship from the Wellcome Trust. The work was supported by the Wellcome Trust and the Howard Hughes Medical Institute.

REFERENCES

1. Worton, R.G., Sutherland, J., Sylvester, J.E., Willard, F.H., Bodrug, S., Dube, I., Duff, C., Kean, V., Ray, N.P. and Schmickel, R.D. (1988) Human ribosomal RNA genes: orientation of the tandem array and conservation of the 5′end. *Science*, **239**, 64–68.
2. Gonzalez, L.I., Wu, S., Li, W., Kuo, A.B. and Svlvester, E.J. (1992) Human ribosomal RNA intergenic spacer sequence. *Nucleic Acids Res.*, **20**, 5846–5847.
3. Sakai, K., Ohta, T., Minoshima, S., Kudoh, J., Wang, Y., De Jong, J.P. and Shimizu, N. (1995) Human ribosomal RNA gene cluster: identification of the proximal end containing a novel tandem repeat sequence. *Genomics*, **26**, 521–526.
4. Gruenbaum, Y., Stein, R., Cedar, H. and Razin, A. (1981) Methylation of CpG sequences in eukaryotic DNA. *FEBS Lett.*, **124**, 67–71.
5. Bird, A.P. (1978) Use of restriction enzymes to study eukaryotic DNA methylation. II: The symmetry of methylated sites supports semi-conservative copying of the methylation pattern. *J. Mol. Biol.*, **118**, 48–60.
6. Bickmore, W. and Bird, A.P. (1992) The use of restriction endonculeases to detect and isolate genes from mammalian cells. *Methods Enzymol.*, **216**, 224–244.
7. Bird, A.P. and Taggart, M.H. (1980) Variable patterns of total DNA and rDNA methylation in animals. *Nucleic Acids Res.*, **8**, 1485–1497.
8. Bird, A.P., Taggart, M.H. and Gehring, C.A. (1981) Methylated and unmethylated ribosomal RNA genes in the mouse. *J. Mol. Biol.*, **152**, 1–17.
9. Dawid, I.B., Brown, D.D. and Reader, R.H. (1970) Composition and structure of chromosomal and amplified ribosomal DNAs of Xenopus laevis. *Mol. Biol.*, **51**, 341–360.
10. Bird, A.P. and Southern, E.M. (1978) Use of restriction enzymes to study eukaryotic DNA methylation. I: The methylation pattern in ribosomal DNA from Xenopus laevis. *J. Mol. Biol.*, **118**, 27–47.
11. Cross, S.H., Charlton, J.A., Nan, X. and Bird, A.P. (1994) Purification of CpG islands using a methylated DNA binding column. *Nature Genet.*, **6**, 236–244.
12. Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. pp 196–198.
13. Cross, S.H. (1989) Isolation and Characterisation of Human Telomeres PhD. Thesis, University of Edinburgh.
14. McClelland, M., Nelson, M. and Raschke, E. (1994) Effect of site-specific modification on restriction endonucleases and DNA modification methyltransferases. *Nucleic Acids Res.*, **22**, 3640–3659.
15. Feinberg, A.P. and Vogelstein, B. (1983) A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.*, **132**, 6–13.
16. Sealey, P.G., Whittaker, P.A. and Southern, E.M. (1985) Removal of repeated sequences from hybridisation probes. *Nucleic Acids Res.*, **13**, 1905–1921.
17. Maden, E.B., Dent, L.C., Farrell, E.T., Garde, J., McCallum, F.S. and Wakeman, A.J. (1987) Clones of human ribosomal DNA containing the complete 18S-rRNA and 28S-rRNA genes. Characterization, a detailed map of the human ribosomal transcription unit and diversity among clones. *Biochem. J.*, **246**, 519–527.
18. Gonzalez, I.L. and Sylvester, J.E. (1995) Complete sequence of the 43-kb human ribosomal DNA repeat: analysis of the intergenic spacer. *Genomics*, **27**, 320–328.
19. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
20. Bernardi, G. (1989) The isochore organization of the human genome. *Annu. Rev. Genet.*, **23**, 637–661.
21. Urano, Y., Kominami, R., Mishima, Y. and Muramatsu, M. (1980) The nucleotide sequence of the putative transcription initiation site of a cloned ribosomal RNA gene of the mouse. *Nucleic Acids Res.*, **8**, 6043–6058.
22. Rothblum, I.L., Reddy, R. and Cassidy, B. (1982) Transcription initiation site of rat ribosomal DNA. *Nucleic Acids Res.*, **10**, 7345–7362.
23. Financsek, I., Mizumoto, K., Y., M. and Muramatsu, M. (1982) Human ribosomal RNA gene: Nucleotide sequence of the transcription initiation region and comparison of three mammalian genes. *Proc. Natl. Acad. Sci. USA*, **79**, 3092–3096.
24. Renalier, M.-H., Mazan, S., Joseph, N., Michot, B. and Bachellerie, J.-P. (1989) Structure of the 5′-external transcribed spacer of the human ribosomal RNA gene. *FEBS Lett.*, **249**, 279–284.
25. Gonzalez, I.L., Chambers, C., Gorski, J.L., Stambolian, D., Schmickel, D.R. and Sylvester, E.J. (1990) Sequence and structure correlation of human ribosomal transcribed spacers. *J. Mol. Biol.*, **212**, 27–35.
26. Bird, A.P. (1986) CpG-rich islands and the function of DNA methylation. *Nature*, **321**, 209–213.
27. Bird, A.P., Taggart, M.H., Nicholls, R.D. and Higgs, D.R. (1987) Non-methylated CpG-rich islands at the human alpha-globin locus: implications for evolution of the alpha-globin pseudogene. *EMBO J.*, **6**, 999–1004.
28. Weiss, A., Keshet, I., Razin, A. and Cedar, H. (1996) DNA demethylation in vitro: involvement of RNA. *Cell*, **86**, 709–718.
29. Dante, R., Percy, M., Baldini, A., Markovic, V., Miller, D., Rocchi, M., Niveleau, A. and Miller, O.J. (1992) Methylation of the 5′ flanking sequences of the ribosomal DNA in human cell lines and a human-hamster cell line. *J. Cell Biochem.*, **50**, 357–362.
30. Qu. H.L, Nicoloso, M. and Bachellerie, J.-P. (1991) A sequence dimorphism in a conserved domain of human 28S rRNA. Uneven distribution of variant genes among individuals, Differential expression of HeLa cells. *Nucleic Acids Res.*, **19**, 1015–1019.
31. MacLeod, D., Charlton, J., Mullins, J. and Bird, A.P. (1994) Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. *Genes Dev.*, **8**, 2282–2292.
32. Brandeis, M., Frank, D., Keshet, I., Siegried, Z., Mendelsohn, M., Nemes, A., Temper, V., Razin, A. and Cedar, H. (1994) Sp1 elements protect a CpG island from de novo methylation. *Nature*, **371**, 435–438.
33. Labhart, P. (1994) Negative and positive effects of CpG-methylation on Xenopus ribosomal gene transcription in vitro. *FEBS Lett.*, **356**, 302–306.