INVESTIGATIONS INTO MOUSE TRINUCLEOTIDE REPEAT ARRAYS AND THEIR PUTATIVE ASSOCIATION WITH CpG ISLANDS.

by

Catherine Anne Auchincloss

A thesis submitted for the degree of
Doctor of Philosophy
At the University of Edinburgh

March 2001

Medical Genetics Section
Department of Medical Sciences
Molecular Medicine Centre
University of Edinburgh

This manuscript is dedicated to Kiara,
with whom we should have been allowed more time.

To Matthew and Mum.

DECLARATION

I hereby declare that the laboratory work described in this volume was performed by myself unless otherwise clearly indicated in the text. The thesis manuscript is an original piece of work entirely composed by myself.

ACKNOWLEDGEMENTS

I would like to thank a great many people for the help, support and guidance I have received during the course of my study. Unfortunately there just isn't enough room to mention every worthy contributor to this period of my life, but I will endeavour to repay each in kind. First and foremost I need to thank my mentor and friend Dr Cathy Abbott, without whom this study would never have reached fruition. She has tirelessly provided a subtle blend of encouragement, support, inspiration, advice, interest, patience and above all faith in my abilities. Cathy has proved to be a kind and enduring supervisor, I have no doubt that a succession of students (not always her own) will benefit from her care and attention. I would also like to thank Dr Ian Jackson who fulfilled his supervisory role with enthusiasm, succour and wisdom. I must also say a huge thank you to Doreen Chambers, who taught me all the "tricks of the trade" and was never too busy to offer advice, or help salvage a disastrous experiment. Doreen was an invaluable resource, brimming with knowledge and fun, she will continue to be sorely missed from our laboratory. I am very grateful to Professor Adrian Bird and Dr Donald Macleod (ICMB, Edinburgh University) who generously donated transgenic constructs, and to Dr Sally Cross (MRC Human Genetics Unit) for gifting the mouse CpG island library. Dr Donald Macleod also deserves many thanks for his enthusiastic collaboration and involvement in the transgenic study. I would like to thank Dr Lucy Rowe and Dr Mary Barter (The Jackson Laboratory) for the BSS backcross panel, and for their assistance in handling the data it produced. Thanks are also due to Brendan Doe for leading me through the complicated process of transgenic mouse production and to all the animal house technicians who provided support and helped maintain the mice. I would like to thank the HGMP resource centre for bioinformatics tools and dedicated user support. Thanks go to Dr Jon Warner for helpful discussions on PCR amplification of expanded human trinucleotide repeats and the accurate sizing of repeat containing PCR products. I would also like to thank Dr Kate Wilson and Professor Nick Hastie for help and consideration through a difficult time. A great debt of gratitude is owed to the friendly faces in the laboratory, Patrick, Dawn, Hannah, Emma, Veronica, Rachel, and many, many more, who have shared in the trials and tribulations of the last few years, and kept my spirits up on "bad lab days". The biggest thank you should go to the friends and family who have listened to my gripes, perked me up, patted me on the back and forgiven my many absences. I owe every achievement to my mother, a gentle angel, who has sacrificed much, is devoted to the nurture of others, and who is proud of all I do. Last, but never least I will always be thankful for Matthew, whose love has lightened my life, and who is my reason for everything.

ABSTRACT

The 'dynamic' or 'expansion' mutation of trinucleotide repeats beyond a normal copy number size range is responsible for a growing number of debilitating, neurological disorders. Much work has been invested in modelling this type of repeat instability in the mouse, but as yet no endogenous expansion mutations have been uncovered. A common feature of trinucleotide repeats which result in human disease is their location within or near CpG islands. This PhD study sought to identify potentially expandable mouse trinucleotide repeats, and to explore the putative association between expansion mutation and CpG islands.

A mouse CpG island library was screened for all 10 classes of trinucleotide repeat. Sequence analysis of 89 positive clones revealed that only 32% represented CpG islands, compared to 67% of randomly derived clones. These data implied that trinucleotide repeats are under represented in mouse CpG islands. Where possible PCR primers were designed to amplify these repeats from the mouse genome. The variability of 51 repeat arrays was assessed by their PCR amplification from a panel of sixteen mouse strains. Trinucleotide repeats that exhibited length variability between C57BL/6J and *Mus spretus* (34/51) were mapped using a relevant interspecific backcross panel. These repeats were then screened by PCR as 'candidates' for causing mouse mutant phenotypes mapping to similar genomic regions. Two complex trinucleotide repeat arrays, which mapped to an identical region of chromosome 7, were found to be expanded in frizzy DNA. Further analysis indicated that neither repeat expansion was the underlying mutation responsible for this phenotype.

A transgenic study was also carried out to explore the putative relationship between CpG islands and trinucleotide repeat instability. An expanded human Myotonic Dystrophy repeat was cloned into two similar transgenic constructs, one known to retain its native CpG island properties, and the other mutated to confer non island transgene status. Once transgenic mice had been produced, the methylation of the transgenes was assessed by PCR and Southern blot. The introduction of the trinucleotide repeat and a small amount of flanking DNA appears to have complicated the predicted methylation of these transgenic constructs. The stability of the trinucleotide repeat arrays was followed through several generations of mice by fluorescent PCR analysis. Moderate trinucleotide repeat instability was observed in the majority of transgenic lines, with a strong bias towards repeat contraction and instability through the female germline. This instability did not appear associated with transgene CpG island status, as defined by methylation.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

ALF™	Automatic Laser Fluorescence
Alu	Interspersed repetitive DNA repeat
Aprt	Adenine phosphoribosyltransferase gene
APS	Ammonium persulphate
AR	Androgen receptor
BAC	Bacterial artificial chromosome
bp	Base pairs
BSS	(C57Bl/6J x *Mus spretus*) x *Mus spretus*
CACNA1A	Calcium channel subunit alpha-1A
Camk2a	Calcium/calmodulin-dependent protein kinase-IIα gene
CBL2	Signal transduction protein-2 proto-oncogene
CFU	Colony forming units
cDNA	Complementary DNA
CIP	Calf intestinal alkaline phosphatase
CJD	Creutzfeldt-Jakob disease
cM	CentiMorgan
CMV	Cytomegalovirus
CNS	Central nervous system
COMP	Cartilage oligomeric matrix protein
CREB	Cyclic adenosine 3',5'-monophosphate response element-binding
CSTB	Cystatin-B gene
ddNTP	Dideoxynucleoside triphosphate
DDT	Dithiothreitol
DIRECT	Direct identification of repeat expansion and cloning technique
DM	Myotonic Dystrophy
DMAHP	DM locus-associated homeodomain protein/Six5
D. melanogaster	*Drosophila melanogaster*
DMPK	Myotonic dystrophy protein kinase gene
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dNTPs	Deoxyribonucleoside triphosphates
DRPLA	Dentatorubral-pallidoluysian atrophy
E. coli	*Escherichia coli*

EGF	Epidermal growth factor
EPM1	Unverricht-Lundborg type progressive myoclonus epilepsy
EMBL	European Molecular Biology Laboratory
EST	Expressed sequence tag
FRDA	Friedreich's ataxia
FEN1	Flap structure-specific endonuclease-1
FISH	Fluorescent in situ hybridisation
FRAXA	Fragile X syndrome type A
FRAXE	Fragile X syndrome type E
FRAXF	Fragile site F
FRA11B	Fragile site 11B
FRA16A	Fragile site 16A
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
glyn	Glutamine
GSD	Gerstmann-Straussler disease
HAP	Huntingtin-associated protein
hCG	Human chorionic gonadotrophin
HD	Huntington's disease
HGMP-RC	Human Genome Mapping Project-Resource Centre
HH	Human homologue
HIP	Huntingtin-interacting protein
HMBD	Recombinant methyl binding domain
HRS	Haw River syndrome
HSA	Human skeletal actin gene
H1	Histone 1
ICMB	Institute of Cell and Molecular Biology
IPTG	Isopropylthio-β-D-galactoside
IU	International units
JAX	The Jackson Laboratory
kb	Kilo base
kD	Kilo Dalton
KLHL1	Kelch like-1
l	Litre
LacZ	β-galactosidase gene
LANP	Leucine-rich acidic nuclear protein

LINE	Long interspersed nuclear elements
LYS2	Lysine-2 gene
M	Molar
Mb	Megabase
MBC	Methyl binding column
MBD	Methyl-CpG binding domain
MED	Multiple epiphyseal dysplasia of the Fairbank type
MGD	Mouse Genome Database
ml	Millilitre
mM	Millimolar
MJD	Machado-Joseph disease
MRI	Magnetic resonance imaging
mRNA	Messenger ribonucleic acid
Mx	Interferon-regulated myxovirus resistance gene
µg	Microgram
µl	Microlitre
µM	Micromolar
neo	Neomycin-resistance gene
NES	Nuclear export signal
NI	Intranuclear inclusions (aggregates)
NIX	Nucleotide identity X
NLS	Nuclear localisation signal
NSE	Neural enolase
OD	Optical density
Oligo	Oligonucleotide
OPMD	Oculopharyngeal muscular dystrophy
ORI	Origin of replication
PABP2	Poly(A)-binding protein-2 gene
pcp-2	Purkinje cell promotor-2
PCR	Polymerase chain reaction
Poly(A)	Polyadenylation
polygln	Polyglutamine
PMSG	Pregnant mare serum gonadotrophin
PP2A	Protein phosphatase-2A
PRP	Prion protein

PSACH	Pseudoachondroplasia
PVP	Polyvinylpyrrolidone
QTL	Quantitative trait loci
RAD	Ultra violet excision repair protein
R band	Early replicating region of chromosome
RED	Repeat expansion detection
RI	Recombinant Inbred (strains)
RT	Room temperature
SAP	Shrimp alkaline phosphatase
SAR	Self-association region
SBMA	Spinobulbar muscular atrophy/ Kennedy disease
SCA	Spinocerebellar ataxia
S. cerevisiae	*Saccharomyces cerevisiae*
SDS	Sodium dodecyl sulphate
SEF2-1	SL3-3 enhancer factor 2 gene
SINE	Short interspersed nuclear elements
SSC	Standard saline citrate
STR	Short tandem repeat
SV40	Simian virus-40
TAE	Tris-acetate
TAF	TATA-binding protein associated factor
Taq	*Thermus aquaticus*
TBE	Tris-borate
TBP	TATA-binding protein
TE	Tris-EDTA
TE	Transposable elements
TEMED	NNN'N' tetramethylenediamide
TNE	Tris-NaCl-EDTA
TR	Trinucleotide repeat
tRNA	Transfer ribonucleic acid
TTB	TATA-binding protein
UTR	Untranslated region
v.	Versus
YFH1	Yeast frataxin homologue-1 gene

CHAPTER 1

INTRODUCTION

1 INTRODUCTION

The introduction to this thesis aims to provide a brief overview of the human disorders caused by trinucleotide repeat expansion, the functional consequences of these expansions which lead to disease and the putative mechanisms by which these repeats expand. The text will also outline the approaches currently available for identifying novel trinucleotide repeats. Finally an insight into the animal models contrived and used to study aspects of repeat expansion will be given, with specific reference to the mouse which bears greatest relevance to the aims of this study.

1.1 Unstable repetitive elements in DNA

1.1.1 Interspersed repetitive DNA

Approximately 45% of the drafted human genome comprises interspersed repetitive DNA sequences which are primarily degenerate copies of transposable elements (TEs) [Consortium, 2001]. These unstable elements are dispersed through out the genome and are capable of migration via transposition or retrotransposition. The most abundant class of TEs in the human genome are short interspersed nuclear elements (SINEs) derived from 7SL RNA, the most common of which is the GC rich, ~280 base pair *Alu* (named after the restriction enzyme used to characterise the repeat). The mouse genome also contains SINEs derived from 7SL, but these are predominantly the shorter B1 (~140 bp) and B2 (~190 bp) species, although a small family similar to human *Alus* does exist at a much lower level. Long interspersed nuclear elements (LINEs/L1s) are derived from a reverse transcriptase and are present in a wide variety of lengths ~5-7 kb in many organisms including humans, mice and plants.

1.1.2 Tandemly repeated DNA

Tandemly repeated DNA can be divided into two subsets, highly repetitive satellite DNA and moderately repetitive microsatellites and minisatellites. Highly repetitive satellite DNA is found in all higher eukaryotes, is comprised of a ~200 bp sequence which is tandemly repeated thousands of times and is located in the heterochromatin of centromeres and telomeres. Moderately repeated microsatellites and minisatellites (otherwise known as short

tandem repeats {STRs}) are also found in all eukaryotes but these are dispersed throughout the euchromatin.

Mononucleotide tracts of A or T are the most abundant microsatellites in the mammalian genome, closely followed by dinucleotides, most frequently CA and GA. With each subsequent addition in repeat unit length (i.e. tri- and tetranucleotides) the microsatellites occur at progressively lower frequencies.

STRs can originate from random mutation in DNA sequences leading to the production of short di- or trinucleotide tracts. Once two or more copies of a repeat unit exist in tandem they become susceptible to unequal pairing, crossing over and DNA polymerase slippage during replication. This can lead to increases and decreases in repeat copy number and allelic polymorphism in subsequent generations. The larger repeat loci are more frequently affected by these mutational processes and as a result they exhibit the highest degree of locus size polymorphism.

Microsatellite tracts can manifest in the genome in several different states. A 'perfect' or 'pure' microsatellite consists of a single uninterrupted repeat embedded in non repetitive DNA sequence. Imperfect microsatellites have two or more runs of the same repeat unit interspersed by short stretches of other sequences. The polymorphic properties of imperfect microsatellites are a function of the longest stretch of perfect repeat within the locus. Microsatellites can also exist in an imperfect or compound state where different 'classes' (same size, different base composition) of repeat occur next to, or interrupting each other. Not infrequently these compound repeats can include microsatellites of different 'types' (i.e. tri- and tetranucleotides).

1.1.3 Trinucleotide repeats

Trinucleotide repeats can be made up from any one of ten possible sequence motifs, however each of these can be written or referred to in six different ways (see figure 1.1). Sutherland and Richards proposed that a convention for describing these repeats should be adopted to prevent confusion and suggested that they be listed in 5' to 3' alphabetical order [Sutherland and Richards, 1995]. This convention was not widely accepted, but expanded repeats are now usually referred to in 5' to 3' gene order, CAG or CTG, CCG or GGC and AAG where relevant.

1.1.3.1 Dynamic mutation

Trinucleotide repeats can be subject to an unusual mutational process termed 'dynamic' or 'expansion' mutation [Richards and Sutherland, 1992]. This is where a repeat tract expands beyond a normal, polymorphic size range often conferring a deleterious phenotype. Dynamic mutation is distinct from conventional mutational mechanisms in that: the product of a dynamic mutation has an increased propensity for further expansion than the original DNA sequence, the probability of this mutation is a direct function of a critical repeat copy number and that once the repeat has expanded beyond the normal size range the process continues through subsequent generations. Other types of repeat have recently been reported to exhibit a similar mutational process, an expanded dodecanucleotide repeat is responsible for the Unverricht-Lundborg variety of progressive myoclonus epilepsy (EPM1) [Lafreniere et al., 1997; Lalioti et al., 1997; Virtaneva et al., 1997] and an expanded pentanucleotide repeat causes spinocerebellar ataxia 10 (SCA10) [Matsuura et al., 2000].

1.2 Expanded trinucleotide repeats

Expanded trinucleotide repeats have now been described at nineteen human loci, fifteen of which result in disease phenotypes. The disease causing repeats can be located in various gene regions including exons, an intron, 5' untranslated regions (UTRs) and 3' UTRs (see figure 1.1). When describing these expansions and diseases in further detail, it has proved useful to use these repeat locations within respective genes as sub categories. To date only three classes of trinucleotide repeat AAG, CAG/CTG and CCG have been reported to undergo dynamic mutation (see figure 1.2).

1.2.1 Anticipation

The diseases caused by trinucleotide repeat expansion can be subject to the phenomenon of genetic anticipation. This is where increases in the pathogenic trinucleotide repeat copy number result in increased disease severity and/or disease penetrance and/or a decreased age of disease onset (see figure 1.3).

1.2.2 Other general features of trinucleotide repeat expansions

The trinucleotide repeat expansions which result in human disease share several distinctive features. The critical trinucleotide repeat copy number, beyond which disease phenotypes ensue is similar (approximately 36-40 repeats) at fourteen out of fifteen loci. The only

Figure 1.1 Location and type of trinucleotide repeat expansions in humans

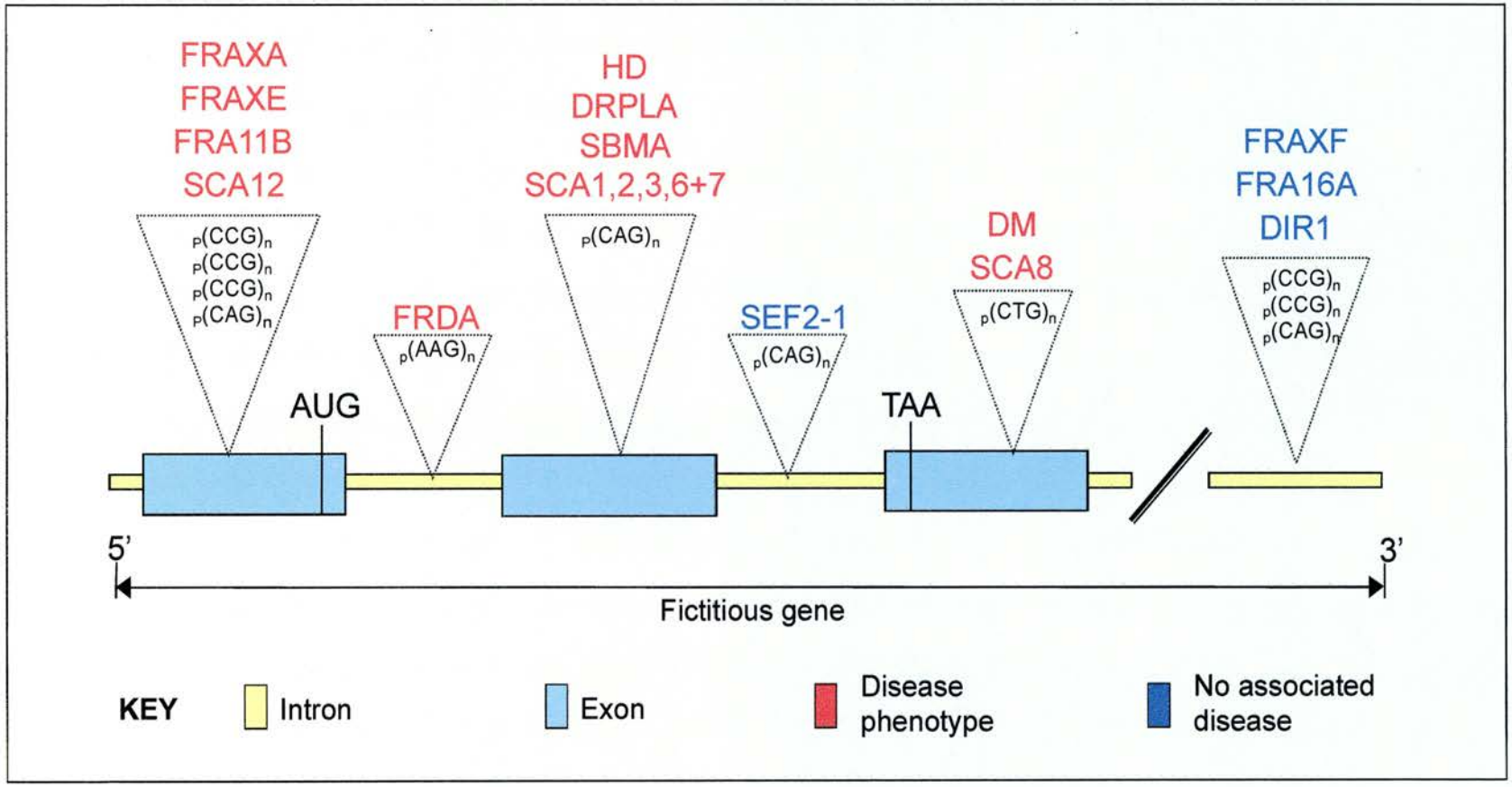


Figure 1.1 Location and type of trinucleotide repeat expansions in humans. This diagram is modified and updated from [Warren, 1996].

Figure 1.2 Trinucleotide repeat sequence motifs

1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
AAC	**AAG**	AAT	ACC	ACG	ACT	**AGC**	AGG	ATC	**CCG**
ACA	**AGA**	ATA	CAC	CGA	TAC	**GCA**	GAG	TCA	**CGC**
CAA	**GAA**	TAA	CCA	GAC	CTA	**CAG**	GGA	CAT	**GCC**
GTT	**TCT**	ATT	GGT	CGT	AGT	**GCT**	CCT	GAT	**CGG**
TGT	**CTT**	TAT	GTG	GTC	TAG	**CTG**	CTC	ATG	**GCG**
TTG	**TTC**	TTA	TGG	TCG	GTA	**TGC**	TCC	TGA	**GGC**

Figure 1.2 Trinucleotide repeat sequence motifs. Shown are the ten classes of trinucleotide repeat and the six possible ways of referring to each class. Highlighted in bold are the three classes of trinucleotide repeat known to undergo expansion.

Figure 1.3 Anticipation in diseases caused by dynamic mutation

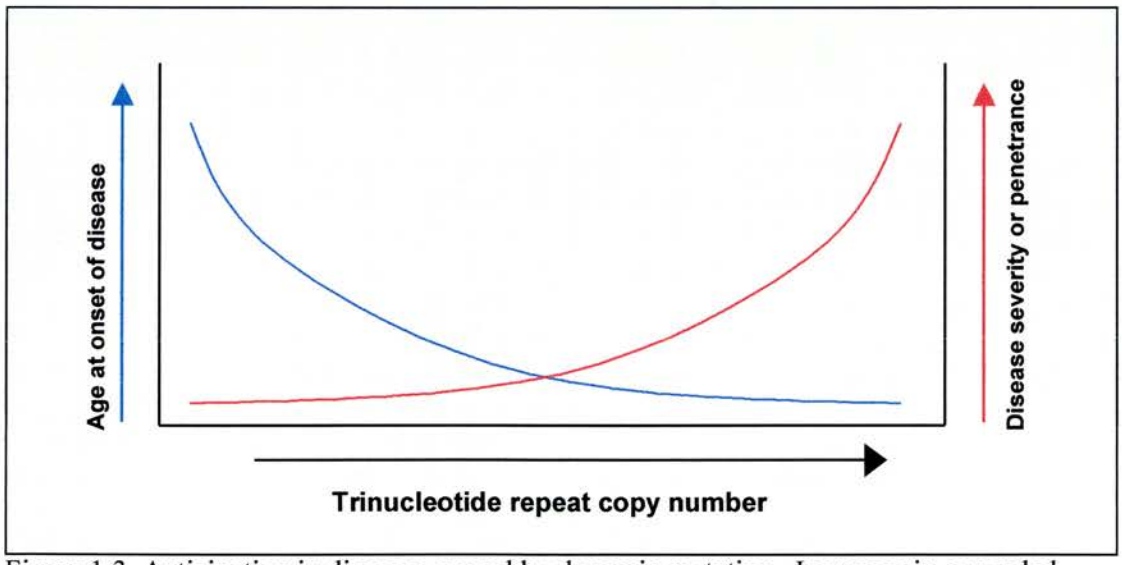


Figure 1.3 Anticipation in diseases caused by dynamic mutation. Increases in expanded repeat copy number are directly related to disease severity and disease penetrance (where penetrance is partial) and inversely correlated with age of disease onset. The figure is reproduced from [Wells and Warren, 1998] and is an overview rather than an accurate curve.

exception to this observation is spinocerebellar ataxia 6 (SCA6), where the phenotype causing copy number is much lower (18 repeats). Recent data implies that SCA8 may also be unusual in that normal alleles containing up to 91 repeats have been reported. A second feature of trinucleotide repeat copy numbers is that the expansions in coding DNA are all considerably smaller than those which occur in noncoding DNA (UTRs and introns).

The trinucleotide repeats causing human disease show distinct preferences for expansion through the germline of one sex or the other. In some instances expansions can only be transmitted by a specific sex, in other cases there is a strong preferential bias and in a few cases the repeats actually contract when transmitted through the other germline. It is interesting that to date all the expansions in coding DNA show increased expansion propensity through the paternal germline and that expansions in noncoding DNA exhibit preferential expansion through the maternal germline. Individuals suffering from many of the disorders caused by repeat expansion also show striking somatic instability of the trinucleotide repeat tracts (for more specific details on these features see section 1.2.3).

1.2.3 Disorders caused by trinucleotide repeat expansion

The human disorders resulting from trinucleotide repeat expansion mutations are summarised in table 1.1. and the repeat copy number size distributions are depicted in figure 1.4.

1.2.3.1 Expansions in 5' UTRs

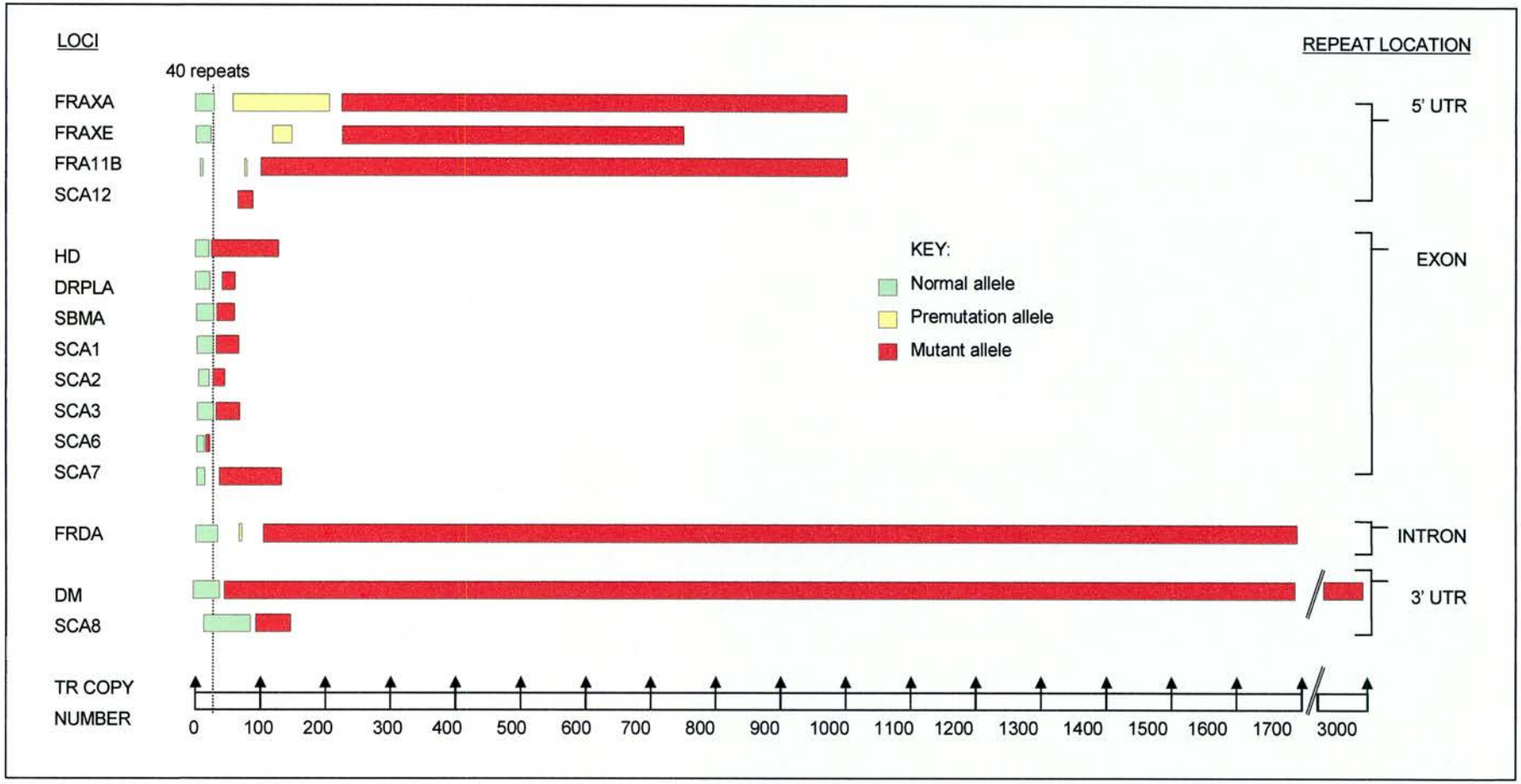
The expansion of trinucleotide repeats in the 5' UTRs of genes causes four disorders, three of which also confer folate-sensitive fragile sites. These sites are cytogenetically visible, non-staining gaps or breaks on metaphase chromosomes, apparent when cells are cultured in the absence of folic acid. The expansion of trinucleotide repeats at these loci exclude nucleosome formation [Wang and Griffith, 1996] which causes hypermethylation of the repeat tract and associated CpG island [Subramanian et al., 1996], resulting in transcriptional silencing of the respective genes. These three expansion loci have distinct 'premutation' alleles [Feng et al., 1995], where the repeat has passed the critical threshold and has an increased propensity for further expansion, but where no disease symptoms or aberrant methylation patterns are observed.

Table 1.1 Summary of expanded human trinucleotide repeats that cause disease

Disease symbol	TR	Affected gene/ location	Gene product	TR site in gene	Inheritance	Disease phenotype	Functional consequences of repeat expansion	Parental exp. bias
FRAXA	CCG	*FMR1* Xq27.3	FMR1: RNA-binding protein	5' UTR	X-linked dominant	Moderate to severe mental retardation, elongated facial features with large ears and macroorchidism	Loss of function -due to transcriptional inactivity caused by CpG island methylation	Maternal
FRAXE	CCG	*FMR2* Xq28 and *FMR3*	FMR2: DNA- binding/ transcription factor	5' UTR	X-linked dominant	Mild to moderate mental retardation	Loss of function -due to transcriptional inactivity caused by CpG island methylation	Maternal
FRA11B	CCG	*CBL2* 11q23.3	CBL2: Proto-oncogene	5' UTR	Autosomal dominant	Severe mental retardation, multiple dysmorphic features and growth retardation	Loss of function -due to transcriptional inactivity caused by CpG island methylation	Not determined
SCA12	CAG	*PPP2R2B* 5q31-33	Ataxin-12: Serine/ threonine phosphatase	5' UTR	Autosomal dominant	Ataxia, dysmetria, dysdiadokinesia, hyperreflexia, nystagmus and dementia.	Loss of function?	Not determined
HD	CAG	*IT15* or *HD* 4p16.3	Huntingtin: Caspase-3 substrate	Exon	Autosomal dominant	Progressive chorea, impairment of cognitive function and dementia	Gain of function-conferred by expanded polyglutamine tract Partial loss of function?	Paternal
DRPLA	CAG	*B37* or *DRPLA* 12p13.31	Atrophin-1	Exon	Autosomal dominant	Myoclonus, epilepsy, choreathetosis, ataxia, impairment of cognitive function and dementia	Gain of function-conferred by expanded polyglutamine tract Partial loss of function?	Paternal
SBMA	CAG	*AR* Xq11-12	AR: Transcription factor	Exon	X-linked recessive	Progressive muscle weakness, atrophy of bulbar muscles and mild androgen insensitivity	Gain of function-conferred by expanded polyglutamine tract Partial loss of function-conferred by expanded polyglutamine tract	Paternal
SCA1	CAG	*SCA1* 6p23	Ataxin-1	Exon	Autosomal dominant	Progressive ataxia, dysarthria, dysmetria and decreased vibration sense	Gain of function-conferred by expanded polyglutamine tract Partial loss of function?	Paternal
SCA2	CAG	*SCA2* 12q24.1	Ataxin-2	Exon	Autosomal dominant	Progressive ataxia and dysarthria	Gain of function-conferred by expanded polyglutamine tract Partial loss of function?	Paternal
SCA3/MJD	CAG	*SCA3* 14q32.1	Ataxin-3	Exon	Autosomal dominant	Progressive ataxia, dystonia and opthalmoplegia	Gain of function-conferred by expanded polyglutamine tract Partial loss of function?	Paternal
SCA6	CAG	*CACNA1A* or *SCA6* 19p13	Ataxin-6: Calcium channel sub unit α_{1A}	Exon	Autosomal dominant	Progressive ataxia, dysarthria, nystagmus and loss of vibratory sense	Gain of function-conferred by expanded polyglutamine tract Partial loss of function?	Not determined
SCA7	CAG	*SCA7* 3p21.1-12	Ataxin-7	Exon	Autosomal dominant	Progressive ataxia, dysarthria and retinal degeneration	Gain of function-conferred by expanded polyglutamine tract Partial loss of function?	Paternal
FRDA	AAG	*X25* or *FRDA* 9q13-21.1	Frataxin: Mitochondrial protein	Intron	Autosomal recessive	Progressive gait and limb ataxia and dysarthria	Partial loss of function -repeat interferes with transcription, suppressing gene expression	Maternal
DM	CTG	*DMPK* 19q13.2-3 and *SIX5* or *DMAHP* and *59* or *DMWD*	DMPK: Serine/threonine kinase DMAHP: homeodomain protein DMWD	3' UTR	Autosomal dominant	Progressive muscle weakness, wasting, myotonia, cardiac conduction defects, cataracts, frontal balding and mental impairment	Partial loss of protein function/RNA gain of function -reduced transcription and defective transcript metabolism	Maternal
SCA8	CTG	*SCA8* 13q21	Ataxin-8: RNA transcript	3' UTR	Autosomal dominant	Ataxia, dysarthria, nystagmus and loss of vibratory sense	Partial loss of function? - reduced transcription and defective transcript metabolism	Maternal

KEY: exp.= expansion, TR= trinucleotide repeat.

Figure 1.4 Distributions of trinucleotide repeat copy number in expansion disorders



1.2.3.1.1 Fragile X type A

Fragile X syndrome type A (FRAXA), the most common cause of hereditary mental retardation [Webb et al., 1986] is caused by a CCG repeat expansion in the 5' UTR of the *FMR1* gene [Kremer et al., 1991; Oberle et al., 1991; Verkerk et al., 1991; Yu et al., 1991] located on Xq27.3. The clinical symptoms of disease include moderate to severe mental retardation, elongated facial features with a prominent jaw, large frontal prominent ears, post pubescent macroorchidism and a high-pitched jocular speech. The CCG repeat size ranges from 6-52 in the normal population, 60-200 in premutation alleles and 230-1000 in affected individuals. The *FMR1* gene codes for an RNA binding protein which is shuttled between the nucleus and cytoplasm and is associated with polysomes [Feng et al., 1997]. The FMR1 protein shares sequence homology with an RNA binding domain of the Nova-2 protein, involved in regulating neuronal RNA metabolism [Lewis et al., 2000]. The FRAXA phenotype is thought to result from perturbed RNA binding by the FMR1 protein. The expanded *FMR1* gene is hypermethylated, transcriptionally silent and exhibits delayed replication timing caused by the altered chromatin structure [Hansen et al., 1997; Hansen et al., 1993; Subramanian et al., 1996; Torchia et al., 1994]. The loss of *FMR1* gene activity can be restored in a cell culture model by demethylating the expanded repeats [Chiurazzi et al., 1998]. The disorder is inherited in an X-linked dominant fashion with penetrance increasing in concert with CCG repeat size. The CCG repeat only expands into the mutant size range on transmission through the female germline [Fu et al., 1991].

1.2.3.1.2 Fragile X type E

Fragile X type E (FRAXE) is a mild form of mental retardation caused by the expansion of a CCG repeat [Knight et al., 1994] in the 5' UTR of the *FMR2* [Gecz et al., 1996; Gu et al., 1996] gene located on Xq28. FRAXE is also inherited in an X-linked dominant fashion with the majority of expansions occurring through the female germline [Knight et al., 1994]. The FRAXE CCG tract contains 7-35 repeats in the normal population, 130-150 repeats in premutation alleles and 230-750 repeats in affected individuals. The *FMR2* gene codes for a nuclear protein with DNA-binding capacity and transcription transactivation potential [Gecz et al., 1997]. In the mammalian brain *FMR2* is predominantly expressed in the neurons of the neocortex, Purkinje cells of the cerebellum and granule cells of the hippocampus [Miller et al., 2000]. The expanded *FMR2* gene is hypermethylated and exhibits delayed replication timing [Subramanian et al., 1996]. A novel gene *FMR3* was recently discovered to be transcribed from the FRAXE CpG island in the opposite direction to *FMR2* [Gecz, 2000].

This gene is also transcriptionally silenced by the FRAXE expansion and could therefore contribute to the FRAXE phenotype.

1.2.3.1.3 Fragile site 11B

Fragile site 11B (FRA11B) is associated with a predisposition to Jacobsen syndrome [Jacobsen et al., 1973; Schinzel et al., 1977], where terminal deletion of 11q23 causes severe mental retardation, multiple dysmorphic features and growth retardation. An expanded CCG repeat in the 5' UTR of the signal transduction-2 (*CBL2*) proto-oncogene triggers this deletion in a subset of affected individuals [Jones et al., 1995]. The expanded repeat and fragile site are inherited in an autosomal dominant fashion. The normal allele contains 11 CCG repeats, the premutation allele contains 80 repeats and expanded alleles contain 100-1000 repeats. The expanded *CBL2* gene is hypermethylated and pathogenicity is thought to result from transcriptional silencing.

1.2.3.1.4 Spinocerebellar ataxia 12

The autosomal dominant disorder spinocerebellar ataxia 12 (SCA12) is caused by an expanded CAG repeat in the 5' UTR of the protein phosphatase (*PPP2R2B*) or *SCA12* gene [Holmes et al., 1999] located on 5q31-3. The disorder is characterised by ataxia, dysmetria, dysdiadokinesia, hyperreflexia, nystagmus and dementia. The CAG is repeated 66-93 times in affected patients. The *SCA12* gene encodes a brain-specific regulatory subunit of the protein phosphatase-2A (PP2A).

1.2.3.2 Expansions in exons/coding regions

The expansion of CAG repeats in coding regions confers a deleterious gain of function to the respective gene products which is caused by an elongated polyglutamine (polygln) tract. This gain of function results in the progressive, selective neuronal cell death underlying each of these disease phenotypes. It has been postulated that the elongated polygln tracts cause excessive protein transglutamination resulting in this neural toxicity. In support of this theory intranuclear inclusions (NIs) or aggregates have been reported in affected brains from all of these disorders. However more recent evidence has revealed that these formations are not required for neurodegeneration [Huynh et al., 2000; Klement et al., 1998; Koyano et al., 1999; Warrick et al., 1998] and may in fact reflect a cellular defence mechanism against this type of neuronal toxicity [Saudou et al., 1998]. Several of the mutated gene products caused by repeat expansion can act as substrates for caspase or apopain which cleaves them into

truncated fragments implicated in cellular toxicity [Ellerby et al., 1999a; Ellerby et al., 1999b; Goldberg et al., 1996b; Wellington et al., 1998]. In a cell culture model expanded polyglns preferentially bind a human TATA- binding protein (TBP)-associated factor ($TAF_{II}130$), inhibiting normal CREB-dependent transcriptional activation thus precipitating cell death [Shimohata et al., 2000]. All the expanded CAG repeat loci that are both transcribed and translated show an increased propensity for expansion when transmitted through the paternal germline (see individual disease synopses below).

1.2.3.2.1 Huntington's disease

Huntington's disease (HD) is caused by a CAG expansion in the first exon of the *IT15* or *HD* gene [Group, 1993] on 4p16.3. The disorder is characterised by progressive chorea, impairment of cognitive function and dementia. The CAG repeat is present 6-36 times in normal alleles and 36-121 times in affected alleles. The *HD* gene encodes huntingtin, a cytoplasmic protein of unknown function which can interact with several other proteins including: huntingtin-interacting protein 1 (HIP1 [Kalchman et al., 1997; Wanker et al., 1997]), huntingtin-interacting protein 2 (HIP2 [Kalchman et al., 1996]) a ubiquitin conjugating enzyme huntingtin-associated protein 1 (HAP1 [Li et al., 1995]), glyceraldehyde-3-phosphate dehydrogenase (GAPDH [Burke et al., 1996]), epidermal growth factor (EGF [Liu et al., 1997]) and calmodulin [Bao et al., 1996]. Huntingtin is also a substrate for both transglutaminase [Kahlem et al., 1998] and caspase-3 [Goldberg et al., 1996b]. This last association has fuelled speculation that HD pathogenesis could be mediated by inappropriate apoptosis. The addition of a nuclear export signal (NES) to mutant huntingtin in a cellular model, revealed that nuclear localisation of huntingtin is a prerequisite for both NI formation and apoptotic neurodegeneration [Saudou et al., 1998]. Work with this cellular system also showed that the suppression of NI formation in the presence of mutant huntingtin results in significantly increased rates of cell death. This implies that NI formation is not required for cell death and may actually reflect a cellular defence mechanism. Huntingtin is widely expressed during development in both the cytoplasm and nucleus of many tissue types, but in adults is predominantly expressed in neurons [De Rooij et al., 1996; Dure et al., 1994; Hoogeveen et al., 1993]. The characteristic HD neurodegeneration emanates from the caudate nucleus and the dorsal putamen [Vonsattel et al., 1985] to progressively include the caudate, putamen, thalamus and cerebral cortex [de la Monte et al., 1988]. HD is inherited in an autosomal dominant fashion and an increased propensity for repeat expansion is observed when the repeat is transmitted through the male germline [Ranen et al., 1995; Trottier et al., 1994].

1.2.3.2.2 Dentatorubral-pallidoluysian atrophy

Dentatorubral-pallidoluysian atrophy (DRPLA) and its allelic variant Haw River syndrome (HRS) are both caused by a CAG expansion in the *B37* or *DRPLA* gene [Burke et al., 1994; Koide et al., 1994; Nagafuchi et al., 1994] on 12p13.31. These disorders are characterised by progressive myoclonus, epilepsy, choreathetosis, ataxia, impairment of cognitive function and dementia. The CAG is repeated 6-35 times in the normal population and 51-88 times in affected individuals. The *DRPLA* gene encodes atrophin-1, a protein of unknown function which interacts with several proteins similar to those reported in association with huntingtin [Burke et al., 1996; Wood et al., 1998] and in addition with an insulin receptor tyrosine kinase substrate [Okamura-Oho et al., 1999]. Like huntingtin the cleavage of atrophin-1 with caspase modulates cytotoxicity [Ellerby et al., 1999a]. Atrophin-1 is predominantly expressed in the cytoplasm of neuronal cells [Yazawa et al., 1995]. The neuropathology of DRPLA patients is characterised by severe atrophy of the dentate nuclei, cerebellar peduncles, globus pallidus externa, subthalamic nuclei and brain stem [Naito and Oyanagi, 1982]. The expanded form of atrophin-1 has been identified in ubiquitinated NIs found in both neurons and glial cells [Hayashi et al., 1998]. The expansion of DRPLA alleles is usually associated with paternal transmission [Koide et al., 1994; Komure et al., 1995; Nagafuchi et al., 1994].

1.2.3.2.3 Spinobulbar muscular atrophy

Spinobulbar muscular atrophy (SBMA) or Kennedy's disease is caused by a CAG expansion in the first exon of the androgen receptor (*AR*) gene [La Spada et al., 1991] on Xq11-12. SBMA is characterised by progressive muscle weakness, atrophy of bulbar muscles and mild androgen insensitivity. The CAG repeat is present 11-33 times in normal alleles and 38-66 times in expanded alleles. The AR is a transcriptional regulatory protein which interacts with a number of proteins (particularly heat shock and GAPDH [Koshy et al., 1996]), binds to ligands, docks with the nuclear transport machinery, associates with the nuclear matrix and interacts with the transcriptional apparatus [Jenster et al., 1991]. The enlarged polygln tract causes a partial functional inhibition of the AR protein which causes the androgen insensitivity. The neurodegenerative aspects of SBMA are assumed to arise through the same type of toxic gain of function seen in the rest of this group of diseases. The AR is highly expressed in many regions of the brain including the hypothalamus, amygdala, medulla, cerebellum and spinal cord [Bingaman et al., 1994; Huang and Harlan, 1994]. The principal areas of neurodegeneration reported in patients are the spinal and bulbar motor

neurons, sensory neurons, spinal cord and brain stem nuclei [Harding et al., 1982]. Like both huntingtin and atrophin-1, cleavage of the expanded AR polygln tract with caspase is a precursor of cytotoxicity [Ellerby et al., 1999b]. The disorder is unusual amongst this group as it is inherited in an X-linked recessive fashion. Increased rates of trinucleotide repeat instability and a bias towards expansion are associated with paternal transmissions [Biancalana et al., 1992; Doyu et al., 1992].

1.2.3.2.4 Spinocerebellar ataxias 1, 2, 3, 6 and 7

This group of spinocerebellar ataxias (SCAs) 1, 2, 3, 6 and 7 have overlapping phenotypes caused by CAG expansions in exons of their respective genes. These disorders are all primarily characterised by progressive ataxia, varying degrees of peripheral neuropathy, pyramidal, extrapyramidal, ocular and cognitive deficits. They are all inherited in an autosomal dominant fashion. The clinical distinction between SCA1, SCA2 and SCA3 is particularly difficult due to the similarity of their phenotypes, but they are now easily resolved with discriminatory polymerase chain reaction (PCR) assays.

SCA1 is caused by a repeat expansion in the *SCA1* gene [Orr et al., 1993] on 6p23. The CAG is repeated 6-39 times in the normal population and 41-81 times in affected individuals. The *SCA1* gene codes for a predicted polypeptide of 87 kD [Banfi et al., 1994] termed ataxin-1. Ataxin-1 has been reported to interact with two proteins GAPDH [Koshy et al., 1996] and the leucine-rich acidic nuclear protein (LANP) which was independently isolated by several research groups and previously assigned a variety of names [Chen et al., 1996; Matilla et al., 1997; Ulitzur et al., 1997; Vaesen et al., 1994]. Ataxin-1 is primarily expressed in the brain and central nervous system but is also present at much lower levels in peripheral tissues, heart, skeletal muscle, liver, pancreas and lung [Servadio et al., 1995a]. The characteristic neuropathology predominantly affects the Purkinje cells of the cerebellum, brainstem neurons, motor neurons and spinal cord fibres [Robitaille et al., 1995]. Ubiquitinated NIs have been identified in SCA1 neurons [Skinner et al., 1997] however the formation of these structures is not a prerequisite for pathogenesis [Klement et al., 1998]. The SCA1 repeat is cryptic in the normal population (interrupted by two ATCs) but pure in affected individuals. Expansion of repeat tracts occur exclusively through the paternal germline [Jodice et al., 1994].

SCA2 is caused by a CAG expansion in the *SCA2* gene [Imbert et al., 1996; Pulst et al., 1996; Sanpei et al., 1996] on 12q24.1, which encodes a 140 kD protein ataxin-2. The CAG

repeat is present 14-31 times in normal alleles and 35-64 times in expanded alleles. Ataxin-2 is a cytoplasmic protein which is widely expressed in the brain (primarily in Purkinje cells and cortical neurons), heart, liver, skeletal muscle and pancreas [Huynh et al., 1999; Imbert et al., 1996; Pulst et al., 1996; Sanpei et al., 1996]. SCA2 neuropathology includes the degeneration of cerebellar Purkinje cells, brainstem neurons, motor neurons and demyelination of spinal cord fibres [Orozco et al., 1989]. The pathogenesis of SCA2 does not require NI formation [Huynh et al., 2000; Koyano et al., 1999]. Large expansions of the SCA2 repeat occur almost exclusively through the male germline [Geschwind et al., 1997; Riess et al., 1997].

SCA3 otherwise known as Machado-Joseph disease (MJD) is caused by a CAG expansion in the *SCA3* gene [Kawaguchi et al., 1994] on 14q32.1. The CAG repeat occurs 12-41 times in the normal population and 35-64 times in affected individuals. The *SCA3* gene encodes ataxin-3, present in both the cytoplasm and nuclei of cells [Paulson et al., 1997; Trottier et al., 1998]. Ataxin-3 can interact with two human homologues of the yeast ultra violet excision repair protein (RAD23 {HHR23A and HHR23B}[Wang et al., 2000]) and is primarily expressed in striatum neurons [Paulson et al., 1997]. SCA3 neurodegeneration affects cerebellar Purkinje cells, brainstem neurons, motor neurons, the subthalamopallidal system and causes demyelination of spinal cord fibres [Takiyama et al., 1994]. Mutant ataxin-3 forms NIs [Evert et al., 1999; Warrick et al., 1998], but these are not required for pathogenesis [Warrick et al., 1998]. Evidence is growing that ataxin-3 misfolding may represent an early step in pathogenesis [Chai et al., 1999; Perez et al., 1999]. The expansion of SCA3 repeats is most prevalent in male germline transmissions [Takiyama et al., 1995]. SCA3 is unusual in that males with similar sized repeats to affected females develop disease symptoms at a significantly earlier age [Kawakami et al., 1995].

SCA6 is caused by a CAG expansion within an exon of the membrane-bound voltage dependent calcium channel subunit alpha-1A (*CACNA1A*) gene [Zhuchenko et al., 1997] or *SCA6* gene on 19p13. SCA6 is unique amongst this group of disorders in that the pathogenic repeat size is below the usual threshold with alleles of 21-26 repeats being pathogenic and normal alleles containing only 7-18 repeats. The expansion of repeats at this locus appear restricted and no anticipation is associated with this disorder. Ataxin-6 is predominantly expressed in Purkinje cells and affected individuals develop non-ubiquitinated aggregates in the cytoplasm of these cells which precede apoptotic cell death [Ishikawa et al., 1999].

SCA7 is caused by a CAG expansion in an exon of the *SCA7* [David et al., 1997; Lindblad et al., 1996; Trottier et al., 1995] gene on 3p12-13. The CAG repeat is present 7-17 times in normal alleles and 38-130 times in affected alleles. The *SCA7* gene encodes a 130 kD protein termed ataxin-7 which has a nuclear localisation signal and is ubiquitously expressed [Kaytor et al., 1999]. Ataxin-7 is proposed as a transcription factor based on the homology of its polyglutamine/polyproline-rich region with similar domains in huntingtin, several homeodomain-containing proteins and other transcription factors [Gerber et al., 1994]. These polyglutamine/polyproline-rich domains are known to be capable of activating transcription *in vitro*. In transgenic mice mutant ataxin-7 accumulates in ubiquitinated NIs which develop in concert with the severe neuronal and photoreceptor degeneration characteristic of SCA7 [Yvert et al., 2000]. The magnitude of SCA7 repeat expansions are considerably larger in paternal germline transmissions [David et al., 1997].

1.2.3.3 Expansion in an intron

1.2.3.3.1 Freidreich's ataxia

Friedreich's ataxia (FRDA) is the most common form of hereditary ataxia and has three unique features amongst the trinucleotide repeat disorders identified to date: it is caused by the expansion of an AAG repeat, the repeat is located within the intron of a gene (*X25* or *FRDA*) and it is inherited in an autosomal recessive fashion [Campuzano et al., 1996]. The disorder is characterised by progressive gait and limb ataxia, a lack of tendon reflexes in the legs, loss of position sense and dysarthria. The AAG repeat occurs 6-34 times in the normal population, 80 times in premutation alleles and 112-1700 times in affected individuals. The *FRDA* gene encodes a mitochondrial protein termed frataxin reputed to be involved in cellular respiration and iron homeostasis [Adamec et al., 2000; Foury and Cazzalini, 1997; Lodi et al., 1999; Wilson and Roof, 1997]. Frataxin is predominantly expressed in the heart, liver, skeletal muscle, pancreas, spinal cord and to a lesser extent in the cerebellum [Campuzano et al., 1996]. The primary sites of FRDA neurodegeneration are the spinocerebellar tracts, dorsal columns, pyramidal tracts, cerebellum and medulla. The expanded AAG repeat tract forms 'sticky DNA' [Sakamoto et al., 1999] which directly interferes with *FRDA* transcription [Bidichandani et al., 1998], reducing the level of *FRDA* gene expression and resulting in frataxin deficiency [Campuzano et al., 1996; Cossee et al., 1997]. Despite being a recessive disorder FRDA exhibits anticipation and expansion of the repeat occurs preferentially through the female germline [Monros et al., 1997; Pianese et al., 1997].

1.2.3.4 Expansions in 3' UTRs

1.2.3.4.1 Myotonic dystrophy

Myotonic dystrophy (DM) is caused by the expansion of a CTG repeat in the 3' UTR of the myotonic dystrophy protein kinase (*DMPK*) gene [Brook et al., 1992], on 19q13.3. The *DMPK* gene encodes a novel serine/threonine phosphatase termed DMPK. DM is a progressive multisystem disorder manifesting variable symptoms of muscle weakness, wasting, myotonia, cardiac conduction defects, cataracts, frontal balding and mental impairment. Congenital cases may present with hypnotic, facial diplegia and severe mental retardation. The CTG repeat is present 5-37 times in the normal population and 50-3000 times in affected individuals. There is data to support several mechanisms by which the CTG expansion results in the disease phenotype and more may yet be elucidated. Transcription of the *DMPK* gene is thought to be repressed by the formation of hyperstable nucleosomes at the expanded CTG repeat tracts, resulting in a partial loss of DMPK function [Amack et al., 1999; Godde et al., 1996; Otten and Tapscott, 1995; Wang et al., 1994; Wang and Griffith, 1995]. Studies on *DMPK* transcription levels initially produced conflicting results [Bhagwati et al., 1996; Fu et al., 1993; Mahadevan et al., 1993; Novelli et al., 1993], however reductions in mature mutant transcripts relative to normal transcripts have now been reported by several groups [Krahe et al., 1995; Wang et al., 1995], suggesting that the CTG expansion also adversely affects transcript metabolism. This is thought to arise through the aberrant binding of proteins to the expanded RNA repeat [Philips et al., 1998]. The complex nature of DM symptoms and early transgenic studies implied that *DMPK* might not be the only gene affected by the CTG repeat expansion [Harris et al., 1996]. Recent experiments have added weight to this hypothesis by revealing that the transcription of two flanking genes is also repressed in an expanded allele-specific manner [Alwazzan et al., 1999; Eriksson et al., 1999; Klesert et al., 1997; Korade-Mirnics et al., 1999; Thornton et al., 1997]. The affected gene located downstream of the trinucleotide repeat is known as the DM locus-associated homeodomain protein (*DMAHP* [Boucher et al., 1995]) gene or *SIX5* and the upstream gene is referred to as *59* or *DMWD*. DM is inherited in an autosomal dominant fashion and large repeat expansions occur almost exclusively through maternal transmission [Harley et al., 1993; Lavedan et al., 1993].

1.2.3.4.2 Spinocerebellar ataxia 8

SCA8 is the product of a CTG repeat expansion in the 3' UTR of a gene of unknown function [Koob et al., 1999] on 13q21. The expandable CTG tract is flanked by a small CTA repeat. The most 5' exon of the *SCA8* gene is also transcribed through the first exon of another gene Kelch like-1 (*KLHL1*), in the opposite orientation [Nemes et al., 2000]. The CTG repeat occurs 15-37 times in the normal population, 37-91 times in a small subset of normal individuals and 100-152 times in affected individuals [Silveira et al., 2000]. This data is not yet well characterised and at least one unaffected individual with 127 repeats has been reported [Worth et al., 2000]. The *SCA8* gene transcript is thought to be an endogenous, antisense RNA that overlaps the transcription and translation start sites as well as the first splice donor sequence of the sense gene. Both these genes are expressed in the cerebellum and thus the pathogenic effect of the CTG expansion may be mediated through one or both of these transcripts. SCA8 is inherited in an autosomal dominant fashion but is distinct from the other spinocerebellar ataxias in that the repeat tract expands more frequently through the female germline [Day et al., 2000; Koob et al., 1999]. Another feature unique to SCA8 alleles is that the expanded repeats can exist in a pure state, or contain one or more CCG, CTA, CTC, CCA, or CTT interruptions [Moseley et al., 2000].

1.2.4 Trinucleotide repeat expansions which confer no phenotype

There are four human trinucleotide repeat expansions which do not elicit disease phenotypes: a CAG expansion in the SL3-3 enhancer factor 2-I (*SEF2-1*) gene [Breschel et al., 1997], a CCG expansion at fragile site F (FRAXF) [Hirst et al., 1993; Parrish et al., 1994], another CCG expansion at fragile site 16A (FRAX16A) [Nancarrow et al., 1994] and a CAG expansion at DIR1 [Ikeuchi et al., 1998; Nakamoto et al., 1997] otherwise known as ERDA. Three of these expansions are not in the vicinity of any known genes which clearly accounts for their lack of effects, the SEF2-1 repeat however is located in the intron of a gene. The differences between this locus in particular and those that cause disease could be key in the understanding of how trinucleotide repeats result in deleterious effects. Conversely it could simply be the nature of this gene which accounts for the lack of an obvious phenotype.

1.2.4.1 SEF2-1

The expansion of a CTG tract within an intron of the *SEF2-1* gene on 18q21.1, was isolated through screening bipolar affective disorder pedigrees but is present in the normal population

and is not associated with any obvious disease phenotype [Breschel et al., 1997]. The *SEF2-1* gene encodes a basic helix-loop-helix DNA-binding protein involved in transcriptional regulation. The CTG repeat exists as stable alleles of 1-37 CTG repeats, unstable alleles of 53-250 CTG repeats and highly unstable expanded alleles of 800-2100 CTG repeats.

1.3 Other diseases associated with repeat sequences

Since the discovery of disorders associated with trinucleotide repeat expansions, other types of repeat have been linked with disease. The expansion of a dodecanucleotide and a pentanucleotide have now been described and it's possible that they share similar mutational mechanisms with the expanded triplet repeats. A second group of smaller 'expansions' and 'contractions' have also been linked with disease (see 1.3.2 for further details). However it is more likely that rather than representing true dynamic mutations these disorders are caused by small duplications or deletions of repeats. The underlying mutational processes involved in these disorders are therefore likely to be different from true repeat expansions.

1.3.1 Expansion of repeats

1.3.1.1 Unverricht-Lundborg progressive myoclonus epilepsy

The Unverricht-Lundborg type of progressive myoclonus epilepsy (EPM1) is most commonly caused by a dodecamer repeat expansion in the 5' UTR of the cystatin-B (*CSTB*) gene on 21q22.3 [Lafreniere et al., 1997; Lalioti et al., 1997; Virtaneva et al., 1997]. EPM1 is characterised by progressive epilepsy, myoclonus, ataxia and dementia, not dissimilar to the SCA phenotypes described in the previous section. EPM1 is inherited in an autosomal recessive fashion and repeats show a preference for expansion when paternally transmitted. The underlying GC rich dodecanucleotide (CCCCGCCCCGCG) is repeated 2-3 times in the normal population, 12-17 times in premutation alleles and 30-75 times in affected individuals. The length of the expanded repeat does not appear to affect disease severity or age of onset. The disease phenotype appears to be caused by a reduction in cystatin-B levels and the targeted disruption of the mouse *Cstb* gene has successfully reproduced all aspects of the human phenotype [Pennacchio et al., 1998]. It is tempting to speculate that the expansion of this GC rich repeat results in the hypermethylation of the region and perhaps an associated CpG island. This could result in the type of transcriptional silencing observed on the expansion of GC rich trinucleotide repeats in the 5' UTRs of the *FMR1* and *FMR2* genes.

1.3.1.2 Spinocerebellar ataxia 10

SCA10 is caused by the expansion of a pentanucleotide repeat (ATTCT) in intron 9 of the *SCA10* gene on 22q13 [Matsuura et al., 2000]. The *SCA10* gene encodes a 475 amino-acid protein of unknown function called ataxin-10, which is widely expressed in human brain. The disorder is inherited in an autosomal dominant fashion and decreasing age of onset is associated with increasing repeat size. The pentanucleotide is repeated 10-22 times in the normal population and is approximately 22.5 kb larger than this in affected individuals. It is possible that this repeat exerts its pathogenic effect by interfering with transcription in a similar way to the FRDA repeat which is also located in an intronic region.

1.3.2 Duplication/deletion of repeats

1.3.2.1 Pseudoachondroplasia/Fairbank multiple epiphyseal dysplasia

Pseudoachondroplasia (PSACH) and its allelic variant multiple epiphyseal dysplasia of the Fairbank type (MED) are both commonly caused by small increases or decreases in a GAC repeat tract [Briggs et al., 1995; Briggs et al., 1998; Delot et al., 1999; Ikegawa et al., 1998]. The GAC repeat which encodes a polyaspartic tract is found in exon 13 of the cartilage oligomeric matrix protein (COMP) gene on 19p13.1. In normal individuals the GAC is tandemly repeated 5 times and in affected patients the repeat has been reported to contract by one triplet, or expand by one or two triplets. These disorders are inherited in an autosomal dominant fashion.

1.3.2.2 Synpolydactyly

Synpolydactyly is an inherited abnormality of the hands and feet and is caused by the duplication of part of a polyalanine stretch in the N terminus of the *HOXD13* gene [Akarsu et al., 1996; Muragaki et al., 1996]. The disorder is inherited in an autosomal dominant fashion and increased alanine tract length is associated with increased penetrance and phenotype severity [Goodman et al., 1997]. The *HOXD13* gene encodes 15 alanines in normal individuals and expanded tracts containing 36, 39 and 45 alanines have been reported in affected individuals. The phenotype is thought to be caused by a deleterious gain of function conferred by the expanded polyalanine tract. A similar duplication within the polyalanine tract of the mouse *Hoxd13* gene causes a homologous phenotype [Johnson et al., 1998]. The mouse gene also encodes 15 alanines in the normal population and a duplicated tract containing 22 alanines was reported in affected mice.

1.3.2.3 Oculopharyngeal muscular dystrophy

Oculopharyngeal muscular dystrophy (OPMD) can be caused by small increases in the number of alanines in a repeat tract found in the N terminus of the poly(A)-binding protein-2 (*PABP2*) gene on 14q11.2-13 [Brais et al., 1998]. The GCG tract which encodes 6 alanines in the normal population can be present as a mutant allele of 7-13 alanines in affected individuals. This repeat is unusual in that size increases can confer a dominant phenotype, modify a dominant phenotype or cause a recessive phenotype. Autosomal dominant OPMD is the result of 8-13 repeats, compound heterozygotes with both 9 and 7 repeats display a more severe phenotype and individuals homozygous for 7 repeats exhibit an autosomal recessive form of the disease. The disorder is characterised by dysphagia and progressive ptosis of the eyelids. The mutated PABP2 protein accumulates as filamentous nuclear inclusions [Tome and Fardeau, 1980; Uyama et al., 1996] and seems most likely to confer a toxic gain of function.

1.3.2.4 Protein infectious agent

Protein infectious agents (prions) lack nucleic acid, but are responsible for a group of neurodegenerative diseases including Creutzfeldt-Jakob disease (CJD), Gerstmann-Straussler disease (GSD), familial fatal insomnia and kuru. Aberrant isoforms of the prion protein (PRP) can be inherited in an autosomal dominant fashion or transmitted by inoculation [Prusiner, 1996]. Pathogenesis results from a conformational change in the normal prion isoform which engenders the formation of insoluble aggregates in the brain. The prion protein gene *PRP* maps to 20pter-p12 and causative mutations include insertions of octapeptide repeats of varying length [Gajdusek, 1991].

1.4 Diseases putatively caused by repeat expansions

The identification of trinucleotide repeat expansion as the molecular basis of the anticipation associated with a number neurodegenerative disorders has fuelled speculation that a similar mechanism might be responsible for any anticipation associated with a broad spectrum of disorders [McInnis, 1996]. The major issue of contention with this hypothesis is that ascertainment and statistical bias can contribute to inaccurate predictions of anticipation. However there is preliminary evidence for anticipation in several psychiatric, neurological, cancers, autoimmune and developmental disorders. Many resources are currently being ploughed into screening the strongest of these contenders for evidence of trinucleotide repeat expansions, particularly the remaining spinocerebellar ataxias, bipolar affective disorder and

schizophrenia. However it should be considered that other factors may play a role in anticipation and that trinucleotide repeat expansion may not be the only molecular mechanism.

1.5 Proposed mechanisms of trinucleotide repeat expansion

Dynamic mutation is thought to be distinct from classical mutational processes such as unequal crossing over, DNA polymerase slippage and gene conversion, as none of these mechanisms can solely account for all the unusual features which include: a critical copy number threshold for expansion, the fact that only three out of ten possible trinucleotide repeat classes appear to have expansion propensity, that large expansions can occur in a single mutational step and the unusual bias towards expansion as opposed to the contraction of repeats.

Perhaps the most salient feature of dynamic mutation is its restriction to the three repeat classes CCG, CAG/CTG and AAG. For this reason much emphasis has been placed on the unusual structural motifs that can be formed by long stretches of these repeats and whether these might be differentially processed during DNA replication and/or repair.

1.5.1 Secondary structures formed by long trinucleotide repeats

Current evidence suggests that GC rich repeats (CCG and CAG) can form stable cruciform (hairpin) structures both *in vitro* and *in vivo* [Chen et al., 1995; Darlow and Leach, 1998a; Darlow and Leach, 1998b; Gacy et al., 1995]. It has recently been established that AAG repeats can also form hairpins and YRY triplex helixes containing non Watson–Crick pairs [Gacy et al., 1998; Suen et al., 1999]. The formation of these secondary structures is dependent on specific repeating nucleotide sequences and a critical threshold repeat length, the longer the sequence the larger and more stable the proposed structures become. These structures have not been found to form or bond so stably within the other classes of trinucleotide repeat, neatly accounting for sequence selectivity of dynamic mutations. Interestingly CTG hairpins are reported to be more stable than CAG hairpins [Gacy et al., 1995; Mitas et al., 1995; Petruska et al., 1996].

The next step in elucidating the mechanism of repeat expansion was to determine whether any of these secondary structures could interfere with normal DNA replication and/or repair.

The three processes which could potentially be affected by these structures to confer instability are DNA slippage, unequal crossing over during recombination and misalignment during excision repair. Of these processes DNA slippage during replication is favoured because the critical repeat length associated with the threshold for expansion corresponds to the size of the Okazaki fragment which is the lagging nascent strand at the replication fork.

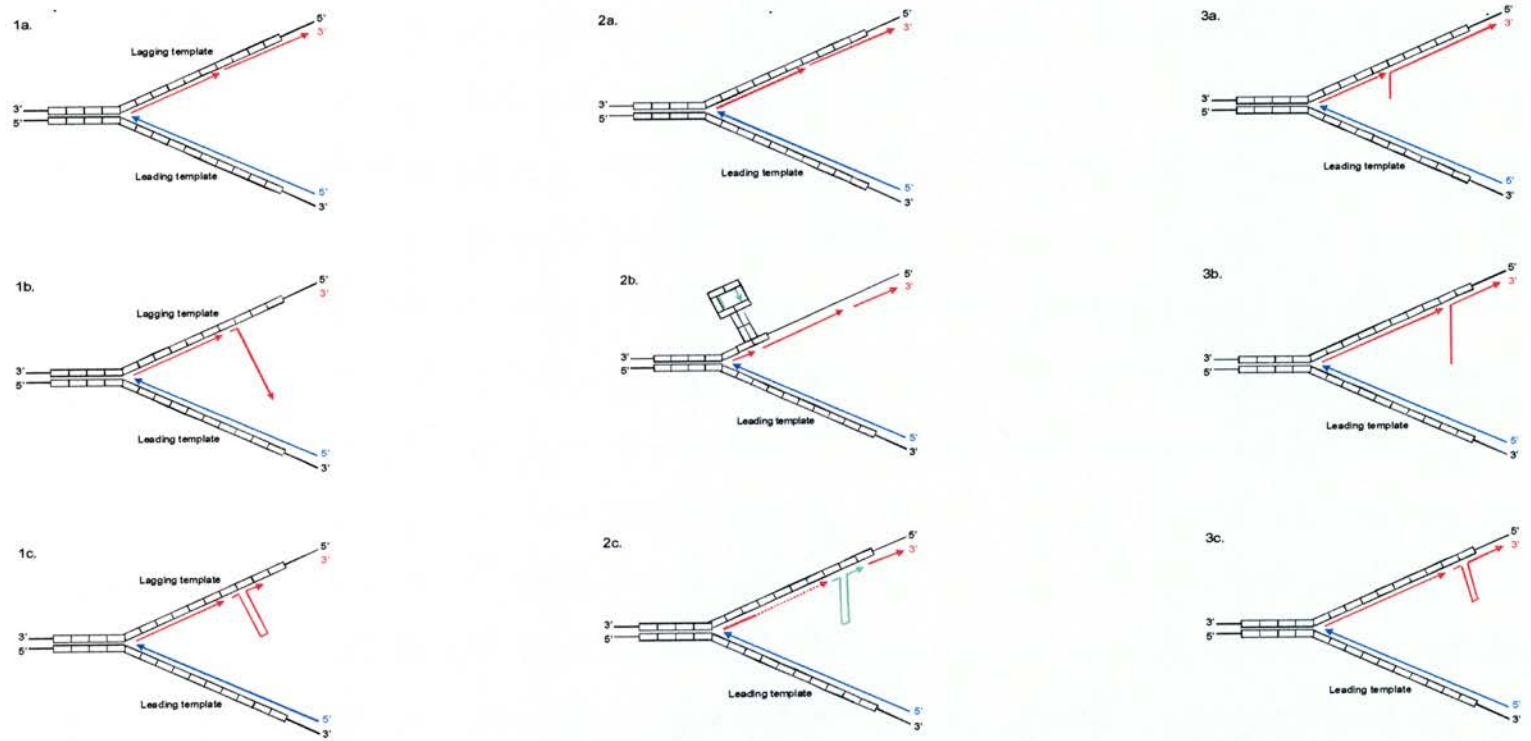
1.5.2 DNA polymerase slippage

The presence of secondary structures in DNA are known to cause DNA polymerase slippage during replication which results in duplications or deletions of genomic DNA [Cooper and Krawczak, 1991; Trinh and Sinden, 1993]. A modified version of this mechanism could potentially account for trinucleotide repeat expansion. In this model the polymerase and its attached nascent strand transiently dissociate from the replication fork complex and secondary structures formed within long triplet repeats stabilise the slippage event by providing equalising energy between the duplex and slipped state. Thus the secondary structures increase the lifetime of the slipped state in a length dependent fashion, allowing large scale slippage or expansion to occur (see figure 1.5 [McMurray, 1995]).

1.5.3 Okazaki fragment slippage

The second mechanism by which secondary structures in DNA could account for repeat expansion is Okazaki fragment slippage [McMurray, 1995; Richards and Sutherland, 1994]. During DNA replication the secondary structures theoretically form in the repeat sequences of the single stranded lagging DNA template, this blocks lagging strand replication at the base of the secondary structure. DNA replication is then primed in the single-stranded region of these structures and as it proceeds along the loop it breaks the hydrogen bonds, relieving the replication block. The loop initiated copy of DNA ends at the preceding Okazaki fragment and becomes ligated there. The normal replication process continuing from the unblocked secondary structure displaces the extra copy of the DNA repeat which bonds to the repetitive template and integrates between the two free ends of adjacent Okazaki fragments (figure 1.5).

Figure 1.5 Putative mechanisms of trinucleotide repeat expansion



1. DNA POLYMERASE SLIPPAGE

1a. Normal DNA replication, repeats shown as boxed regions.

1b. The lagging strand transiently dissociates from the replication fork.

1c. Secondary structures increase the lifetime of the slipped state, allowing a large scale slippage or expansion to occur.

2. OKAZAKI FRAGMENT SLIPPAGE

2a. Normal DNA replication, repeats shown as boxed regions.

2b. Secondary structures form in the lagging strand, normal replication is blocked by the structure, replication is primed within the single stranded loop which breaks secondary structure bonds releasing the block.

2c. Normal replication proceeds and the extra loop initiated repeat is incorporated as an expansion.

3. OKAZAKI FRAGMENT DISPLACEMENT

3a. Normal DNA replication, repeats shown as boxed regions.

3b. 5' end of okazaki fragments displaced by preceding okazaki replication.

3c. Secondary structures form in displaced okazaki flap, these structures are resistant to FEN1 which normally removes flaps, extra repeat containing flaps are incorporated as expansions, or cause replication to stall resulting in double stranded DNA break which when end-joined causes expansion.

1.5.4 Okazaki fragment displacement

The 5' flaps of Okazaki fragments are displaced during normal replication when the DNA polymerase encounters the 5' end of a downstream Okazaki fragment. Usually the flap structure-specific endonuclease-1 (FEN1) removes these flaps prior to Okazaki fragment ligation. However in this model the trinucleotide repeats in the Okazaki flaps form secondary structures which have been shown to be resistant to FEN1 activity [Henricksen et al., 2000]. The unexcised flaps either ligate into the lagging strand to cause expansion directly (see figure 1.5), or cause replication to be stalled resulting in a double stranded DNA break. As this break would occur in a region of repeated DNA, repair by end-joining could easily lead to an expansion [Gordenin et al., 1997]. Replication stalling by long trinucleotide repeats has now been demonstrated *in vivo* and is dependent on repeat orientation, length and purity [Samadashwily et al., 1997].

1.5.5 Replication repair

Studies on trinucleotide repeat instability in *Escherichia coli* (*E. coli*) suggest that small triplet slippages occur frequently during lagging strand replication but are removed by mismatch repair mechanisms [Wells et al., 1998]. However this repair is compromised if the lagging strand DNA contains repeats that can form stable secondary structures and inaccurate repair can ensue [Jaworski et al., 1995; Schmidt et al., 2000]. Experiments in *Saccharomyces cerevisiae* (*S. cerevisiae)* show that inaccurate loops of CCG and CAG/CTG repeats are inefficiently repaired during meiotic recombination [Moore et al., 1999]. It was hypothesised that inefficient mismatch repair may be one of several molecular mechanisms involved in trinucleotide repeat instability. However a recent study on the somatic instability of expanded trinucleotide repeats in mismatch repair deficient mice (*Msh2*-/-) revealed that functional *Msh2* is actually a requirement for instability [Manley et al., 1999].

1.5.6 Proximity to *cis* acting factors

The majority of trinucleotide repeat expansions occur within or near CpG islands [Brock et al., 1999; Gourdon et al., 1997a], certainly the most mutable repeats are all located in these regions. It is as yet unclear whether CpG islands directly influence the expansion potential of trinucleotide repeats, or whether this effect is conferred by other factors commonly associated with CpG islands, or if this observation is purely coincidental. CpG islands are often initiation sites of DNA replication and transcription [Antequera and Bird, 1999;

Delgado et al., 1998] and this is a strong contending factor for influencing repeat instability because although all CpG islands are origins of replication (ORIs), not all ORIs are CpG islands. This relationship could explain the few instances of trinucleotide repeat expansions which do not occur in CpG islands. Origin of replication and transcription effects have been shown to modify triplet repeat instability in model organisms [Bowater et al., 1997; Kang et al., 1995; Miret et al., 1998]. CpG islands are also almost invariably associated with the promoter region and first exon of genes and are distinct from bulk DNA in that they are unmethylated. Therefore differential methylation, affiliated binding proteins and chromatin structure at CpG islands could all potentially modify trinucleotide repeat instability.

1.6 Detection of trinucleotide repeat loci

As the number of human disorders caused by trinucleotide repeat expansion continues to rise, the search for novel repeat sequences has broadened and intensified. Researchers are predominantly interested in identifying repeat sequences associated with genes, as these are the most likely to confer disease phenotypes. However the search has been complicated by the discovery that pathogenic expanded repeats can be located in various genic regions (translated, transcribed but not translated, or untranscribed), that an increasing number of repeat classes can underlie disease phenotypes (AAG, CAG/CTG, CCG) and most recently that other types of expanded repeat (pentanucleotides and dodecanucleotides, see 1.3.1) can also be implicated in disease. To begin with the search for trinucleotide repeats followed conventional methods for identifying novel sequences, more recently however specific methods for detecting repeats or their protein products have been devised. The following section of text briefly describes the most common techniques that have been implemented by researchers in the quest for novel expanded repeat sequences.

1.6.1 cDNA and genome screens

cDNA and genomes can be screened for all classes and sizes of trinucleotide repeats using complementary probes. The repeats and flanking DNA can then be sequenced and studied for polymorphisms or expansions in genomic DNA. The drawback of this approach is that the majority of repeats identified may not be polymorphic, expandable, or associated with any phenotype. Therefore many repeats must be processed with no guarantee of identifying interesting loci. By screening libraries considered the most likely to yield expandable repeats, this task can be somewhat reduced and to date this remains a practical approach for identifying novel repeats.

The majority of expandable trinucleotide repeats initially identified were located in transcribed DNA. For this reason cDNA libraries have been extensively screened with specific oligonucleotide probes, to identify human [Bulle et al., 1997; Jiang et al., 1995; Li et al., 1993; Margolis et al., 1997; Margolis et al., 1995a; Margolis et al., 1995b; Neri et al., 1996; Riggins et al., 1992] and to a lesser extent mouse [Chambers and Abbott, 1996] trinucleotide repeats. The pathogenic trinucleotide repeats underlying DRPLA, SCA3 and SCA6 were all identified using strategy. The disadvantages of this approach are that cDNA libraries tend to underrepresent repeats, intronic repeats can not be identified and 5' repeats may not be present either depending on how the library was constructed. To avoid these limitations the search for large or expanded trinucleotide repeats has now encompassed screening complete genome libraries, single chromosome libraries and contigs spanning regions linked to putative expansion mutation diseases [Fu et al., 1992; Pulst et al., 1996]. Other methods adapted to screen the genome for trinucleotide repeats include genomic differential display [Broude et al., 1997] and fluorescent *in situ* hybridisation (FISH) [Haaf et al., 1996].

1.6.2 Repeat expansion detection

Repeat expansion detection (RED) detection was developed to determine whether a certain class of expanded trinucleotide repeat was the cause of a specific disease phenotype [Schalling et al., 1993]. This was achieved by producing trinucleotide repeat concatomers in a ligation-chain reaction which are unique to affected patient DNA. The major restriction of this technique is that even if the result is positive it cannot be used to locate the causative trinucleotide repeat locus from genomic or cloned DNA.

In a RED reaction DNA is denatured and repeat-specific oligonucleotides allowed to anneal to target sequences. When two or more of these oligonucleotides anneal to adjacent DNA bases in the target DNA they are joined together by a thermostable ligase. The reaction is then cycled through this linear process several hundred times to produce a pool of multimers, where the longest product corresponds to the longest repeat in the template. The multimers are visualised as a ladder of bands by gel electrophoresis followed by automated fluorescence detection (if the oligos contained a fluorescent residue), or by Southern blot analysis and hybridisation to a labelled, repeat-specific probe. Unique, or larger bands which segregate through a disease pedigree with the phenotype indicate that the repeat is causative. This process can also be used to accurately size the repeat alleles when run against suitable controls. RED was successfully used to determine that CAG repeat

expansions were responsible for SCA7 [Lindblad et al., 1996] and SCA12 [Holmes et al., 1999]. This technique has also been used as a selection tool to identify repeat containing DNA fragments when constructing DNA libraries [Koob et al., 1998; Nakamoto et al., 1997].

1.6.3 Direct identification of repeat expansion and cloning technique

The direct identification of repeat expansion and cloning technique (DIRECT) permits the detection of large or expanded repeats from genomic DNA and their subsequent cloning from the relevant genomic fragment [Sanpei et al., 1996].

In the DIRECT approach genomic target DNA is fragmented with suitable enzymes and separated into two aliquots, the first of which is resolved by agarose gel electrophoresis and Southern blotted. The Southern blot is then hybridised to a long (expanded), labelled, repeat-specific probe. The hybridisation conditions employed are so stringent that the probe will only bind to target DNA containing a repeat of equivalent size, or larger. Expanded trinucleotide repeat containing fragments are identified as positively hybridising bands on the blot, against a suitable set of controls. The second aliquot of target DNA is then separated on a fresh gel and the appropriate fragment (estimated against size standards) cut from the gel, electroeluted and cloned into λ-phage vectors. The resulting phage genomic library is then screened using the original probe and hybridisation conditions. This technique was successfully employed to clone the causative gene for SCA2, from a Southern blot of an affected pedigree [Sanpei et al., 1996].

1.6.4 Recognition of expanded polyglutamine tracts via monoclonal antibodies

Monoclonal antibodies have recently been used to identify proteins encoding pathogenic length polygIns and to clone the genes which encode large polygIns. The expansion of polygIns beyond a normal length range is responsible for the largest subset of expanded trinucleotide repeat associated disorders.

Polygln stretches are present in many eukaryotic proteins, particularly transcription factors [Gerber et al., 1994], the largest non-pathogenic tract identified being the human TATA-binding protein (TBP) with a rare upper allele length of 38 glns [Gostout et al., 1993; Imbert

et al., 1994]. A monoclonal antibody (1C2) raised to TBP was found to specifically recognise the polygln stretch [Lescure et al., 1994] and can be used to selectively identify proteins containing expanded polyglns in Western blot analyses. This assay also provides an indication of the polygln length, as the positive signal strength intensifies directly with increased polygln length. This technique was used to identify the pathogenic proteins causing SCA7 and SCA2 [Trottier et al., 1995] from patient lymphoblastoid cell extracts. The antibody has also been used successfully to clone candidate genes encoding large polyglns [Imbert et al., 1996].

Obviously once the entire human genome has been sequenced and compiled, these detection techniques will become redundant for identifying large novel repeat sequences as genome scans are likely to be performed solely by computerised database searches. However several of the techniques described here will remain useful for determining whether an expanded repeat underlies specific disease phenotypes.

1.7 Model systems employed to study trinucleotide repeat expansions and their consequences

Given the inherent limitations of studying this genetic mutation and its consequences in humans, there are obvious benefits to utilising model organisms. In these model systems mechanisms can be determined, mutations induced, phenotypes studied and followed through numerous generations in a relatively short space of time. Since the first trinucleotide repeat expansion was found to underlie a human disorder a combination of these approaches has brought us a greater understanding of this mutational mechanism and its consequences.

The optimum system in which to study trinucleotide repeat expansion would obviously be a genetically controllable higher eukaryote such as the mouse. However bacteria and yeast are suitable for providing relatively rapid answers to many basic mechanistic questions. The genomes of such organisms are relatively easy to manipulate and a host of mutants are already available for screening the effects of DNA replication, repair and recombination processes on trinucleotide repeat stability.

1.7.1 *Escherichia coli*

The simple prokaryotic organism *Escherichia coli* (*E. coli*) has proved invaluable in studying the mechanisms of trinucleotide repeat instability. Used as a model system *E. coli* was key

in determining that secondary structures could be formed *in vivo* by long trinucleotide repeat tracts [Darlow and Leach, 1995; Darlow and Leach, 1998a; Darlow and Leach, 1998b; Ohshima et al., 1996b; Sanpei et al., 1996]. The behaviour of trinucleotide repeat tracts in this organism also revealed that instability occurred in an orientation dependent fashion [Hirst and White, 1998; Ohshima et al., 1996a], a finding which lead to the proposal that instability occurs at the lagging strand during DNA replication. This model has also been used to identify modifiers of repeat instability including repeat length and purity [Hirst and White, 1998; Sanpei et al., 1996], mismatch repair [Jaworski et al., 1995; Pan and Leach, 2000; Parniewski et al., 2000; Schmidt et al., 2000; Schumacher et al., 1998] and DNA polymerase proof-reading mutants [Iyer et al., 2000].

Expanded polyglutamine tracts are cytotoxic to *E. coli* in a length dependent manner which correlates with the toxicity observed in humans [Onodera et al., 1996]. Therefore expression of polyglutamine proteins in *E. coli* may help in further defining the pathogenic effects and be useful in preliminary screens of potentially therapeutic compounds.

1.7.2 *Saccharomyces cerevisiae*

The lower eukaryotic yeast *Saccharomyces cerevisiae* (*S. cerevisiae*) has also been used as a model system to prove that long trinucleotide repeat tracts can form secondary structures which escape DNA repair *in vivo* [Moore et al., 1999]. Likewise, it was utilised to confirm that instability occurred in an orientation dependent fashion [Balakumaran et al., 2000; Freudenreich et al., 1997; Maurer et al., 1996; Miret et al., 1998] increasing the weight of evidence for lagging strand involvement in repeat instability. This system also implicated more specific replication processes in repeat instability including okazaki fragment maturation [Schweitzer and Livingston, 1998] and double stranded DNA break repair [Jankowski et al., 2000].

The putative function of frataxin was elucidated by knocking out the yeast homologue and studying the phenotypic effects [Babcock et al., 1997; Foury, 1999; Foury and Cazzalini, 1997; Koutnikova et al., 1997; Wilson and Roof, 1997] and by identifying an interacting yeast protein [Branda et al., 1999].

Yeasts have also played an indirect role in unravelling the function of the human genes affected by trinucleotide repeat expansions. The yeast two hybrid screen has been extensively used to identify proteins which interact with the normal and expanded gene products including huntingtin [Boutell et al., 1998; Faber et al., 1998; Wanker et al., 1997]

atrophin-1 [Okamura-Oho et al., 1999] and expanded polyglutamines [Shimohata et al., 2000; Waragai et al., 1999].

1.7.3 Cultured mammalian cells

Cultured mammalian cells have been used to study various aspects of trinucleotide repeat instability and mechanisms of expansion pathogenicity. This was the first system which successfully recreated the expansion biased instability of CTG repeats seen in human DM patients [Wohrle et al., 1995]. This system also revealed that the methylation of CCG repeats directly affected their instability [Glaser et al., 1999; Wohrle et al., 1993].

These studies have clarified the effects of expanded repeats on transcription [Beilin et al., 2000; Ohshima et al., 1998], mutant gene products [Abdullah et al., 1998; Matsuyama et al., 1999; Shimohata et al., 2000] and cellular pathology [Usuki and Ishiura, 1998; Usuki et al., 1997].

Much recent investigation was focused on the NI observed in expanded polyglutamine disorders and attempts to reveal whether these aggregates cause neuronal cell toxicity or are a consequence of it. Cellular models were involved in resolving this issue and showing that NIs are not required for neurodegeneration [Kazantsev et al., 1999; Sato et al., 1999a; Yasuda et al., 1999] and may even be a cellular defence mechanism to protect against it [Saudou et al., 1998]. Other cell line studies have shown that cleavage of mutant polyglutamine proteins with caspase, produces truncated fragments which may mediate cellular toxicity [Ellerby et al., 1999a; Ellerby et al., 1999b; Goldberg et al., 1996b; Wellington et al., 1998].

A crucial observation identified in cell culture was that expanded CUG containing transcripts inhibit myogenic cell differentiation [Amack et al., 1999] a pathogenic feature of DM which had proved particularly difficult to model.

1.7.4 Drosophila melanogaster

The fruit fly *Drosophila melanogaster* (*D. melanogaster*) has primarily been used to study trinucleotide repeat expansion pathogenicity via the creation of transgenic lines. The majority of transgenic flies created produce expanded polyglutamine tracts either alone [Marsh et al., 2000] or in partial context of the human genes [Jackson et al., 1998; Warrick et al., 1998]. These flies have successfully modelled the NIs and neural degeneration observed

in humans, implying that this disease pathogenicity is conserved and can therefore be studied in this invertebrate.

It was commonly accepted that *D. melanogaster* DNA is not methylated [Urieli-Shoval et al., 1982] and as trinucleotide repeat expansion itself may be associated with methylation and the pathogenicity of several disorders results from hypermethylation, *D. melanogaster* may not be considered the ideal model in which to study all aspects of this phenomenon. However the methylation status of *D. melanogaster* DNA has currently been thrown into contention as recent experiments using highly sensitive assays have revealed evidence of methylation in this species [Lyko et al., 2000; Warrick et al., 1998].

1.8 The mouse as a model organism in which to study trinucleotide repeat expansion

1.8.1 Endogenous trinucleotide repeat loci

To date no endogenous mouse trinucleotide repeat expansion loci have been identified. The mouse genome has only been screened to a limited extent for large, variable trinucleotide repeat loci, the emphasis being placed on identifying just one class of repeats (CAG/CTG) [Abbott and Chambers, 1994; Chambers and Abbott, 1996; Kim et al., 1997; King et al., 1998].

1.8.2 Mouse homologues of expanded human trinicleotide repeat loci

Of the fifteen human expanded trinucleotide repeats, the mouse homologues of eleven have been described (see table 1.2). The mouse trinucleotide repeats are all much smaller than their human counterparts, in several instances only one or two trinucleotides exist. In the cases where the repeats are translated as polyglutamines, the polyglutamine stretch is often conserved to a certain degree, but is coded for by interrupted CAGs (by CAAs which also encode glutamine). As the length and purity of trinucleotide repeat tracts directly affects instability, it is unlikely that any of the mouse homologues are capable of expansion. Only one of the expanded repeat homologues *Sca7* {Auchincloss unpublished data} was found to show polymorphism in different mouse strains and this tract only varies by 3 repeats in total.

Table 1.2 Mouse homologues of human trinucleotide repeat expansion loci

Disease symbol	Affected gene/ location	Repeat length human normal	Repeat length mouse*	Sequence similarity	Polymorphic	Map	Reference or sequence accession
FRAXA	*FMR1* Xq27.3	CCG_{6-52}	CCG_6	89% similar DNA 94% similar amino acid	-	X chrom, 24.5 cM Conserved synteny with human Xq27	[Ashley et al., 1993; Faust et al., 1992; Laval et al., 1992]
FRAXE	*FMR2* Xq28	CCG_{7-35}	CCG_4, TGTGCA, CCG_4	85% similar DNA 88% similar amino acid	-	X chrom	[Chakrabarti et al., 1998]
HD	*HD* 4p16.3	CAG_{6-36} 6-36 polyglutamines	CAG_2, CAA, CAG_4 7 polyglutamines	86% similar DNA 91% similar amino acid	No	Chrom 5, 20 cM Conserved synteny with human 4p16.3	[Barnes et al., 1994; Nasir et al., 1994]
DRPLA	*DRPLA* 12p13.31	CAG_{6-35} 6-35 polyglutamines	CAG_3 3 polyglutamines	87% similar DNA 92% similar amino acid	-	Chrom 6, 60.2 cM Conserved synteny with human 12p13	[Ansari-Lari et al., 1998; Oyake et al., 1997]
SBMA	*AR* Xq11-12	CAG_{11-33} 11-33 polyglutamines	CAG_3, CAA_2, CAG_3 8 polyglutamines	83% similar DNA 90% similar amino acid	-	X chrom, 36 cM Conserved synteny with human Xq11	[Amar et al., 1988] NM_013476
SCA1	*SCA1* 6p23	CAG_{6-39} 6-39 polyglutamines	CAG_2 2 polyglutamines	86% similar DNA 89% similar amino acid	No	Chrom 13, 28 cM	[Banfi et al., 1996; Servadio et al., 1995b]
SCA2	*SCA2* 12q24.1	CAG_{14-31} 14-31 polyglutamines	CAG_1 1 polyglutamine	89% similar DNA 91% similar amino acid	-	-	[Nechiporuk et al., 1998]
SCA3/MJD	*SCA3* 14q32.1	CAG_{12-41} 12-41 polyglutamines	CAG, CAA, CAG 3 polyglutamines	88% similar DNA 94% similar amino acid	-	Chrom 3, 47 cM	[Schmitt et al., 1997]
SCA6	*SCA6* 19p13	CAG_{7-18} 7-18 polyglutamines	CAG_2 2 polyglutamines	87% similar DNA 92% similar amino acid	-	-	NM_007578
SCA7	*SCA7* 3p21.1-12	CAG_{7-17} 7-17 polyglutamines	CAG_{5-8} 5-8 polyglutamines	85% similar DNA	Yes	Chrom 7 Conserved synteny with human 3p21.1	{Auchincloss, unpublished data}
DM	*DMPK* 19q13.2-3 *SIX5* 19q13.2-3 *DMWD* 19q13.2-3	CTG_{5-37}	CTG_2	80% similar DNA 93% similar amino acid	No	Chrom 7, 4 cM Chrom7, 4 cM Conserved synteny with human 19q	[Cavanna et al., 1990; Mahadevan et al., 1993]

*The repeats described are all from *Mus musculus* species homologues with the exception of the *SCA3* repeat which is from *Rattus norvegicus*

1.8.3 Transgenic studies

1.8.3.1 To model aspects of disease phenotypes

Transgenic mice generated to model the human trinucleotide repeat phenotypes have been widely successful and have helped increase our knowledge of the function of affected genes, the consequences of expansion mutations and the underlying pathogenic mechanisms. Perhaps most notably mouse models have inferred that the DM phenotype has a multigenic origin [Benders et al., 1997; Berul et al., 1999; Klesert et al., 2000; Mankodi et al., 2000; Reddy et al., 1996; Sarkar et al., 2000] and that NIs are not a cause of neurodegeneration, but are more likely a product of it [Klement et al., 1998]. Some of the most useful developments in this area of research have been the knock-outs of homologous mouse genes [Berul et al., 1999; Consortium, 1994; Cossee et al., 2000; Duyao et al., 1995; Jansen et al., 1996; Klesert et al., 2000; Nasir et al., 1995; Reddy et al., 1996; Sarkar et al., 2000; Zeitlin et al., 1995] and the knock-ins of expanded repeats into homologous genes [Lorenzetti et al., 2000; Shelbourne et al., 1999].

The following text will summarise all the mouse models created to study dynamic mutation disorder phenotypes, the most salient features are also presented in table 1.3.

1.8.3.1.1 FRAXA

In FRAXA the disease phenotype is thought to arise from gene silencing, in an attempt to recreate a similar phenotype in the mouse the *FMR1* homologue (*Fmr1)* was knocked out by homologous recombination [Consortium, 1994]. Homozygous knockout mice do not express any *Fmr1* protein and exhibit learning deficits, hyperactivity, increased anxiety and macroorchidism resulting from increased Sertoli cell proliferation [Slegtenhorst-Eegdeman et al., 1998]. This model implies that FRAXA symptoms are due to a loss of FMR1 protein function. The brains of FRAXA humans are known to contain abnormally sized structures (hippocampus and ventricular system) which can be detected by magnetic resonance imaging (MRI). Imaging of knockout mouse brains however did not reveal any parallel observations [Kooy et al., 1999]. Subsequently transgenic mice were created to rescue the knockout phenotype. This construct took the form of a yeast artificial chromosome (YAC) containing the entire *FMR1* gene and 400 kb of flanking sequence, coupled to two auxotrophic marker genes: neomycin-resistance (*neo*) and lysine-2 (*LYS2)* [Peier et al., 2000]. The *LYS2* selectable marker was used to identify successfully retrofitted YACS and the *neo* marker was

Table 1.3 Transgenic mice used to model expanded trinucleotide repeat disease phenotypes

Disease symbol	Promoter	Transgenic construct 5'-3'	Repeat	Relevant human symptoms/expression pattern	Mouse symptom/expression pattern	References
FRAXA	-	Knockout- *Fmr1*	-	Mental retardation, aberrant behaviour (hyperactive, anxious) macroorchidism	Learning deficits, aberrant behaviour (hyperactive, anxious) macroorchidism	[Consortium, 1994]
FRAXA	*FMR1*	Entire *FMR1* and 400 kb flanking sequence (YAC) - *LYS2 and neo* markers	CGG_{20} Normal	Expressed in brain (neurons of hippocampus) and testis (spermatogonia)	Aberrant behaviour (anxious)/overexpressed *FMR1* in brain (neurons of hippocampus) and testis (spermatogonia)	[Peier et al., 2000]
FRAXA	*FMR1*	2.8 kb *FMR1* promoter- *LacZ* marker	CGG_{20} Normal	Expressed in brain (neurons of hippocampus) and testis (spermatogonia)	Expressed in brain (neurons of hippocampus) and testis (spermatogonia)	[Hergersberg et al., 1995]
HD	-	Knockout- *Hdh*	-	Progressive chorea, impaired cognitive function and dementia	Heterozygotes- ↑ motor activity and impaired cognitive function Homozygotes- embryonic lethal/no *Hdh* expression	[Duyao et al., 1995; Nasir et al., 1995; Zeitlin et al., 1995]
HD	*HD*	Entire *HD* (YAC) and >350 kb flanking sequence	CAG_{18} Normal	-	Rescues embryonic lethality/expressed ubiquitously	[Hodgson et al., 1996]
HD	CMV	Entire *HD* cDNA and 1 kb flanking sequence- CMV expression vector	CAG_{44} Expanded	Progressive chorea, impaired cognitive function and dementia	No phenotype/a frameshift mutation introduced into construct prevented *HD* expression	[Goldberg et al., 1996a]
HD	*HD*	Promoter and exon 1 of *HD*	$CAG_{113\text{-}149}$ Expanded	Progressive chorea, impaired cognitive function, ↓ muscle bulk, ↓ brain weight, NIs, neurodegeneration	Progressive chorea, impaired cognitive function, ↓ muscle bulk, ↓ brain weight, NIs, neurodegeneration/expressed ubiquitously	[Davies et al., 1997; Lione et al., 1999; Mangiarini et al., 1996]
HD	CMV	Entire *HD* cDNA- CMV expression vector including SV40 enhancer	CAG_{48+89} Expanded	Progressive chorea, aggression, neurodegeneration, NIs	Progressive motor dysfunction, aggression, neurodegeneration, NIs/expressed ubiquitously	[Reddy et al., 1998]
HD	*Hdh*	Knockin	CAG_{72+80} Expanded	Aggression, neurodegeneration	Aggression, no neurodegeneration/expressed ubiquitously	[Shelbourne et al., 1999]
HD	*Camk2a*	Conditional knockout- Cre/*loxP* (postnatal forebrain and testis)	-	Progressive neurodegeneration	Progressive neurodegeneration and reduced male fertility/expression reduced in postnatal forebrain and testis	[Dragatsis et al., 2000]
DRPLA	*DRPLA*	SV40- *neo* marker- entire *DRPLA* and 23 kb flanking seq (cosmid)	CAG_{76} Expanded	Myoclonus, epilepsy, ataxia and impaired cognitive function/expressed in brain	No phenotype at 12 months/expressed in brain	[Sato et al., 1999b]

KEY: ↑= increased, ↓= decreased.

Table 1.3 Transgenic mice used to model expanded trinucleotide repeat disease phenotypes (continued)

Disease symbol	Promoter	Transgenic construct 5'-3'	Repeat	Relevant human symptom/expression pattern	Mouse symptom/expression pattern	Reference
SBMA	*NSE* *Mx*	*NSE* promoter- entire *AR* cDNA- β-globin tag *Mx* promoter- entire *AR* cDNA- β-globin tag	CAG_{45} Expanded CAG_{45} Expanded	Progressive muscle weakness, atrophy and mild androgen insensitivity	No phenotype/expressed CNS as directed by promoter No phenotype/expressed CNS as directed by promoter	[Bingham et al., 1995]
SBMA	*AR*	Entire *AR* and 450 kb flanking sequence (YAC)	CAG_{45} Expanded	Progressive muscle weakness, atrophy and fasciculations	No phenotype/no expression YAC *AR* sequence fragmented in all lines	[La Spada et al., 1998]
SCA1	*pcp-2*	*pcp-2* promoter- entire *SCA1* cDNA- SV40 poly(A)	CAG_{82} Expanded	Progressive ataxia, dysarthria, dysmetria and neurodegeneration	Progressive ataxia and neurodegeneration/expressed in Purkinje cells as directed by promoter	[Burright et al., 1995; Clark et al., 1997]
SCA1 SCA1	*pcp-2* *pcp-2*	*pcp-2* promoter- entire *SCA1* cDNA with a mutated NLS *pcp-2* promoter- entire *SCA1* cDNA with a deleted SAR	CAG_{82} Expanded CAG_{77} Expanded	Progressive ataxia, dysarthria, dysmetria and neurodegeneration	No phenotype/expressed in Purkinje cells as directed by promoter, localised in the cytoplasm Progressive ataxia and neurodegeneration, no NIs/expressed in Purkinje cells as directed by promoter, localised in nucleus	[Klement et al., 1998]
SCA1	*Sca1*	Knockin- expanded repeat	CAG_{78} Expanded	Progressive ataxia, dysarthria, dysmetria and neurodegeneration	Motor incoordination, no obvious neurodegeneration/expressed in brain	[Lorenzetti et al., 2000]
SCA7	Rhodopsin *pcp-2*	Rhodopsin promoter- entire *SCA7* cDNA, 3' UTR– SV40 poly(A) *Pcp-2* promoter- entire *SCA7* cDNA, 3' UTR- SV40 poly(A)	CAG_{90} Expanded CAG_{90} Expanded	Progressive ataxia, dysarthria, retinal degeneration and neurodegeneration	Progressive retinal degeneration, NIs/overexpressed in photoreceptors Progressive ataxia, NIs/overexpressed in Purkinje cells	[Yvert et al., 2000]
FRDA	-	Knockout- *Frda*	-	Progressive gait and limb ataxia and dysarthria	Early embryonic lethality without iron accumulation	[Cossee et al., 2000]
DM	-	Knockout- *Dmpk*	-	Progressive muscle weakness, wasting, myotonia, cardiac conduction defects, cataracts and mental impairment	Progressive muscle weakness and cardiac conduction defects/no *Dmpk* expression	[Berul et al., 1999; Jansen et al., 1996; Reddy et al., 1996]
DM	*DMPK*	Entire *DMPK* and flanking sequence (14 kb cosmid)	CTG_{20} Normal	Progressive muscle weakness, wasting, myotonia, cardiac conduction defects, cataracts and mental impairment	Hypertrophic cardiomyopathy and increased neonatal mortality/overexpressed *DMPK*	[Jansen et al., 1996]
DM	-	Knockout- *Dmahp*	-	Progressive muscle weakness, wasting, myotonia, cardiac conduction defects, cataracts and mental impairment	Cataracts/no *Dmahp* expression	[Klesert et al., 2000; Sarkar et al., 2000]

included for selection purposes in potential cell culture experiments. The resulting mice exhibited increased anxiety and expressed FMR1 protein in a cell and tissue specific manner at 10-15 times the normal human levels. When crossed with the knockout mice the YAC transgene completely rescued the learning deficits and macroorchidism but intriguingly compounded the anxious behaviour.

1.8.3.1.2 HD

To investigate the normal function of huntingtin, three research groups independently inactivated the murine homologue *Hdh* by homologous recombination [Duyao et al., 1995; Nasir et al., 1995; Zeitlin et al., 1995]. Homozygosity proved to be embryonic lethal, with readsorption by day 8.5. Afflicted embryos display abnormal gastrulation and do not proceed to the formation of somites or to organogenesis. Interestingly these embryos show increased apoptotic cell death in regions where *Hdh* would normally be expressed, suggesting that huntingtin may normally be involved in repressing an apoptotic pathway [Zeitlin et al., 1995]. Heterozygous mice were indistinguishable from wildtype in two studies, but showed increased motor activity and impaired cognitive function in the third. *Hdh* is therefore considered critical in early embryonic development, but as inactivation does not mimic adult HD neuropathology, the disease is most likely conferred by a toxic gain of function. To rescue the embryonic lethality of null *Hdh* homozygotes, transgenic mice were generated from YAC constructs containing the entire HD and more than 350 kb of flanking sequence [Hodgson et al., 1996]. These mice expressed human huntingtin in a tissue specific manner and were bred with heterozygous *Hdh* knockout mice to obtain homozygous null *Hdh* mice expressing the YAC. The YAC construct successfully rescued the embryonic lethality.

In the first attempt to model the HD phenotype, mice were generated from a construct containing the full length *HD* cDNA with 316 bp of 5' UTR and 619 bp of 3' flanking sequence, in a cytomegalovirus (CMV) expression vector [Goldberg et al., 1996a]. The construct was created by end ligation of multiple cDNA clones coupled to the expression vector. Unfortunately during this process an undetected frameshift mutation was introduced which prevented the expression of the transgene. However as the transgenic mice were normal this experiment implies that translation of this expanded trinucleotide repeat is required for HD pathogenesis.

In a second transgenic study mice were generated from a genomic *HD* fragment containing the native promoter and first exon of the gene with an expanded $CAG_{\sim 130}$ repeat [Mangiarini et al., 1996]. Surprisingly this fragment of the *HD* gene was sufficient to cause progressive chorea, reduced muscle bulk, a 20% reduction in brain weight and epileptic seizures in transgenic animals. Further studies of these mice revealed a progressive impairment of cognitive function [Lione et al., 1999], NIs of amyloid-like protein aggregates in the brain [Davies et al., 1997; Scherzinger et al., 1997] and evidence of nonapoptotic neurodegeneration [Turmaine et al., 2000].

A further group also produced transgenic mice with expanded HD repeats ($CAG_{48 \text{ and } 89}$) by end ligation of *HD* cDNA PCR products which were cloned into a CMV expression vector with a simian virus-40 (SV40) enhancer [Reddy et al., 1998]. These mice manifest progressive behavioural abnormalities, locomoter dysfunction and selective neuronal cell loss. Homozygous individuals exhibited earlier symptom onset at approximately eight weeks prior to heterozygous mice. The mutant huntingtin was found to be ubiquitously expressed at higher than normal human levels.

To assess the effects of expanded repeats in genomic context, $CAG_{70 \text{ and } 82}$ were knocked-in to the murine *Hdh* gene by homologous recombination [Shelbourne et al., 1999]. These mice express the mutant huntingtin ubiquitously but at lower than normal levels and exhibit aggressive behaviour, which is sometimes apparent in HD patients.

To assess the role of huntingtin in adult mice, Cre/*loxP* conditional mutants lacking the *Hdh* promoter and first exon were created [Dragatsis et al., 2000]. A calcium/calmodulin-dependent protein kinase-IIα (*Camk2a*) promoter was employed to express Cre in the adult mouse forebrain. Expression of mutant huntingtin was thus reduced in the forebrain and testis of adult transgenic mice. These mice display a progressive degenerative neuronal phenotype and reduced male fertility. This experiment indicates that components of the HD phenotype may be due to a partial loss of *HD* function.

1.8.3.1.3 DRPLA

Transgenic mice were created from a construct containing the entire *DRPLA* gene, with an expanded repeat tract (CAG_{78}) and 23 kb of flanking sequence assembled in a cosmid vector [Sato et al., 1999b]. Analysis of mice from three lines, at up to 12 months of age revealed no phenotype despite high levels of transgene expression in the brain.

1.8.3.1.4 SBMA

Transgenic mice were produced from two constructs, both of which contained the entire *AR* cDNA with an expanded repeat (CAG_{45}) coupled to a segment of mouse β-globin 3' UTR. The first construct was driven by a neural enolase (*NSE*) promoter and the second by an interferon-regulated myxovirus resistance (*Mx*) promoter [Bingham et al., 1995]. These constructs were both expressed at low levels in the central nervous system (CNS) of transgenic mice as directed by the promoters, but no obvious phenotypes were reported.

In a separate endeavour to model SBMA, transgenic mice were generated from a YAC containing the entire *AR* with an expanded repeat (CAG_{45}) and 450 kb of flanking sequence [La Spada et al., 1998]. Unfortunately no expression of human *AR* was detected in transgenic mice, nor was any discernible phenotype observed. The lack of expression was most probably due to the fragmentation of the *AR* gene in two transgenic lines and the disruption of the 3' UTR sequence in the third.

1.8.3.1.5 SCA1

To create a model for SCA1, transgenic mice were produced from a construct containing a Purkinje cell-specific-2 (*pcp-2*) promoter, the entire *SCA1* cDNA with an expanded repeat (CAG_{82}) and an SV40 polyadenylation signal (poly(A)) [Burright et al., 1995]. These transgenic mice developed ataxia in concert with Purkinje cell degeneration which implies that prior cellular dysfunction must underlie *SCA1* pathogenesis [Clark et al., 1997]. To determine whether nuclear localisation of mutant ataxin-1 was required for pathogenesis this group modified their original construct by mutating the nuclear localisation signal (NLS) and produced more transgenic mice. These mice manifest no symptoms demonstrating that nuclear localisation of mutant ataxin-1 is a prerequisite for *SCA1* pathogenesis. The group then mutated their construct a second time by destroying the self-associating or aggregating region (SAR) of mutant ataxin-1 and generated a third group of transgenic mice. These mice showed the characteristic ataxia and Purkinje degeneration associated with *SCA1* without producing NIs. Thus demonstrating that it is the nuclear localisation and not the aggregation of mutant ataxin-1 which mediates cell toxicity [Klement et al., 1998].

In an alternative approach to modelling *SCA1*, an expanded repeat (CAG_{78}) tract was knocked into the mouse *Sca1* gene by homologous recombination [Lorenzetti et al., 2000]. Mutant ataxin-1 was expressed throughout the brain but heterozygous mice displayed no

discernible phenotype. Homozygous individuals exhibited mild motorincoordination without any evidence of neurodegeneration, at up to 18 months of age.

1.8.3.1.6 SCA7

Transgenic mice were generated from two constructs, both containing the entire *SCA7* cDNA with an expanded repeat (CAG_{90}) and 720 bp of 3' UTR coupled to an SV40 poly(A). The first construct was driven by a rhodopsin promoter and the second by a *pcp-2* promoter [Yvert et al., 2000]. The first construct overexpressed mutant ataxin-7 in the photoreceptors of transgenic mice eliciting progressive retinal degeneration. The second construct overexpressed mutant ataxin-7 in the Purkinje cells of transgenic mice which resulted in a progressive ataxic phenotype. In both types of cell expressing mutant ataxin-7 the N-terminal fragment of the protein was found to accumulate in ubiquinated NIs and recruit a distinct group of chaperone/proteasome subunits. This observation implicates proteolytic cleavage of mutant ataxin-7 in *SCA7* pathogenesis.

1.8.3.1.7 FRDA

To determine the function of the *FRDA* gene, the mouse homologue *Frda* was inactivated by homologous recombination [Cossee et al., 2000]. Null homozygotes die a few days after implantation, demonstrating that *Frda* is crucial during early embryonic development. This mouse phenotype is far more severe than that of affected FRDA humans which suggests that humans either maintain some residual frataxin function or that the FRDA symptoms reflect more than just a simple loss of gene function. Interestingly there was no evidence of iron accumulation in *Frda-/-* mouse embryos prior to their readsorption. This was surprising as the yeast *FRDA* homologue-1 (*YFH1*) knock-out exhibits mitochondrial iron accumulation [Babcock et al., 1997; Foury, 1999; Foury and Cazzalini, 1997; Koutnikova et al., 1997; Wilson and Roof, 1997] as do human FRDA tissues [Lamarche et al., 1993; Rotig et al., 1997].

1.8.3.1.8 DM

DM was originally thought to be due to just the reduced transcription of the *DMPK* gene which contains the expanded repeat. To investigate the role of DMPK in DM pathogenesis the mouse *Dmpk* gene was knocked out by homologous recombination [Reddy et al., 1996]. Expression of *Dmpk* was obliterated in resulting mice and homozygous individuals display progressive muscle weakness and cardiac conduction abnormalities [Benders et al., 1997;

Berul et al., 1999]. In depth analysis of knockout skeletal myocytes revealed a deficiency in sodium channel gating [Mounsey et al., 2000]. This first transgenic study elevated suspicions that the complicated DM phenotype reflects multigenic trinucleotide repeat expansion effects.

A second research group also knocked out *Dmpk* and reported a similar phenotype in homozygous mice [Jansen et al., 1996]. This team also produced transgenic lines overexpressing the entire *DMPK* gene and flanking sequence with a normal repeat (CTG_{20}) tract. These transgenic mice manifest hypertrophic cardiomyopathy and increased neonatal mortality

Research suggested that the trinucleotide repeat expansion in *DMPK* also reduced the transcription of an adjacent gene *DMAHP*, otherwise known as *SIX5* [Klesert et al., 1997; Thornton et al., 1997]. To determine whether *DMAHP* deficiency contributes to the DM phenotype the mouse homologue *Six5* was knocked out independently by two research groups [Klesert et al., 2000; Sarkar et al., 2000]. Both heterozygous and homozygous knockout mice have an increased propensity for cataract development, implying that DM does represent a multigenic disorder.

At this point in time all the mice generated to model the DM phenotype had failed to replicate the key disease characteristics of myotonia and myopathy. Research had demonstrated that mutant *DMPK* and *DMAHP* transcripts are aberrantly retained in the nucleus of cells [Davis et al., 1997; Taneja et al., 1995]. To determine whether expanded mRNA transcripts and their altered nuclear localisation were implicated in the DM phenotype a mouse model expressing a mutant CUG was generated [Mankodi et al., 2000]. The transgenic construct contained an expanded repeat (CTG_{250}) that was out of DM gene context, but aimed to be expressed in a similar manner in the 3' UTR of the human skeletal actin (*HSA*) gene. The transgenic mice expressed the mutant transcripts in skeletal muscle and the expanded mRNAs were found to be abnormally retained in discrete nuclear foci. These mice developed progressive myotonia and myopathy, thus implying that the expression of expanded CUG transcripts, even out of context is sufficient to confer a myotonic phenotype. The only similarity between the mutant *DMPK* gene and the transgenic *HSA* construct was the expanded CUG repeat, which suggests that this sequence is sufficient to confer nuclear retention of mature mRNA transcripts. Therefore an RNA gain of function is implicated in DM pathogenesis.

1.8.3.2 To model trinucleotide repeat instability

Transgenic mice which model expanded trinucleotide repeat phenotypes were rapidly produced with varying degrees of success as described in the previous section of text. However the extent of transmissional repeat instability observed in humans has proved a far more challenging phenomenon to reproduce. Often attempts have resulted in little or no instability, lower rates of instability, a smaller magnitude of size changes and mutational bias towards contraction rather than expansion. In fact only the recent study of Seznec et al [Seznec et al., 2000] where a massive DM repeat ($CTG_{>300}$) embedded in 300 kb of contextual sequence was used as a construct, has come anywhere close to replicating this instability. As yet even they have not reported any large, single-step expansions which in humans would frequently take this size of repeat into the congenitally affected size range.

A few general observations about these models for repeat instability can be made (for specific details see the models described below). Somatic repeat instability is only reported in transgenic mice also showing transmissional instability. When displayed both transmissional and somatic repeat instability increase (in both rate and size range when noted) with the age of the mouse. The level of somatic instability exhibited by transgenes does not appear correlated with highly proliferative cells or tissues as might have been predicted. Finally that when repeat instability is described, a bias for an elevated rate or magnitude of change is often associated with one parental sex.

Details of all transgenic mouse models created to study repeat instability are described in the following text and summarised in table 1.4.

1.8.3.2.1 FRAXA

In humans FRAXA premutation alleles ($CGG_{50\text{-}100}$) expand through the maternal germline into the pathogenic size range in almost 100% of transmissions. Three separate transgenic experiments have attempted to replicate this repeat instability in mice [Bontekoe et al., 1997; Lavedan et al., 1998; Lavedan et al., 1997]. The constructs generated each contained the native *FMR1* promoter and a premutation repeat ($CCG_{76,\ 81,\ 88\ and\ 120}$) region. In two of these studies the *FMR1* sequence was coupled to a β-galactosidase (*LacZ*) reporter and in the third to a *neo* selection gene. However no repeat instability was observed in transgenic mice generated from any of these constructs in a total of 691 progeny that were analysed.

Table 1.4 Transgenic mice used to model expanded trinucleotide repeat instability

Disease symbol	Promoter	Transgenic construct 5'-3'	Repeat size	n Of mice	% of ♀ transmissions with a change in repeat size/ size of changes (size range)	% of ♂ transmissions with a change in repeat size/ size of changes (size range)	Somatic changes	Notes	Reference
FRAXA	*FMR1*	2.8 kb of *FMR1* promoter and repeat- *LacZ* reporter	CGG_{81} Premutant	263	0%	0%	No	-	[Bontekoe et al., 1997]
FRAXA	-	ORI- *neo* marker- 264 bp of *FMR1* repeat region	CGG_{88} Premutant	342	0%	0%	No	-	[Lavedan et al., 1997]
FRAXA	*FMR1*	557 bp of *FMR1* promoter and repeat- *LacZ* reporter	CGG_{76} CGG_{120} Premutant	24 62	0% 0%	0% 0%	-	-	[Lavedan et al., 1998]
HD	CMV	Entire *HD* cDNA and 1 kb flanking sequence- CMV expression vector	CAG_{44} Expanded	65	<1% / +1 (1)	<1% / +1 (1)	-	A frameshift mutation prevented *HD* expression	[Goldberg et al., 1996a]
HD	*HD*	Promoter and exon 1 of *HD*	$CAG_{113\text{-}149}$	363	?%/ -2 to +1 (3) n=32	?% / -3 to +12 (15) n=331	Yes	↑ in transmission changes with ↑ age Gender of embryo affects instability	[Kovtun et al., 2000; Mangiarini et al., 1997]
HD	CMV	Entire *HD* cDNA- CMV expression vector, SV40 enhancer	CAG_{48} CAG_{89} Expanded	26 24	0% 0%	0% 0%	-	-	[Reddy et al., 1998]
HD	-	3.6 kb of *HD*, exon 1 5' of ATG	CAG_{48} Expanded	102	2%	2%	-	-	[Wheeler et al., 1999]
HD	*Hdh*	Knockin- expanded repeat	CAG_{72+80} Expanded	219	27% / -4 to +2 (6)	22% / +1 to +8 (8)	Yes	↑ in somatic changes with ↑ age	[Kennedy and Shelbourne, 2000; Shelbourne et al., 1999]

KEY: ↑= increase, n= number.

Table 1.4 Transgenic mice used to model expanded trinucleotide repeat instability (continued)

Disease symbol	Promoter	Transgenic construct 5'-3'	Repeat size	n of mice	% of ♀ transmissions with a change in repeat size/ size of changes (size range)	% of ♂ transmissions with a change in repeat size/ size of changes (size range)	Somatic changes	Notes	Reference
HD	*Hdh*	Knockin- expanded repeat	CAG_{48}	155	4% / -1 to -2 (1) n=47	4% / -1 to +1 (2) n=108	Yes	↑ in somatic changes with ↑ age	[Wheeler et al., 1999]
			CAG_{90}	84	69%/ -6 to +3 (9) n=42	41%/ -1 to +1 (2) n=42			
			CAG_{109} Expanded	95	87%/ -8 to +2 (10) n=31	66% / -2 to +2 (4) n=64			
DRPLA	*DRPLA*	SV40- *neo* marker- entire *DRPLA* and 23 kb flanking seq (cosmid)	CAG_{76} Expanded	426	27% / -3 to -1 (3) n=215	7.6% / -2 to +1 (3) n=211	Yes	↑ in transmission and somatic changes with ↑ age	[Sato et al., 1999b]
SBMA	*NSE*	*NSE* promoter- entire *AR* cDNA- β-globin tag	CAG_{45} Expanded	14	0%	0%	No	-	[Bingham et al., 1995]
	Mx	*Mx* promoter- entire *AR* cDNA- β-globin tag	CAG_{45} Expanded	62	0%	0%	No	-	
SBMA	*AR*	Entire *AR* and 450 kb flanking sequence (YAC)	CAG_{45} Expanded	178	16% / -3 to +1 (4) n=73	5% / -20 to +1 (21) n=105	No	↑ in transmission changes with ↑ age, *AR* not expressed	[La Spada et al., 1998]
SCA1	*pcp-2*	*pcp-2* promoter- entire *SCA1* cDNA- SV40 poly(A)	CAG_{82} Expanded	125	0%	0%	No	-	[Burright et al., 1995]
SCA1	*pcp-2*	*pcp-2* promoter- entire *SCA1* cDNA- SV40 poly(A) small t intron	CAG_{82} Expanded	299	67%/ -9 to –1 (9) n=213	3% n=86	No	↑ in transmission changes with ↑ age	[Kaytor et al., 1997]
SCA1	*Sca1*	Knockin- expanded repeat	CAG_{78} Expanded	200	70% / -5 to +1 (6) n=91	9% / -4 to +2 (6) n=109	-	↑ in transmission changes with ↑ age	[Lorenzetti et al., 2000]
DM	*DMWD* *DMPK* *DMAHP*	Entire *DMWD*, *DMPK* and *DMAHP* genes (45 kb cosmid)	CTG_{55} $CTG_{>300}$ Expanded	205 262	7% / -1 to +6 (7) 94% / -30 to +30 (60)	7% / -1 to +6 (7) 94% / -25 to +60 (85)	Yes	↑ in ♂ transmission and somatic changes with ↑ age ↑ 300 repeats strong expansion bias	[Gourdon et al., 1997b; Lia et al., 1998]
DM	-	1.2 kb of *DMPK* 3' UTR including repeat region	CTG_{162} Expanded	241	61% / -7 to +2 (9) n=134	34% / -11 to +7 (18) n=107	Yes	↑ in tissue specific somatic changes with ↑ age	[Fortune et al., 2000; Monckton et al., 1997]

KEY: ↑= increase, n= number.

1.8.3.2.2 HD

In the first attempt to model HD, mice were generated from a construct containing the full length *HD* cDNA with 316 bp of 5' UTR and 619 bp of 3' flanking sequence, in a cytomegalovirus (CMV) expression vector [Goldberg et al., 1996a]. A frameshift mutation was accidentally introduced into this construct which prevented transgene expression but the expanded CAG_{44} repeat did exhibit a very low level of transmissional instability (<1%).

In a second study transgenic mice were generated from a genomic *HD* fragment containing just the promoter and first exon of the gene, carrying an expanded $CAG_{\sim 130}$ repeat [Mangiarini et al., 1996]. These repeats displayed instability on transmission ranging from -2 to +1 repeats on maternal transmission and -3 to +12 repeats on paternal inheritance. The repeats also exhibited somatic instability, predominantly in the brain, kidney and liver which was found to increase with age [Mangiarini et al., 1997]. More in depth analysis of instability in these mice revealed that the gender of the embryo also contributes to instability, with expansion more common in males and contractions more prevalent in female progeny [Kovtun et al., 2000]. The affect of Msh2 on somatic repeat instability was determined by crossing these mice with *Msh2-/-* individuals, no somatic instability was apparent in the offspring produced, indicating that Msh2 is required for somatic instability [Manley et al., 1999].

A further group generated transgenic mice with expanded HD repeats ($CAG_{48 \text{ and } 89}$) by end ligation of *HD* cDNA PCR products which were cloned into a CMV expression vector [Reddy et al., 1998]. These mice were reported not to exhibit any transmissional repeat instability, but only a relatively small number of offspring were actually analysed (50).

Another group used two approaches to generate mouse models for HD [Wheeler et al., 1999]. They first used a 3.6 kb fragment of the human *HD* gene with an expanded repeat (CAG_{48}) to produce transgenic mice. The construct included sequence 5' of the transcription initiation site and exon 1 of the *HD* gene. The repeat in this transgene showed a low level of instability on transmission (2%). They then produced a more extensive study where several expanded repeats ($CAG_{48, 90 \text{ and } 109}$) were knocked into the mouse *Hdh* locus. These repeats exhibited 4% instability (with a size range of -1 to -2 repeats), 69% instability (-6 to +3 repeats) and 87% instability (-8 to +2 repeats) respectively when transmitted through the female germline. When inherited paternally they showed 4% instability (-1 to +1 repeats), 41% instability (-1 to +1 repeats) and 66% instability (-2 to +2 repeats) respectively. These

repeats were all somatically unstable and the most marked variability was observed in the transgenic mouse kidney and brain.

A second group also generated mice by knocking in expanded repeats ($CAG_{72 \text{ and } 80}$) to the murine *Hdh* locus [Shelbourne et al., 1999]. These repeats displayed size variation in 27% of female germline transmissions (-4 to +2 repeats) and in 22% of male germline transmissions (+1 to +8 repeats). The mice exhibited dramatic somatic instability particularly in the striatum of the brain. The somatic instability increased with mouse age to a point where repeats were present at triple their germline size [Kennedy and Shelbourne, 2000]. As neurons are terminally differentiated it is tempting to speculate that somatic instability may arise through a different mechanism than germline instability.

The increased rates of instability observed in the knock-in mouse models are probably a reflection of contextual *cis* elements present in and around the homologous *Hdh* mouse gene and a lack of positional integration effects.

1.8.3.2.3 DRPLA

To model DRPLA, transgenic mice were created from a cosmid construct containing the entire *DRPLA* gene with an expanded repeat (CAG_{78}) and 23 kb of flanking sequence cloned into an SV40 expression vector [Sato et al., 1999b]. When transmitted maternally the repeat contracted by 1 to 3 triplets in 27% of progeny and when inherited paternally the repeat varied by -2 to +1 repeats in 8% of offspring. The repeat was also somatically unstable and both types of instability were found to increase in line with age. The percentages of size changes reported were not dissimilar to those observed in humans, but the magnitude of size changes were considerably smaller in the mouse.

1.8.3.2.4 SBMA

One research group produced transgenic mice from two constructs, both of which contained the entire *SBMA* cDNA with an expanded repeat (CAG_{45}) tract and 665 bp of flanking sequence. The first construct was driven by an *NSE* promoter and the second was driven by an *Mx* promoter [Bingham et al., 1995]. No transgene repeat instability was observed in mice that inherited either of these constructs despite the fact that this size of repeat would be unstable in 25% of human transmissions.

A second transgenic study employed a YAC construct containing the entire *AR* with an expanded repeat (CAG_{45}) and 450 kb of flanking sequence [La Spada et al., 1998]. This transgenic repeat varied in 16% of maternal transmissions (-3 to +1 repeats) and in 5% of paternal transmissions (-20 to +1 repeats). The percentage of size changes observed increased with the age of the transmitting parent but no somatic instability was identified.

1.8.3.2.5 SCA1

Transgenic mice were generated from a construct with a *pcp-2* promoter, the entire *SCA1* cDNA with an expanded repeat (CAG_{82}) and an SV40 poly(A) signal [Burright et al., 1995]. This trinucleotide repeat was inherited stably in all the transgenic progeny analysed and no somatic size variation was detected. This group modified their original transgenic construct by including a small t intron within the SV40 poly(A) signal and generated further transgenic lines [Kaytor et al., 1997]. The repeats in these transgenes were unstable in 67% of maternal transmissions (-9 to –1 repeats) and in 3% of paternal transmissions. The rate of size changes reported increased considerably with the age of the transmitting parent but no somatic instability was evident. The researchers did not attribute the repeat variability observed in these mice to the minor transgenic construct modification, but to analysing sufficient progeny, specifically through the maternal germline and at increasing ages. Size variations comparable to those displayed by the progeny of these mice were identified in unfertilised oocytes, suggesting that maternal instability occurs after meiotic replication but prior to fertilisation.

To create a contextual SCA1 model an expanded repeat tract (CAG_{78}) was knocked into the mouse *Sca1* gene [Lorenzetti et al., 2000]. The trinucleotide repeats of knock-in mice displayed instability in 70% (-5 to +1 repeats) of progeny when transmitted maternally and in 9% (-4 to +2 repeats) when transmitted paternally. The percentage of offspring exhibiting repeat size changes increased with the age of the transmitting parent.

1.8.3.2.6 DM

In order to create a mouse model for DM, transgenic mice were created from a cosmid construct containing the entire *DMWD*, *DMPK* with an expanded repeat (CTG_{55}) and *DMAHP* genes [Gourdon et al., 1997b]. Resulting mice manifest repeat instability in 7% (-1 to +6 repeats) of transmissions through both the male and female germline. Somatic instability of repeats was also reported and both types of instability increased with age [Lia

et al., 1998]. The somatic mutation rate observed in these mice did not appear to correlate with the rate of tissue proliferation or differences in transcriptional levels of the three genes. This group subsequently generated mice from a similar construct containing a much larger repeat tract ($CTG_{>300}$) [Fortune et al., 2000]. These mice are the only model to date which show rates of transmissional instability on a par with those observed in humans carrying similar sized repeats (94% of transmissions). They also display a comparable bias towards repeat expansion (>90% of changes expansions) and larger size increases on male germline transmission (a mean of +12 repeats more). Somatic instability in these mice was found to increase with progenitor repeat size and the age of the mice. However as yet no vast, single-step expansions leading to the equivalent of congenital DM repeat sizes have been observed.

A model produced solely to study DM repeat instability was generated from a 1.2 kb construct spanning most of the *DMPK* 3' UTR and including an expanded repeat tract (CTG_{162})[Monckton et al., 1997]. No coding DNA or promoter elements were contained in this transgene as it was only intended to address repeat instability. Resulting transgenic mice showed repeat instability in 61% (-7 to +2 repeats) of offspring when transmitted maternally and in 34% (-11 to +7 repeats) of offspring when transmitted paternally. The transgenic repeats were also somatically unstable and this instability increased in a tissue specific manner with the age of the mice [Fortune et al., 2000]. The most dramatic somatic changes were observed in the kidney, which is not a highly proliferative tissue.

1.9 CpG islands

CpG islands are short, approximately 1 kb regions of DNA which contain relatively abundant CpG residues that are resistant to methylation. CpG islands are often associated with the 5' end of genes, their promoters and origins of replication and transcription [Antequera and Bird, 1999; Delgado et al., 1998]. The models proposed to account for the lack of methylation at CpG islands include: island DNA being an unsuitable substrate for methyltransferases, that following DNA replication a protein complex blocks the access of methyltransferases to island DNA, that islands are actively demethylated during early embryogenesis and that gene promoters generate methylation free footprints by recruiting replication complexes in germ cells and that these footprints are transmitted through cell division by a maintenance methyltransferase [Antequera and Bird, 1999]. The chromatin of CpG islands is distinct from bulk DNA in that it has an 'active' or 'open' structure, the nucleosomes are histone 1 (H1) depleted, histones H3 and H4 are hyperacetylated and that nucleosome free regions exist [Tazi and Bird, 1990].

1.9.1 Putative association with expanded trinucleotide repeat loci

As previously mentioned (see 1.5.6) the majority of expanded trinucleotide repeats are located within or near CpG islands [Brock et al., 1999; Gourdon et al., 1997a]. Brock *et al* clarified this anecdotal, putative association by comparing the relative 'expandability' of all the CAG/CTG repeat expansion loci with the GC content of their flanking DNA (see table 1.5). They performed this analysis by collating the published data on intergenerational repeat instabilities and by calculating the GC contents of the flanking DNA sequences. The degree of 'expandability' was determined by calculating the mean size of repeat changes as a factor of the starting repeat size. This eliminated the documented bias towards increased rate and magnitude of size changes which correlate with increasing progenitor repeat size [Ashley and Warren, 1995]. They also dissected this data into male and female transmissions, to avoid biasing the results with the known sex of founder effects. When the loci are ranked on highest flanking GC content versus highest repeat expandability through the paternal line (the most expandable sex in the majority of these loci) the association is positive and statistically significant ($p<0.01$), see table 1.5. This relationship is striking. The most expandable repeats are all located within CpG islands, the less expandable loci are flanked by CpG islands and the least variable of the loci are not associated with any CpG island. It is also of note that the DNA immediately flanking the repeats (within 100 bp) seems to exert a stronger effect on expandability than that over a longer range (500 bp).

It remains to be determined whether CpG islands themselves are directly or indirectly associated with repeat expandability. Certainly the discovery that two expanded repeat loci SEF2-1 and DIR1 are not associated with CpG islands precludes the possibility that they are a prerequisite of expansion. It is possible that the association of CpG islands with trinucleotide repeat expansions actually represents a link with origins of replication and transcription [Antequera and Bird, 1999; Delgado et al., 1998]. This could explain the expansion potential of the repeats that don't fall within islands but may still be associated with ORIs. There are several factors unique to CpG islands that could potentially account for their association with repeat mutability: they are unmethylated, have an 'active' chromatin structure, unusual DNA conformation and it is possible that more crucial differences from bulk DNA have yet to be discovered.

Table 1.5 CAG/CTG trinucleotide repeat expandability and flanking sequence analyses

Locus	% GC- 100 bp flanking the repeat	% GC- 500 bp flanking the repeat	Repeat within or flanking a CpG island*	Estimated expandability (95% confidence interval)			Accession number
				Male	Female	Sperm	
DM	69.5%	66%	Within CpG island	**4.81** (3.46-7.22)	**7.64** (5.16-10.87)	**4.34** (4.17-4.50)	X84813 and 100727
SCA7	83.5%	71.5%	Within CpG island	**1.30** (0.80-1.65)	**0.04** (0.27-0.56)	**7.80** (7.52-7.98)	AF032102
SCA2	77%	79%	Within CpG island	**0.97** (0.65-1.33)	**0.45** (0.25-0.64)	n/a	AC004085
HD	74.5%	71%	Within CpG island	**0.29** (0.21-0.43)	**0.09** (0.00-0.17)	**0.98** (0.84-1.0)	Z68756
DRPLA	63.5%	66%	Flanking CpG island	**0.19** (0.14-0.24)	**0.04** (-0.05-0.14)	n/a	U47924
SCA1	66%	67.2%	Flanking CpG island	**0.14** (0.00-0.24)	**0.00** (-0.05-0.04)	**0.26** (0.19-0.35)	AC002326
SBMA	65%	59%	Flanking CpG island	**0.08** (0.00-0.22)	**0.00** (0.00-0.00)	**0.13** (0.08-0.13)	X78592 and M27423
SCA3	36.5%	38.5%[a]	n/a	**0.07** (0.05-0.09)	**0.02** (0.01-0.03)	n/a	AJ000501
DIR1	38.5%	43%	No CpG island	**-0.01** (-0.07-0.02)	**0.00** (-0.03-0.05)	n/a	AC004108
SEF2-1	45%	n/a	No CpG island	n/a	n/a	n/a	U75701

KEY: n/a= not available, [a]= only 435 bp of sequence available on the 5' flank. This table is reproduced from [Brock et al., 1999].

*The presence of a CpG island was predicted using the calculation of [Gardiner-Garden and Frommer, 1987](see 3.2.3, method 2 for more details).

1.10 Aims of PhD study

The aim of this PhD study was to screen the mouse genome for any large, endogenous trinucleotide repeats and to assess their size variability. This was with the objective of identifying the first expanded repeats, which could potentially underlie any mouse mutant phenotype. The other research goal was to determine whether CpG islands have any effect on trinucleotide repeat instability.

1.10.1 Survey of endogenous mouse trinucleotide repeats from CpG islands

Most expanded human trinucleotide repeat loci are found within or near CpG islands. Therefore mouse CpG islands seemed the most likely genomic regions to yield expanded repeat loci if any were to be found. A mouse CpG island library [Cross et al., 1997] was screened for all 10 classes of trinucleotide repeat, this process and the positive clones obtained are described in chapter 3. A human CpG island library [Cross et al., 1994] has been systematically sequenced by the Sanger Centre and the nucleotide data is available to researchers. These nucleotide sequences were also screened for trinucleotide repeats to provide data which could be compared with the mouse screen. The results of this screen, analyses and comparisons of the mouse and human data are also detailed in chapter 3.

Once sequence information of mouse trinucleotide repeat loci was obtained, where possible polymerase chain reaction (PCR) primers were designed to amplify these repeats from mouse species. The variability of these repeats was then assessed by determining their levels of polymorphism in a panel of mouse species and strains. Where species size variability permitted these repeats were also mapped using a mouse interspecific backcross [Rowe et al., 1994]. The variability of mouse trinucleotide repeats, their PCR amplification conditions and where relevant their map positions are described in chapter 4.

1.10.2 Trinucleotide repeats as candidates for causing mouse mutant phenotypes

The mapped trinucleotide repeat loci were assessed as candidates for causing previously mapped mouse mutant phenotypes. This was accomplished by amplifying the repeats from mutant mice and normal individuals with the same genetic background. Two trinucleotide repeats were found to be expanded in a sample of frizzy mouse mutant DNA and were

investigated further as the genetic mutations responsible for the phenotype. The screening of trinucleotide repeats as candidates for causing mutant phenotypes and the investigation of the two expanded repeat loci are detailed in chapter 4.

1.10.3 Transgenic study to address the putative relationship between trinucleotide repeat instability and CpG islands

To date the majority of trinucleotide repeat expansions have been found within or near CpG islands. Many of the initial transgenic studies of trinucleotide repeats (often out of genomic context) failed to replicate the degree of instability observed in humans. We aimed to assess whether the location of a trinucleotide repeat within a CpG island directly affected its instability on transmission. To achieve this goal we cloned an expanded human myotonic dystrophy repeat into two transgenic constructs one of which contained a well characterised CpG island and the other in which the island properties were negated. These constructs were identical bar three point mutations in the Sp1 binding site of the mouse adenine phosphoribosyltransferase (*Aprt*) gene, which is known to destroy the native CpG island conformation in transgenes [Brandeis et al., 1994; Macleod et al., 1994]. As trinucleotide repeat orientation relative to the origin of replication is known to effect instability, the repeat was cloned into both constructs in both possible orientations. The generation of transgenic mice containing these constructs, the preliminary analyses of repeat stability in these lines and their CpG island status (as determined by transgene methylation) is described in chapter 5.

A brief discussion of overall project achievements, conclusions and a premise of possible future research directions are presented in chapter 6 of this thesis.

CHAPTER 2

MATERIALS AND METHODS

2 MATERIALS AND METHODS

2.1 Materials

2.1.1 Chemicals and reagents

All chemicals for general use were supplied by Sigma (Sigma-Aldrich Ltd) or BDH (Merck Ltd) unless otherwise stated, and were of molecular biology grade. Bacterial media were supplied by Difco U.K. Ltd. Hormones were supplied by Intervet UK Ltd, and anaesthetics were supplied by Roche Diagnostic Ltd, and Janssen Pharmaceutical Ltd.

2.1.2 Radiochemicals

The radioactive isotopes $\{\alpha^{32}P\}$ dCTP at a specific activity of 3000 Ci/mmol and $\{\gamma^{32}P\}$ ATP at a specific activity of 1415 Ci/mmol were supplied by Amersham Life Science. The four Redivue™ ^{33}P labelled dideoxynucleotide terminators at a specific activity of 1500 Ci/mmol (450 μCi/ml) each, were also supplied by Amersham Life Science.

2.1.3 Enzymes

Restriction enzymes, T4 polynucleotide kinase and DNA ligase were supplied either by Gibco BRL Life Technologies or New England Biolabs. Klenow DNA polymerase was provided with the random primed DNA labelling kit (Boehringer). Desiccated proteinase K and RNAase-A were supplied by Sigma. Taq polymerase was supplied at a concentration of 5 U/μl by Gibco BRL Life Technologies. Thermo-Sequenase DNA polymerase was supplied with the Thermo-Sequenase radiolabelled terminator cycle sequencing kit (Amersham Life Science) and AmpliTaq FS DNA polymerase was supplied with the ABI PRISM® BigDye™ terminator cycle sequencing kit (Perkin Elmer Applied Biosystems).

2.1.4 Vectors and markers

pUC18™ phagemid vector was obtained from Stratagene. Epicurican Coli® JM109 competent cells were also purchased from Stratagene. The 1 kb ladder DNA marker was obtained from Gibco BRL Life Technologies. The fluorescent 50-500 bp sizer™ was purchased from Amersham pharmacia biotech.

2.1.5 Solutions and buffers

Unless otherwise stated, solutions and buffers were prepared using distilled and deionised water and were stored at room temperature (i.e. between 15 and 25°C). Sterilisation was achieved by autoclaving at 15 lbs. psi 121°C (30 min). The components of general solutions and buffers are described in appendix 2.A.1.

2.1.6 Mouse CpG island library

The mouse CpG island library was provided courtesy of Dr Sally Cross [Cross et al., 1994] and Professor Adrian Bird from the Institute of Cell and Molecular Biology (ICMB) at Edinburgh University.

2.1.7 Mouse bacterial artificial chromosome (BAC) library

The mouse BAC library (mouse ES, release II), which was gridded onto nylon filters, was purchased from Incyte Genomics, St. Louis, USA.

2.1.8 Mice and DNA samples

2.1.8.1 DNA from mouse strains, inbred strains and colonies

DNA from the mouse strains detailed in table 2.1 was purchased from The Jackson Laboratory (JAX), Bar Harbor, Maine.

2.1.8.2 The Jackson Laboratory C57BL/6J x *Mus spretus* (BSS) backcross

The mouse interspecific backcross was established at the Jackson laboratory. C57BL/6J (B) females were crossed with *Mus spretus* (S) males, and the resulting F_1 females were then backcrossed to *Mus spretus*. DNA from the 94, N_2 progeny of this cross were made available to the research community as a tool for mapping mouse loci [Rowe et al., 1994].

2.1.8.3 DNA from mouse/hamster somatic cell hybrids

The DNA from a mouse/hamster somatic cell hybrid panel was provided by the Human Genome Mapping Project-Resource Centre (HGMP-RC) at Hinxton, Cambridge, UK. The 22 hybrid clones were characterised by the polymerase chain reaction (PCR) and fluorescent in situ hybridisation (FISH) [Williamson et al., 1995], and made available for chromosome assignment.

Table 2.1 DNA from mouse strains, inbred strains and colonies

DESIGNATION	DESCRIPTION OF STRAIN OR COLONY
129/J	Agouti, standard inbred strain derived from English coat colour stocks and a chinchilla stock (Dunn 1928)
A/J	Albino, standard inbred strain derived from outbred albino stocks (Strong 1921)
AKR/J	Albino, standard inbred strain derived from a high leukaemia strain (JAX 1940)
BALB/c	Albino, standard inbred strain (Bagg 1913)
C3H/HeSnJ	Agouti, standard inbred strain derived from BALB/c crossed with DBA (Strong 1920)
C57BL/6J	Black, standard inbred substrain of C57BL, derived from Lathrop stock, 57 crossed with 52 (Little 1921)
CAST/Ei	Agouti, *Mus musculus castaneus*, inbred strain derived from wild progenitors (Marshall to Eicher 1971)
CBA/J	Agouti, standard inbred strain derived from BALB/c crossed to DBA (Strong 1920)
DBA/2J	Dilute brown, standard inbred strain derived from coat colour stocks (Little 1909)
FVB	Albino, recently derived inbred strain from outbred Swiss mice (National Institute for Health {NIH} 1975)
IS/Cam Ei	Agouti, inbred strain derived from *Mus musculus praetextus* crossed to *Mus musculus domesticus* (Wallace to Eicher 1961)
MOLF/Ei	Agouti, *Mus musculus molossinus*, inbred strain derived from wild progenitors (Potter to Eicher ~1975)
Mus caroli	*Mus caroli*, colony derived from wild Southeast Asian stock, distant member of the genus *Mus*
NZB/B1NJ	Black, inbred strain derived from outbred mice (Imperial Cancer Research Fund {Bielschowsky 1948})
PERA/CamEi	Recently derived inbred strain from wild Peruvian progenitors (Atteck to Eicher 1961)

DESIGNATION	DESCRIPTION OF STRAIN OR COLONY
PERU/Ei	Agouti, recently derived inbred strain from wild Peruvian progenitors (Wallace 1972)
Rattus norvegicus	*Rattus norvegicus*, colony derived from wild Norway or common rat stock, sister of the genus *Mus*, from the family Murinae
SJL/J	Albino, recently derived inbred strain, from outbred Swiss mice (JAX 1955)
SPRET/Ei	Agouti, standard inbred strain derived from *Mus spretus*, western Mediterranean short-tailed mouse (Eicher 1988)

2.1.8.4 DNA from mutant strains of mice

DNA from mutant mouse strains (described in Table 2.2) was purchased from The Jackson Laboratory, Bar Harbor, Maine, USA.

2.1.8.5 Mice

All live mouse stock used in this study (CD1, F_1 CBA x C57BL/6J), were purchased from Charles River Laboratories UK Ltd, Margate, Kent.

2.1.9 Transgenic constructs

The transgenic constructs pABS and pAZM2 were provided courtesy of Dr Donald Macleod [Macleod et al., 1994] and Professor Adrian Bird from the ICMB.

Table 2.2 DNA from mouse mutants

MUTANT	SYMBOL	STRAIN	GENOTYPE
leaden	*ln*	V/Le	*a/a fz ln/fz ln*
dreher	*dr*	B6C3Fe-a/a-*dr*^{J}/+	*dr/dr*
vacuolated lens	*vl*	C3H/HeSn-*vl*/+	*vl/vl*
griege	*ge*	DBA/2J-*ge*	*ge/ge*
sepia	*sea*	C57BL/6J-*sea*	*sea/sea*
ichthyosis	*ic*	IC/Le-*ic*/+	*ic/ic*
fidget	*fi*	STOCK *stb + a/+ fi a*	*fi+/fi+*
rachiterata	*rh*	STOCK *rh/+*	*rh/rh*
ulnaless	*Ul*	B6EiC3H-ta/*Ul A*	*Ul/+*
epilepsy 2	*El2*	ABP.EL-*El2<e>*	*El2/+*
droopy ear	*de*	B6EiC3H-*a/A-de<H>*	*de<H>/de<H>*
light ear	*le*	C57BL/6J-*Pdeb<rd1> le*	*rd le/rd le*
buff	*bf*	C57BL/6J-*bf*	*bf/bf*
jagged tail	*jg*	JGBF/Le	*jg +/jg -*
cerebellar deficient folia	*cdf*	C3H/HeSnJ-*cdf*	*cdf/cdf*
truncate	*tc*	STOCK *tc/tc*	*tc/tc*
reduced pigmentation	*rp*	C57BL-*rp/+*	*rp/rp*
quivering	*qv*	C3FeB6-*A/A<w-J>-qv<J>*	*qv<J>/ qv<J>*
claw paw	*clp*	C57BL/6J-*clp*	*clp/clp*
hydrocephaly with hop gait	*hyh*	B6C3Fe-*a/a-hyh*	*hyh/hyh*
frizzy	*fr*	FS/Ei	*fr/fr*
long hair	*lgh*	A/J-*lgh*	*lgh/lgh*
motor neuron degeneration	*mnd*	B6.KB2-*mnd*	*mnd/mnd*
nervous	*nr*	C3HeB/FeJ-*nr*	*nr/nr*
adrenocortical dysplasia	*acd*	DW/J-*acd*/+	*acd/acd*
proportional dwarf	*pdw*	DBA/2J-*pdw*	*pdw/pdw*
oligosyndactylism	*Os*	C57Bl/6J*Os +/+ tgla*	*Os +/+ tg<la>*
hydrocephalus-3	*hy3*	B6CBACa-A<W-J>/A-hy3/+	*hy<3>/hy<3>*
ashen	*ash*	C3H/HeSn-*ash*/+	*ash/ash*
flailer	*flail*	B6CBACa-*A<W-J>/A-flail*	*flail/flail*
tail kinks	*tk*	TKDU/Dn	*tk +/tk +*
fur deficient	*fd*	STOCK-*a/a d fd/+ +*	*d fd/d fd*
epilepsy 1	*El1*	ABP.EL-*El1 El4*	*El1/+*

MUTANT	SYMBOL	STRAIN	GENOTYPE
Hypotransferineamia with hemochromatosis	*hpx*	BALB/cJ-*hpx*	*hpx/+*
ducky	*du*	TKDU/Dn	*+ du/+ du*
spinner	*sr*	C57BL/6J-*sr/+*	*sr/sr*
downless	*dl*	GL/Le	*dl<J>/+ dl<J>/-*
grey lethal	*gl*	GL/Le	*+gl/-gl*
kidney disease	*kd*	CBA/CaH-*kd*	*kd/kd*
waltzer	*v*	V/Le	*v/v*
Jackson circler	*jc*	C57BL/6J-*jc*	*jc/jc*
Juvenile congenital polycystic kidney disease	*jcpk*	C57BL/6J-*jcpk*	*jcpk/jcpk*
Ames waltzer	*av*	B6.BKs-*av<J>/+*	*av<J>/av<J>*
Jittery (previously hesitant)	*ji (hes)*	C3HeB/FeJ-*ji<hes>*	*hes/hes*
silver	*si*	B6C3Fe-*a/a* F1xSTOCKaTyrp1*<b>si*	*a b si/a b si*
altrichosis	*at*	ATEB/Le	*at +/at ?*
eye blebs	*eb*	ATEB/Le	*+ eb/? eb*
Myosin XV (previously shaker2)	*Myo15 (sh2)*	STOCK*myo15^{sh2}/+*	*sh2-/sh2+*
juvenile cystic kidney	*jck*	C57BL/6J-*jck*	*jck/jck*
Phosphatidylinositol transfer protein (previously vibrator)	*Pitpn (vb)*	B6C3Fe *a/a-pitpn^{vb}*	*vb +/vb -*
neurofibromatosis	*Nfi*	C57BL/6-*Nf1<tm1Fcr>*	*+/-*
bare skin	*Bsk*	C57BL/6J-*Bsk*	*Bsk/+ a/a*
crinkled	*cr*	B6C3Fe-*a/a-cr*	*cr/cr*
satin	*sa*	SB/Le	*sa bg/sa bg*
congenital hydrocephalus	*ch*	CHMU/Le	*+ mu/+ mu*
shimmy	*shmy*	B6C3Fe-*a/a-shmy/+*	*shmy/shmy*
waved coat	*Wc*	B6C3Fe-*a/a-Wc*	*Wc/+*
gunmetal	*gm*	C57BL/6J-*gm/+*	*gm/gm*
under white	*uw*	C57BL/6J-*uw*	*uw/uw*
progressive ankylosis	*ank*	C3FeB6-A/A*<w-J>-ank*	*ank/ank*
dorsal dark stripe	*dds*	C3H/HeSnJxSTOCK*dds/+*	*dds/dds*
harlequin	*Hq*	B6CBACa-A*<W-J>/A-Hq*	*Hq/y*
Proteolipid protein (myelin)	*Plp (rh)*	STOCK *rh/+*	*rh/rh*

2.2 Methods

2.2.1 Standard DNA protocols

2.2.1.1 Preparation of genomic DNA from general mouse tissue

Tissues were removed from mice and frozen immediately in liquid nitrogen. A small amount of tissue was then removed with a scalpel, and homogenised in a sterile Treff™ homogeniser in 3 ml of TNE buffer (see 2.A.1). Proteins were broken down by the addition of 100 µg/ml proteinase K and incubation at 55°C overnight. Cell debris was removed by salt precipitation, where an equal volume of 2.6 M NaCl was added and the mixture shaken vigorously and the debris pelleted by centrifugation at 12 000 xg for 10 minutes. The supernatant was removed from the pellet and the DNA was precipitated with ethanol. The DNA was washed twice in 70% ethanol and allowed to air dry. The DNA was then resuspended in an appropriate volume of TE buffer (see 2.A.1).

2.2.1.2 Extraction of DNA from mouse tail tips

Tail tips were incubated overnight at 55°C in 500 µl of tail tip DNA extraction buffer containing 50 µg proteinase K. Proteins and cell debris were removed from the mixture by a phenol: chloroform: isoamyl alcohol extraction, and excess phenol was removed by a subsequent chloroform: isoamyl alcohol extraction. The nucleic acids were precipitated with 1/10 of the volume sodium acetate (3 M) and an equal volume of ice-cold isopropanol. The nucleic acids were pelleted by centrifugation at 12 000 xg, and were washed twice in 70% ethanol. The DNA was allowed to air dry and resuspended in 50 µl TE.

2.2.1.3 Small scale preparation (miniprep) of plasmid DNA (S.N.A.P.™)

A single colony was used to inoculate 10 ml of LB-broth (see 2.A.1) containing 20 µg/ml ampicillin, and was incubated with vigorous shaking (225 rpm), at 37°C overnight. From this culture a 1.5 ml aliquot was transferred into a 1.5 ml micro-centrifuge tube (Eppendorf™) and the cells were pelleted by centrifugation at 12 000 x*g* for 5 min. The cells were resuspended in 150 µl of S.N.A.P.™ buffer 1*, and lysed with 150 µl of buffer 2. Lysis was allowed to proceed for 3 minutes at room temperature, then 150 µl ice-cold buffer 3 was added and the tube inverted 6-8 times. The mixture was centrifuged at 12 000 x*g* for 5 minutes and the gelatinous pellet containing host chromosomal DNA and cell debris was discarded. The supernatant containing the plasmid DNA was mixed with 600 µl of binding

buffer and pipetted onto a S.N.A.P.™ miniprep column, which was then placed into a 2 ml collection tube. The plasmid DNA was bound to the column by centrifuging the supernatant through it. The column was then washed with 500 µl of wash buffer, followed by 900 µl of final wash buffer. The column was then dried by brief centrifugation, and the plasmid DNA eluted in 60 µl of TE buffer.

*For all S.N.A.P.™ buffer components refer to 2.A.2.2.

2.2.1.4 Phenol: chloroform: isoamyl alcohol extraction

Phenol and chloroform were routinely used to remove proteinaceous material from nucleic acid solutions. An equal volume of phenol: chloroform: isoamyl alcohol (25: 24: 1) was added to the nucleic acid sample in a polypropylene tube. The contents of the tube were mixed well (by gentle inversion), until an emulsion was formed. The organic and aqueous phases were then separated by centrifugation at 12 000 xg for 5 minutes. The aqueous layer (usually the top phase, identifiable by the yellow colour of hydroxyquinoline added during phenol equilibration) was then carefully removed with a large-bore pipette, and the interface and organic phase discarded. The phenol: chloroform: isoamyl alcohol extraction was repeated until no protein (cloudy, white precipitate) was visible at the interface. To remove traces of phenol, the nucleic acid solution was then extracted with an equal volume of chloroform: isoamyl alcohol (24: 1) in exactly the same manner. The nucleic acid was then precipitated from the final aqueous phase using a combination of salt and ethanol.

2.2.1.5 Ethanol precipitation

The volume of the nucleic acid solution to be precipitated was measured, and the salt concentration adjusted by adding 1/4 volume of 10 M ammonium acetate (final concentration 2.5 M). After mixing well, 2 volumes of ice-cold 100% ethanol were added and the solution was mixed again. When small amounts of DNA were to be precipitated, 5-10 µg tRNA was then added to increase nucleic acid recovery. The DNA was then allowed to precipitate at 0°C (on ice) for 30 minutes. Depending on the amount of nucleic acid present, it was then spooled out (visible amounts) using a sterile glass loop, or pelleted (smaller amounts) by centrifugation at 12 000 xg for 15 minutes. The nucleic acid was then washed twice with 70% ethanol and allowed to air dry. Finally the DNA was resuspended in an appropriate volume of TE buffer.

2.2.1.6 Spectrophotometric quantitation of DNA

The concentration of DNA was determined by spectrophotometric measurement at 260 nm (OD_{260}). The spectrophotometer was set to zero with TE buffer and the optical density (OD) of a diluted DNA solution determined. This reading was then used to calculate the DNA concentration. An OD of 1 is equivalent to approximately 50 μg/ml of double stranded DNA, 40 μg/ml for single stranded DNA and 20 μg/ml for single-stranded oligonucleotides. The DNA purity was determined by taking a further reading at 280 nm, anything with an OD_{260}/OD_{280} ratio of less than 1.8 was considered contaminated (with protein or phenol) and could not be accurately quantitated by this method.

e.g. DNA concentration ng/μl=(OD_{260})(50 μg/ml)(dilution factor {D})

for genomic DNA

2.2.1.7 Restriction enzyme digestion of DNA

Restriction endonuclease digestions were performed under the appropriate conditions as recommended by the manufacturers, and with the buffers supplied with the enzymes. In general, 10 μg of genomic DNA was digested with 5 U of enzyme, in a total reaction volume of 40 μl (made up with sterile water) and incubated at the optimal enzyme temperature for 16 hours. Restriction digestion of plasmid DNA and PCR product was performed with 1 U of enzyme per 2-5μg of DNA, in a volume of 20 μl for 6-8 hours. If the DNA was to be used in a subsequent digestion or manipulation, the enzyme was heat inactivated where possible, or removed by phenol: chloroform: isoamyl alcohol extraction and the nucleic acid ethanol precipitated.

2.2.1.8 Agarose gel electrophoresis

DNA was visualised and size fractionated by agarose gel electrophoresis. The gels were prepared to concentrations ranging from 0.8 to 6% agarose (weight/volume) in the appropriate volume of 1x TAE or 0.5x TBE buffer (see 2.A.1). The higher percentage gels were used to visualise low molecular weight DNA, and to compare small size differences. The low percentage gels were used to visualise high molecular weight DNA and to differentiate large size differences. TBE buffer was used preferentially as the buffering capacity was higher, TAE was used only when the DNA was to be subsequently extracted from the gel matrix. Ethidium bromide was added at a concentration of 0.5 μg/ml to molten agarose before each gel was poured. Prior to electrophoresis, DNA samples were combined

with one-tenth volume loading buffer (see 2.A.1) and loaded into the wells alongside an appropriate size marker. The gels were then submerged under buffer and the samples electrophoresed through them at 50-150 volts, until the DNA had run a sufficient distance. The DNA was then visualised by viewing under ultra violet light (on a transilluminator) and the image captured using a digital imager, Syngene®, produced by Genetic Research Instrumentation (GRI), Essex.

2.2.1.9 Recovery of DNA fragments from agarose gels

2.2.1.9.1 Glass wool

The DNA band of interest was excised from the agarose gel using a sterile scalpel and any excess buffer removed by blotting dry on 3MM™ (Whatmann) filter paper. A small amount of siliconised glass wool (approximately 0.01 g) was then loosely packed into the bottom of a 0.5 ml micro-centrifuge tube. The bottom of the micro-centrifuge tube was then pierced with a fine bore needle (25 G x 5/8) and placed within a 1.5 ml micro-centrifuge tube. The agarose containing the DNA was then placed on top of the glass wool and centrifuged at 3000 xg for 5 minutes. The DNA and buffer were deposited into the 1.5 ml tube, while the agarose remained on top of the glass wool. The DNA needed no further purification if used as a hybridisation probe, but for use in ligations the DNA was ethanol precipitated.

2.2.1.9.2 Glass milk

This extraction method was used for higher molecular weight, more fragile DNA. The DNA band was excised from the agarose gel, and its volume determined by its overall weight. The agarose slice was then dissolved in 3 volumes of NaI solution (see 2.A.1) by heating at 65°C for 5 minutes and gently inverting the mixture. The DNA was then allowed to bind to 10 µl of glass milk (Silica, 325 Mesh {a powdered flint glass obtainable from ceramic stores}, mixed with TE and stored as a 50% slurry), by incubation at 0°C (on ice) for 15 minutes. The glass milk and bound DNA was then pelleted by centrifugation at 12 000 xg for 5 minutes, and the agarose containing supernatant was removed and discarded. The glass milk was then washed twice in ice-cold ethanol to remove any traces of agarose and NaI solution. The final pellet was then resuspended in 20 µl TE and the DNA allowed to dissociate from the glass milk by heating the mixture to 50°C for 10 minutes. The glass milk was pelleted once more and the supernatant containing the eluted DNA removed.

2.2.1.10 Treatment of DNA for cloning

Vector and construct DNA were digested with restriction endonucleases, modified when necessary (see below), and purified from agarose gels using glass milk.

When the sticky ends created by digestion could not be used for cloning, the overhangs were converted to blunt ends using the 3'-5' exonuclease activity of T4 DNA polymerase. This modification was carried out in the completed restriction digestion reaction, by adding 2 mM of each dNTP, 1 U of T4 DNA polymerase and incubating at 12°C for 15 minutes. The enzyme was then heat inactivated by heating to 68°C for 10 minutes.

When the vector carried identical sticky or blunt ends, it was treated with phosphatase to prevent possible self-ligation. Reactions were performed in 1x Calf intestinal alkaline phosphatase (CIP) buffer (see 2.A.1) containing 1 U of CIP at 37°C for 30 minutes.

2.2.1.11 Ligation of DNA

Once digested with the appropriate restriction enzymes, and modified when necessary, DNA fragments were ligated together in 20 µl reactions. A 3-fold molar excess of insert DNA: vector DNA was used for most cloning experiments. Reactions were performed in 1x T4 DNA ligase buffer (see 2.A.1) containing 1 U of T4 DNA ligase at 14°C overnight.

2.2.1.12 Preparation of electro-competant *Escherichia coli*

The host cells used for cloning purposes were *E. coli* strain JM109 (Epicurean Coli®) purchased from Stratagene. This strain of *E. coli* was restriction endonuclease and recombination deficient, ensuring the stability of any inserts successfully cloned within the system. These host cells also contain the F′ episome, incorporating ampicillin resistance and fragments of the bacterial β-galactosidase (*LacZ*) gene, permitting ampicillin selection and blue (no insert)/white selection for plasmids containing inserts.

A culture of JM109 cells were grown for 16 hours at 37°C (to lag phase), in 10 mls of LB-broth. This whole culture was used to inoculate 1 L of sterile LB-broth, which was incubated at 37°C with vigorous shaking, until the optical density reached 0.5-1 when read at a wavelength of 600 nm. The culture was then chilled on ice for 30 minutes and then pelleted by centrifugation at 4000 xg for 15 minutes. The supernatant was carefully removed and the cell pellet resuspended in 1 L of ice-cold sterile dH_2O. The cells were pelleted again by centrifugation (as before), the supernatant removed and the cells washed in 500 ml of ice-

cold sterile dH_2O. The cells were pelleted and washed again in 20 mls of dH_2O, and finally 2 mls of ice-cold 10% glycerol. The cells were snap frozen (in liquid nitrogen) in 45 µl aliquots and stored at -70°C until required.

2.2.1.13 Transformation of competent *Escherichia coli*

For transformation, generally up to 1-2 µl of DNA or ligation reaction (50-100 ng) was added to the 45 µl aliquot of competent cells and cooled on ice for 5 minutes. The gene-pulser apparatus was then set to 25 µF, 2.5kV and the controller to 200 Ω. The cells were then placed in a chilled 0.1 cm electroporation cuvette and pulsed with a time constant of 5 msec. The cuvette was removed from the gene pulser and 1 ml of LB-broth without ampicillin (to allow for pre-expression of the antibiotic resistance) was immediately mixed with the cells, which were then cultured at 37°C for one hour. The cells (100 µl, 200 µl and 500 µl) were then plated onto fresh LB-agar (see 2.A.1) containing 20 µg/ml ampicillin and cultured at 37°C for 16 hours.

The transformation efficiency was tested by transforming with 10 ng of undigested vector DNA. The efficiency of the CIP reaction was examined by transforming with dephosphorylated vector.

2.2.1.14 Southern blotting

Individual fragments of DNA were transferred from agarose gels onto nylon membrane to allow subsequent detection with specific radioactively labelled probes. This transfer was performed according to the method originally described by Southern [Southern, 1975]. Prior to DNA transfer, the gels were immersed in depurination solution (see 2.A.1) for 15 minutes (or until the bromophenol blue loading dye changed from blue to green) to break down larger fragments of DNA, permitting their efficient transfer. The gels were then soaked in denaturing solution (see 2.A.1) for 30 minutes, followed by neutralising solution (see 2.A.1) for 1 hour. The DNA was transferred by capillary action from the gel to the membrane on an apparatus consisting of a reservoir of 20x SSC (the transfer buffer, see 2.A.1), 3MM filter paper wicks, a blotting platform and paper towels. Gels were placed on the blotting platform and the area around the gel occluded with cling film to prevent the buffer flow through the gel being short-circuited. A Hybond-N^+ nylon membrane was pre-cut to the exact size of the gel, pre-soaked in transfer buffer, placed on top of the gel, and all air bubbles excluded. Three pieces of 3MM paper were then placed on top of the membrane, followed by a 10 cm stack of absorbent paper towels, a glass plate and a 200 g weight. The DNA was allowed to

transfer for at least 16 hours. After transfer, the membrane was rinsed briefly in 2x SSC and allowed to air dry. The DNA was fixed to the membrane by baking at 80°C for 2 hours and cross-linking by exposure to ultraviolet light in a UV Stratalinker® 2400 transilluminator (Stratagene).

2.2.1.15 Radio-labelling DNA probes

Probes were labelled using the random hexanucleotide priming method, first described by Feinberg and Vogelstein [Feinberg and Vogelstein, 1983]. The Random primed DNA labelling kit* purchased from Boehringer was used with $\{\alpha^{32}P\}$ dCTP (3000 Ci/mMol) as the radioisotope. The DNA probe (25 ng) was made up to 20 µl volume using distilled deionised water and denatured by boiling for 5 minutes. The probe was then added to a mixture of Klenow enzyme (5 U), 5 µl 5x OLB mix (containing random 6mer primers and a buffered solution of cold dATP, dGTP, and dTTP) and 3 µl (30 µCi) $\{\alpha^{32}P\}$ dCTP. The labelling reaction was allowed to proceed at room temperature for 1 hour and 30 minutes. The labelled probe was then ethanol precipitated, and unincorporated $\{\alpha^{32}P\}$ dCTP removed in two 70% ethanol washes. The labelled probe was resuspended in 100 µl TE buffer, denatured by boiling for 10 minutes and added to the hybridisation mix.

*For more in depth details of kit refer to manufacturer.

2.2.1.16 Hybridisation of radio-labelled probes to DNA bound on membranes

Hybond-N+™ nylon membranes were pre-wet in 2x SSC, and pre-hybridised in glass bottles containing 20 mls of Church and Gilbert's solution (see 2.A.1) at 68°C for 1 hour. Radioactively labelled DNA probes were denatured and added directly to the pre-hybridisation solution and allowed to hybridise at 68°C for 16 hours.

2.2.1.17 Post-hybridisation washing and radioactive signal detection

Following hybridisation, residual and non-specifically bound probe was removed by washing for 2 x 30 minutes in 2x SSC/0.1%SDS, and 1 x 30minutes in 0.2x SSC/0.1%SDS at 68°C. For radioactive signal detection each membrane was sealed in a plastic bag and exposed to autoradiographic film (Curix RP1 X-ray film, produced by Agfa-gevaert Ltd. Belgium) in light-proof cassettes with intensifying screens for between 16 hours and 14 days.

2.2.1.18 Removal of radio-labelled probes

To remove radio-labelled probes for subsequent hybridisation experiments, membranes were submerged in boiling 0.1x SSC/0.1% SDS in a plastic tray, and allowed to cool to room temperature. Membranes were exposed to autoradiographic film as previously described to ensure complete removal of probe. If removal was incomplete the whole process was repeated.

2.2.1.19 The polymerase chain reaction

2.2.1.19.1 Oligonucleotide primer design

Oligonucleotide primers were designed using the Primer3 software (developed at the Whitehead Institute in 1998 by Steve Rozen and Helen Skaletsky), which selects the best pair of primers to amplify a given target within the desired sequence. The software considers many parameters including; primer size, primer T_m, product T_m, primer GC content, primer self complementarity, primer 3' stability, avoidance of single polynucleotide stretches, GC clamps, and pair complementarity. The program also screens all potential primers against an organism specific mispriming library (repeat sequences, i.e. ALUs, LINEs, etc.). When it was not possible to use this computer program, oligonucleotide primers were designed by eye, taking into account as many of the aforementioned parameters as was feasible. Oligonucleotide primers were purchased from Sigma-Genosys UK Ltd., which were supplied as lyophilised products at a synthesis concentration of 0.3 μM. Primers were initially assayed between 55°C and 60°C to determine their optimum annealing temperature.

2.2.1.19.2 Polymerase chain reaction conditions

Polymerase chain reaction (PCR) reactions were carried out in a Hybaid Omnigene Thermal Cycler, in either 0.5 ml tubes or 0.5 ml, 96 well microtitre plates. The pipettes, tips, tubes and reagents were sterile and kept separately from those used in other experiments. PCR reactions were set up in a designated room, where PCR products were prohibited. These combined precautions minimised the possibility of extraneous DNA contaminating the PCR reactions. PCR was initially carried out in 1x PCR buffer (see 2.A.1), with 200 μM of deoxynucleotide triphosphates (dNTPs), 30 ng of each primer, 0.1 U of *Taq* polymerase, and template DNA (50-100 ng of genomic DNA, or 0.1-1 ng of plasmid DNA), in a volume of 25 μl. PCR reactions were overlaid with 25 μl of mineral oil (Sigma) to prevent evaporation and therefore alteration of the buffering conditions. After an initial denaturation step of

95°C for 3 minutes, the PCR conditions were 95°C for 1 minute, the appropriate primer annealing temperature for 1 minute, and primer extension at 72°C for 1 minute (if the product was <500 bp), or 2 minutes (if the product was >500 bp). The overall number of cycles was dependent on the type of DNA template; 25 cycles for plasmid DNA, or 35 cycles for genomic DNA. If the desired PCR product was weak, or not obtained using these conditions the annealing temperature was lowered, betaine (Sigma) was added to a final concentration of 2 M or the reaction magnesium chloride concentration was titrated from 1 mM to 4 mM (in 0.5 mM increments). If non-specific, or multiple PCR products were obtained using these conditions the annealing temperature was increased, dimethyl sulfoxide (DMSO) was added to a final volume of 10%, or again the reaction magnesium concentration was titrated.

2.2.1.20 Sequencing

The nucleotide sequence of relevant DNA was determined using variations of the Sanger end terminator sequencing technique [Sanger et al., 1977]. This process is based on the principle that 2', 3'-dideoxynucleoside triphosphates (ddNTPs) can be incorporated into a growing DNA chain through their 5' triphosphate, but are unable to form phosphodiester bonds with the next deoxynucleotide triphosphate, thus specifically terminating DNA synthesis.

The DNA was sequenced using the Thermo-Sequenase radiolabelled terminator cycle sequencing kit (Amersham Life Science {for kit component details see 2.A.2.3}), or the BigDye™ terminator sequencing kit (PE Applied Biosystems). The automated fluorescent sequencing method (BigDye™ {for kit component details see 2.A.2.1}) was more efficient for sequencing large numbers of samples, but was less reliable than manual (radiolabelled) sequencing when the template DNA contained long stretches of trinucleotide repeats.

2.2.1.20.1 Manual sequencing: Thermo-Sequenase kit

2.2.1.20.1.1 Preparation of DNA template for sequencing

Plasmid DNA was ready for direct sequencing once it had been extracted (see small scale preparation of plasmid DNA 2.2.1.3) and appropriately diluted. PCR products however, were first treated with a combination of exonuclease I and shrimp alkaline phosphatase (SAP). The exonuclease I was used to remove residual single-stranded primers and any extraneous single stranded DNA produced in the PCR. The shrimp alkaline phosphatase removed the remaining dNTPs from the PCR mixture, thus preventing their interference with

the subsequent sequencing reactions. This reaction comprised: 5 µl of PCR product (approximately 2.5 µg), 1 µl/10 U exonuclease I, and 1 µl/2 U shrimp alkaline phosphatase (SAP). The reaction was buffered by the PCR buffer constituents, and allowed to proceed at 37°C for 15 minutes. The enzymes were then heat inactivated at 80°C for 15 minutes.

2.2.1.20.1.2 Sequencing reactions

The termination mixes were prepared for each {α^{33}P} labelled ddNTP (dATP, dCTP, dGTP and dTTP) by combining 2 µl of termination master mix (usually dGTP, but dITP was used to eliminate compression artefacts in sequence of high GC content) and 0.5 µl of each radioactively labelled ddNTP. Reaction mixtures were then prepared comprising: 2 µl of 10x reaction buffer, 50-500 ng of template DNA, 30 ng of sequencing primer, sterile H_2O (to adjust total volume to 20 µl), and 2 µl (8 U) of Thermo-Sequenase DNA polymerase. A quarter of the reaction mixture (4.5 µl) was then transferred into each termination tube (labelled A, C, G, and T), and overlaid with 25 µl of mineral oil.

2.2.1.20.1.3 Cycle sequencing conditions

Cycle sequencing reactions were carried out in a Hybaid Omnigene Thermal Cycler, in either 0.5 ml tubes or 0.5 ml, 96 well microtitre plates. Standard cycle sequencing reactions using dGTP termination mix were cycled 60 times through: 95°C for 30 seconds, 55°C for 30 seconds, and 72°C for 1 minute. When using the dITP termination mix the sequencing reactions were cycled 60 times through: 95°C for 30 seconds, 50°C for 30 seconds, and 60°C for 10 minutes. After completion of the sequencing cycles, 4 µl of stop solution was added to each termination reaction. The samples were then denatured at 80°C for 10 minutes and 2.5 µl of each was loaded immediately onto a glycerol tolerant sequencing gel.

2.2.1.20.1.4 Separation of sequencing products using polyacrylamide gel electrophoresis (PAGE)

Polyacrylamide gels for electrophoresis of sequencing products were prepared using a 38 x 50cm-sequencing rig (Bio-rad). The acrylamide gel mix was prepared as a 950 ml stock comprising: 7 M urea, 6% acrylamide/*N, N'*-methylenebisacrylamide (19: 1, Acrylogel mix 5- Sigma) and sterile H_2O. Prior to pouring, 95 mls of this acrylamide mix was combined with 5 mls of glycerol tolerant buffer (see 2.A.1). To polymerise the gel matrix, 100 µl of NNN'N' tetramethylenediamide (TEMED) and 150 µl of 10% ammonium persulphate

(APS) were also added to the mix. The gels were prepared to a thickness of 0.4 mm and allowed to polymerise for 1 hour at room temperature.

Sequencing gels were pre-run for 20 minutes in 1x glycerol tolerant buffer, at 90 watts to allow the gel to reach 45°C, before the sequencing products were loaded. Once denatured (see *2.2.16.1.4*) the sequencing reactions were loaded into wells created by a 72-well, or 88-well sharks-tooth comb, and run at 90 watts for the desired length of time.

2.2.1.20.1.5 Drying of polyacrylamide gels and radioactive signal detection

After the run, the sequencing rig was disassembled, and the gel transferred onto 3MM filter paper. The topside of the gel was then covered with cling film, and the whole thing dried under a vacuum at 80°C for 1 hour (Bio-rad gel drier). The cling film was then removed and the gel exposed to autoradiographic film (Curix RP1 X-ray film) in light-proof cassettes for between 16 and 40 hours.

2.2.1.20.2 Automated sequencing: BigDye™ terminator sequencing kit.

2.2.1.20.2.1 Preparation of DNA template

Plasmid DNA was ready for direct sequencing once it had been extracted (see small scale preparation of plasmid DNA 2.2.1.3) and appropriately diluted. PCR products however, first had to be cleaned using a QIAquick™ (Qiagen) PCR purification kit to remove extraneous single stranded DNA, unincorporated dNTPs, enzyme and buffer. To each PCR reaction 200 µl of buffer PB* was added and mixed thoroughly before loading directly onto a QIAquick™ spin column. The PCR product was bound to the column by centrifugation at 12 000 xg for 1 minute, whilst the other reaction components passed through and were discarded. The DNA bound to the column was washed by adding 750 µl of buffer PE followed by brief centrifugation and discarding the flow through. The columns were then dried by brief centrifugation. The PCR product was eluted in 20 µl of buffer EB, which was collected from the column during the final centrifugation.

*For components of QIAquick™ kit refer to manufacturers.

2.2.1.20.2.2 Sequencing reactions

The sequencing reactions were set up in 0.5 ml PCR tubes and contained; 1x terminator ready reaction mix, 3.2 pmol of sequencing primer, the appropriate concentration of template DNA (15-45 ng of PCR product, 25-50 ng of single-stranded DNA, or 100-250 ng of double-stranded DNA), and deionised water to a total reaction volume of 10 µl. Each reaction was mixed well, collected by brief centrifugation and overlaid with 25 µl of mineral oil.

2.2.1.20.2.3 Sequencing conditions

The sequencing process was carried out in a Perkin Elmer Thermal Cycler, by 25 cycles of the following conditions: a rapid thermal ramp (1°C/sec) to 96°C, 96°C for 30 sec, a rapid thermal ramp to 50°C, 50°C for 15 sec, and a rapid thermal ramp to 60°C, 60°C for 4 min. The sequencing products were then purified.

2.2.1.20.2.4 Purification of sequencing products

The sequencing reactions were separated from the mineral oil by pipetting onto the top of a Parafilm M (Sigma) slope. The oil stuck to the film at the top of the slope and the separated reaction could be collected from the bottom. The sequencing products were then purified by ethanol precipitation (see 2.2.1.5). The final pellet containing the sequencing products was dried for 1 minute in a heat block at 90°C, and resuspended in 3 µl of ABI/ALF gel loading buffer (see 2.A.1). The samples were then denatured at 95°C for 5 minutes prior to loading onto the sequencing gel

2.2.1.20.2.5 Analysis of sequencing products using PAGE and automated fluorescence detection

Polyacrylamide gels for electrophoresis of sequencing products were prepared using the Applied Bio-Systems (ABI) casting rig for the Prism™ 377 DNA sequencer. The gel solution was prepared to 4% acrylamide/*N, N'*-methylenebisacrylamide (19: 1, Acrylogel mix 5 in 1x TBE buffer, 6M Urea, and 1.25% weight/volume amberlite (ABI). This solution was de-gassed and polymerisation induced by the addition of 55 µl TEMED and 400 µl of 10% APS. The gel was quickly poured and allowed to polymerise for at least two hours at room temperature. The gel was attached to the 377 ABI sequencer and the gel checked with the running software. The gel was pre-run (run program: seq run 48E-1200) in 1x TBE until it had heated up to the optimum running temperature of 50°C. The denatured samples were then loaded (1-2 µl) and electrophoresed at for 10 hours (run program: seq run 48E-1200). The sequences were then analysed and edited using the Genescan™ software (ABI).

2.2.2 CpG island library screen

2.2.2.1 Plating out the CpG island library

The CpG island library had been prepared from male mouse (MF1 and 129) genomic DNA by Dr Sally Cross in the laboratory of Professor Adrian Bird. The library was titrated to 10 000 colony-forming units (CFUs) in 2 ml of LB broth (to give a confluent plate with a colony size of approximately 1 mm), which was spread onto a 20 x 20cm Hybond-N™ hybridisation membrane which had been laid onto suitable selection media (LB agar containing Ampicillin) and grown for sixteen hours at 37°C.

2.2.2.2 Replication of CpG island colonies

This master membrane was then removed and placed face up onto 3 x 3MM filter papers. A fresh duplicate membrane was wetted on an LB agar plate, blotted on 3MM to remove excess moisture, and laid square onto the master membrane. A stack of 3 x 3MM filter papers (22 x 22cm) were laid on top of this membrane 'sandwich' and a clean glass plate (25 x 25cm) was laid on top of this. The stack was compressed by even pressure on the glass plate to transfer the colonies. The glass plate and filter papers were carefully removed (to prevent smearing or spreading of colonies) and alignment holes were made through each corner of the membrane sandwich with a sterile needle and waterproof ink. The master and duplicate membranes were then separated and placed colony side up on fresh LB agar and allowed to recover for 30 minutes at 37°C. This entire process was repeated until two replica membranes had been taken from each master. The master was then stored at 4°C until positive colonies were picked. Once screened the two identical replica membranes were directly compared to minimise false positive results.

2.2.2.3 Processing of CpG island colony lifts

All the replica membranes were then processed in the same way. They were first placed colony side up on a stack of 3MM paper soaked in denaturing solution (see 2.A.1) for 7 minutes, and then transferred onto two subsequent stacks of 3MM soaked in neutralising solution (see 2.A.1) for 5 minutes each. The membranes were then washed vigorously in 2x SSC to remove any remaining bacterial colonies, and then allowed to air dry on clean 3MM filter paper. The DNA was then fixed to the membranes by baking at 80°C for two hours, and processing DNA side down in a Stratagene, UV Stratalinker™, on auto-cross-link.

2.2.2.4 End-labelling of oligonucleotide probes

Synthetic 15mer oligonucleotides of each trinucleotide repeat class (AAT, ACT, AAG, ACC, CCG, ACG, AGC, ATC, AAC and $AGG)_5$ were radio actively labelled at their unphosphorylated 5' end in a standard end-labelling reaction [Richardson, 1965]. Each synthetic oligonucleotide (10 pmol) was labelled with 30 μCi $\{\gamma^{32}P\}$ ATP by the action of T4 polynucleotide kinase (5 U), in a 45 minute reaction incubated at 37°C. The reaction was buffered with 50 mM Tris Cl (pH 7.6), 10 mM $MgCl_2$, 5 mM dithiothreitol, 100 μM spermidine HCl and 100 μM EDTA (pH 8.0).

2.2.2.5 Hybridisation of colony lifts to radio-labelled oligonucleotide probes

Membranes were pre-wetted in 2x SSC and pre-hybridised in Church and Gilbert's solution for 2 hours. Hybridisation was performed at 5°C below the melting temperature (T_m) of each oligonucleotide probe, where:

$$T_m = 81.5\text{-}16.6(\log_{10}\{Na^+\}+0.41(\%G+C)\text{-}(600/N)$$

N = Oligonucleotide chain length (15)

This equation, originally used to calculate the relationship between G+C content, ionic strength of hybridisation solution, and the T_m of long DNA molecules was found by E. Fritsch [Sambrook et al., 1989] to reliably predict the T_m of oligonucleotides between 14 and 70 bases in length.

Oligonucleotide	T_m °C	5°C below T_m
AAT	37	32
AAC	50	45
AAG	50	45
ACT	50	45
ATC	50	45
ACC	64	59
ACG	64	59
AGC	64	59
AGG	64	59
CCG	77	72

2.2.2.6 Post-hybridisation washing and radioactive signal detection

Following hybridisation residual and non-specifically bound probe was removed by washing for 2 x 15 minutes in 4x SSC/0.1%SDS, at the same temperature as the hybridisation step. For radioactive signal detection each membrane was sealed in a plastic bag and exposed to autoradiographic film in light-proof cassettes with intensifying screens for between 1 and 16 hours.

2.2.2.7 Identification of positive trinucleotide repeat containing clones

The autoradiographs from each two replica membranes were overlaid to identify positive hybridisation. The signals were clearly marked and then the autoradiograph was laid underneath the master membrane on a light box and the alignment holes used to match their positions. Positive clones were then picked and grown up in 5 ml of LB-broth with ampicillin for 16 hours. The plasmid DNA was extracted from each of these cultures (see 2.2.1.3) and a glycerol stock (15% glycerol, 85% culture, snap frozen) stored at -70°C for future reference.

2.2.3 Production of transgenic mice

2.2.3.1 Preparation of construct DNA for micro-injection

The transgenic constructs were excised from their respective plasmids by restriction endonuclease digestion (see 2.2.1.7). The plasmid and construct fragments were separated on a 0.8% agarose gel, in 1x TAE gel. Once sufficient distance lay between the fragments, they were visualised under ultra violet light (for as short a time as possible to minimise DNA breakage) and the construct band removed with a sterile scalpel. The construct DNA was then extracted from the agarose using glass milk (see 2.2.1.9.2), and eluted in 60 µl of TE buffer. The DNA was then gently (to prevent DNA fragmentation) phenol: chloroform: isoamyl alcohol extracted, and then chloroform: isoamyl alcohol extracted.

To remove any possible dust contamination* in the DNA sample, it was mixed with 260 µl of 0.1 mM EDTA/1 mM Tris and passed through a microcon-30 column (Millipore). On centrifugation at 12 000 xg for 8 minutes the DNA became concentrated on the column membrane, and was washed twice with 500 µl of 0.1 mM EDTA/1 mM Tris pH 7.4. The column was then inverted and 20 µl of 0.1 mM EDTA/1 mM Tris pH 7.4 was added to the membrane. The DNA was then eluted from the column by centrifugation at 12 000 xg for 3 minutes. The DNA was then diluted 1 in 10 in injection buffer (see 2.A.1), and its

concentration determined by running an aliquot on a mini agarose gel against known DNA standards. The DNA was then diluted to 3 ng/µl in injection buffer, and any remaining dust removed by passing it through a Spinex 0.22 µm column (Costar). The DNA was then stored at -20°C until required.

* To minimise further dust contamination in this final clean-up stage all tips, eppendorfs and gloves were rinsed in sterile dH_2O, and all solutions were filter sterilised.

2.2.3.2 Preparation of donor mice*

Three days prior to micro-injection at 2pm, sufficient donor mice (4-8 week old, F_1 CBA x C57BL/6J) were injected intraperitoneally with 10 international units (IU) of pregnant mare serum gonadotrophin (PMSG).

At midday the day before embryo collection and micro-injection, the donor mice were injected intraperitoneally with 10 IU of human chorionic gonadotrophin (hCG). That evening each donor female was set up to mate with one proven stud male (8 weeks-8months old, F_1 CBA x C57BL/6J).

*Donor mice were prepared by animal house technicians.

2.2.3.3 Collection of fertilized oocytes from donor mice

On the morning of micro-injection the donor females were checked for vaginal plugs* as an indication of successful mating. Those that had mated were sacrificed by cervical dislocation. The mice were then positioned on their backs and their abdomens cleaned with 70% ethanol. Their abdominal skin was opened and pulled aside, a horizontal incision was then made in their peritoneum and their ovaries were located. The oviducts were dissected out attached to a small portion of uterus, and placed in a watch-glass of sterile saline pre-heated to 37°C**. The oviducts were then transferred one at a time into a watch-glass containing M2 culture media (see 2.A.6) also pre-heated to 37°C. Under a microscope the swollen ampulla of each oviduct containing the mass of eggs and cumulus cells, was torn open with a set of fine tipped forceps. Once all the eggs had been released into the media, a couple of drops of hyaluronidase enzyme (approximately 400 µl of 1 mg/ml) were added to detach the cumulus (follicle) cells from the oocytes. As quickly as possible (as hyaluronidase will eventually harm the embryos) the oocytes were pipetted into a second watch glass containing M2 media, to wash away residual enzyme. The embryos were then further washed by pipetting through three consecutive drops (200 µl) of M2 media, in a small

petri dish covered with paraffin oil (embryo tested). At this stage the oocytes were sorted, and any misshapen or unfertilised embryos were discarded. The oocytes were then processed through three consecutive drops of M16 culture media (see 2.A.6) which had been covered in paraffin oil and incubated in 5% CO_2 (in a CO_2 incubator) for 16 hours. The embryos were stored in the third drop of media, in the CO_2 incubator until required for micro-injection.

* Donor mice were checked for vaginal plugs by animal house technicians.

** All media used during micro-injection and embryo culturing, was pre-heated to 37^0C.

2.2.3.4 Injection of construct DNA into mouse embryo pronuclei

A large drop of M2 media (approximately 200 µl) was set up on a sterile depression slide and overlaid with paraffin oil. The first batch of embryos to be micro-injected was removed from the CO_2 incubator, passaged through three drops of M2 media and placed in the centre of the depression slide media. A micro-injection needle (Femtotips, Eppendorf) was loaded with 5 µl of freshly thawed construct DNA, and fixed to the mechanical injection micro-manipulator (Narishige M151). The injection needle was also attached to a Narishige IM300 pressure injector system, which produced a continuous flow of DNA from the needle tip. A holding capillary (Vacutips, Eppendorf) was back filled with M2 media and attached to the mechanical holding micro-manipulator (also Narishige M151). The injection needle and holding capillary were then lowered into the depression slide drop of media and under inverse stereo microscopy (Ziess Axiovert 100) were used to micro-inject construct DNA into the pronuclei (preferably the male as this is usually larger and nearer the outside of the cell) of all the available embryos.

2.2.3.5 Culture of injected embryos

Once injected the embryos were passaged once more through three fresh drops of M16 media (set up as before) and cultured in 5% CO_2 at 37°C for 16 hours. Those embryos, which successfully matured to the two-cell stage after micro-injection, were selected for transfer into recipient female mice to continue gestation.

2.2.3.6 Preparation of recipient mice*

The evening before injected embryos were expected to be transferred, sufficient recipient female mice (8 weeks-6 month old CD1 {albino}) were set up, one to one, to mate with

vasectomised males (8 weeks-8months old CD1). The following morning females were checked for vaginal plugs, and those that had mated** were used to receive the injected two cell embryos.

* Recipient mice were set up to mate and subsequently plug checked by animal house technicians.

** If recipient females were checked for oestrus prior to this mating, approximately 50% would be expected to successfully plug.

2.2.3.7 Preparation of transfer pipettes

The two cell embryos were washed in a watch-glass containing M2 media and loaded in the smallest amount of media possible into a sterile loading pipette (with a diameter just wider than the embryos themselves, approximately 150 μm). It was preferable to load a small air bubble into the loading pipette either side of the embryos, as this made their successful introduction into the oviducts of recipient mice far easier to visualise. For single sided oviduct transfers up to 30 embryos were loaded, for double sided transfers up to 50 embryos in total were used. Approximately 25-30% of transferred embryos were born, therefore double-sided transfers were preferable as litters of less than ten offspring were often rejected by the foster mothers.

2.2.3.8 Transfer of embryos into oviducts of recipient mice

Prior to the transfer procedure all surgical instruments were sterilised in ethanol. Each recipient mouse (approximately 40 g) was anaesthetised by a 400 μl intraperitoneal injection containing; 600 μg fluanisone, 80 μg midazolam and 20 μg fentanyl citrate (see 2.A.7). Successful anaesthesia was determined 5 minutes later, if no reflex action occurred in response to moderate pressure applied to the tip of the tail. The mouse was placed prone onto the lid of a petri dish and its back cleaned with 70% ethanol (not too much as this would lower the body temperature of the mouse). A small incision was made just below the ribcage and horizontal to the spinal column, through the skin. The incision was then manipulated to locate the fat pad, which lies, directly over the reproductive tract, an incision was then made through the peritoneum just large enough to provide access. The fat pad was grasped with fine forceps and pulled through the incision. The ovary, oviduct and uterus, which were attached to the underside of the pad were then exposed by gently rotating the pad and weighting it in place with a serrefine clamp. Under stereomicroscopy the opening to the

oviduct (infundibulum) was located amongst the oviduct coils, and a small tear was made in the ovarian bursa (a transparent vascularised layer of tissue). The entrance to the infundibulum was then explored gently with the watchmaker's forceps to clear any debris and the area swabbed dry of excess fluid or blood. The transfer pipette was then carefully inserted into the infundibulum to the end of the first coil and the embryos blown gently in. Successful transfer was assumed if the small air bubbles flanking the embryos could be seen moving through the oviduct. The clamp was then released from the fat pad and the body wall incision manipulated to encourage it and the reproductive tract gently back into the body cavity. The incision was then closed with one or two surgical knots, and the same procedure followed to transfer embryos into the other oviduct. Finally the skin was closed with two small wound clips and the mouse placed back into a warm cage to recover.

2.2.3.9 Identification of transgenic mice

All resulting progeny born to recipient females could be identified as those from successful transfers rather than incomplete vasectomy by virtue of being a different coloured mouse strain (F_1 CBA x C57BL/6J, agouti or black). All offspring from recipient female mice were tail-tipped at weaning (3-4 weeks of age) to determine if the construct DNA had integrated into their own. The DNA was extracted from each tail tip using the protocol previously described (2.2.1.2).

2.2.3.10 Size analysis of trinucleotide repeat in transgenic mice

One of the primer pair used to amplify the human repeat region in transgenic mice (DMtrans see chapter 5) was coupled to a (FAM) fluorescent residue, which enabled the PCR products to be sized accurately by being run out on a polyacrylamide gel. The PCR products were mixed 50: 50 with ABI/ALF loading dye containing two size standards (fluorescently labelled PCR products of a known size) one predicted to be smaller, and the other larger than the transgenic PCR product. The samples were run on a 6% acrylamide gel in 1x TBE on an Automatic Laser Fluorescence (ALF™) system (Amersham pharmacia biotech). They were run at a temperature of 40°C and 50 watts power for 3 hours, alongside a 50-500 bp size marker. The size of the transgenic PCR products were accurately sized using both the internal size standards and size marker as references for the Allelelinks software program (version 1.01-Amersham Pharmacia Biotech).

2.2.4 Bioinformatics

Allele-link software: version 1.01, Amersham pharmacia biotech.

BLAST at Sanger Centre: www.sanger.ac.uk/HGP/blast_server.shtml

BSS mouse interspecific backcross dataset at The Jackson Laboratory:
www.jax.org/resources/documents/cmdata/bkmap/BSS.html

Chi-Square Calculator: www.fourmilab.ch/rpkp/experiments/analysis/chiCalc.html

Cutter: www.medkem.gu.se/cutter/

GCG-HGMP-RC (clustal, gap, pileup, CpG plot) at:
www.hgmp.mrc.ac.uk/Registered/Option/gcg.html

Genescan™ software: Applied Biosystems.

LALIGN-HGMP-RC at: www.hgmp.mrc.ac.uk/Registered/Webapp/lalign/

Mouse chromosome committee reports at The Jackson Laboratory:
www.informatics.jax.org/ccr/searches/index.cgi?year=1999

Mouse Genome Database (MGD) at The Jackson Laboratory: www.informatics.jax.org

Mouse phenotype descriptions at The Jackson Laboratory:
www.informatics.jax.org/searches/marker_form.shtml

NIX-HGMP-RC at: www.hgmp.mrc.ac.uk/Registered/Webapp/nix/

Primer3 software at: www-genome.wi.mit.edu/cgi-bin/primer/primer3

Web Chi Square Calculator: www.georgetown.edu/cball/webtools/web_chi.html

2.A Materials and methods appendices

2.A.1 General solutions and buffers

ABI/ALF gel, sample loading buffer	95% deionised formamide, 20 mM ethylenediamine tetra acetic acid (EDTA), 10 mg/ml dextran blue
Acrylamide	30% Stock (19: 1) 28.5% Acrylamide, 1.5% *N, N'*-methylenebisacrylamide
Acrylamide gel, sample loading buffer	2x Stock 98% deionised formamide, 10 mM EDTA pH 8.0, 0.025% xylene cyanol FF, 0.025% bromophenol blue
Agarose gel, sample loading buffer	10x Stock 0.4% bromophenol blue, 50% glycerol
10% Ammonium persulphate (APS)	10% weight: volume APS: dH_2O Store aliquots at -20°C, defrost once only.
Betaine	5.5 M in dH_2O Store aliquots at -20°C, defrost once only.
Calf intestinal alkaline phosphatase (CIP) buffer	10x STOCK 0.5 M Tris-HCl (pH 9.0), 10 mM $MgCl_2$, 1 mM $ZnCl_2$, 10 mM spermidine
Chloroform: isoamyl alcohol	24: 1, Fluka
Church and Gilbert hybridisation solution	7% SDS, 0.5 M $NaPO_4$ (pH 7.2), 1 mM EDTA

Denaturing solution	0.5 M NaOH, 1.5 M NaCl
deoxyribonucleoside triphosphates (dNTPs)	10 nM/µl of each dATP, dCTP, dGTP and dTTP in dH_2O Store aliquots at -20°C, defrost once only.
Depurination solution	250 mM HCl
Dimethyl sulfoxide (DMSO)	100% Stock Store aliquots at -20°C, defrost once only
Ethidium bromide	10 mg/ml in dH_2O Store in darkness at room temperature
Glycerol tolerant buffer	10x STOCK 1.8 M Tris, 1.7 M Taurine , 10 mM EDTA
Injection buffer	10 mM Tris, 0.1 mM EDTA, pH to 7.4 Autoclave and filter sterilise
Isopropylthio-β-D-galactoside (IPTG)	100 mM in dH_2O Filter sterilise to 0.22 microns, store at -20°C
3-{N-Morpholino}propane-sulphonic acid (MOPS)	20x STOCK 0.2 M MOPS, 0.05 M Sodium acetate, 0.01 M EDTA
NaI solution	6 M NaI, 160 mM Na_2SO_3 Filter- solution should be saturated, store in dark
Neutralisation solution	0.5 M Tris pH 7.5, 1.5 M NaCl
Phenol: chloroform: isoamyl alcohol	25: 24: 1, Fluka

Polymerase chain reaction buffer	10x STOCK 750 mM Tris HCl pH9.0, 200 mM Ammonium sulphate, 0.1% weight: volume Tween
Sodium dodecyl sulphate (SDS)	10% weight: volume SDS: dH_2O Heat to dissolve, pH to 7.2 with HCl
Standard saline citrate (SSC)	20x Stock 3 M NaCl, 0.3 M trisodium citrate, pH to 7.0
T4 DNA ligase buffer	5x Stock 250 mM Tris HCl pH7.6, 50 mM $MgCl_2$, 5 mM ATP, 5 mM dithiothreitol (DTT), 25% (w/v) polyethylene glycol 8000
Tail tip DNA extraction buffer	10 mM Tris pH 7.0, 1 mM EDTA, 1% SDS, 0.3 M sodium acetate pH 7.0
Tris*-acetate (TAE) ***(Hydroxymethyl) aminomethane**	10x Stock 0.4 M Tris-HCL pH 8.0, 0.2 M sodium acetate, 10 mM EDTA, 9 ml/l glacial acetic acid
Tris-borate (TBE)	20x Stock 1.8 M Tris, 1.8 M boric acid, 40 mM EDTA pH 8.0
Tris-EDTA (TE)	10 mM Tris HCl pH 7.6, 1 mM EDTA pH 8.0
Tris-NaCl-EDTA (TNE)	10 mM Tris, 400 mM NaCl, 2 mM EDTA
5-Bromo-4-chloro-3-indolyl-β-D-galactoside (X-gal)	20 mg/ml in dimethylformamide Store in darkness at -20°C

2.A.2 Kit contents

2.A.2.1 Big dye™ terminator cycle sequencing ready reaction kit

Termination ready reaction mix	Dye terminators: ddATP labelled with dichloro{R6G}, ddCTP dichloro{ROX}, ddGTP dichloro{R110} and ddTTP dichloro{TAMRA} Deoxynucleoside triphosphates (dATP, dCTP, dITP, dUTP) AmpliTaq DNA polymerase, FS, with thermally stable pyrophosphatase $MgCl_2$ Tris-HCl buffer, pH 9.0

2.A.2.2 S.N.A.P.™ miniprep kit

Buffer 1, resuspension buffer	50 mM Tris pH 8.0, 10 mM EDTA, 100 µg/ml RNAase A
Buffer 2, lysis buffer	0.2 M NaOH, 1% SDS
Buffer 3, precipitation salt	3 M potassium acetate pH 5.2
Binding buffer	7 M Guanidine-HCl
Wash buffer	5 M Guanidine-HCl, 50 mM MOPS, pH 7.0
Final wash buffer	100 mM NaCl in 70% ethanol

2.A.2.3 Thermosequenase radiolabeled terminator cycle sequencing kit

Thermo-sequenase DNA polymerase 4 U/µl	Buffered in 50 mM Tris-HCl pH 8.0, 1 mM dithiothreitol (DDT), 0.1 mM EDTA, 0.5% Tween™-20, 0.5% Nonidet™ P-40, 50% glycerol

Reaction buffer	10x STOCK 260 mM Tris-HCl pH 9.5, 65 mM $MgCl_2$
dGTP nucleotide master mix	7.5 μM dATP, dCTP, dGTP, dTTP
dITP nucleotide master mix	7.5 μM dATP, dCTP, dTTP, 35.7 μM dITP
Stop solution	95% formamide, 20 mM EDTA, 0.05% bromophenol blue, 0.05% xylene cyanol FF

2.A.3 Growth media for bacterial cultures

Luria-Bertani (LB)-broth	10 g/l Bacto-tryptone, 5 g/l Bacto-yeast extract, 5 g/l NaCl, pH to 7.0 with NaOH
Luria-Bertani (LB)-agar	15 g/l Bacto-agar, 10 g/l Bacto-tryptone, 5 g/l Bacto-yeast extract, 5 g/l NaCl, pH to 7.0 with NaOH

2.A.4 Enzymes

Hyaluronidase	Stock 1 mg/ml in PBS, 230 mM polyvinylpyrrolidone (PVP)	Reaction 100 mg/ml
Proteinase K	Stock 50 mg/ml in dH_2O	Reaction 50-100 μg/ml

2.A.5 Antibiotics

Ampicillin	Stock 50 mg/ml in dH_2O	Media 20 μg/ml
Penicillin	Stock 60 mg/ml in dH_2O	Media 60 μg/ml
Streptomycin	Stock 50 mg/ml in dH_2O	Media 50 μg/ml

2.A.6 Culture media for fertilised oocytes

M2 manipulation medium	95 mM NaCl, 5 mM KCl, 5 mM D-glucose, 0.2% Na lactate, 3 mM sodium pyruvate, 2 mM $NaHCO_3$, 2 mM $CaCl_2 2H_2O$, 2 mM hepes, 1 mM $MgSO_4 7H_2O$, 1 mM KH_2PO, 60 μg/ml penicillin, 50 μg/ml streptomycin, pH to 7.4
M16 culture medium	95 mM NaCl, 5 mM KCl, 5 mM D-glucose, 0.2% Na lactate, 3 mM sodium pyruvate, 4 mM $NaHCO_3$, 2 mM $CaCl_2 2H_2O$, 1 mM $MgSO_4 7H_2O$, 1 mM KH_2PO, 60 μg/ml penicillin, 50 μg/ml streptomycin, 10 μg/ml phenol red, pH to 7.6 (changes to pH 7.4 in CO_2 incubator)

2.A.7 Anaesthetic

Hypnorm™	Stock	Working concentration
Janssen Pharmaceutical Ltd	10 mg/ml fluanisone, 0.315 mg/ml fentanyl citrate	1.5 mg/ml fluanisone, 50μg/ml fentanyl citrate

Hypnovel®	Stock	Working concentration
Roche	5 mg/ml midazolam	200 μg/ml midazolam

Anaesthetics were purchased separately as Hypnorm™ and Hypnovel®, and were mixed together in sterile dH_20 to give the final combined working concentrations indicated above.

2.A.8 Hormones

Pregnant mare serum gonadotrophin (PMSG), obtained from Intervet as Folligon™.
Human chorionic gonadotrophin (hCG), obtained from Intervet as Chorulon™.
Hormones were diluted to a concentration of 50 IU/ml in sterile dH_20 and aliquots frozen at -20°C until use.

CHAPTER 3

A SURVEY OF TRINUCLEOTIDE REPEATS IN MOUSE CpG ISLANDS

3 A survey of trinucleotide repeats in mouse CpG islands

3.1 Introduction

The rationale behind screening CpG islands for trinucleotide repeats in the mouse was several fold. The majority of expandable trinucleotide repeats known to cause disease in humans lie within or near CpG islands [Brock et al., 1999; Gourdon et al., 1997a]. Although it is as yet unclear whether this association relates directly to the CpG islands themselves, or their known association with origins of replication (ORIs) [Antequera and Bird, 1999] and transcription [Delgado et al., 1998], this portion of the mouse genome might be considered the most likely to yield expandable trinucleotide repeats, if any were to be found. CpG islands are also associated with the 5' end of genes, which means that any trinucleotide repeats identified from this survey would also be associated with genes. This was important because repeats associated with genes could potentially cause a phenotype if they underwent expansion, making them the most relevant type of repeats to identify for this study. Until such time as the entire mouse genome has been sequenced and can be screened with computer programs to detect predicted coding regions, screening DNA associated with CpG islands, cDNA, and mRNA remains the best way to scan for genes amongst bulk DNA. From a more practical perspective the CpG island clones were a good resource to screen because they were contiguous with genomic DNA, whilst lacking the majority of intronic regions which are potentially difficult to work with. The CpG island library also provided relatively short, manageable clone inserts, most of which were seqencable in a single pass.

Previous studies of endogenous trinucleotide repeats in mouse cDNA [Abbott and Chambers, 1994; Chambers and Abbott, 1996] have revealed extensive size variation amongst inbred strains, which encouraged us to undertake more extensive screening of coding regions for variable and possibly expandable repeats.

3.1.1 Distribution of CpG islands in mammalian genomes

CpG islands are short genomic stretches of approximately 1 kb which are distinct from the bulk of DNA in that their abundant CpG moieties remain unmethylated. CpG islands are usually associated with the 5' end of genes and are concentrated in the early replicating

(R band) regions of the genome [Craig and Bickmore, 1994]. There are approximately 50 000 CpG island sequences per haploid human genome [Consortium, 2001], which constitutes roughly 1.5% of the genome as a whole. However a considerable percentage of these sequences (40%) represent GC-rich repeat elements which are unlikely to function as true islands.

3.1.2 CpG islands in the mouse genome

The number of CpG islands predicted per haploid mouse genome is 37 000 [Antequera and Bird, 1993], a figure at least 20% lower than that described in humans. Analysis of homologous genes reveals that 20% of those associated with an island in humans appear to have no corresponding island in the mouse. When directly compared, mouse CpG islands are almost always less distinct (with lower CpG densities) [Matsuo et al., 1993]. It is believed that CpG islands are being lost in both species over evolutionary time by *de novo* methylation and CpG mutation, but that the process is more rapid in rodents. This 'erosion' most commonly occurs through the methylation of CpGs followed by the mutation of 5-methylcytosine (CpG) to thymine (TpG) or adenine (CpA) through deamination [Barker et al., 1984; Sved and Bird, 1990].

3.1.3 Mouse CpG island library

A method for purifying predominantly intact CpG islands from genomic DNA [Cross et al., 1994] was used to generate a mouse CpG island library [Cross et al., 1997]. The library (courtesy of Dr Sally Cross and Professor Adrian Bird, ICMB) was prepared from 175 μg of male MF1 and 129 mouse DNA. The CpG island portions were generated by complete digestion of the genomic DNA with the restriction enzyme *Mse*I (recognition sequence TTAA), which cuts rarely in CpG islands but frequently in other DNA regions. The bulk of DNA which was methylated at the CpG residues was removed from the restriction mixture by virtue of its strong binding affinity when passed through a methyl binding column (MBC).

The methyl binding column was constructed by coupling a recombinant methyl binding domain (HMBD) produced from rat MeCP2 protein, to an agarose matrix. The recombinant protein was produced by amplifying rat methyl binding domain (MBD) DNA by PCR, and cloning it into a bacterial expression vector which contained a histidine tag upstream of the cloning site. The expressed recombinant protein (see figure 3.1) was then purified and attached to a nickel-agarose matrix via the histidine tags at its N-termini.

Figure 3.1 Sequence of the methyl-CpG binding domain

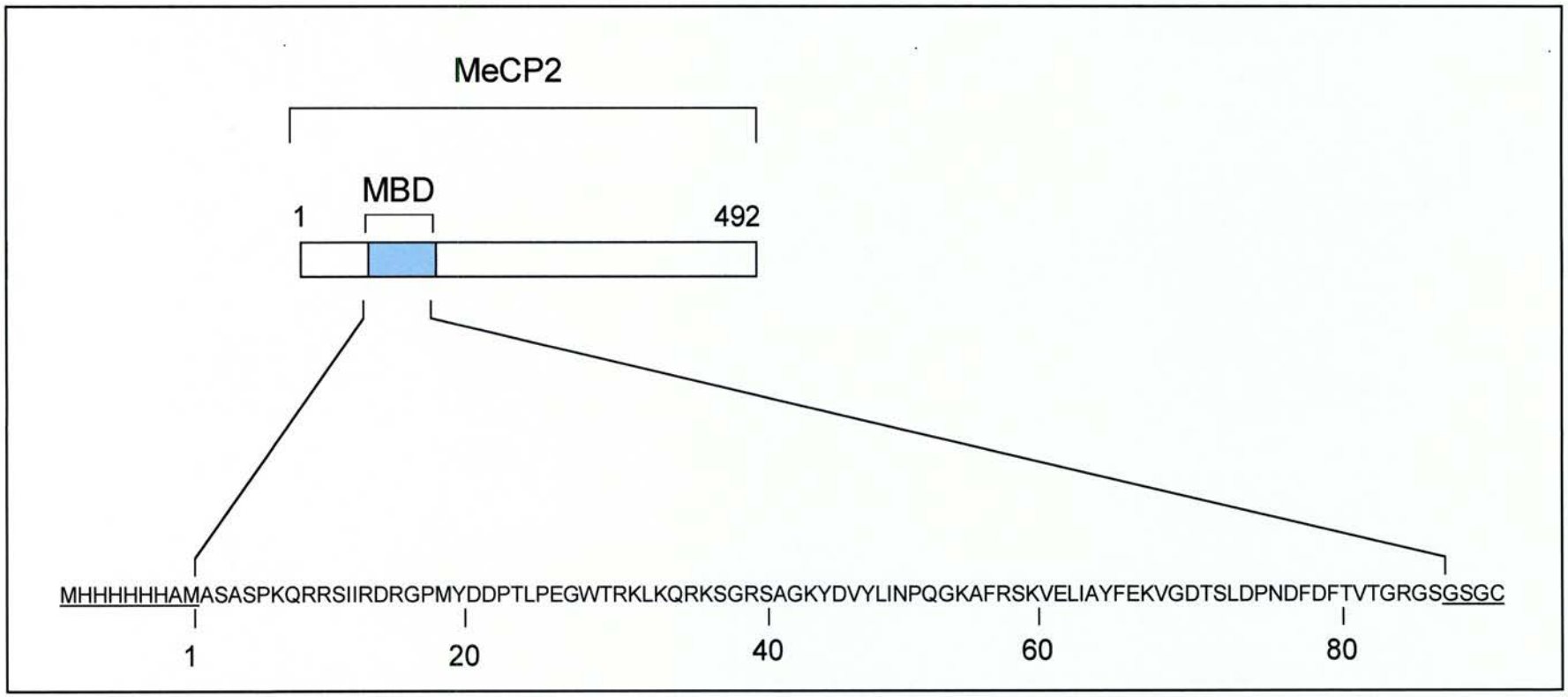


Figure 3.1 Sequence of the methyl-CpG binding domain. Schematic diagram of rat MeCP-2 showing the position of the methyl-CpG binding domain (MBD) as a blue box. The sequence of the recombinant HMBD protein is shown below. Amino acids derived from the plasmid vector sequences are underlined. HMBD has a molecular weight of 11.4 kD. Diagram taken from [Cross et al., 1994].

The run through DNA mixture containing the CpG islands was then methylated at all CpG residues using CpG methylase (New England Biolabs {NEB}). The CpG islands were then selected from the mixture a second time (by exploiting their much increased binding affinity) as fragments which bound strongly when passed through the MBD column. Catch linkers were attached to these purified fragments, and the DNA was amplified 30 times using the polymerase chain reaction (PCR). The catch linkers were then removed by digestion with *Mse*I and the CpG islands were cloned into the *Nde*I site of the pGEM-5Zf(-) (Promega) vector and transformed into SURE bacteria (Stratagene).

3.1.4 Human CpG island library

A human CpG island library had previously been constructed [Cross et al., 1994] using the same methodology. The library was made widely available to the scientific research community, and large numbers of the clones were systematically sequenced by the Sanger Centre. The sequence generated from these *Mse*I fragments provided comparable trinucleotide repeat data from human CpG islands.

3.1.5 Distribution of trinucleotide repeats in mammalian genomes

The distribution of trinucleotide repeat classes in mammalian genomes does not conform to predictions based on the known frequencies of particular nucleotide sequences and their combinations in the genome. The distributions are non random in all species studied with some evidence of trinucleotide repeat clustering, suggesting that the repeats may not arise totally independently. A similar type of repeat clustering has been observed for other types of microsatellite repeats with *Alu* repetitive elements [Beckman and Weber, 1992]. Generally speaking the classes of trinucleotide repeat are distributed in similar ratios throughout different species, for example the AGC repeat is the most common in both mouse and human sequence. This repeat is also restricted to similar genomic regions in both species, being conspicuously absent from introns [Stallings, 1994]. A few obvious exceptions to this rule include the relative abundance of ACC repeats in rodent genomes, compared to human. However, surveys of homologous genes containing trinucleotide repeat tracts reveals a very low degree of conservation between mammalian species. Most frequently a trinucleotide repeat present in one species is not found at all in another, or if it is present it is not in an orthologous position and is often of a different length. For example the AGC repeat array in the rat and human androgen receptor genes are located in distinctly different regions of the gene [Lubahn et al., 1988; Tan et al., 1988] and although the AGC

repeat tract found in the mouse pim-1 proto-oncogene was found in a similar location to the human repeat, it was considerably smaller [Stallings, 1994]. These anomalies between species are even true of repeats within coding regions, indicating that conservation of these arrays is often not critical for protein function.

The relative frequencies of trinucleotide repeats in mouse and man predicted from sequence databases [Stallings, 1994] will be presented along side the CpG island data from this study in the results section of this chapter. Their differences and relevance will be considered in greater detail in the discussion section of this chapter.

Information recently generated from the draft human sequence [Consortium, 2001] is not included in this analysis as no comparable mouse data is currently available.

3.1.6 Trinucleotide repeats in the mouse genome

The naturally occurring phenomenon of dynamic mutation has only been observed in humans to date. Several previous studies have identified extensive polymorphic variation of trinucleotide repeats between inbred strains, but no pathogenic expansions have been reported [Abbott and Chambers, 1994; Chambers and Abbott, 1996; Kim et al., 1997; King et al., 1998]. The mouse homologues of genes which contain expandable repeats in humans, invariably contain significantly smaller repeat tracts that are often disrupted and inherently more stable (discussed in greater depth in chapter 1). Many transgenic studies of large trinucleotide repeats artificially introduced into mice have been undertaken. The early results of these studies indicated that although the repeats underwent small incremental changes in size (biased towards contraction), there was no evidence for the large, dynamic size changes which frequently occurred with this type of repeat in man. At the beginning of my PhD study it certainly looked possible that trinucleotide repeat expansion and its underlying mechanisms may not occur naturally, or even be feasible in the mouse. More recently mice containing a large portion of the Myotonic Dystrophy protein kinase gene with 320 repeats have been produced [Seznec et al., 2000], and the large size changes observed in single generations prove that the phenomenon of dynamic mutation is at least reproducible using the mouse as a model organism.

3.2 Results

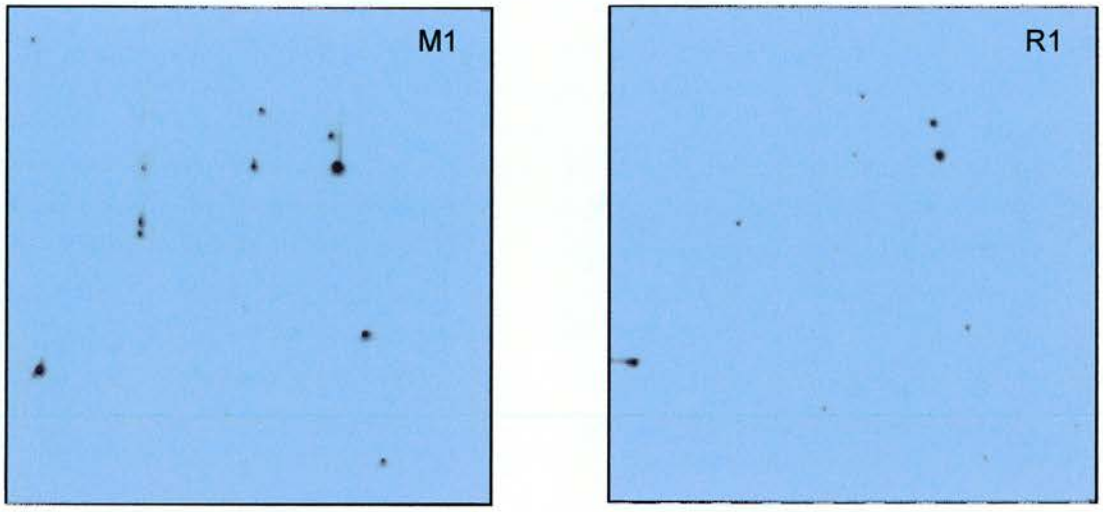
3.2.1 Mouse CpG island library screen for all classes of trinucleotide repeat

Preliminary mouse CpG island library screens were performed to obtain positive controls, and to estimate the relative frequencies of each trinucleotide repeat (TR) class. In order to obtain reasonable numbers of positive clones for each class of trinucleotide repeat, the following number of colony forming units (CFUs) were screened:

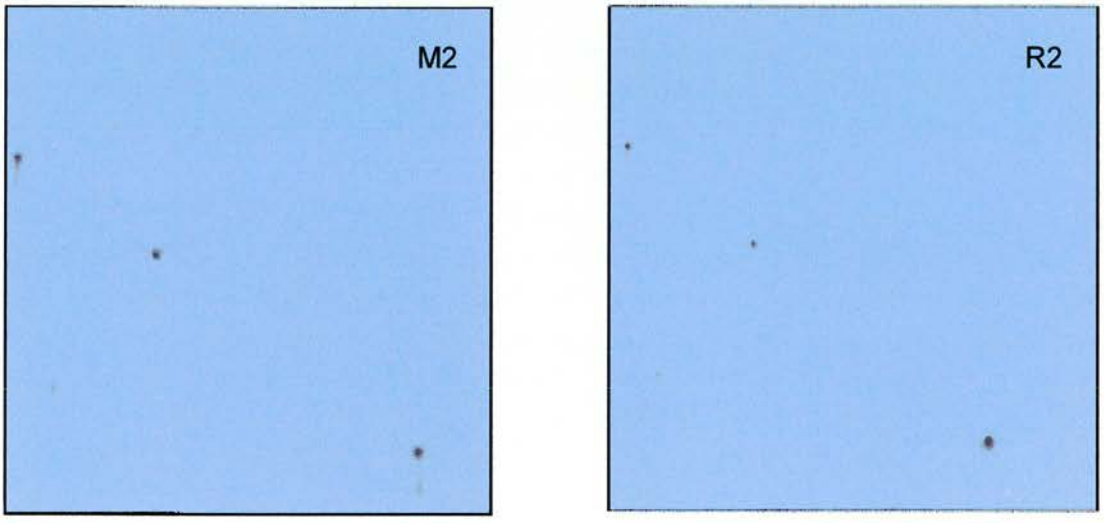
TR Class	Number of CFUs screened
AAT	100 000
AAC	20 000
AAG	100 000
ACT	100 000
ATC	100 000
ACC	40 000
ACG	40 000
AGC	40 000
AGG	20 000
CCG	40 000

The mouse CpG island library was screened with 15mer oligonucleotides from each trinucleotide repeat class, using the methodology described in 2.2.2. Examples of the positively hybridised filters are shown in figure 3.2. Positive clones were picked, cultured and their DNA extracted for further analysis.

Figure 3.2 Autoradiographs of duplicate filters from mouse CpG island library screen for trinucleotide repeats.



a. Autoradiographs of duplicate filters (M1=master1, R1=replica1) showing positive hybridisation to radioactively labelled trinucleotide repeat probes $(CCG)_5$.



b. Autoradiographs of duplicate filters (M2=master2, R2=replica2) showing positive hybridisation to radioactively labelled trinucleotide repeat probes $(AAT)_5$.

3.2.2 Sequence of mouse trinucleotide repeat containing clones

The sequence of each positive clone insert was obtained by sequencing (see 2.2.1.20) with pGEM vector (CpG island library vector was pGEM-5Zf{-}) forward and reverse sequencing primers:

pGEM forward	5' CGG CCG CCT GCA GGT CGA CCT TAA 3'
pGEM reverse	5' AAC GCG TTG GGA GCT CTC CCT TAA 3'

The sequence was read and the size of each trinucleotide repeat noted. Only clones containing trinucleotide repeats of four or more were considered to be positive. The sequence was then run through the NIX program (HGMP-RC) which: searches for CpG islands, promoter regions, and predicted exons; predicts the protein sequence and compares this to known sequences (Swissprot); compares the nucleotide sequence to known mRNAs ESTs and sequences (EMBL); notes poly A tracts, repeat sequences (SINEs) and vector sequences. The size of trinucleotide repeats obtained from the CpG island library screen and any significant homologies or predictions from NIX are shown in Tables 3.1.1-3.1.9. The sequence of each positive clone insert was deposited in EMBL (see # numbers in tables 3.1-3.9).

Identical clones were obtained several times from every class of trinucleotide repeat. This indicates that the library was screened to saturation and that further screening was unlikely to yield significant numbers of unique clones.

3.2.3 CpG status of mouse trinucleotide repeat containing clones

The exact definition of a CpG island remains somewhat contentious, but the lack of DNA methylation of such regions is probably the most significant factor. Directly determining the methylation status of the DNA in these clones would therefore have been the most accurate means of CpG island prediction. However, this would have involved either digesting the DNA with methylation sensitive restriction enzymes and southern blotting, or bisulphite genomic sequencing. Both these methods were considered impractical for analysing the large number of clones isolated in this study. For this reason the CpG status of each trinucleotide repeat containing clone was predicted indirectly, using CpG dinucleotide frequency as an indicator of island presence. In vertebrate DNA the CpG dinucleotide

occurs at only 0.2 to 0.25 % of the frequency expected from the overall base composition, only when the frequency is significantly higher than this can the region be considered a CpG island. Three slightly different methods were employed to determine CpG island status, in an attempt to include several salient island features.

1. The first and arguably most relevant method focuses on the observation that a CpG dinucleotide count higher than 30 per 500 bp (>6 per 100 bp) indicates that the centre of a DNA region will remain unmethylated (shown by the methylation status of restriction sites [Matsuo et al., 1993]). The window of at least 500 bp is critical to this observation, suggesting that CpG islands of a smaller size may not remain unmethylated and therefore can not be considered true islands. So for the purposes of this analysis a stretch of DNA of 500 bp or more, with an observed/expected CpG ratio (CpG score) of 0.6 or higher was considered to be a CpG island. Any clone containing an island of less than 500 bp, but where the island extended to either end, and therefore potentially beyond the *Mse*I fragment (51 bases from the beginning of the sequence, 48 from the end) was also considered an island, as the true extent of the island in its genomic context could not be determined.

2. The second method devised by Gardiner-Garden and Frommer [Gardiner-Garden and Frommer, 1987], defines a CpG island as a stretch of DNA with a % G+C (GC score) of greater than 50 and again an observed/expected CpG ratio (CpG score) of 0.6 or more:

$$\text{Observed/Expected CpG} = \frac{\text{Number of CpG}}{\text{Number of C x number of G}} = \text{x } N$$

N = the total number of nucleotides in the sequence being analysed

This analysis was performed by the NIX program (available HGMP-RC website, www.hgmp.mrc.ac.uk/Registered/Webapp/nix/).

3. The third and possibly the crudest method of identifying a CpG island in this survey, was simply the presence of *Bst*UI restriction sites. The recognition sequence for this restriction enzyme (CGCG) is rare in non island DNA (~1 site/5-10 kb) but frequent within CpG islands (~1 site/125 bp) [Cross et al., 1994]. Therefore any clone containing *Bst*UI sites (sequence screened for restriction sites at www.medkem.gu.se/cutter/) was considered to be a potential CpG island.

Using this method, far fewer than the expected 61% of clones [Cross et al., 1997] were found to contain predicted CpG islands:

CLONES CONTAINING PREDICTED CpG ISLANDS	
Random clones	**TR containing clones**
30 (61%) n = 49	41 (46%) n = 89

To determine if this was a reflection of the overall mouse CpG island library, or specific to trinucleotide repeat containing clones, a random selection of clones from the library were sequenced and their CpG status was also assessed. These clones are described in further detail in Table 3.1.10.

3.2.4 CpG status of random mouse CpG island library clones v. those containing trinucleotide repeats

The percentage of random clones from this library containing CpG islands was compared to that of the trinucleotide repeat containing clones using each of the three determination methods previously described. The data and its significance is presented in Tables 3.1.11-3.1.16

These comparisons show that trinucleotide repeat containing clones from the mouse CpG island library contain significantly fewer CpG islands than random clones, whichever method of determination is used. This suggests that trinucleotide repeats are selected against or depleted in mouse CpG island clones. When breaking these results down further, it is apparent that even when the trinucleotide repeat itself is 66% GC (which if large enough would itself constitute a CpG island), the clones still contain significantly fewer than expected CpG islands (with the exception of *Bst*UI site determination).

Tables 3.1 Clones from the mouse CpG island library

Table 3.1.1 AAT repeat containing clones

Name, accession number & size	Repeat sequence	Sequence homology or prediction	CpG status	*Bst*UI sites
AAT1 **AJ409638** **504 bp**	$(AAT)_{15}$	SINE/B2 (B2_Mm2) & LTR/MaLR (MTC) Predicted gene	No island	None
AAT2 - **153 bp**	$(AAT)_{12}$	-	No island	None
AAT3 **AJ409639** **449 bp**	$(AAT)_{9}$	-	No island	None
AAT4 **AJ409640** **645 bp**	$(AAT)_{8}$ $(AAC)_{6}$ $(ATCC)_{8,7,5,3}$ $(AACC)_{5}$ $(AAAC)_{4}$	SINE/*Alu* (B1F) Predicted gene	No island	None

NIX CpG islands of 500 bp or more = 0/4 0%

NIX any CpG island = 0/4 0%

*Bst*UI CpG islands = 0/4 0%

Accession numbers are only provided for sequences which were successfully amplified by PCR in standard mouse strains.

Table 3.1.2 AAC repeat containing clones

Name, accession number & size	Repeat sequence	Sequence homology or prediction	CpG status	*Bst*UI sites
AAC1 **AJ409619** **304 bp**	$(AAC)_{13}$ $(AGC)_5$	-	No island	None
AAC2 **AJ409620** **875 bp**	$(AAC)_{12}$ $(AGC)_5(AAC)_3$	100% identity with: mouse mRNA for fatso protein 96% identity with: fragments of human chromosome 16 clone RP11-515G11 (htg)	CpG 362-827 bp (465 bp) CpG score 0.81 GC score 54.39	7 sites
AAC3 **AJ409621** **527 bp**	$(AAC)_9$		No island	1 site
AAC4 **AJ409622** **630 bp**	$(AAC)_8$	SINE/*Alu* (B1-F) & LINE/L1 (L1)	No island	None
AAC5 **AJ409623** **547 bp**	$(AAC)_3(AAT)_1(AAC)_8(ACT)_3$	-	No island	None
AAC6 - **147 bp**	$(AAC)_7$	-	No island	None
AAC7 - **371 bp**	$(AAC)_6$	Predicted gene	No island	1 site
AAC8 - **351 bp**	$(AAC)_5$	100% identity with: mouse alpha (1,2) fucosyltransferase gene, complete cds	No island	None
AAC9 **AJ409624** **638 bp**	$(AAC)_5$	95% identity with: human Bac clone GS1-119p5 7q21, and 92% identity with: mouse cDNA IMAGE clone 616327	CpG 51-267 bp (216 bp) CpG score 0.83 GC score 59.18	1 site

AAC10 **AJ409625** **619 bp**	$(AAC)_5$	SINE/*Alu* (B1-MM)	No island	None
AAC11 - **315 bp**	$(AAC)_5$	98% identity with: human PAC clone 528L19 on chromosome 6q23, which contains a polymorphic CA repeat	No island	None
AAC12 **AJ409626** **407 bp**	$(AAC)_5$ poly T	SINE/*Alu* (B1F) & 100% identity with: mouse hepatocyte growth factor-like protein receptor (Ron) gene, complete cds	No island	None
AAC13 **AJ409627** **313 bp**	$(AAC)_5$ poly T	SINE/*Alu* (B1_MM)	No island	1 site
AAC14 **AJ409628** **375 bp**	$(AG)_{17}(AGG)_1(AAC)_5$	SINE/B2 (B3A) & LINE/L1 (Lx9)	No island	None
AAC15 **AJ409629** **625 bp**	$(AAC)_5(AAAC)_4$	Predicted gene	No island	2 sites
AAC16 **AJ409630** **408 bp**	$(AAC)_3(AGC)_3(AAC)_5$	Predicted gene	No island	1 site
AAC17 - **373 bp**	$(AAC)_4(ATC)_1(AAC)_3$ $(GTCT)_7$	100% identity with: mouse cDNA IMAGE clone 479310, and genomic clone RPCI-23-352K24 95% identity with: fragments of human clone RP11-34F13 on chromosome 15	No island	None
AAC18 - **694 bp**	$(AAAC)_3(AAC)_4$ $(AGCG)_6$	85% identity with: fragments of human clone RP11-555K12 on chromosome 4	No island	None
AAC19 - **465 bp**	$(AAC)_4$	Predicted gene	No island	None

NIX CpG islands of 500 bp or more = 2/19 10.5%

NIX any CpG island = 2/19 10.5%

*Bst*UI CpG islands = 7/19 37%

Table 3.1.3 AAG repeat containing clones

Name, accession number & size	Repeat sequence	Sequence homology or prediction	CpG status	*Bst*UI sites
AAG1 **AJ409631** **392 bp**	$(AAG)_{35, 3}$ $(AGG)_{13, 3}$	Predicted gene	No island	None
AAG2 - **602 bp**	$(AAG)_{35, 3,}$ $(AGG)_{8, 6, 5, 4, 4, 3, 3, 3, 3, 3}$ $(AAGAGG)_6$ $(AAGAGGAGG)_6$	Predicted gene	No island	None
AAG3 - **362 bp**	$(AAG)_{29}$	90% identity with: fragments of mouse clone RP21-247L16 on chromosome 10 Predicted gene	No island	1 site
AAG4 **AJ409632** **351 bp**	$(AAG)_{26}$	SINE/B2 (B2_Mm1) Predicted gene	No island	None
AAG5 **AJ409633** **583 bp**	$(AAG)_{25}$	92% identity with: mouse mRNA for ribosomal protein L6 and mouse clone RP22-395M5 on chromosome 11 85% identity with: human mRNA for ribosomal protein L6 and human clone RP11-263014 on chromosome 18	No island	None
AAG6 **AJ409634** **244 bp**	$(AAG)_{25, 4, 3, 3}$ $(AGG)_{6, 3, 3, 3}$	Predicted gene	No island	None
AAG7 - **400 bp**	$(AAG)_{24, 14}$ $(AG)_{32, 9, 5, 4}$ $(AGGG)_{4, 4, 4}$ $(AACGAAG)_7$	Predicted gene	No island	None
AAG8 - **197 bp**	$(AAG)_{23}$	Predicted gene	No island	None
AAG9 **AJ409635** **670 bp**	$(AAG)_{20}(GA)_1(GAAGGA)_7$	100% identity with: mouse clone RP23-310A8	No island	None

AAG10 **AJ409636** **312 bp**	$(AAG)_{18}$ $(AGG)_{12,3,3}$	Predicted gene	No island	None
AAG11 - **810 bp**	$(AAG)_{17,17,}$ $(AGG)_6$ $(AG)_{12}$ $(AGGG)_{11,3}$ $(AGGAAG)_7(AAGAGG)_4$	LINE/L1 (Lx), 91% identity with: mouse chromosome X BAC B178A13 and mouse T cell receptor gamma locus	No island	None
AAG12 **AJ409637** **1095 bp**	$(AAG)_{11,11,10,9,7,7,3,3,3,3,3,3,3}$ $(AGG)_7$	SINE/*Alu* (B1_MM) 100% identity with: mouse cDNA IMAGE clone 642165 Predicted gene	CpG 490-848 bp (358 bp) CpG score 0.90 GC score 49.84	5 sites

NIX CpG islands of 500 bp or more = 1/12 8%

NIX any CpG islands = 1/12 8%

*Bst*UI CpG islands = 2/12 17%

Table 3.1.4 ACG repeat containing clones

Name, accession number & size	Repeat sequence	Sequence homology or prediction	CpG status	*Bst*UI sites
ACG1 - **834 bp**	$(ACG)_5$	100% identity with: mouse cDNA IMAGE clone 1908172 and 1907713 90% identity with: human cDNA IMAGE clone 664913 and cerebellum II cDNA clone EST29620 Predicted gene	CpG 245-615 bp (370 bp) CpG score 1.11 GC score 64.76	7 sites
ACG2 **AJ409647** **830 bp**	$(ACG)_5$ $(AC)_4(AA)_1(AC)_7$	100% identity with: mouse cDNA clone 1180507 and genomic clone RPCI-23-473D16.TV Predicted gene	CpG 180-441 bp (261 bp) CpG score 1.02 GC score 51.77	None

NIX CpG islands of 500 bp or more = 0/2 0%

NIX any CpG island = 2/2 100%

*Bst*UI CpG islands = 1/2 50%

Table 3.1.5 ATC repeat containing clones

Name, accession number & size	Repeat sequence	Sequence homology or prediction	CpG status	*Bst*UI sites
ATC1 **AJ409662** **522 bp**	$(ATC)_6$	LINE/L1 (L1M4) & LINE/L1 (Lx7) Predicted gene	No island	None
ATC2 - **573 bp**	$(GTT)_5(ATT)_1(ATC)_6$	-	No island	None
ATC3 **AJ409663** **565 bp**	$(ATC)_5$	100% identity with: mouse HS1-associating protein (mHAX-1s) mRNA and pseudogene mRNA 90% identity with: human clone RP11-137P24 and RP11-214A22	No island	1 site
ATC4 **AJ409664** **343 bp**	$(ATC)_5A(ATC)_3$	LINE/L2 (L2)	No island	None
ATC5 **AJ409665** **907 bp**	$(ATC)_5(ACC)_3$	SINE/B2 (B3) & LINE/L1 (L1)	No island	None
ATC6 - **388 bp**	$(ATC)_4$	-	No island	None
ATC7 - **757 bp**	$(ATC)_4$	94% identity with: mouse cDNA IMAGE clone 1907592 & brain cDNA clone MNCb-1213 predicted gene	CpG 51-500 bp (449 bp) CpG score 0.98 GC score 64.82	6 sites
ATC8 - **521 bp**	$(ATC)_4$	SINE/*Alu* (PB1D10) & SINE/B4 (B4A)	No island	None
ATC9 - **450 bp**	$(ATC)_4$	98% identity with: fragments of mouse clone RP23-15K8 on chromosome 6	No island	None

NIX CpG islands of 500 bp or more = 1/9 11%

NIX any CpG island = 1/9 11%

BstUI CpG islands = 2/9 22%

Table 3.1.6 ACC repeat containing clones

Name, accession number & size	Repeat sequence	Sequence homology or prediction	CpG status	*Bst*UI sites
ACC1 **AJ409641** **346 bp**	$(ACC)_{15}$	-	No island	None
ACC2 **AJ409642** **607 bp**	$(AC)_6(ACC)_{10}$	96% identity with: mouse clone RP23-75M14 and 90% identity with clone ct&-596k8 on chromosome 6, similar to mouse zinc-finger protein gb:L16904 mRNA	No island	1 site
ACC3 **AJ409643** **704 bp**	$(ACC)_6$	-	CpG 195-418 bp (223 bp) CpG score 0.70 GC score 56.78	1 site
ACC4 **AJ409644** **809 bp**	$(ACC)_6(GCC)_3$	100% identity with: mouse cDNA IMAGE clone 837282	CpG 431-761 bp (330 bp) CpG score 0.74 GC score 59.16	5 sites
ACC5 **AJ409645** **585 bp**	$(ACC)_5$	-	No island	None
ACC6 - **512 bp**	$(ACC)_4$	100% identity with: mouse cDNA IMAGE clone 514609, similar to mouse notch-1 mRNA gb:Z11886	No island	1 site
ACC7 **AJ409646** **764 bp**	$(ACC)_4(GCC)_1(ACC)_4$	90% identity with: human clone 372K1 on chromosome 6q24, contains EST, STS, GSS and CpG island 65% identity with: basic helix-loop-helix protein	No island	2 sites
ACC8 - **589 bp**	$(ACC)_4$ $(GCC)_3$	SINE/*Alu* (B1F) 100% identity with: mouse cDNA clone 758218 90% identity with: fragments of human clone RP11-2C24	CpG 51-410 bp (359 bp) CpG score 0.90 GC score 60.69	3 sites

NIX CpG islands of 500 bp or more = 2/8 25%

NIX any CpG islands = 3/8 37.5%

*Bst*UI CpG islands = 6/8 75%

Table 3.1.7 AGC repeat containing clones

Name, accession number & size	Repeat sequence	Sequence homology or prediction	CpG status	*BstUI* sites
AGC1 **AJ409648** **331 bp**	$(AGC)_{23}$	100% identity with: mouse cDNA IMAGE clones 334556 and 1246452 90% identity with: human HSPC283 mRNA	No island	None
AGC2 **AJ409649** **483 bp**	$(AGC)_{20}$	SINE/*Alu* (PB1D10) Predicted gene	No island	None
AGC3 **AJ409650** **727 bp**	$(AGC)_{19}(AAC)_{11}$	LINE/L1 (Lx7), 100% identity with: mouse mRNA for J domain protein 1 (Jdp1)	CpG 437-679 bp (242 bp) CpG score 0.68 GC score 45.72	1 site
AGC4 - **772 bp**	$(AGC)_{10}$	100% identity with: mouse mRNA for GD1 alpha synthase	No island	3 sites
AGC5 **AJ409651** **518 bp**	$(ACCACCACT)_5(ACC)_1(AGC)_8$ $(CT)_{29}(GT)_{11}$	-	No island	None
AGC6 - **262 bp**	$(AGC)_8$	95% identity with: human CpG island clone 133e4, cosmid clone 34a5 on chromosome 9p22 Predicted gene	No island	2 sites
AGC7 **AJ409652** **483 bp**	$(AGC)_7$	SINE/*Alu* (PB1D10) Predicted gene	No island	None
AGC8 - **992 bp**	$(AGC)_{7,3,3,3}$	100% identity with mouse cDNA IMAGE clone 1764543 95% identity with: fragments of human clone RP11-43816 on chromosome 18 65% similar to human TED protein	CpG 233-649 bp (416 bp) CpG score 0.74 GC score 65.17	4 sites
AGC9 - **693 bp**	$(AGC)_5$ $(AAC)_3$	-	No island	None

AGC10 **AJ409653** **955 bp**	$(AGC)_5$	**SINE/B2 (B3A)**	**CpG 293-529 bp (236 bp) CpG score 0.65 GC score 57.53**	**1 site**

NIX CpG islands of 500 bp or more = 1/10 10%

NIX any CpG island = 3/10 30%

*Bst*UI CpG islands = 5/10 50%

Table 3.1.8 AGG repeat containing clones

Name, accession number & size	Repeat sequence	Sequence homology or prediction	CpG status	*Bst*UI sites
AGG1 **AJ409654** **1013 bp**	$(AGG)_{16}$ $(AGGC)_3$	95% identity with: fragments of human clone RP11-118M12 on chromosome 2 Predicted gene	No island	None
AGG2 **AJ409655** **786 bp**	$(AGG)_{13}$	LTR (RMER15) 90% identity with: mouse mRNA for ribosomal-like protein L41 and clone RP23-331I17 on chromosome 11	No island	None
AGG3 **AJ409656** **338 bp**	$(AGG)_{12,3}$ $(AC)_{6,4,3}$	Predicted gene	No island	None
AGG4 **AJ409657** **1198 bp**	$(ACT)_6(CCT)_8$	Predicted gene	CpG 115-779 bp (664 bp) CpG score 0.87 GC score 64.28	7 sites
AGG5 **AJ409658** **874 bp**	$(AGG)_8$ $(CGG)_4$	100% identity with: mouse cDNA IMAGE clone 2373246 and 2649207 90% identity with: fragments of human clone CTD-2347K12 on chromosome 5	CpG 51-694 bp (643 bp) CpG score 0.78 GC score 66.16	7 sites
AGG6 - **552 bp**	$(AGG)_7$	SINE/MIR (MIR)	No island	None
AGG7 - **337 bp**	$(AGG)_5$	100% identity with: mouse clone RPCI-23-367F3 and RPCI-23-1N1.TJ Predicted gene	No island	None
AGG8 **AJ409659** **727 bp**	$(AGG)_5$	LTR/MaLR (MTC) 100% identity with: mouse mRNA for carbonic anhydrase-related polypeptide	CpG 51-305 bp (254 bp) CpG score 0.72 GC score 65.62	5 sites
AGG9 - **587 bp**	$(AGG)_5$	Predicted gene	No island	2 sites
AGG10 **AJ409660** **653 bp**	$(AGG)_5$	88% identity with: fragments of human clone RP11-560I19 67% similar to zinc finger protein FEZ	CpG 213-605 bp (392 bp) CpG score 0.74 GC score 54.03	5 sites

AGG11 **AJ409661** **566 bp**	$(AGG)_{5,4,4,4}$	Predicted gene	No island	None
AGG12 - **391 bp**	$(AGG)_3(AAG)_1(AGG)_5$	96% identity with: fragments of human cDNA clones 729432, RP11-302L10 and RP11-301K23 on chromosome 6	No island	1 site
AGG13 - **610 bp**	$(AGG)_5$ $(CGG)_3$	95% identity with: fragments of human BAC clone CIT987SK on chromosome 16	CpG 85-320 bp (235 bp) CpG score 0.66 GC score 62.53	1 site
AGG14 - **818 bp**	$(AGG)_4$	100% identity with: mouse mRNA for X-linked PEST-containing transporter (Xpct)	CpG 71-366 bp (295 bp) CpG score 0.75 GC score 66.03	2 sites

NIX CpG islands of 500 bp or more = 4/14 29%

NIX any CpG islands = 6/14 43%

*Bst*UI CpG islands = 8/14 57%

Table 3.1.9 CCG repeat containing clones

Name, accession number & size	Repeat sequence	Sequence homology or prediction	CpG status	*Bst*UI sites
CCG1 **AJ409666** **597 bp**	$(CCG)_8$	100% identity with: fragments of mouse clone CT7-378P20 on chromosome 6, similar to mRNA for 043691 repressor protein	CpG 51-504 bp (453 bp) CpG score 0.90 GC score 66.3	3 sites
CCG2 - **796 bp**	$(CCG)_6$	94% identity with: fragments of human cosmid clones 7H3, 14D7, C1230, 11E7, F1096, A12197, 12G8, A09100 on chromosome Xq28 Predicted gene	No island	3 sites
CCG3 **AJ409667** **445 bp**	$(CCG)_5$	-	CpG 111-397 bp (286 bp) CpG score GC score 56.55	2 sites
CCG4 - **775 bp**	$(CCG)_5$ $(CCT)_3$	100% identity with: mouse cDNA IMAGE clone 944887, similar to human Sua1 92% identity with: fragments of human clone RP11-124P12 on chromosome 19	CpG 163-641 bp (478 bp) CpG score 0.97 GC score 65.37	4 sites
CCG5 - **769 bp**	$(CCG)_4$	95% identity with: fragments of human clone RP11-435P9 Predicted gene	No island	4 sites
CCG6 - **925 bp**	$(CCG)_4$	92% identity with: fragments of human clone RP11-383C5 Predicted gene	CpG 246-877 bp (631 bp) CpG score 0.81 GC score 61.51	8 sites
CCG7 - **615 bp**	$(CCG)_4$	95% identity with: fragments of human clone RP11-138018	CpG 51-306 bp (255 bp) CpG score 0.77 GC score 63.68	1 site
CCG8 - **597 bp**	$(CCG)_4$	-	CpG 157-549 bp (392 bp) CpG score 0.99 GC score 57.38	7 sites
CCG9 - **903 bp**	$(CCG)_4$ $(CCA)_3$	rRNA (5S) 100% identity with: mouse cDNA clone UI-M-AQ1-ady-c-06-0-UI 92% identity with: fragments of a novel human gene on chromosome 13, mRNA for the BRCA2 region and PAC 248015 on 13q12-q13	CpG 467-761 bp (294 bp) CpG score 0.74 GC score 59.05	None

CCG10 **AJ409669** **429 bp**	$(CCG)_{4,3}$	-	CpG 76-341 bp (265 bp) CpG score 0.68 GC score 64.56	1 site
CCG11 **AJ409669** **760 bp**	$(CCG)_{4,4,4,4,4,3,3,3,3}$	100% identity with: fragments of mouse cDNA IMAGE clone 2645531	CpG 51-645 bp (594 bp) CpG score 0.96 GC score 72.12	2 sites

NIX CpG islands of 500 bp or more = 6/11 54%

NIX any CpG islands = 9/11 82%

*Bst*UI CpG islands = 10/11 91%

Average clone length 578 bp

Average CpG island length 372 bp

Table 3.1.10 Random mouse CpG island library clones

Name, accession number and size	Clone sequence homology or prediction	CpG status	*BstUI* sites
1 - **550 bp**	-	CpG 69-315 bp (246 bp) CpG score 0.74 GC score 62.19	4 sites
2 - **455 bp**	SINE/ B4 (B4A)	-	None
3 - **429 bp**	100% identity with: mouse mRNA for external transcribed spacer B2 element and 18S ribosomal RNA	CpG 51-381 bp (330 bp) CpG score 0.79 GC score 52.66	1 site
4 - **760 bp**	-	-	None
5 - **441 bp**	100% identity with: mouse cDNA IMAGE clone 1362103	-	None
6 - **600 bp**	-	CpG 297-532 bp (235bp) CpG score 0.65 GC score 51.05	1 site
7 - **604 bp**	-	CpG 339-556 bp (217 bp) CpG score 0.55 GC score 48.26	2 sites
8 - **548 bp**	-	CpG 51-264 bp (213 bp) CpG score 0.82 GC score 69.74	3 sites
9 - **613 bp**	-	-	

10 - **588 bp**	73% similar to putative binding-protein-dependant transport systems inner membrane protein	CpG 51-451 bp (400 bp) CpG score 1.00 GC score 54.96	5 sites
11 - **315 bp**	100% identity with: human clone RP11-103E2 on chromosome 15	CpG 51-267 bp (216 bp) CpG score 0.71 GC score 45.36	3 sites
12 - **505 bp**	51% similar to DRPLA protein	CpG 101-457 bp (356 bp) CpG score 0.98 GC score 48.32	1 site
13 - **357 bp**	100% identity with: mouse cDNA IMAGE clone 892409, similar to mouse NADH-ubiquinone oxidoreductase chain 1	-	None
14 - **632 bp**		CpG 299-584 bp (285 bp) CpG score 0.71 GC score 53.30	2 sites
15 - **430 bp**	LINE/L1 (L1)	-	None
16 - **554 bp**	65% similar to 2-keto-4-hydroxyglutarate aldolase	CpG 126-506 bp (380 bp) CpG score 0.94 GC score 49.26	5 sites
17 - **495 bp**	% identity with: mouse histone H4 gene	CpG 196-447 bp (251 bp) CpG score 0.85 GC score 52.73	4 sites
18 - **867 bp**	-	CpG 106-450 bp (344 bp) CpG score 0.84 GC score 62.41	3 sites
19 - **547 bp**	100% identity with: mouse clone RP23-240H6 on chromosome 5, and mouse putative hepatic transcription factor (Wbscr14) mRNA	-	None
20 - **573 bp**	100% identity with: mouse 45s pre rRNA gene	CpG 81-477 bp (396 bp) CpG score 1.08 GC score 57.51	2 sites

21 - **869 bp**	90% identity with: fragments of human clone RP5-837M10	-	4 sites
22 - **686 bp**	tRNA (tRNA-ala-GCG)	-	2 sites
23 - **473 bp**	60% similarity to amino acid ABC transporter, permease protein	CpG 51-425 bp (374 bp) CpG score 0.80 GC score 48.77	2 sites
24 - **574 bp**	100% identity with: mouse 45S pre rRNA gene	CpG 98-478 bp (380 bp) CpG score 1.06 GC score 58.3	2 sites
25 - **539 bp**	-	-	None
26 - **831 bp**	100% identity with: mouse cDNA IMAGE clone 1889991	CpG 52-783 bp (731 bp) CpG score 0.87 GC score 59.05	6 sites
27 - **848 bp**	-	CpG 350-800 bp (450 bp) CpG score 0.76 GC score 58.55	3 sites
28 - **493 bp**	100% identity with: mouse type II keratin submit protein mRNA	CpG 79-323bp (244bp) CpG score 0.89 GC score 69.14	None
29 - **922 bp**	100% identity with: mouse Dlx-2 gene	CpG 51-759 bp (708 bp) CpG score 0.69 GC score 62.70	2 sites
30 - **625 bp**	100% identity with: mouse cDNA IMAGE clone 2192526	CpG 51-577 bp (526 bp) CpG score 0.81 GC score 60.76	6 sites
31 - **507 bp**	100% identity with: mouse mRNA for testican	-	1 site

32 - **658 bp**	-	CpG 341-610 bp (269 bp) CpG score 0.68 GC score 56.20	1 site
33 - **902 bp**	-	-	5 sites
34 - **396 bp**	100% identity with: mouse G1 cyclin-Cdk protein kinase inhibitor p27 mRNA	CpG 123-348 bp (225 bp) CpG score 0.68 GC score 44.76	1 site
35 - **834 bp**	100% identity with: mouse p27 promoter region and partial 5' UTR	CpG 86-336 bp (250 bp) CpG score 0.68 GC score 54.65	5 sites
36 - **584 bp**	100% identity with: mouse cDNA IMAGE clones 3153983 and 337291	CpG 83-336 bp (253 bp) CpG score 0.87 GC score 60.13	1 site

NIX CpG islands of 500 bp or more = 17/36 47%

NIX any CpG islands = 24/36 67%

*Bst*UI CpG islands = 27/36 75%

Average clone length 588 bp

Average CpG island length 345 bp

Table 3.1.11 Number of random clones from mouse CpG island library containing CpG islands of more than 500 bp (CpG score 0.6+)

	Number/% of clones containing CpG islands	Number/% of clones not containing CpG islands
n = 36	17/47%	19/53%

Table 3.1.12 Number of trinucleotide repeat containing clones from mouse CpG island library containing CpG islands of more than 500 bp (CpG score 0.6+)

Compared to random clones (Table 3.1.11 above)

% GC of TR	CpG island containing clones	Non CpG island containing clones	Chi-square value	*p*	*Q*	Signif
100% **n = 11**	6/55%	5/45%	0.1808	1	0.6707	**No**
66% **n = 34**	7/21%	27/79%	5.5053	0.025	0.01896	**Yes**
33% **n = 40**	4/10%	36/90%	13.1280	0.001	0.0003	**Yes**
0% **n = 4**	0/0%	4/100%	3.2850	0.1	0.0700	**Yes**
Total **n = 89**	17/19%	72/81%	10.2564	0.01	0.0014	**Yes**

p = Probability of occurrence by chance

Q = Probability of non chance occurrence

Significance = to one degree of freedom

Table 3.1.13 Number of random clones from mouse CpG island library containing CpG islands (as determined by NIX)

	Number/% of clones containing CpG islands	Number/% of clones not containing CpG islands
n = 36	24/67%	12/33%

Table 3.1.14 Number of trinucleotide repeat containing clones from mouse CpG island library containing CpG islands (as determined by NIX)

Compared to random clones (Table 3.1.13 above)

% GC of TR	CpG island containing clones	Non CpG island containing clones	Chi-square value	*p*	*Q*	Signif
100% **n = 11**	9/82%	2/18%	0.9248	1	0.3362	**No**
66% **n = 34**	14/41%	20/59%	4.5782	0.05	0.0324	**Yes**
33% **n = 40**	4/10%	36/90%	26.1476	0.001	3.1641	**Yes**
0% **n = 4**	0/0%	4/100%	6.6667	0.01	0.0098	**Yes**
Total **n = 89**	27/32%	62/70%	14.0062	0.001	0.0002	**Yes**

Table 3.1.15 Number of random clones from mouse CpG island library containing CpG islands (as determined by presence of *Bst*UI restriction sites)

	Number/ % of clones containing CpG islands	Number/ % of clones not containing CpG islands
n = 36	27/75%	9/25%

Table 3.1.16 Number of trinucleotide repeat containing clones from mouse CpG island library containing CpG islands (as determined by presence of *Bst*UI restriction sites)

Compared to random clones (Table 3.1.15 above)

% GC of TR	CpG island containing clones	Non CpG island containing clones	Chi-square value	*p*	*Q*	Signif
100% **n = 11**	10/91%	1/9%	1.2732	1	0.2592	**No**
66% **n = 34**	20/59%	14/41%	2.0741	0.2	0.1498	**No**
33% **n = 40**	11/30%	29/70%	17.1	0.001	0.0000	**Yes**
0% **n = 4**	0/0%	4/100%	9.2308	0.01	0.0024	**Yes**
Total **n = 89**	41/46%	48/54%	8.6495	0.01	0.0032	**Yes**

3.2.5 Human CpG island library screen for all classes of trinucleotide repeat

To determine if trinucleotide repeats were also selected against in human CpG islands, a comparable human CpG island library needed to be screened for all 10 classes of trinucleotide repeat. A human CpG island library was previously constructed using the same methodology as the mouse library [Cross et al., 1994]. The Sanger centre had systematically sequenced a number of clones from this human CpG island library and the sequence was available at their website (www.sanger.ac.uk/HGP/blast_server.shtml) and had also been deposited at EMBL. Instead of screening the human library itself for each class of trinucleotide repeat, all the currently sequenced clones were screened by BLAST analysis for each 15mer trinucleotide repeat using the Sanger website.

Many of the human CpG island library clones appeared much smaller than those from the mouse library. To assess whether the Sanger Centre had sequenced fragments of clones or entire clone inserts, each sequence was analysed further. Where possible the homologous sequence to each clone (generated by the human genome sequencing endeavour) along with flanking sequence was put through a restriction endonuclease mapping program (www.medkem.gu.se/cutter/) for *Mse*I. The resulting 'virtual' *Mse*I fragments were then compared to the sequences produced by the Sanger centre. In the majority of cases the entire clone had been sequenced, when this wasn't so, the entire virtual fragments were used instead in the following analyses.

3.2.6 Sequence of human trinucleotide repeat containing clones

As with the mouse library, only clones containing trinucleotide repeats of four or more were considered to be positive. The sequence from these clones was also run through the NIX program (HGMP-RC). The size of trinucleotide repeats obtained from this 'dry' library screen, and any significant homologies or predictions from the NIX analysis are shown in Tables 3.2.1-3.2.9.

3.2.7 CpG status of human CpG island library clones

The trinucleotide repeat containing clones from the human CpG island library had their CpG status assessed using the same three criterion used on the mouse clones. A random selection of clones from the human library (the first clones sequenced from it [Cross et al., 1994], see

Tables 3.2 Clones from the human CpG island library

Table 3.2.1 AAT repeat containing clones

Name, number and size	Repeat sequence	Sequence homology or prediction	CpG status NIX	CpG status *Bst*UI sites
AAT1 CpG152e6 200 bp	$(AAT)_7$	SINE/*Alu* (*Alu*Sg/x)	No island	1 site
AAT2 CpG53g12 210 bp	$(AAT)_6$	100% identity with: human chromosome X, complete sequence	No island	4 sites
AAT3 CpG202a 390 bp	$(AAT)_5$	SINE/Alu (*Alu*Sg)	No island	None
AAT4 CpG219a7 283 bp	$(AAT)_5$	SINE/*Alu* (*Alu*Sq) 100% identity with: human clone RP11-46H1 on chromosome 2	No island	None
AAT5 CpG75g5 1508 bp	$(AAT)_5$	SINE/*Alu* (*Alu*Jo) 100% identity with: human clone RP11-46C6 on chromosome 19	CpG 774-1370 bp (596 bp) CpG score 0.74 GC score 68.74	4 sites

NIX CpG islands of 500 bp and more = 1/5 20%

NIX any CpG island = 1/5 20%

*Bst*UI CpG islands = 3/5 60%

Table 3.2.2 AAC repeat containing clones

Name, number and size	Repeat sequence	Sequence homology or prediction	CpG status NIX	CpG status *Bst*UI sites
AAC1 CpG14c4 294 bp	$(AAC)_9$	LTR/Retroviral (LTR64) SINE/*Alu* (*Alu*Sg/x)	No island	None
AAC2 CpG162c10 210 bp	$(AAC)_7$	SINE/*Alu* (*Alu*Sc)	No island	None
AAC3 CpG12g8 323 bp	$(AAC)_7$	SINE/*Alu* (*Alu*Sc)	No island	2 sites
AAC4 CpG30c3 285 bp	$(AAC)_7$	95% identity with: human clone CTD-3098H1 on chromosome 19	No island	2 sites
AAC5 CpG13b12 318 bp	$(AAC)_6$	SINE/*Alu* (AluY)	No island	1 site
AAC6 CpG235a5 222 bp	$(AAC)_6$	100% identity with: fragments of human IMAGE clone 1626799	No island	None
AAC7 CpG13c2 242bp	$(AAC)_6$	SINE/*Alu* (*Alu*Sg/x)	No island	None
AAC8 CpG63c2 233 bp	$(AAC)_6$	SINE/*Alu* (*Alu*Sq)	No island	None

AAC9 **CpG99f6** **259 bp**	$(AAC)_5$	SINE/*Alu* (*Alu*Sg/x)	No island	None
AAC10 **CpG282a4** **251 bp**	$(AAC)_5$	100% identity with: human clone RP11-256B12 on chromosome 3	No island	None
AAC11 **CpG96b8** **316 bp**	$(AAC)_{4,4,4,4}$	SINE/*Alu* (*Alu*Sg/x)	No island	None
AAC12 **CpG165d10** **257 bp**	$(AAC)_4$	LINE/L1 (L1)	No island	None
AAC13 **CpG205a3** **700 bp**	$(AAC)_4$	100% identity with: human clone RP5-1132F1	CpG 278-485 bp (207 bp) CpG score 0.79 GC score 60.4	1 site

NIX CpG islands of 500 bp or more = 0/13 0%

NIX any CpG island = 1/13 8%

*Bst*UI CpG islands = 4/13 31%

Table 3.2.3 AAG repeat containing clones

Name, number and size	Repeat sequence	Sequence homology or prediction	CpG status NIX	CpG status *Bst*UI sites
AAG1 CpG 46c4 202 bp	$(AAG)_7$	100% identity with: human clone RP11-386F9 on chromosome 17	No island	4 sites
AAG2 CpG194f12 260bp bp	$(AAG)_4$	100% identity with: human cosmid clone R30102:R29350:R27740 on chromosome 19 contains MEF2B	No island	1 site
AAG3 CpG70c5 275 bp	$(AAG)_4$	SINE/*Alu* (*Alu*Sp) 100% identity with: human clone CTB-77M18 on chromosome 5	No island	1 site
AAG4 CpG32f1 279 bp	$(AAG)_4$	100% identity with: human clone RP11-409K20	No island	1 site
AAG5 CpG121b8 244 bp	$(AAG)_4$ $(TG)_{10}$	98% identity with: human clone RP11-94F16	No island	None
AAG6 CpG77c3 275 bp	$(AAG)_4$	100% identity with: human clone RP11-113N11 on chromosome 3	No island	1 site
AAG7 CpG190d10 369 bp	$(AAG)_4$	100% identity with: human sonic hedgehog protein (SHH)	CpG 112-321 bp (209 bp) CpG score 0.74 GC score 49.67	2 sites

NIX CpG islands of 500 bp and more = 1/7 14%

NIX any CpG island = 1/7 14%

*Bst*UI CpG islands = 6/7 86%

Table 3.2.4 ACT repeat containing clones

Name, number and size	Repeat sequence	Sequence homology or prediction	CpG status NIX	CpG status *Bst*UI
ACT1 CpG18c12 207 bp	$(ACT)_4$	-	No island	None

NIX CpG islands of 500 bp and more = 0/1 0%

NIX any CpG island = 0/1 0%

*Bst*UI CpG islands = 0/1 0%

Table 3.2.5 ATC repeat containing clones

Name, number and size	Repeat sequence	Sequence homology or prediction	CpG status NIX	CpG status *Bst*UI sites
ATC1 CpG30b12 473 bp	$(ATC)_4(CTC)_3(ATC)_3$	100% identity with: human clone RP11-619L19 on chromosome 18	CpG 151-427 bp (276 bp) CpG score 0.83 GC score 57.05	2 sites
ATC2 CpG121b11 1020 bp	$(ATC)_4$	SINE/*Alu* (*Alu*Jo/FRAM) 100% identity with: human clone RP11-415G10 on chromosome 11	CpG 321-971 bp (650 bp) CpG score 0.85 GC score 69.65	13 sites
ATC3 CpG19c7 712 bp	$(ATC)_4$	SINE/*Alu* (*Alu*Sx) SINE/MIR (MIR) 100% identity with: human clone RP11-186B13 on chromosome 18	No island	None
ATC4 CpG258c3 339 bp	$(ATC)_4$	LINE/L1 (L1P)	No island	None
ATC5 CpG185b 279 bp	$(ATC)_4$	100% identity with: human clone RP11-139021	No island	None

NIX CpG islands of 500 bp or more = 2/5 40%

NIX any CpG island = 2/5 40%

*Bst*UI CpG islands = 2/5 40%

Table 3.2.6 ACC repeat containing clones

Name, number and size	Repeat sequence	Sequence homology or prediction	CpG status NIX	CpG status *Bst*UI
ACC1 CpG278h9 668 bp	$(ACC)_4$	100% identity with: human HOX A1 homeodomain protein (HOXA1)	CpG 199-620 bp (421 bp) CpG score 0.66 GC score 49.64	3 sites
ACC2 CpG252g9 398 bp	$(ACC)_4$	100% identity with: human alternatively spliced interferon receptor (IFNAR2) gene	No island	None
ACC3 CpG75a10 205 bp	$(ACC)_4$	SINE/MER1_type (MER1B)	No island	1 site
ACC4 CpG177g5 309 bp	$(ACC)_4$	100% identity with: human clone RP11-571B6 on chromosome 3	No island	None
ACC5 CpG73e10 229 bp	$(ACC)_4$	LINE/L1 (L1M4)	No island	None

NIX CpG islands of 500 bp or more = 1/5 20%

NIX any CpG island = 1/5 20%

*Bst*UI CpG islands = 2/5 40%

Table 3.2.7 AGC repeat containing clones

Name, number and size	Repeat sequence	Sequence homology or prediction	CpG status NIX	CpG status *Bst*UI
AGC1 **CpG174d12** **1021 bp**	$(AGC)_{10}$	100% identity with: human clone RP11-515017 on chromosome 17	CpG 400-980 bp (580 bp) CpG score 0.82 GC score 64.18	6 sites
AGC2 **CpG1d2** **1080 bp**	$(AGC)_9$ $(CCG)4$	100% identity with: human clone RP11-149G19 on chromosome 11	CpG 51-1032 bp (981 bp) CpG score 0.91 GC score 76.13	26 sites
AGC3 **CpG72d8** **1900 bp**	$(AGC)_6(ATC)_1(AGC)_2$	100% identity with: human clone RP11-558P14 on chromosome Xp21.3-22.13	CpG 485-1742 bp (1257 bp) CpG score 0.85 GC score 69.03	16 sites
AGC4 **CpG99b3** **432 bp**	$(AGC)_{4,3}$	100% identity with: human clone RP1-121N10	No island	1 site
AGC5 **CpG94d10** **687 bp**	$(AGC)_4$	DNA/MER1_type (MER20) 100% identity with: human clone RP11-401H23	CpG 146-536 bp (390bp) CpG score 0.88 GC score 65.43	3 sites
AGC6 **CpG56g5** **1081 bp**	$(AGC)_4$	100% identity with: human clone RP11-161F11 on chromosome 1	CpG 51-692 bp (641 bp) CpG score 0.96 GC score 65.88	8 sites

NIX CpG islands of 500 bp or more = 4/6 67%

NIX any CpG island = 5/6 83%

*Bst*UI CpG islands = 6/6 100%

Table 3.2.8 AGG repeat containing clones

Name, number and size	Repeat sequence	Sequence homology or prediction	CpG status NIX	CpG status *Bst*UI sites
AGG1 CpG65f7 280 bp	$(AGG)_8$	-	No island	1 site
AGG2 CpG3d7 319 bp	$(AGG)_8$	-	No island	None
AGG3 CpG74g6 1026 bp	$(AGG)_8$ $(AAG)_4$	100% identity with: human clone CTD-3245B9 on chromosome 11q	CpG 382-978 bp (596 bp) CpG score 0.81 GC score 55.72	3 sites
AGG4 CpG71d10 941 bp	$(AGG)_7$	SINE/*Alu* (*Alu*Jb) 100% identity with: human clone RP11-156M10	CpG 119-396 bp (277 bp) CpG score 0.71 GC score 71.44	3 sites
AGG5 CpG173h12 1298 bp	$(AGG)_6$	100% identity with: human clone RP11-78C11	CpG 245-541 bp (296 bp) CpG score 0.83 GC score 70.76	6 sites
AGG6 CpG167h2 1449 bp	$(AGG)_{5,5}$	100% identity with: human clone RP11-541M19 on chromosome 15	CpG 661-1330 bp (669 bp) CpG score 0.83 GC score 63.63	8 sites
AGG7 CpG234b11 269 bp	$(AGG)_5$	100% identity with: human clone RP11-573G7 on chromosome 15	No island	3 sites
AGG8 CpG21h4 816 bp	$(AGG)_5$	SINE/MIR (MIR) 100% identity with: human clone RP11-430H10 on chromosome 11	No island	None

AGG9 **CpG76h2** **1135 bp**	$(AGG)_5$	LINE/L2 (L2), SINE/*Alu* (*Alu*Sq/x), SINE/*Alu* (*Alu*), LINE/L2 (L2), LTR/MaLR (MLT1K) and SINE/*Alu* (*Alu*Jb) 100% identity with: human clone RP11-361K8 on chromosome 17	No island	None
AGG10 **CpG73d6** **485 bp**	$(AGG)_5$ $(AG)_{8,5}$	98% identity with: human chromosome RP11-444B4 on chromosome 2	CpG 91-330 bp (239 bp) CpG score 0.84 GC score 53.77	1 site
AGG11 **CpG78c5** **970 bp**	$(AGG)_{4,3}$	100% identity with: human clone RP11-489G11 on chromosome 4	No island	2 sites
AGG12 **CpG54f5** **590 bp**	$(AGG)_4$	100% identity with: human clone CTD-2532L16 on chromosome 7	CpG51-386 bp (335 bp) CpG score 0.90 GC score 59.14	4 sites

NIX CpG islands of 500 bp or more = 3/12 25%

NIX any CpG island = 6/12 50%

*Bst*UI CpG islands =9/12 75%

Table 3.2.9 CCG repeat containing clones

Name, number and size	Repeat sequence	Sequence homology or prediction	CpG status NIX	CpG status *Bst*UI
CCG1 CpG258g9 281 bp	$(CCG)_7$	LINE/L1 (L1P)	No island	3 sites
CCG2 CpG18f10 412 bp	$(CCG)_6$	100% identity with: human clone RP11-593A16	CpG 51-274 bp (223 bp) CpG score 0.76 GC score 64.11	1 site
CCG3 CpG172b4 229 bp	$(CCG)_6$	98% identity with: human IMAGE clone 664601 similar to human DNA binding protein SATB1	No island	1 site
CCG4 CpG279b6 213 bp	$(CCG)_5$	100% identity with: human sequence tagged site UT1116 on chromosome 5	No island	2 sites
CCG5 CpG34a12 845 bp	$(CCG)_5$	100% identity with: human clone RP11-80E12 on chromosome 5	CpG 587-812 bp (225 bp) CpG score 0.84 GC score 50.32	2 sites
CCG6 CpG43c11 1334 bp	$(CCG)_5$	100% identity with: human clone RP11-284021	No island	None
CCG7 CpG243h5 342 bp	$(CCG)_5$	100% identity with: human clone RP11-546M21 on chromosome 17	CpG 51-302 bp (251 bp) CpG score 0.83 GC score 59.46	4 sites
CCG8 CpG249d4 1340 bp	$(CCG)_5$	100% identity with: human clone RP11-171A24	CpG 126-1312 bp (1186 bp) CpG score 0.80 GC score 61.48	10 sites
CCG9 CpG196f6 1334 bp	$(CCG)_5$	100% identity with human clone RP11-284021	No island	None

CCG10 **CpG2g10** **244 bp**	$(CCG)_4$ $(CCG)_3$	100% identity with: human clone RP11-109E12 on chromosome 2	No island	3 sites
CCG11 **CpG169b11** **486 bp**	$(CCG)_4$	100% identity with: human clone RP11-15H22 on chromosome 14	CpG 51-452 bp (401 bp) CpG score 0.74 GC score 61.85	3 sites

NIX CpG islands of 500 bp or more = 5/11 45%

NIX any CpG island = 5/11 45%

*Bst*UI CpG islands = 9/11 82%

Table 3.2.10 Random human CpG island library clones

Name, accession number and size	Sequence homology or prediction	CpG status NIX	CpG status *BstU*I sites
1 **X76662** **418 bp**	LINE/L1 (L1MC1)	No island	None
2 **X76663** **664 bp**	100% identity with: human 8D6 antigen mRNA AF161254, and human clone CTD-3020H12 on chromosome 19	CpG 67-518 (451bp) CpG score 0.91 GC score 68.24	9 sites
3 **X76665** **228 bp**	100% identity with: human 39 kD NADH-ubiquinone oxidoreductase sub unit gene HSCPGUORS, and human clone RPC111-500M8 on chromosome 12p13.3	No island	None
4 **X76666** **534 bp**	SINE/*Alu* (*Alu*Y) and SINE/*Alu* (*Alu*Sx)	No island	1 sites
5 **#76667** **693bp**	100% identity with: human clone RP11-13J10 on chromosome 2	CpG 51-586 bp (535 bp) CpG score0.83 GC score 65.46	5 sites
6 **X7668** **472 bp**	100% identity: with human clone RP11-538I12 on chromosome 16	No island	1 sites
7 **X76670** **934 bp**	-	CpG 346-653 bp (307 bp) CpG score 0.69 GC score 70.98	2 sites
8 **X76671** **893 bp**	100 % identity with: human clone RP11-38M7 on chromosome 10	CpG 117-845 bp (728 bp) CpG score 0.94 GC score 61.99	9 sites

9 **X76672** **731 bp**	100% identity with: human clone RP11-361M10 on chromosome 15	CpG 248-539 bp (291 bp) CpG score 0.80 GC score 62.64	1 sites
10 **X76673** **431 bp**	96% identity with: human clone 2365010	CpG 89-383 bp (294 bp) CpG score 0.66 GC score 52.78	2 sites
11 **X76674** **571 bp**	96% identity with: human chromosome RP11-494M19 on chromosome 1	CpG 55-506 bp (451 bp) CpG score 0.74 GC score 59.97	5 sites
12 **X76675** **862 bp**	100% identity with: human clone RP11-660M5 on chromosome 4	CpG 306-691 bp (385 bp) CpG score 1.01 GC score 61.91	7 sites
13 **X76676** **472 bp**	100% identity with: human clone RP11-646F1 on chromosome 17	No island	None
14 **X76677** **543 bp**	-	No island	2 sites
15 **X76694** **311 bp**	100% identity with: human IMAGE clone 111622	No island	1 sites
16 **X76695** **377 bp**	97% identity with: human IMAGE clone 2905781	No island	1 sites

NIX CpG islands of 500 bp or more = 3/16 19%

NIX any CpG island = 8/16 50%

*Bst*UI CpG islands = 13/16 81%

Table 3.2.10) were also assessed and used as controls for comparison. The CpG status data and its significance is presented in Tables 3.2.11-3.2.16.

These comparisons show that trinucleotide repeat containing clones from the human CpG island library do not contain significantly fewer CpG islands than random clones. However, the percentage of TR containing clones containing CpG islands was less than random clones by two of the methods used (NIX and *Bst*UI), but not in a statistically relevant way.

3.2.8 Comparison of CpG islands in random mouse and human library clones

The relative frequencies of human and mouse random clones containing CpG islands are represented and statistically compared in Table 3.3.1. This data shows that there was no significant difference in CpG status between random clones from the mouse and human CpG island library. The percentage of clones containing CpG islands was higher in the mouse library when assessed by CpG percentage and NIX, but higher in the human library for *Bst*UI sites.

3.2.9 CpG status of random human CpG island library clones v. those containing trinucleotide repeats

The relative frequencies of human and mouse TR containing clones with CpG islands are shown and compared in Table 3.3.2. The percentage of mouse clones containing islands is always less than that of the human library. This difference is only statistically relevant when the CpG status is determined by the presence of *Bst*UI sites.

The relative CpG and GC scores of all mouse and human CpG islands identified in this study are plotted in Figure 3.3. The only obvious trends apparent from this data are that the GC score is usually highest in human CpG islands, but that the CpG score was often higher in mouse islands.

Table 3.2.11 Number of random clones from human CpG island library containing CpG islands of more than 500 bp (CpG score 0.6+)

	Number/% of clones containing CpG islands	Number/% of clones not containing CpG islands
n = 16	4/25%	12/75%

Table 3.2.12 Number of trinucleotide repeat containing clones from human CpG island library containing CpG islands of more than 500 bp (CpG score 0.6+)

Compared to random clones (Table 3.2.11 above)

% GC of TR	CpG island containing clones	Non CpG island containing clones	Chi-square value	*p*	*Q*	Signif.
100% **n = 11**	5/45%	6/55%	1.2273	0.05	0.2679	**No**
66% **n = 23**	8/35%	15/65%	0.4239	0.05	0.5150	**No**
33% **n = 26**	3/12%	23/88%	1.2923	0.05	0.2556	**No**
0% **n = 5**	1/20%	4/80%	0.0525	0.05	0.8744	**No**
Total **n = 65**	17/26%	48/74%	0.0089	0.05	0.9248	**No**

p = Probability of occurrence by chance

Q = Probability of non chance occurrence

Signif. = Significance to one degree of freedom

Table 3.2.13 Number of random clones from human CpG island library containing CpG islands (as determined by NIX)

	Number/% of clones containing CpG islands	Number/% of clones not containing CpG islands
n = 16	8/50%	8/50%

Table 3.2.14 Number of trinucleotide repeat containing clones from human CpG island library containing CpG islands (as determined by NIX)

Compared to random clones (Table 3.2.13 above)

% GC of TR	CpG island containing clones	Non CpG island containing clones	Chi-square value	*p*	*Q*	Signif.
100% **n = 11**	5/45%	6/55%	0.5390	0.05	0.4626	**No**
66% **n = 23**	12/52%	11/48%	0.0178	0.05	0.6728	**No**
33% **n = 26**	4/15%	22/85%	5.8154	0.025	0.0159	**Yes**
0% **n = 5**	1/20%	4/80%	1.4000	0.05	0.2367	**No**
Total **n = 65**	22/34%	43/66%	1.4367	0.05	0.2307	**No**

Table 3.2.15 Number of random clones from human CpG island library containing CpG islands (as determined by presence of *Bst*UI restriction sites)

	Number/% of clones containing CpG islands	Number/% of clones not containing CpG islands
n = 16	13/81%	3/19%

Table 3.2.16 Number of trinucleotide repeat containing clones from human CpG island library containing CpG islands (as determined by presence of *Bst*UI restriction sites)

Compared to random clones (Table 3.2.15 above)

% GC of TR	CpG island containing clones	Non CpG island containing clones	Chi-square value	*p*	*Q*	Signif.
100% **n = 11**	9/82%	2/18%	0.0014	0.05	0.9702	**No**
66% **n = 23**	17/74%	6/26%	0.2861	0.05	0.5927	**No**
33% **n = 26**	12/46%	14/54%	5.0638	0.025	0.0244	**Yes**
0% **n = 5**	3/60%	2/40%	0.9483	0.05	0.3302	**No**
Total **n = 65**	41/63%	24/37%	1.9082	0.05	0.1672	**No**

Table 3.3.1 Number of random clones containing CpG islands; human v. mouse CpG island library

	Human		Mouse		Chi-square value	*p*	*Q*	Signif.
	CpG	Non	CpG	Non				
500+ NIX	4 25%	12 75%	17 47%	19 53%	2.2721	0.05	0.1317	**No**
Any NIX	8 50%	8 50%	24 67%	12 33%	1.3	0.05	0.2542	**No**
***Bst*UI**	13 81%	3 19%	27 75%	9 25%	0.2437	0.05	0.6215	**No**

Table 3.3.2 Number of TR containing clones which also have CpG islands; human v. mouse CpG island library

	Human		Mouse		Chi-square value	*p*	*Q*	Signif.
	CpG	Non	CpG	Non				
500+ NIX	17 26%	48 74%	17 20%	72 80%	1.086	0.05	0.2973	**No**
Any NIX	22 34%	43 66%	27 32%	62 68%	0.2132	0.05	0.6443	**No**
***Bst*UI**	41 63%	24 37%	41 48%	48 52%	4.3658	0.05	0.0367	**Yes**

3.2.10 Relative frequency of trinucleotide repeat classes in mouse CpG island library clones

Once all the clone inserts had been sequenced and their CpG status assessed, the relative frequency of each trinucleotide repeat class was calculated. This data is summarised in Table 3.4.1. The most abundant trinucleotide repeat in the mouse CpG island library clones was CCG (34%), followed closely by AGG (23%). The trinucleotide repeats AGC, ACG, ACC and AAC each represent approximately 10% of the overall number. AAG and ATC each constitute 1%, ACT and AAT were not represented at all.

3.2.11 Relative frequency of trinucleotide repeat classes in different genomic regions in the mouse

The data from this study on trinucleotide repeat frequency in mouse CpG islands was compared to data on their known frequency in cDNA [Chambers and Abbott, 1996] and bulk DNA [Stallings, 1994] (Table 3.4.2). The most apparent differences in trinucleotide repeat frequencies are as follows: a depletion of AAT and AAG in CpG islands, CCG in cDNA, and a relative abundance of AAT and AAC in cDNA, AGC in bulk DNA and CCG in CpG islands.

3.2.12 Relative frequency of trinucleotide repeat classes in different genomic regions; mouse v. human

The data compiled in this study on trinucleotide repeat frequencies in human and mouse CpG islands, was compared to their respective frequencies in cDNA and total genomic DNA [Stallings, 1994] in table 3.4.3. The most striking findings were; a comparative abundance of AAC in mouse cDNA (33%-1.5%), and a depletion of AGC (12%-23%, 3%-53.5%) in mouse CpG island and cDNA clones.

3.2.13 Size distribution of mouse trinucleotide repeats

The size distribution of trinucleotide repeats found in the mouse CpG island library clones is displayed in figure 3.4., the average size of repeat from each class of array is shown in figure 3.6. The widest distribution of repeat array size was seen with AAG and AGC, and the narrowest distribution with repeats ATC, ACG and ACT. All the classes of trinucleotide repeat had a bias towards the smaller end of the size range, with the exception of AAG repeats which were skewed towards larger repeats. These factors are reflected in the average

Table 3.4.1 Relative frequency of trinucleotide repeat classes in mouse CpG island library clones

	Clones from CpG island library		Clones containing CpG islands*	
TR class	TR containing clones per 100 000 CFUs	% of overall TR containing clones	TR containing clones per 100 000 CFUs	% of overall TR containing clones
AAT	3	2%	0	0%
AAC	30	20%	5	8%
AAG	6	4%	1	1%
ACT	0	0%	0	0%
ATC	6	4%	1	1%
ACC	17.5	12%	7.5	12%
ACG	5	3%	5	8%
AGC	15	10%	7.5	12%
AGG	40	27%	15	23%
CCG	27.5	18%	22.5	35%
Total	150	100%	64.5	100%

*As predicted by NIX

Table 3.4.2 Relative frequency of trinucleotide repeat classes in different genomic regions in the mouse

RELATIVE FREQUENCIES OF TRINUCLEOTIDE REPEATS IN MOUSE DNA			
TR class	Mouse genome n = 54	CpG islands* n = 23	cDNA n = 181
AAT	11%	0%	22%
AAC	9%	8%	33%
AAG	11%	1%	8%
ACT	0%	0%	0%
ATC	5.5%	1%	6%
ACC	15%	12%	6%
ACG	0%	8%	0%
AGC	26%	12%	3%
AGG	17%	23%	22%
CCG	5.5%	35%	0%
Total	100%	100%	100%

Table 3.4.3 Relative frequency of trinucleotide repeat classes in mouse v. human DNA

TR class	Total genome		CpG islands*		cDNA	
	Mouse n = 54	Human n = 51	Mouse n = 23	Human n = 22	Mouse n = 181	Human n = 209
AAT	11%	20%	0%	4.5%	22%	1.5%
AAC	9%	17.5%	8%	4.5%	33%	1.5%
AAG	11%	4%	1%	4.5%	8%	0%
ACT	0%	0%	0%	0%	0%	0%
ATC	5.5%	8%	1%	9%	6%	1%
ACC	15%	0%	12%	4.5%	6%	3%
ACG	0%	0%	8%	0%	0%	0%
AGC	26%	23%	12%	23%	3%	53.5%
AGG	17%	10%	23%	27%	22%	1.5%
CCG	5.5%	17.5%	35%	23%	0%	38%
Total	100%	100%	100%	100%	100%	100%

*As predicted by NIX

Figure 3.4 Size distributions of trinucleotide repeats isolated from mouse CpG island library clones

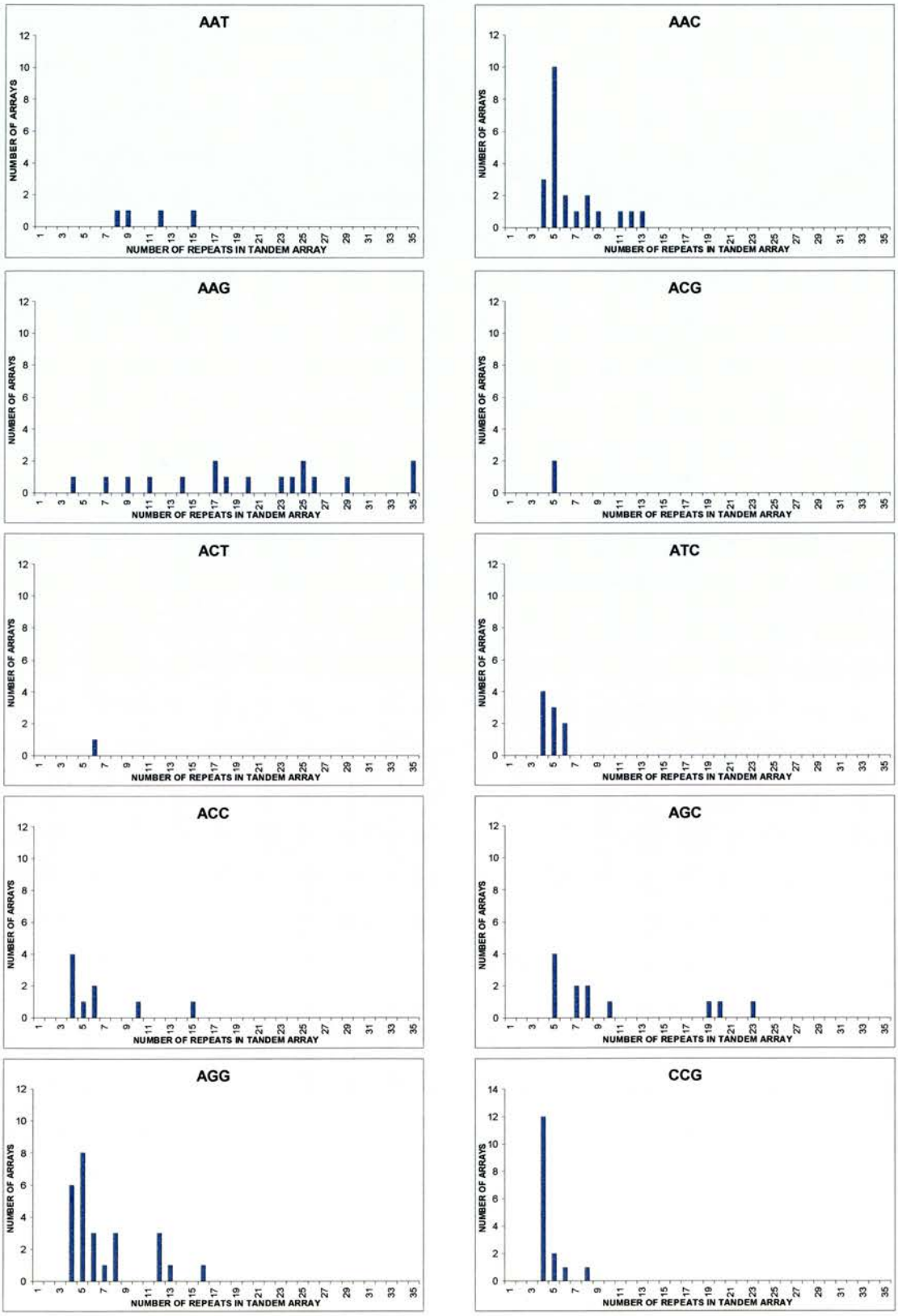


Figure 3.5 Size distributions of trinucleotide repeats in mouse CpG islands and non island regions

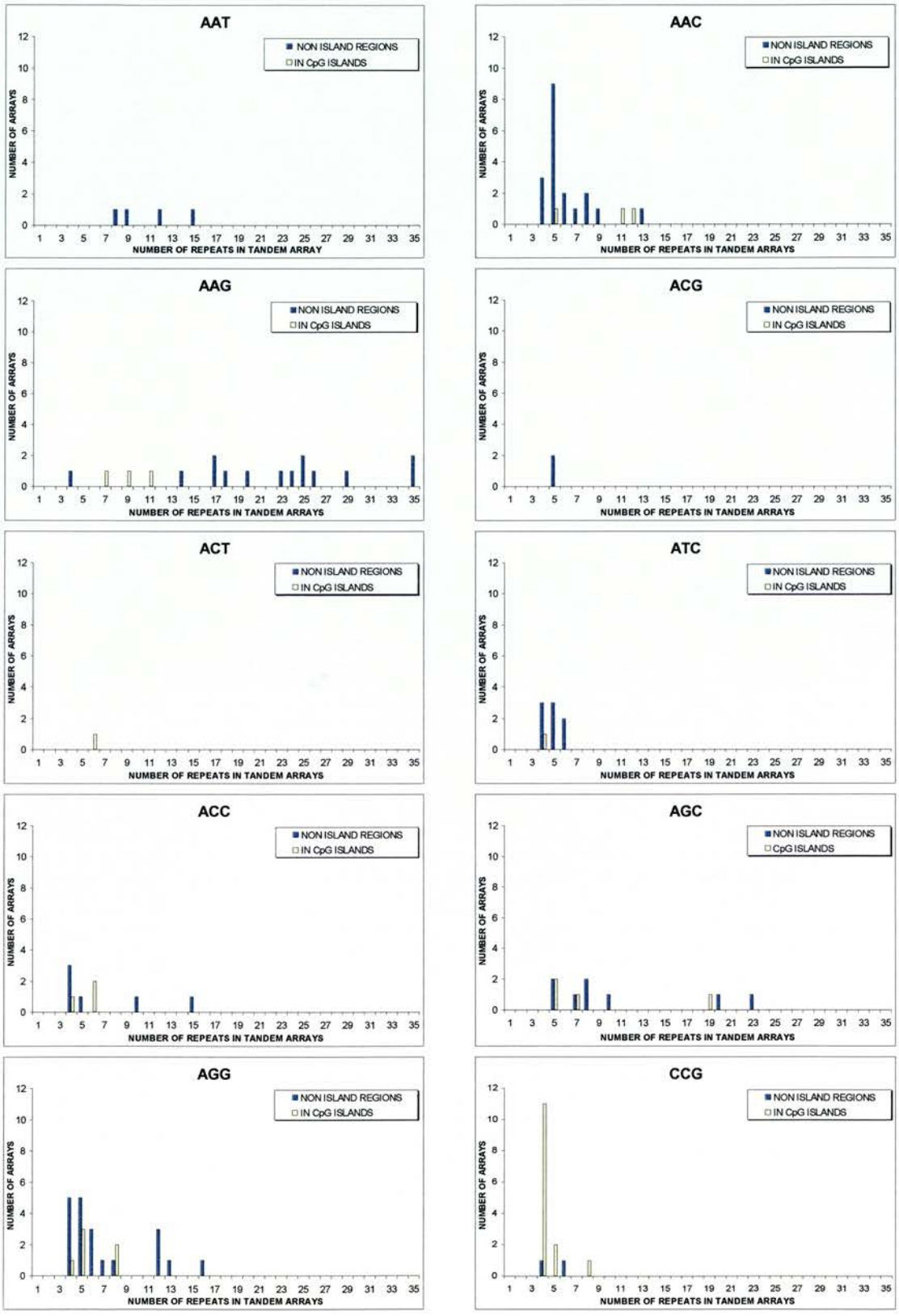


Figure 3.6 Average number of trinucleotide repeats in mouse arrays identified in this study

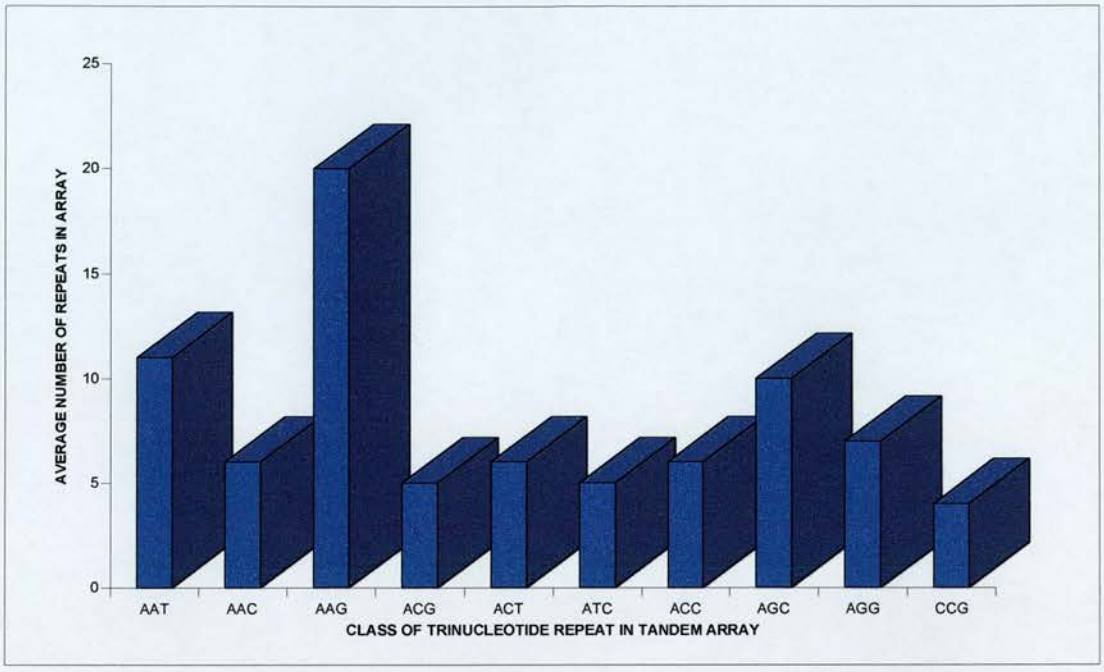
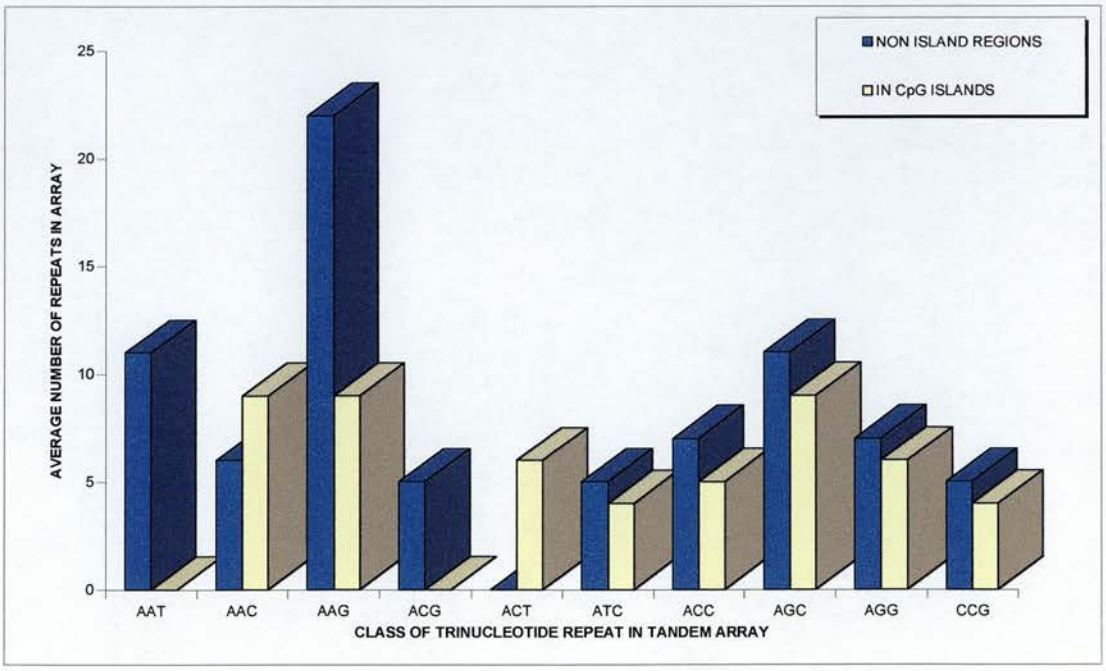


Figure 3.7 Average number of trinucleotide repeats in arrays found in mouse CpG islands and non island regions



size of repeat array; AAG being the largest by far, followed by AAT, AGC and AGG. The largest pure trinucleotide repeats identified were two clones containing tracts of $(AAG)_{35}$. The majority of trinucleotide repeats were pure, but in some cases the repeat sequence was complex and interrupted. A number of clones contained more than one class of trinucleotide repeat. It was apparent that certain classes of repeat appeared to cluster together frequently, for example AAG and AGG, AGC and AAC. Several clones also contained other classes of tandem repeat including dinucleotides, tetranucleotides, hexanucleotides, heptanucleotides, and nonanucleotides. In several instances these repeats occur many times throughout the clone in varying lengths.

This data was broken down further into size distributions in CpG island* and non island clones and depicted in Figure 3.5. The average repeat sizes in island and non island clones is displayed in figure 3.7. This figure clearly shows that the average size of repeat array is nearly always subtly smaller in CpG island clones, with the exception of AAC. The most notable average size reduction in repeat array is a drop of 12 repeats in AAG between non island and island clones.

*CpG status as determined by NIX analysis.

3.3 Discussion

3.3.1 Trinucleotide repeats are depleted in mouse CpG islands

Whichever method of predicting CpG island status you consider most relevant, the results of this survey clearly show that trinucleotide repeats are significantly depleted, or selected against in these regions. The first method of identifying CpG islands used in this survey is probably the most relevant, as the lack of DNA methylation is arguably the most critical function of an island. However all these methods were merely tools for predicting CpG islands, none actually determined the methylation status of the DNA directly.

The depletion of trinucleotide repeats in the mouse CpG island library could easily be considered an artefact of the PCR amplification step used in constructing the library. PCR amplification notoriously struggles when amplifying GC rich DNA, and short tandem repeats. During the amplification of a mixture of islands it feasible that DNA without trinucleotide repeats (especially GC rich ones) would be preferentially amplified, biasing the resulting library. This effect could have been minimised by using betaine in the PCR reaction. Betaine improves the amplification of both GC rich [Henke et al., 1997] and trinucleotide repeat regions [Papp et al., 1996]. However as the trinucleotide repeats were not significantly depleted in the human CpG island library which was constructed in the same way, the depletion of trinucleotide repeats in mouse CpG islands is likely to be real.

It is currently hypothesised that trinucleotide repeats are more likely to expand when located in CpG islands [Brock et al., 1999; Gourdon et al., 1997a]. This is because the majority of trinucleotide repeats which have undergone expansion in the human genome were located within or flanking CpG islands. The only exceptions to this rule include trinucleotide repeat expansions of a relatively modest magnitude [Brock et al., 1999]. As CpG islands comprise such a small percentage of the overall genome (~1.5%), this association seems unlikely to be the product of chance. CpG islands are known to be associated with origins of replication (ORIs) [Antequera and Bird, 1999] and transcription [Delgado et al., 1998]. It is possible that it is the location of a trinucleotide repeat near an origin of replication, or transcription that is the true prerequisite for potential expansion. The fact that not all ORIs are located in CpG islands could well explain the few cases of expansion which are known to occur outside island regions. If we use CpG islands as a general indicator that an ORI may be present, then the depletion of trinucleotide repeats in mouse CpG islands is highly significant. The

fact that trinucleotide repeats are depleted in the regions of known expansion potential, could be the key to the lack of dynamic mutation observed in the mouse.

It should be noted that the majority of trinucleotide repeat expansions identified in humans have been found within or near genes. Likewise, most efforts to identify large potentially expandable trinucleotide repeats in mice have concentrated on genic regions. However, it does not necessarily follow that most large repeats and potential expansions are located in these regions, it is more likely a result of ascertainment bias. The bias has been created because most screening endeavours have focused on finding disease causing expansions, and to date we assume that a repeat has to be associated with a gene to cause a disease phenotype. Therefore genic regions have been the priority for screening efforts, and most expandable repeats have been reported to occur there. The only screening studies to date which are not biased in this way are based on the repeat expansion detection (RED) technique [Schalling et al., 1993]. This method can detect expanded trinucleotide repeats anywhere in the genome. Once the complete human and mouse sequence is available the overall frequencies and genomic locations of all large, expandable trinucleotide repeats will finally be elucidated.

3.3.2 A comparison of the CpG islands from the mouse and human libraries

The percentage of mouse trinucleotide repeat containing clones with CpG islands was always less than the percentage of human clones. This may be a reflection of the fact that the mouse library was screened for trinucleotide repeats to saturation, but that the human library was not, possibly omitting many larger repeats which would have surely affected this outcome. As it is, the overall distribution of trinucleotide repeat size was much smaller in the human clones. It is therefore possible that the constraints on these repeats to be in non CpG island regions was not as high as for the larger repeats in the mouse.

3.3.3 The relative frequencies of different trinucleotide repeat classes identified in mouse CpG island clones

The most abundant trinucleotide repeat in the mouse CpG island library clones was CCG (34%), followed closely by AGG (23%). The presence of such a high percentage of CCG repeats is most likely to be a direct result of the base composition of island regions. CpG islands contain a much higher than average G+C content, which directly increases the chance

of GC rich repeats occurring. It could also be said that a large CCG repeat, if large enough would itself constitute a CpG island. However the CCG repeats identified in this survey were all too small to have such an effect on the clone. The absence of any AAT repeats in CpG island containing clones is again likely to be a direct effect of CpG island sequence composition. A+T rich sequences are known to be exceedingly rare in CpG islands, for example the restriction enzyme *Mse*I (recognition site TTAA) used to produce this library is rare in island DNA (~1 per 1000 bp), but frequent in bulk DNA (~1 per 140 bp). Also A+T rich trinucleotide repeat tracts of sufficient length could by their very nature, theoretically change or disrupt CpG islands. This could potentially bias against the number of clones containing both AT rich repeats and CpG islands. A schematic diagram detailing the theoretical insertion of AAT tracts of varying length into a well defined mouse CpG island is shown in figure 3.8. This analysis was purely theoretical, based on predictions made by the bioinformatics program NIX. In the examples shown the AAT repeats apparently disrupt a portion of the CpG island larger than just the repeat tracts themselves (e.g. a repeat of 45 bp, destroys 200 bp of island). In this survey no A+T rich trinucleotide repeats were identified in CpG islands, but NIX analysis of the DNA revealed that none of these clones would have constituted CpG islands even without the repeat sequences.

3.3.4 Relative trinucleotide repeat frequencies in different portions of the mouse genome (total DNA, cDNA and CpG islands)

Trinucleotide repeats, as with all microsatellite repeats (mononucleotides- hexanucleotides) are known to be enriched in eukaryotic noncoding DNA [Cox and Mirkin, 1997; Hancock, 1995]. More specifically this means that they are more abundant than would be predicted from the random association of nucleotides. However in coding DNA only trinucleotides and hexanucleotides remain significantly enriched. This implies that it is possible for trinucleotides to exist and mutate successfully in coding DNA because they do not result in potentially deleterious frameshift mutations [Metzgar et al., 2000]. Therefore the expansion of trinucleotide repeats are not subject to differential selective pressures in coding and noncoding regions, to the same extent as other repeats. This means that significant differences in trinucleotide frequency between these regions is likely to be the result of another type of selective influence.

The distribution of trinucleotide repeat classes among the mouse CpG island clones in most cases parallels that described previously from surveys of total mouse DNA in databases [Stallings, 1994] and cDNA library clones [Chambers and Abbott, 1996]. The most apparent

Figure 3.8 Predicted effects of AAT rich repeats on CpG islands

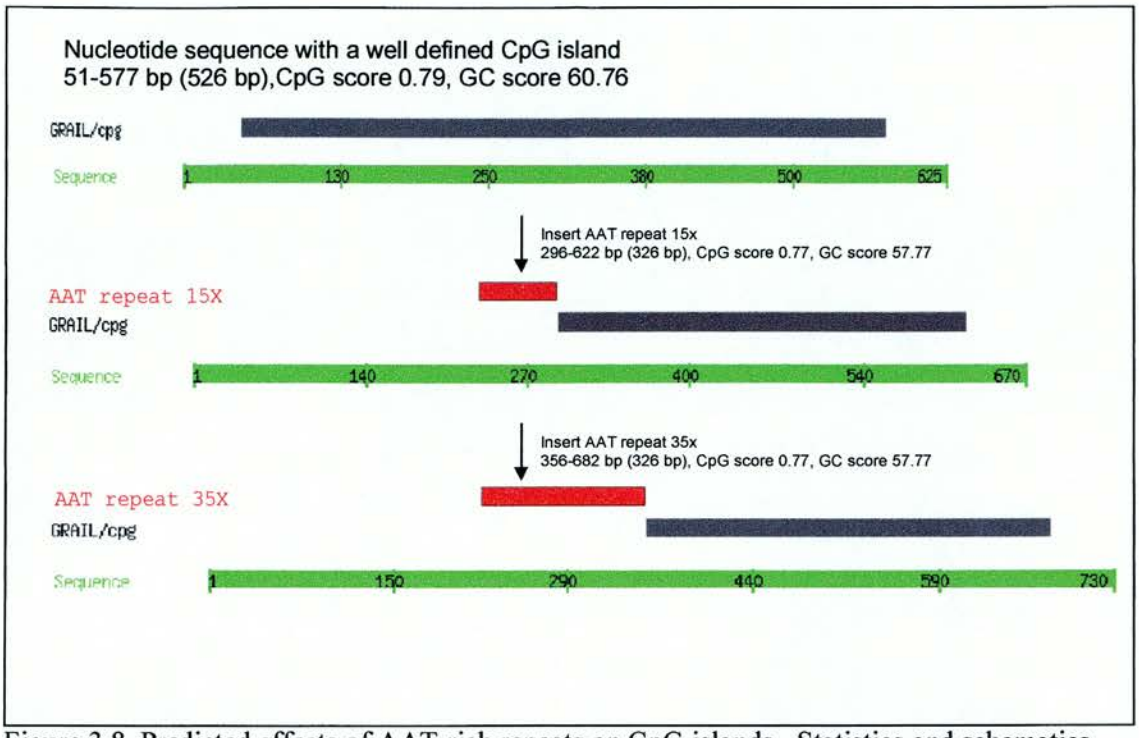


Figure 3.8 Predicted effects of AAT rich repeats on CpG islands. Statistics and schematics predicted by NIX analysis.

differences in trinucleotide repeat frequencies are as follows: a depletion of AAT and AAG in CpG islands, and CCG in cDNA, and a relative abundance of AAT and AAC in cDNA, AGC in bulk DNA and CCG in CpG islands.

The complete lack of AAT repeats, and abundance of CCG repeats in CpG islands can be explained as a direct result of sequence composition (see 3.3.3). In the same way it can be postulated that the depletion of AAG repeats is also due to the repeat composition itself, as any large AAG repeat (only 33% G+C) would by its very nature disrupt the properties of a CpG island. This is actually reflected in the reduction in frequency observed for this whole group of repeats (AAC, AAG, and ATC {no ACT repeats were reported}). The fact that the mouse AAG repeats identified in this study were often large in size, could be why AAG in particular was the most noticeably depleted. However as large, expanded AAG repeats can be deleterious in man, perhaps they have been selected against in the mouse, especially in coding (CpG island) regions. The lack of CCG repeats in cDNA was probably due to the absence of 5' untranslated (UTR) sequences which this class of repeats are often associated with [Riggins et al., 1992; Stallings, 1994]. The cDNA library was depleted for 5' UTRs because it was constructed by oligo(dT)-priming. The abundance of AAT and AAC repeats in cDNA clones is probably a result of their association with short interspersed elements (SINEs). These repetitive elements were not present in such large percentages in the CpG island clones because they would be likely to be methylated in the genome and therefore removed with bulk DNA from this library. The AGC repeat was the most noticeably reduced in coding DNA compared to bulk DNA, which was unexpected. Perhaps as this is the most commonly reported disease causing repeat (in humans), it has been effectively selected against in mouse coding regions.

3.3.5 Comparison of trinucleotide repeat frequencies; mouse v. human

The most striking observations from this comparison of data is the overrepresentation of AAC repeats in mouse cDNA, and the lack of AGC repeats in mouse transcribed regions (CpG islands 12-23%, and cDNA 3-53%) when compared to human data. The large proportion of AAC repeats in mouse cDNA is probably a reflection of their association with SINEs (see 3.3.4) and because most of these repetitive elements contain a poly A tract which would have been primed from in the making of the mouse library. Their relative lack in the human cDNA data is most likely due to this information coming from human genes in database searches, rather than specifically oligo(dT) primed libraries which bias the survey

towards repetitive AAC containing elements. The elevated levels of AGC repeats in transcribed human DNA is unexpected. Expansion of the AGC repeat in human coding sequence is the most common cause of dynamic mutation disorders, its depletion in mouse coding regions may be a significant factor in the absence of any trinucleotide repeat expansion phenotypes in the mouse.

3.3.6 Summary of size distribution of mouse trinucleotide repeat arrays

The widest distribution of trinucleotide repeat sizes was seen with AAG and AGC, and the narrowest distribution with ATC, ACG and ACT. The reasons for the large distribution in the size of AAG and AGC repeats is something of an enigma, but is perhaps highly significant. These two classes of repeat undergo the most extreme size range of expansion in humans, perhaps there is something inherently more unstable about these repeat arrays. This speculation is reflected in the observation that all classes of repeat distributions are skewed towards small arrays with the exception of the AAG repeat tract which is more frequently large than small. The largest repeats identified in this survey were two pure tracts of AAG_{35}. The narrow distribution of ATC, ACG and ACT repeats is most likely a result of their extreme rarity, but perhaps this alone suggests a selective mechanism working against their existence.

A number of clones contained more than one class of trinucleotide repeat, several classes of which appear to cluster together frequently, including; AAG and AGG, AGC and AAC. Other clones contained other classes of tandem repeat including dinucleotides, tetranucleotides, hexanucleotides, heptanucleotides, and nonanucleotides in addition to the trinucleotide repeats. This clustering suggests that in some cases the repetitive species may not be totally independent, as has been reported with microsatellites and *Alu* elements [Beckman and Weber, 1992]. For example one repeat in a pure tract could by the mutation of a single base change repeat class (AAG to AGG), this single repeat could in turn expand resulting in a compound or clustered repeat. Another possible explanation, most specifically where two classes of repeat appear in conjunction frequently, is that they are both involved in coding for a conserved and frequent protein domain.

The average size of trinucleotide repeat array was nearly always subtly smaller in the CpG island clones. The most notable size reduction is an average drop of 12 AAG repeats between non-island and island containing clones. From these results we cannot postulate whether this reduction in repeat size range is due to coding region constraints, or more

specifically to ORI region selection pressure. If these distributions were smaller than those observed in cDNA, this would certainly add more weight to the ORI versus CpG island relevance question.

The human CpG island library was not screened for trinucleotide repeats to saturation. For this reason it is not possible to directly compare the relative sizes of the repeats found with those from the mouse library. However the comparisons between relative repeat class frequency and CpG status are likely to be significant, because a representative sample is the most important factor. The most critical question that remains is how representative of the real genomes are the CpG island libraries.

CHAPTER 4

SIZE VARIATION AND MAPPING OF MOUSE TRINUCLEOTIDE REPEAT ARRAYS

4 Size variation and mapping of mouse trinucleotide repeat arrays

4.1 Introduction

The trinucleotide repeat arrays isolated from the mouse CpG island library screen were where possible amplified by PCR in a variety of inbred mouse strains and stocks. The number of different alleles identified for each trinucleotide repeat, and their overall size range, were used to assess their variability. The polymorphic trinucleotide repeats were then mapped using an interspecific backcross panel. The variable trinucleotide repeats were also screened as 'candidates' for causing mouse mutant phenotypes located in the same genomic region.

4.1.1 Mice, the *Mus* species group and the laboratory mouse

The order *Rodentia* includes five different families (Heteromyidae, Gliridae, Seleviniidae, Zapodidae, and Muridae) which all contain animals that are commonly referred to as mice. However, it is more specifically within the family Muridae, subfamily Murinae that the 300 plus species of Old World mice and rats are placed. The subfamily Murinae includes the genus *Mus*, which is further divided into four subgenera, one of which also carries the name *mus*. It is this subgenus of *Mus* which contains all the 'true Old World mice' including the house mouse *Mus musculus*, the progenitor of most laboratory mouse strains. It is this group of animals which will commonly be referred to as mice in this thesis.

4.1.2 Origin and generation of inbred mouse strains

Inbred strains are animals which result from 20 or more sequential generations of sibling (brother-sister) matings. The original 'standard' or 'classical' inbred strains of mice were derived from fancy mice purchased from pet mouse breeders at the beginning of the 20th century. These inbred strains are mosaic lines containing each of the four house mouse subspecies, although *Mus musculus domesticus* is the predominant component. These particular inbred strains are not actually representative of any naturally occurring population, but are still widely used for many biological studies. However, because of this mixed and close relationship there is an extremely high level of non-polymorphism observed at the majority of loci examined in these strains. More recently, newer inbred strains have been

generated directly from wild mice representing true inbred lines of the *Mus musculus* species *Mus musculus domesticus* (WSB/Ei, ZALENDE/Ei), *Mus musculus musculus* (CZECH II/Ei), and *Mus musculus castaneus* (CAST/Ei). *Mus musculus musculus* mice have also been inbred with a couple of the more distantly related *Mus* species, *Mus spicilegus* (PANCEVO/Ei) and *Mus spretus* (SPRET/Ei) which are capable of producing viable, fertile offspring.

The use of such standard inbred strains in the laboratory makes it possible to reduce, or eliminate the effect of genetic variability on experimental results.

4.1.3 Standardised nomenclature of inbred mouse strains

The naming of the original classical inbred strains of mice at the beginning of the century was a rather random process, but newer strains are named with a capital letter, followed by other descriptive capital letters or numbers, in as concise a way as possible. When two mice have the same inbred strain name, for example C57BL/6 (B6), it means they can both trace their lineage back through sibling matings to the same founder pair of animals. However, when mice from the same inbred strain have been maintained independently for a sufficient period of time the strain will gradually drift apart genetically, and the resulting differences will become fixed. This is how new substrains arise, particularly when: (1) the branches of an inbred strain have been separated before F_{40}; (2) a branch has been maintained independently for at least 100 generations; (3) genetic differences from other branches are identified. Substrains determined by these criteria are indicated by adding a slash (/) to the strain symbol followed by an appropriate substrain symbol, for example DBA/1 and DBA/2. Independently maintained inbred strains are assigned different Laboratory registration codes following the standard name, to highlight the possible genetic differences, for example the full name of the B6 mice maintained by The Jackson Laboratory (JAX) are C57BL/6J.

4.1.4 Relatives of *Mus mus musculus* species and interspecific hybrids

All of the *Mus* species have the same standard karyotype comprising 40 acrocentric chromosomes. The closest relatives of the *Mus musculus* group are the species *Mus spretus*, *Mus spicilegus* and *Mus macedonicus*. Although these species share some common and overlapping geographical distributions, they are not capable of producing interspecific offspring in their natural habitats. However under artificial, laboratory conditions *Mus*

musculus group mice have been successfully mated with each of these three species to produce viable interspecific F_1 hybrids [Bonhomme et al., 1984]. These breeding endeavours revealed that although *Mus spretus*, *Mus spicilegus* and *Mus macedonicus* males could be crossed with *Mus musculus* group females, the resulting male F_1 hybrids were all sterile (Haldane's rule [Haldane, 1922]). The female progeny however were fertile and could be used in further breeding protocols.

4.1.5 Interspecific backcross mapping

Female interspecific F_1 hybrids can be successfully backcrossed to either parental male, resulting in viable and fertile N_2 progeny. The F_1 hybrid females will be uniformly heterozygous, and the N_2 offspring (usually 100-1000 animals) provide a segregating population. The DNA obtained from the N_2 animals, referred to as an 'interspecific mapping panel' can be typed at previously assigned loci (usually characterised genes or sequences, proviral loci and microsatellites) and 'anchored' onto consensus genetic linkage maps. The remaining DNA can then be used by investigators to map loci of interest [Avner et al., 1988; Copeland and Jenkins, 1991; Robert et al., 1985], relative to the common anchor sites. The larger the number of N_2 offspring produced and analysed in an interspecific backcross, the more accurately any test loci can be positioned.

The generation of interspecific backcrosses was a significant breakthrough in mouse genetics as it permitted multilocus linkage analysis with molecular and biochemical markers [Avner et al., 1988; Bonhomme et al., 1979; Bonhomme et al., 1982; Guenet et al., 1990].

4.1.6 The BSS mapping panel

The interspecific backcross (see figure 4.1) DNA panel used to map the trinucleotide repeat loci in this study was established by Lucy Rowe and her co-workers [Rowe et al., 1994] at The Jackson Laboratory. The cross (C57BL/6J x SPRET/Ei) F_1 female x SPRET/Ei male, or 'BSS' comprised DNA from the parents of the F_1 hybrid and 94 N_2 animals, permitting loci to be mapped to a resolution of approximately 1.1 centimorgans. This panel map was anchored by 49 short sequence length polymorphism loci, 43 proviral loci and 60 gene sequence loci. To date 4986 loci have been mapped by linkage to this resource (see www.jax.org/resources/documents/cmdata/bkmap/BSS.html).

Figure 4.1 The BSS interspecific backcross

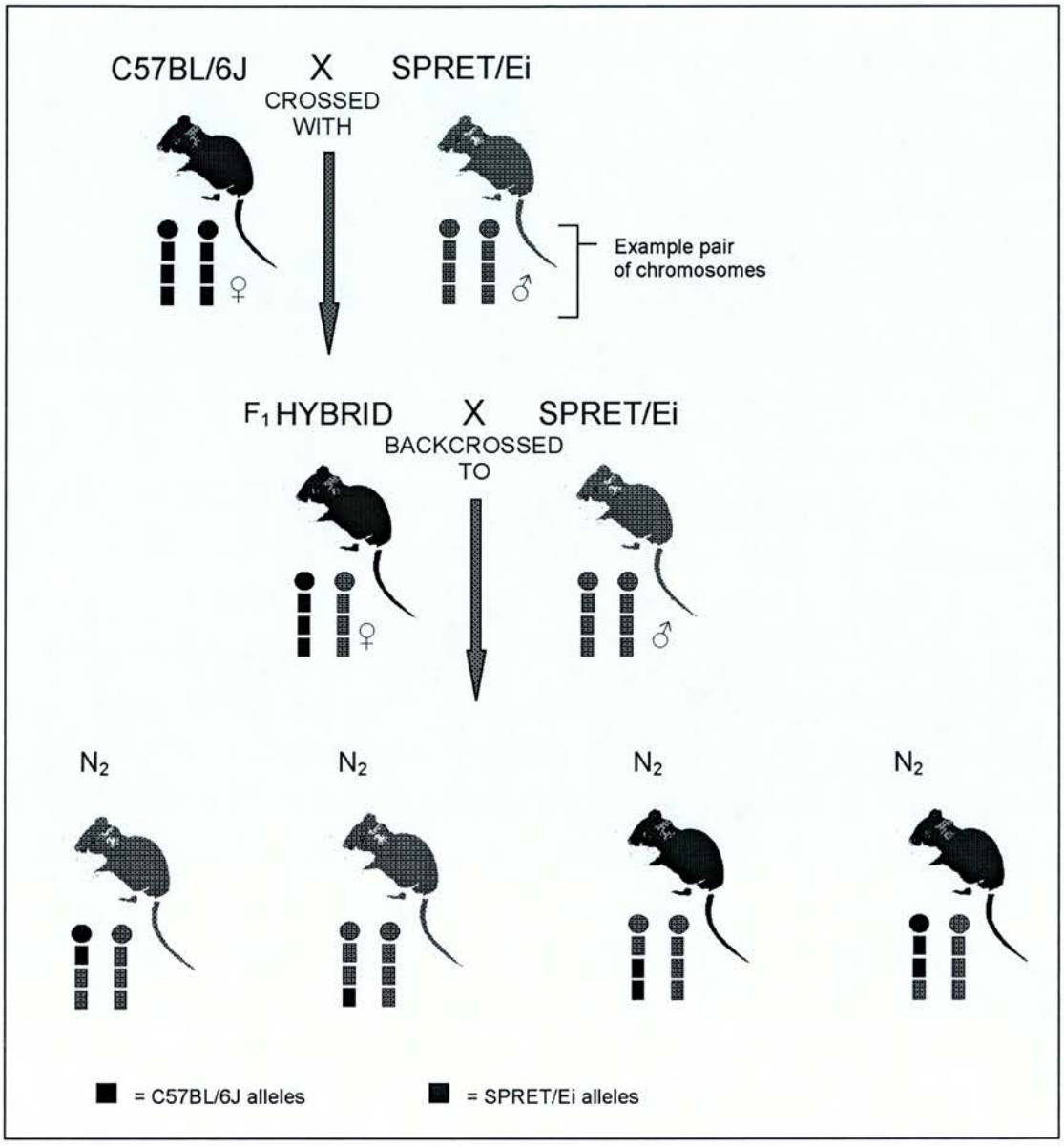


Figure 4.1 The BSS interspecific backcross. The breeding scheme used to generate the BSS mouse interspecific backcross mapping panel. The recombination events scored in the N_2 backcross offspring occurred in the F_1 parent. 94 N_2 offspring were produced from this backcross. Only one pair of example chromosomes are depicted for each mouse.
The information in this figure was adapted from [Copeland and Jenkins, 1991].

4.2 Results

4.2.1 Size variation of trinucleotide repeats

The variability of the trinucleotide repeat arrays isolated in this study was assessed by their PCR amplification in a panel of mouse DNA. Where possible primers were designed (2.2.1.19.1) to flank each trinucleotide repeat and the PCR amplification conditions were optimised (Table 4.1).

It was not possible to design PCR primers to amplify several of the trinucleotide repeat arrays, because insufficient 'unique' sequence was available in the respective clones. This was a consequence of their location directly flanking the plasmid polylinkers, or being embedded in repetitive DNA. Attempts were made to isolate the longest of these repeat sequences (and thus the flanking genomic DNA) from a mouse bacterial artificial chromosome (BAC) library, but they proved unsuccessful. Therefore despite these sequences including some of the largest trinucleotide repeats identified in this survey (AAT2, AAG3, AAG8, ATC2) they could not be analysed further.

Although relatively arbitrary the mouse DNA panel (see below) included mice from a wide range of evolutionary relationships, from closely related inbred strains (derived from the subspecies *Mus musculus*), to more distant species (*Mus spretus* and *Mus caroli*) and as far as different genera (*Rattus norvegicus*). This was in an attempt to determine the broadest extent of size variations that could be observed in the trinucleotide repeat arrays. The choice of mice also encompassed the parental strains of the BSS backcross, and several standard recombinant inbred (RI) lines, to include a range of possible mapping resources. A schematic diagram of the relationships between these mice is shown in Figure 4.2.

Mouse DNA panel:

1. C57BL/6J
2. SPRET/Ei
3. C3H/HeSnJ
4. AKR/J
5. CAST/Ei
6. IS/CamEi
7. NZB/B1NJ
8. PERA/CamEi
9. 129/J
10. SJL/J
11. MOLF/Ei
12. A/J
13. CBA/J
14. DBA/2J
15. *Mus caroli*
16. *Rattus norvegicus*

Table 4.1 PCR primers and conditions for the amplification of mouse trinucleotide repeat arrays

Clone name	Primer sequences	Annealing temperature	Other reagents	Product size*
AAT1	5' CGATGGAACACGAGTCTGTC 3' 5' GTGGAACGGTAGTCATGGTG 3'	58°C	-	246 bp
AAT3	5' TGGAAGGTTTATGCCCACA 3' 5' CACACATTGTTTGCTTTCCAA 3'	54°C	-	261 bp
AAT4	5' CCAAATGTCACAGGGGAGAG 3' 5' TTAGCCAGAGCGACATAGCA 3'	60°C	-	498 bp
AAC1	5' CAGCAGCAACAACGACAGTAA 3' 5' AGCCAGAAACGCAGCTGATA 3'	58°C	-	198bp
AAC2	5' AACTAATCGACAATCAGCAGCA 3' 5' AGAATTGGCAACCCTAAGTCC 3'	55°C	-	198 bp
AAC3	5' CCACGATGGTTTTCATTTCC 3' 5' CAGCTGCATCCTTTCCTTTC 3'	58°C	-	202 bp
AAC4	5' CAACCTTGGCTACACAGCAA 3' 5' GGGGTCCATTCACGTACAAC 3'	60°C	-	203 bp
AAC5	5' GACACTAGTTCAGGAATATCAACAACA 3' 5' CGTTCGCAATAAGAAGAGCC 3'	55°C	-	291 bp
AAC9	5' TGAAATGGGGTCAAGTCACA 3' 5' CCACACAGGTAGCAGAAGCA 3'	58°C	-	188 bp
AAC10	5' CTAGCCTGCACTGTGTTCCA 3' 5' GGAGCCTCCTTGTCCTATCC 3'	58°C	-	292 bp
AAC12	5' TTTGAGAGATGGAAGCAGGG 3' 5' AAACCCAGGAGCTCACACAC 3'	60°C	-	180 bp
AAC13	5' CAGCATCCTTTCACTTCTGC 3' 5' TTCAGGACAACCTGGGTGAT 3'	55°C	-	184 bp
AAC14	5' CTTGTAACCCCAGCCTCAAA 3' 5' CGGCTTCTAACTCTCCAGCA 3'	58°C	-	184 bp
AAC15	5' GAGGCTTCTTCACCTTTCCA 3' 5' TAACGCTCAGGAAATGCAGG 3'	55°C	-	224 bp
AAC16	5' CACAGGAGCCCAATGGAAC 3' 5' CACCAAAATGCAAGCAGAGA 3'	58°C	-	166 bp
AAG1	5' GTCCAATCCTAGCATCCCAA 3' 5' TTGCAAGTCATTGTTGGCAG 3'	55°C	-	329 bp

AAG2	5' TATTTACCAATGCAAAGAAGAAC 3' 5' TTCGTACACGGTGGTACGG 3'	No product**	-	561 bp
AAG4	5' GCTCCTGTCCAGCCATCTAT 3' 5' GGGACCTGAACTCTGCACTT 3'	58°C	-	213 bp
AAG5	5' GTCAGGTACTCCCGATCTGC 3' 5' TCCCCACCAACTGTTTTTGT 3'	55°C	-	200 bp
AAG6	5' CTGGCTGGAATACAGAGGGA 3' 5' AACTAGTAAAACTGCCTCCTCCTC 3'	60°C	-	236 bp
AAG7	5' GAAATATTGGTAAAAGAGAGAGA 3' 5' AAATATTTTCTTTCTTTCTT 3'	Not specific***	-	400 bp
AAG9	5' TCAATCAATCAATTATGTGACCA 3' 5' TTGCTAATCACCACTGGGAA 3'	55°C	-	261 bp
AAG10	5' CTCCTGTCCAGCCTCTATTTATCT 3' 5' AAAAGTCAGATCATCAACCTCCTC 3'	64°C	-	215 bp
AAG11	5' TGAGGGAAATGCAAAATGAA 3' 5' TAACTCAGAGGGAGCCCAGT 3'	Not specific	-	480 bp
AAG12	5' TCAGAGTACAGTCTGAGATG 3' 5' GCAGGAACTGGAAAGATGCTTGCTTC 3'	58°C	-	506 bp
ACG1	5' CTTTCCCACAGCCCAAGAT 3' 5' CTCCTGCACGTCCGCATC 3'	No product	-	178 bp
ACG2	5' CCTTAACACGCAACACATGC 3' 5' GCGGGGGAATGTAGGTAGTG 3'	60°C	2 M BETAINE	323 bp
ATC1	5' TTTTCCCTCTCCCAACTCCT 3' 5' CTGATGCATGAGGGTGTCTC 3'	59°C	-	216 bp
ATC3	5' CATAGCTCTCTCGACCCCAC 3' 5' GCTGCCCTTGTGTTGATTTT 3'	58°C	-	179 bp
ATC4	5' TCGTGCATCCTGTTTACTGC 3' 5' TGGTGATGATGCTGGTAGTGA 3'	55°C	-	236 bp
ATC5	5' CGACCCTAGGATCTTGGAGA 3' 5' AGGTCCAGCTGTTCTCCTCA 3'	58°C	-	200 bp
ACC1	5' CACGACAGTGATGTGCTTCC 3' 5' TTTGGAGTTTTGGAGCTTCG 3'	60°C	-	293 bp
ACC2	5' GCCAGGGGAGAAAGTAGG 3' 5' TAATGCGGAAAGTCCGAT 3'	55°C	-	292 bp
ACC3	5' GCATCCAGACCTGGAACCTA 3' 5' AGGCTAGGCTCACAACCAAA 3'	55°C	-	190 bp

ACC4	5' TGACAAACTAGTAAGTGTAGCC 3' 5' TTGTGTTCCTGCTCTGCATC 3'	55°C	-	184 bp
ACC5	5' GAGCCCTCAGAGGAAGCAC 3' 5' ATCCCAGGGGATCTCATACC 3'	58°C	-	209 bp
ACC7	5' ATCTCCGGACATGGATGAGA 3' 5' CCGACAGCTGCTTCTTGATT 3'	58°C	-	197 bp
AGC1	5' GTGGCGTCCCACTCTTGG 3' 5' CACAGAGGACTGGCAAAGC 3'	58°C	-	190 bp
AGC2	5' GAAGGATAACAGTGAAGAG 3' 5' GATTATTCACCAGAGCAG 3'	55°C	-	285 bp
AGC3	5' AATGAAACAAAAAGTAAAGAGCAA 3' 5' GATCTGTAGTTTGGGTTTGTTGTTG 3'	55°C	-	216 bp
AGC4	5' CACAAAAGAGGAGCCCAAAA 3' 5' GGACTGCTCTCCACAAGCTG 3'	No product	-	207 bp
AGC5	5' GTTACTCCTGGAGGGCAACA 3' 5' GAACGTGCTGACAGGGATAC 3'	55°C	-	262 bp
AGC6	5' GCTGATTGGCTCTTTCGAG 3' 5' TCCCATAAACTCACCCCAAA 3'	No product		205 bp
AGC7	5' GTCACCTGAGTCCCCTGA 3' 5' TCCTCCGAGGTCTGTTTC 3'	55°C	-	185 bp
AGC8	5' CTTCTCTCTTGTTTTTCACAGTCC 3' 5' AGGAGGGCTCCTCGTGTC 3'	Not specific	-	212 bp
AGC10	5' GAGCAACGGAAAGAGAAAGC 3' 5' AATCAGGCTTGGGTTGTGAT 3'	58°C	-	195 bp
AGG1	5' CCTTCTCAAGTCCCGACT 3' 5' TCAACCTCACCAACCTTC 3'	58°C	10% DMSO	205 bp
AGG2	5' TCTTCTCCTCTTTGCTCCCC 3' 5' GCACCTTTGGAGTTCTGGAC 3'	59°C	-	206 bp
AGG3	5' AGTGAGCAAGGGGAGGTCTT 3' 5' TGTGTAGCTCCACGTCTGGT 3'	55°C	-	256 bp
AGG4	5' GTTCCGAGGTGTCAGCTCTC 3' 5' TTCTCCAGACCTGGTGACACT 3'	55°C	-	248 bp
AGG5	5' TCCGAGCCTAGCTCGTAGTC 3' 5' GGGAAAAGGAAAACCTGCTC 3'	60°C	2 M BETAINE	201 bp
AGG6	5' AGAGAGTCTGGGGTGTGGAT 3' 5' AAATCAATAGCCACGCCATC 3'	No product	-	180 bp
AGG8	5' ATGGCTGACCTGAGCTTCAT 3' 5' CAGGTGCGAACAGCACTC 3'	55°C	-	191 bp

AGG10	5' TCTTGGCTGCGTTATTGACA 3' 5' ACAGACCTGCGGAGAGGAC 3'	55°C	-	200 bp
AGG11	5' TGCCAGAGAGGGACTAGTTGA 3' 5' GTTTCAGCGAGGCTTTTCAG 3'	58°C	-	198 bp
CCG1	5' CCCCACACTCAATGACTGAA 3' 5' CAGACGTCCTCACCCTCCT 3'	60°C	2 M BETAINE	192 bp
CCG3	5' CAGTCCTTCAGCACAAACCA 3' 5' ACTTTTCCATCCGGGACTCT 3'	60°C	2 M BETAINE	165 bp
CCG10	5' TCTTGCCTAATCCCTTGACG 3' 5' GGCTTTGAGAGACCGAACAG 3'	60°C	2 M BETAINE	202 bp
CCG11	5' CTCAAGTTCGGTGGCTCAG 3' 5' TTCACTGTCCGAATCCGAGT 3'	60°C	2 M BETAINE	507 bp

*PCR product size in mouse CpG island library clone (either MF1/J or 129/J DNA)

**No product: no PCR product of the appropriate size amplified under any conditions used

***Not specific: too many PCR products were amplified under all the conditions tried

Figure 4.2 Relationships of the mouse strains, species, colonies and genera found in the DNA panel

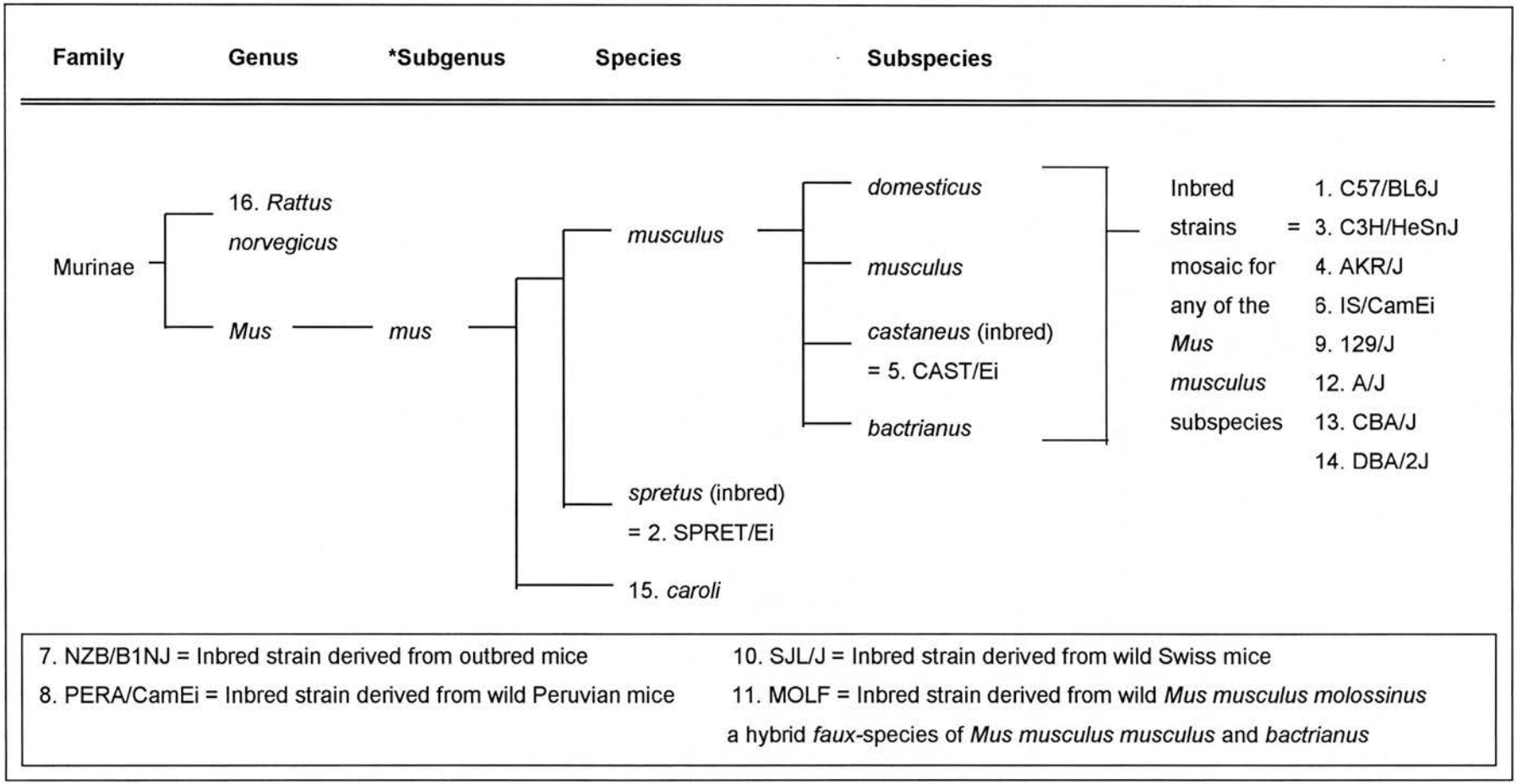


Figure 4.2 Relationships of the mouse strains, species, colonies and genera found in the DNA panel. A phylogenetic tree showing the relationships between the mouse strains, species, colonies and genera found in the DNA panel. The information represented in this figure was obtained from [Silver, 1995]. *The subgenus is not referred to in the formal name.

The PCR products were electrophoresed through high percentage (4-6%) agarose gels alongside size standards. The approximate sizes of the PCR products were used to estimate* the size of the trinucleotide repeat alleles in the mouse DNA panel. The number of different sized alleles observed in the DNA panel and the overall size range which they spanned was used as an indicator of the variability of each repeat. It proved possible to assess this variability in 51 of the 89 trinucleotide repeat arrays isolated in this study (see Table 4.2.).

*The actual size of the trinucleotide repeat was not determined directly in every case. However sequencing of the PCR products from a random selection of panels, revealed that the size changes were caused by the contraction or expansion of the trinucleotide repeat arrays. Where compound repeats or several classes of repeat occur in one clone the picture is likely to be more complex. However for the purposes of this study it has been assumed that size changes in the largest trinucleotide repeat in each clone were responsible for the overall allele size changes observed.

4.2.2 A 'variation score' to reflect the degree of size variation

In an attempt to quantify the variability of the trinucleotide repeats a method was devised to assign each one a representative score. A point was given for each trinucleotide repeat in the overall size range of alleles. This score was then multiplied by the number of *alleles that were different from that observed in the CpG island clone (MF1 or 129 DNA). Multiplication was used here because this score reflected the number of mutational events which occurred, which was considered more important than the overall size of the change.

e.g. The variation score for a repeat which had four alleles that differed in size from that in the CpG island library clone, over a range of approximately 18 bp would score 24 as follows:

$4 \times 6 = 24$

Examples of trinucleotide repeat arrays showing different degrees of variability are shown in Figure 4.3. The variation scores for each trinucleotide repeat array are described in detail in Table 4.2.

*This did not include the instances where no PCR product was observed. These results were not considered as different alleles because although changes in the primer target sequence were most likely responsible for this phenomenon, it could not be determined whether the repeat itself had changed or not.

Figure 4.3 Trinucleotide repeat size variability

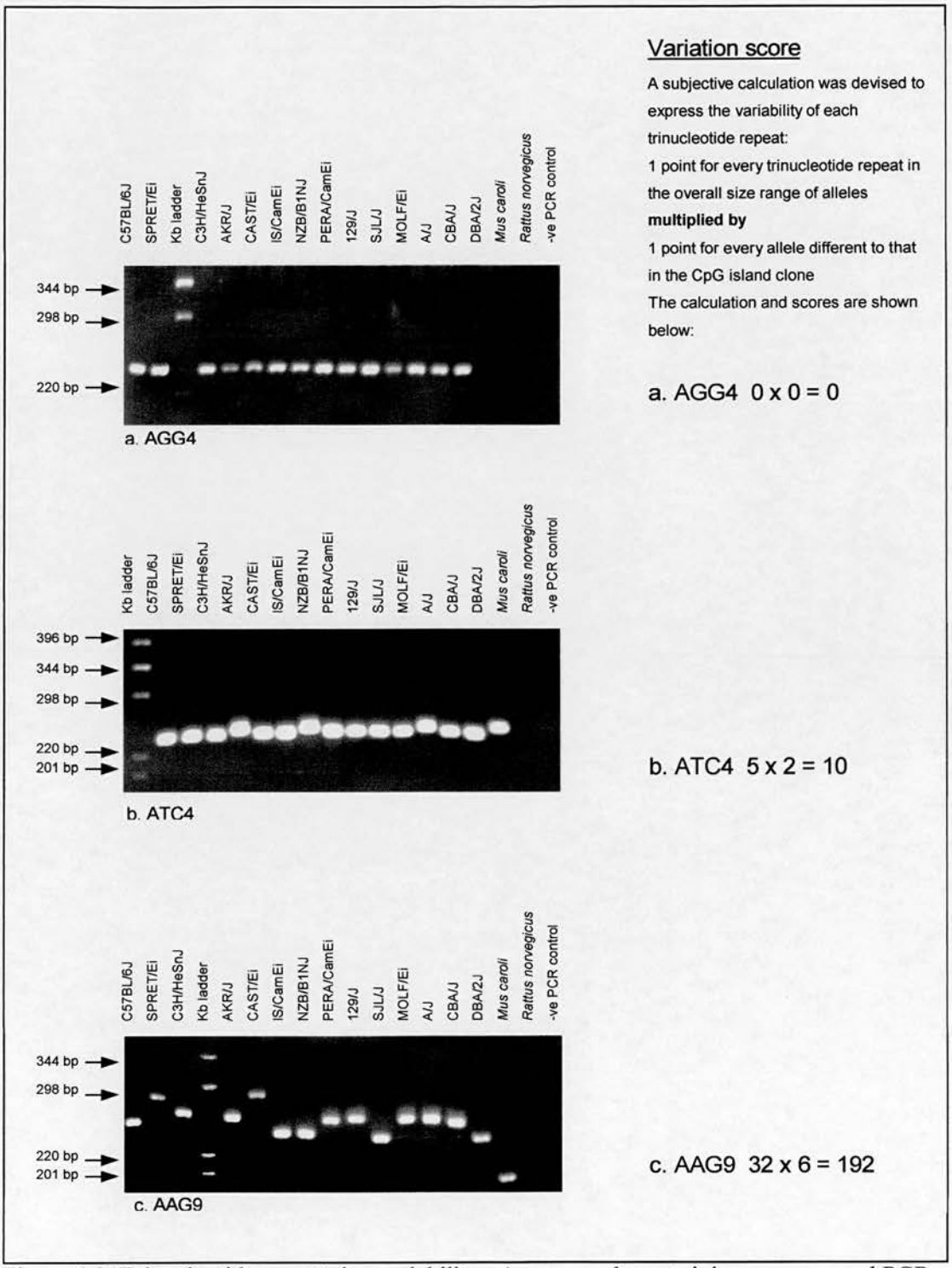


Figure 4.3 Trinucleotide repeat size variability. Agarose gels containing mouse panel PCR products which exhibit different degrees of trinucleotide repeat variability. The instances where no PCR product was observed were not considered as different alleles in the calculation of variation score.

Table 4.2 Variability of mouse trinucleotide repeat arrays

Clone name	Repeat sequence	Number of alleles*	Allele size range**	Variability score***
AAT1	$(AAT)_{15}$	6	24 bp/8 repeats	40
AAT3	$(AAT)_9$	5	21 bp/7 repeats	28
AAT4	$(AAT)_8$ $(AAC)_6$ $(ATCC)_{8,7,5,3}$ $(AACC)_5$ $(AAAC)_4$	4	21 bp/7 repeats	21
AAC1	$(AAC)_{13}$ $(AGC)_5$	7	30 bp/10 repeats	60
AAC2	$(AAC)_{12}$ $(AGC)_5(AAC)_3$	8	36 bp/12 repeats	84
AAC3	$(AAC)_9$	1	-	0
AAC4	$(AAC)_8$	3	18 bp/6 repeats	12
AAC5	$(AAC)_3(AAT)_1(AAC)_8(AAC)_3$	4	9 bp/3 repeats	9
AAC9	$(AAC)_5$	2	3 bp/1 repeat	1
AAC10	$(AAC)_5$	4	9 bp/3 repeats	9
AAC12	$(AAC)_5$ poly T	2	6 bp/2 repeats	2
AAC13	$(AAC)_5$ poly T	3	12 bp/4 repeats	8
AAC14	$(AG)_{17}(AGG)_1(AAC)_5$	4	9 bp/3 repeats	15
AAC15	$(AAC)_5(AAAC)_4$	4	15 bp/5 repeats	15
AAC16	$(AAC)_3(AGC)_3(AAC)_5$	3	12 bp/4 repeats	8
AAG1	$(AAG)_{35,3}$ $(AGG)_{13,3}$	8	54 bp/18 repeats	126
AAG4	$(AAG)_{26}$	3	15 bp/5 repeats	10
AAG5	$(AAG)_{25}$	8	66 bp/22 repeats	154
AAG6	$(AAG)_{25,4,3,3}$ $(AGG)_{6,3,3,3}$	5	39 bp/13 repeats	52
AAG9	$(AAG)_{20}(GA)_1(GAAGGA)_7$	6	96 bp/32 repeats	192
AAG10	$(AAG)_{18}$ $(AAG)_{12,3,3}$	3	24 bp/8 repeats	16

AAG12	$(AAG)_{11.11,10,9,7,7,3,3,3,3,3,3,3}$ $(AGG)_7$	4	252 bp/84 repeats	336
ACG2	$(ACG)_5$ $(AC)_4(AA)_1(AC)_7$	2	3 bp/1 repeat	1
ATC1	$(ATC)_6$	2	3 bp/1 repeat	1
ATC3	$(ATC)_5$	2	6 bp/2 repeats	2
ATC4	$(ATC)_5A(ATC)_3$	3	15 bp/5 repeats	10
ATC5	$(ATC)_5(ACC)_3$	3	9 bp/3 repeats	6
ACC1	$(ACC)_{15}$	6	96 bp/32 repeats	160
ACC2	$(AC)_6(ACC)_{10}$	6	21 bp/7 repeats	35
ACC3	$(ACC)_6$	2	3 bp/1 repeat	1
ACC4	$(ACC)_6(GCC)_3$	3	30 bp/10 repeats	20
ACC5	$(ACC)_5$	2	3 bp/1 repeat	1
ACC7	$(ACC)_4(GCC)_1(ACC)_4$	2	6 bp/2 repeats	2
AGC1	$(AGC)_{23}$	5	63 bp/21 repeats	84
AGC2	$(AGC)_{20}$	4	54 bp/18 repeats	54
AGC3	$(AGC)_{19}(AAC)_{11}$	5	57 bp/19 repeats	76
AGC5	$(ACCACCACT)_5(ACC)_1(AGC)_8$ $(CT)_{29}(GT)_{11}$	10	96 bp/32 repeats	288
AGC7	$(AGC)_7$	4	18 bp/6 repeats	18
AGC10	$(AGC)_5$	4	12 bp/4 repeats	12
AGG1	$(AGG)_{16}$ $(AGGC)_3$	4	21 bp/7 repeats	21
AGG2	$(AGG)_{12}$	4	12 bp/4 repeats	12
AGG3	$(AGG)_{12,3}$ $(AC)_{6,4,3}$	4	21 bp/7 repeats	21
AGG4	$(AGG)_8(AGT)_6$	1	-	0
AGG5	$(AGG)_6$ $(CGG)_4$	4	12 bp/4 repeats	12
AGG8	$(AGG)_5$	1	-	0
AGG10	$(AGG)_5$	1	-	0

AGG11	$(AGG)_{5,4,4,4}$	3	24 bp/8 repeats	16
CCG1	$(CCG)_8$	3	6 bp/2 repeats	4
CCG3	$(CCG)_5$	3	12 bp/4 repeats	8
CCG10	$(CCG)_{4,3}$	1	-	0
CCG11	$(CCG)_{4,4,4,4,4,3,3,3,3}$	2	3 bp/1 repeat	1

*Number of alleles = number of alleles different to those observed in the library clones, not including lack of PCR product as a separate allele

**Size range of alleles = an approximate over all size range of the different alleles observed

***Variation score = a subjective calculation to reflect the variability of each repeat (see 4.2.2)

4.2.3 Variation scores and their relation to: repeat size; CpG status; repeat class and; other flanking repetetive elements

The variation score of each trinucleotide repeat was plotted against the repeat array size* in figures 4.4.1 and 4.4.2. In the case of compound, and cryptic repeats, variation score was plotted against the size of the largest trinucleotide repeat in the array. From these figures it is apparent that in general repeat variability increases in line with the size of the repeat array. The most notable exceptions to this pattern are two relatively small repeats (AGC5 and AAG12, circled on the scatter plots) which exhibit the highest variability (variation scores 288 and 336), and a very large repeat (AAG4) with a low variability score of 10. The other striking feature of these data is that the majority of arrays analysed contained 10 repeats or less, and were inherently stable.

The variation score versus repeat size plot was broken down further into CpG island and non CpG island clones (figure 4.4.3). The majority of highly variable trinucleotide repeats, with variation scores of 100 or more were found in non CpG island DNA regions. Most repeat arrays in CpG island regions tended to be small, comprising of less than 10 tandem repeats, and were very stable with variation scores of less than 25.

The next scatter plot (figure 4.4.4) shows the data separated into the different classes of trinucleotide repeat arrays. This figure shows that although no major trend emerges, AGG repeats remained relatively stable despite size increases, and that AGC and AAG exhibited the widest range of variability.

The final scatter plot (figure 4.4.5) shows whether the trinucleotide repeat arrays were flanked by any other type or class of repeat. Most repeats exhibiting variability above and below the best-fit region, were accompanied by other repetitive elements in the clone.

*The repeat array size could arguably have been the size of the most common allele in the DNA panel, as this may well have been the 'founding' allele from which the rest arose. However, as the size of each array was only measured directly (by sequencing) in the CpG island library clone (MF1 or 129 DNA), this was the 'repeat size' referred to in the rest of the analysis. Figures 4.4 and 4.5 show that the overall pattern of distribution is not significantly affected whichever 'repeat size' is used.

Figure 4.4.1 Scatter plot showing variation score v. trinucleotide repeat array size (in CpG island library clone)

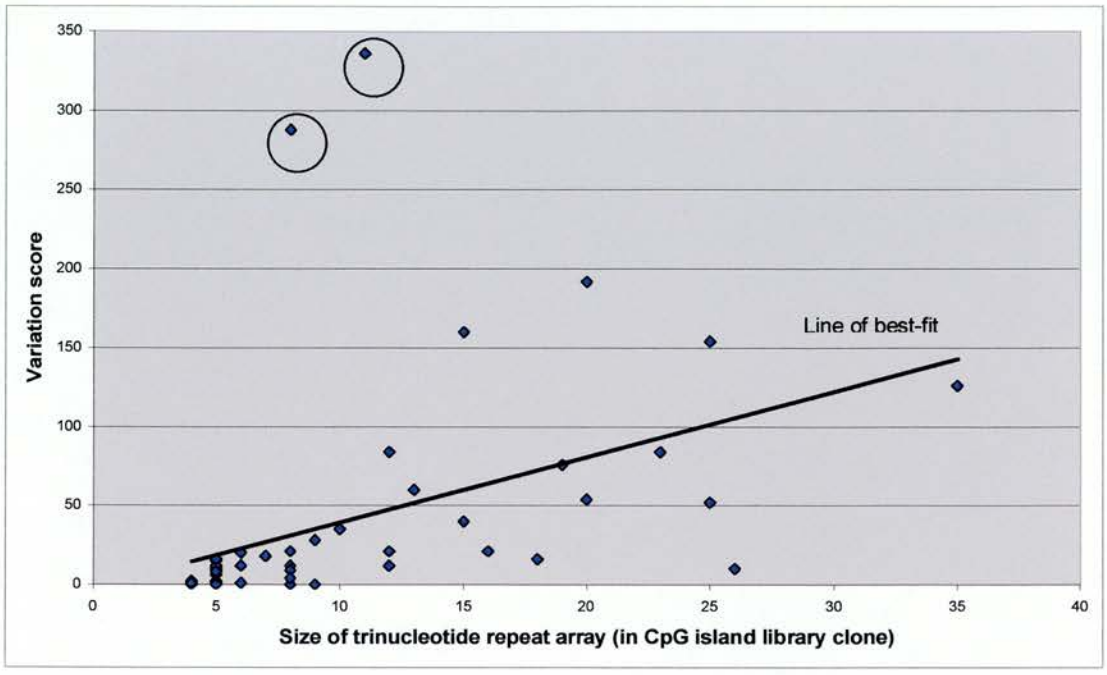


Figure 4.4.2 Scatter plot showing variation score v. trinucleotide repeat array size (most common allele in DNA panel)

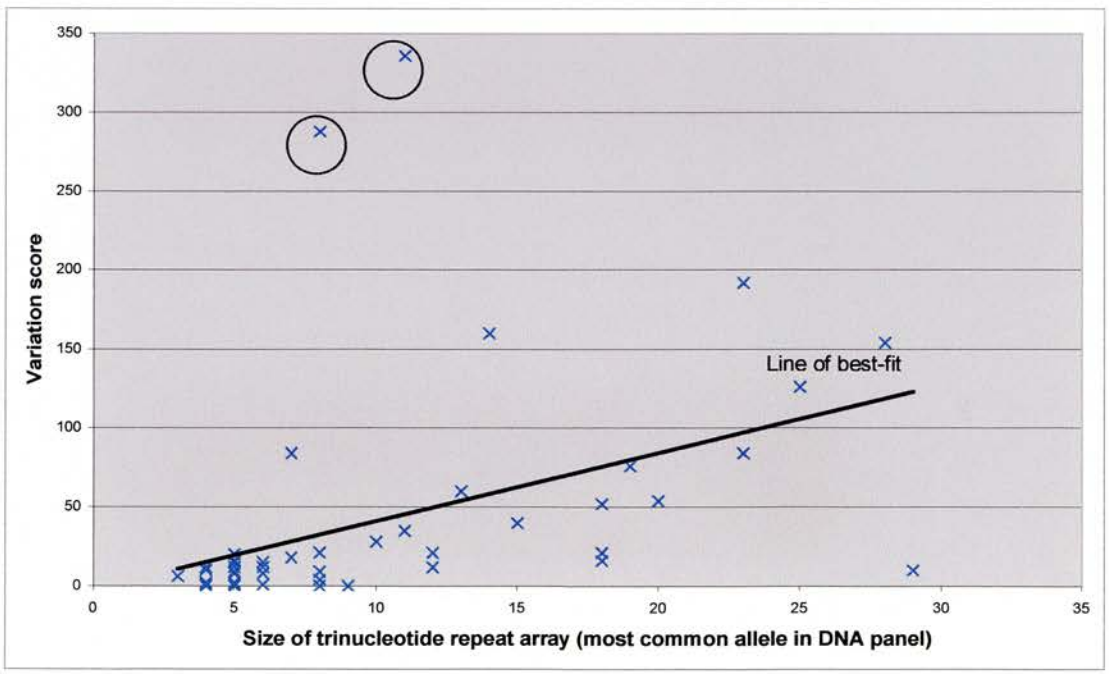


Figure 4.4.3 Scatter plot showing variation score, trinucleotide repeat array size, and clone CpG status (determined by NIX)

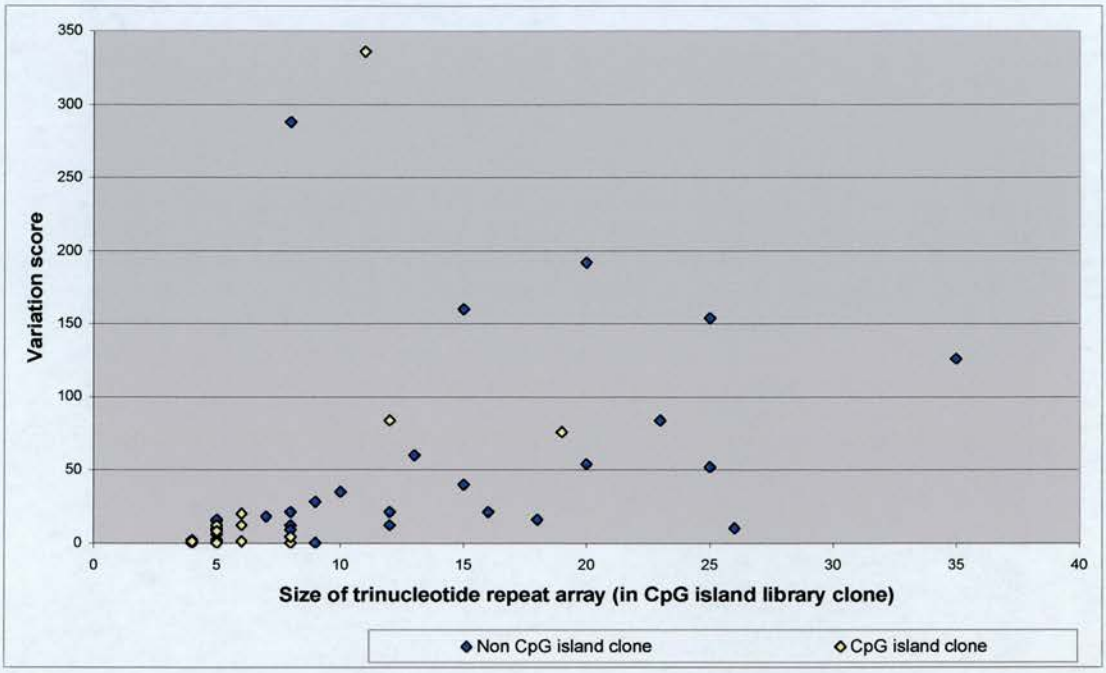


Figure 4.4.4 Scatter plot showing variation score, trinucleotide repeat array size, and class

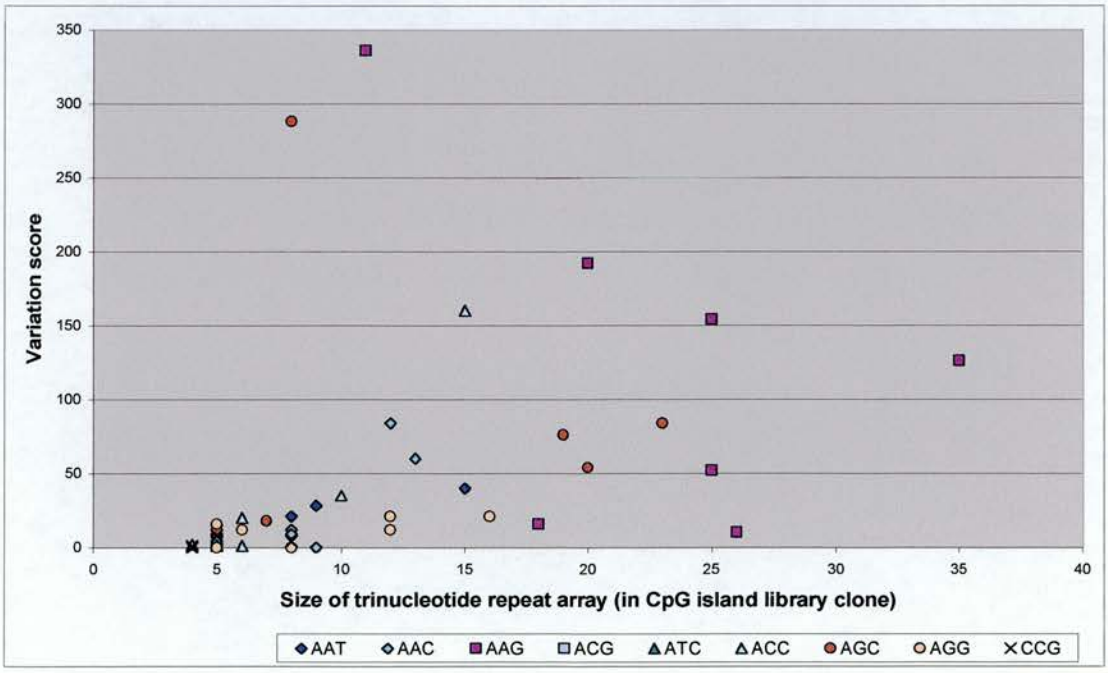
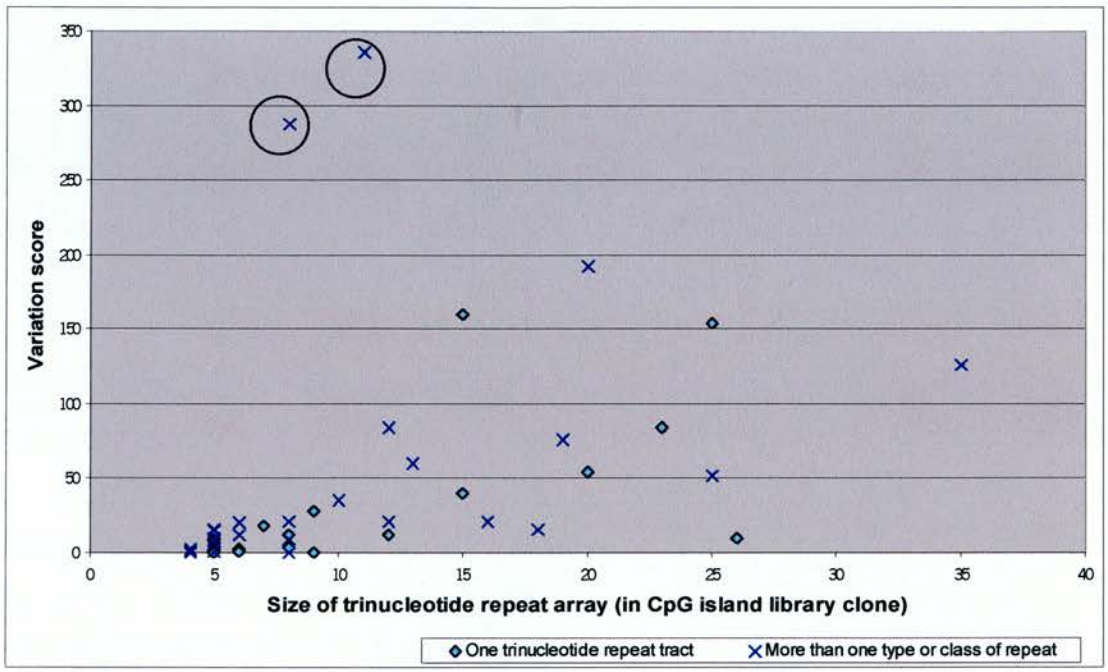


Figure 4.4.5 Scatter plot showing variation score, trinucleotide repeat array size, and clustering with other types or classes of repeat



4.2.4 Interspecies variation of repeat arrays

Out of the 51 repeat arrays analysed, only 5 (10%) did not vary in size at all amongst the panel of mouse species. From this total number, 44 (86%) showed size variation between two or more of the standard inbred strains (each derived from *Mus musculus* subspecies). In pairwise comparisons of repeat array size between C57BL/6J (*Mus musculus* subspecies) and *Mus spretus,* 26 out of 39* (67%) loci exhibited differences. Out of 32* repeat arrays, 7 (22%) showed size variability between one or more of the standard inbred strains and *Rattus norvegicus*.

The data from this study on mouse interspecies size variation of trinucleotide repeat arrays was compared to previous data from two smaller trinucleotide repeat surveys, and that from a large survey of dinucleotide repeats in tables 4.3.1-4.3.4. The first survey included nine CAG (≥7) trinucleotide repeats identified from mouse database cDNA nucleotide sequences [Abbott and Chambers, 1994]. The other trinucleotide repeat data came from a mouse brain cDNA library screen for all classes repeat, and comprised 25 repeat arrays (≥7) [Chambers and Abbott, 1996]. The dinucleotide repeat study included 300 repeat arrays (≥15) isolated from random regions of the mouse genome [Dietrich et al., 1992].

The variation between the standard inbred strains observed in this study (86%) was significantly higher that that observed in both the other trinucleotide repeat surveys (55% and 36%) and the dinucleotide repeat study (50%). The level of variation observed between C57BL/6J (*Mus musculus* subspecies) and *Mus spretus* was not significantly different from this study in either of the trinucleotide repeat datasets (67%, 78% and 64%), but was significantly lower than that seen in the dinucleotide repeat study (90%). The number of repeat arrays that were completely stable in all the mouse species tested was much higher in the survey of trinucleotide repeats from a brain cDNA library, than from the other two studies.

*At a number of trinucleotide repeat loci the PCR primers failed to amplify an allelic product in one or more of the mouse species. This was almost certainly due to an interspecific polymorphism in one of the primer target sequences. When this occurred in one of the species included in pairwise comparisons, the repeat locus was excluded from the dataset, as it was impossible to determine whether the repeat tract had changed in size.

TABLES 4.3 INTERSPECIES VARIATION

Table 4.3.1 Interspecies size variation of mouse trinucleotide repeat arrays (≥4 repeats) identified in this study

Species comparisons	n	Number/% of repeats showing variation	Number/% of repeats not varying in size
Mus musculus inbred strains	51	44/86%	7/14%
Mus musculus, *Mus spretus*	39	26/67%	13/33%
All mouse species in panel	51	46/90%	5/10%

Table 4.3.2 Interspecies size variation of mouse CAG trinucleotide repeat arrays (≥7 repeats) from cDNA sequences [Abbott and Chambers, 1994],compared to data from this study (table 4.3.1)

Species comparisons	n	variation	No variation	X^2	p	Q	Sig.
Mus musculus inbred strains	9	5/55%	4/45%	4.822	0.05	0.0281	**Yes**
Mus musculus, *Mus spretus*	9	7/78%	2/22%	0.4202	0.05	0.5168	**No**
All mouse species in panel	9	8/89%	1/11%	0.0145	0.05	0.9041	**No**

X^2 = Chi-square value

p = Probability of occurrence by chance

Q = Probability of non chance occurrence

Sig. = Significance to one degree of freedom

Table 4.3.3 Interspecies size variation of mouse trinucleotide repeat arrays (≥5 repeats) from a brain cDNA library [Chambers and Abbott, 1996], compared to data from this study (table 4.3.1)

Species comparisons	n	variation	No variation	X^2	*p*	*Q*	Sig.
Mus musculus inbred strains	25	9/36%	16/64%	20.0917	0.001	0.00001	**Yes**
Mus musculus, *Mus spretus*	25	16/64%	9/36%	0.0480	0.05	0.8265	**No**
All mouse species in panel	25	18/72%	7/28%	4.1775	0.05	0.0410	**Yes**

Table 4.3.4 Interspecies size variation of random mouse dinucleotide repeat arrays (≥15 repeats) [Dietrich et al., 1992], compared to data from this study (table 4.3.1)

Species comparisons	n	variation	No variation	X^2	*p*	*Q*	Sig.
Mus musculus inbred strains	300	150/50%	150/50%	23.2007	0.001	0.000001	Yes
Mus musculus, *Mus spretus*	300	270/90%	30/10%	16.9659	0.001	0.00003	Yes

X^2 = Chi-square value

p = Probability of occurrence by chance

Q = Probability of non chance occurrence

Sig. = Significance to one degree of freedom

4.2.5 Map locations of polymorphic trinucleotide repeat arrays

Where a polymorphic repeat size difference (including non amplification of an allelic PCR product) existed between C57BL/6J and *Mus spretus* mice, the trinucleotide repeat loci were mapped by PCR amplification of the BSS mouse interspecific backcross DNA panel (for more information see 4.1.6).

An example typing of the trinucleotide repeat locus AGC5 using the backcross panel is depicted in figures 4.5 and 4.6. How this information was incorporated into the current dataset, resulting in a map position is summarised in figure 4.7. This analysis was carried out by Lucy Rowe at The Jackson Laboratory, using the Map Manager program [Manly, 1993]. In essence the typings were grouped with identical, or similar datasets produced from the 4986 loci already incorporated into this map. The distance from flanking groups was estimated by the number of recombination events separating the two groups of data. For example, one recombination event out of 94 typings (1.06%) was considered equivalent to a relative distance of 1.06 cM ± 1.06, and three recombination events (3.19%) would predict a relative distance of 3.19 cM ± 1.81 from flanking markers.

Of the trinucleotide repeat arrays identified in this survey, 34 were successfully mapped using this protocol. In each instance all 94 N_2 offspring were successfully typed, ensuring correct map position assignment. The map positions of the loci are summarised in figure 4.8, and described in further detail in table 4.4. The complete BSS mapping datasets can be directly accessed at: www.jax.org/resources/documents/cmdata/bkmap/BSS.html.

The majority of the trinucleotide repeat arrays (approximately 50%) clustered on chromosomes 1, 6, 7 and 8. Two sets of repeats (AAC4/ *D1Abb2* and AGG1/*D1Abb3*, AAG1/ *D7Abb2* and AGG11/ *D7Abb3*) were assigned to indistinguishable map regions. Aside from these obvious groupings, the repeat arrays were distributed relatively evenly throughout the genome, although non were mapped to chromosomes 4, 16, 18 and 19.

Figure 4.5 PCR amplification of the BSS interspecific backcross DNA panel with AGC5 primers

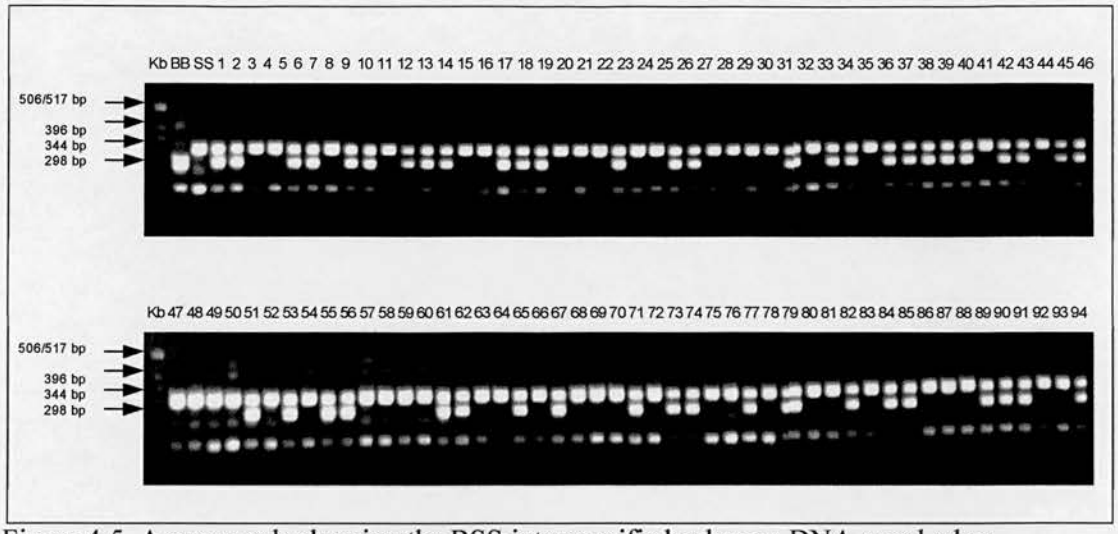


Figure 4.5 Agarose gels showing the BSS interspecific backcross DNA panel when amplified with trinucleotide repeat AGC5 PCR primers.

BB = C57BL/6J parent; SS = SPRET/Ei parent; 1-94 = N_2 backcross offspring animals 1-94.

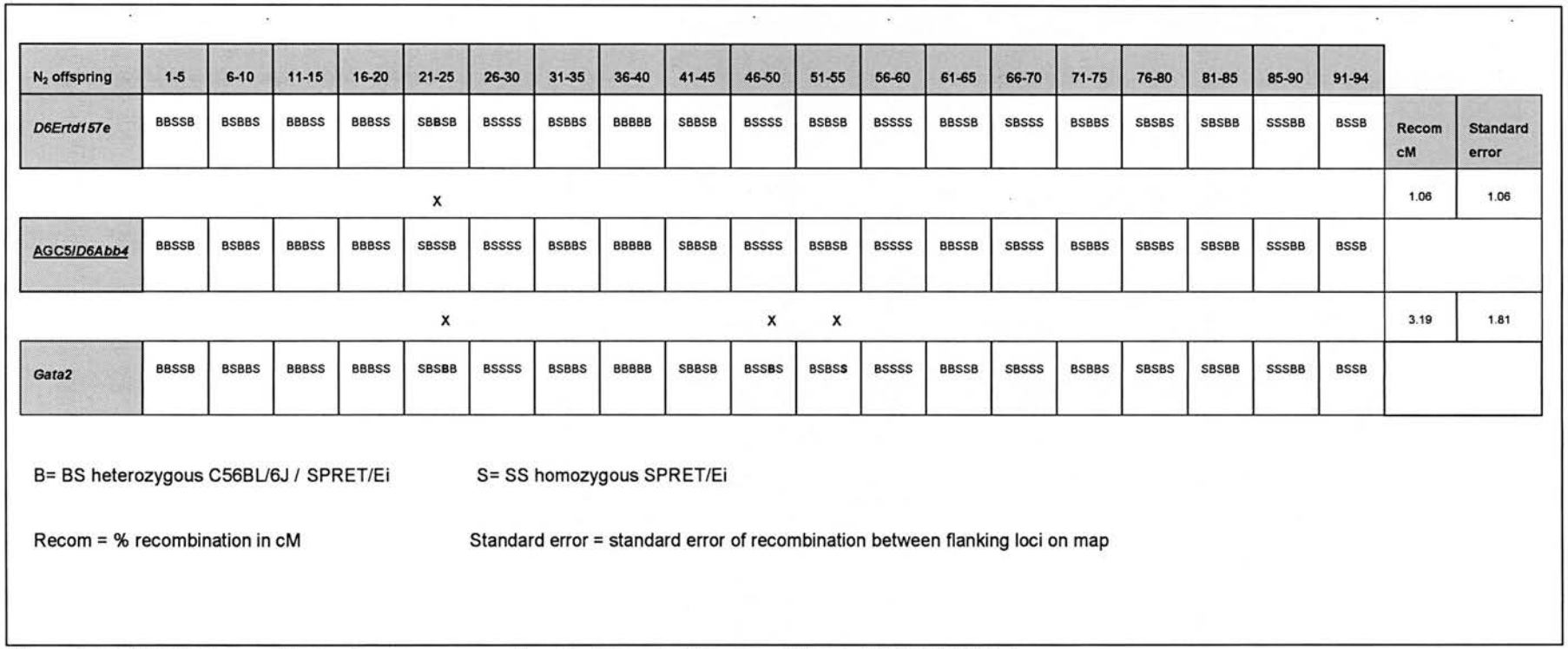
Figure 4.6 Typing of the BSS interspecific backcross DNA panel for the trinucleotide repeat locus AGC5

BB	SS	1 BS	2 BS	3 SS	4 SS	5 BS	6 BS	7 SS	8 BS	9 BS	10 SS
11 BS	12 BS	13 BS	14 SS	15 SS	16 BS	17 BS	18 BS	19 SS	20 SS	21 SS	22 BS
23 SS	24 SS	25 BS	26 BS	27 SS	28 SS	29 SS	30 SS	31 BS	32 SS	33 BS	34 BS
35 SS	36 BS	37 BS	38 BS	39 BS	40 BS	41 SS	42 BS	43 BS	44 SS	45 BS	46 BS
47 SS	48 SS	49 SS	50 SS	51 BS	52 SS	53 BS	54 SS	55 BS	56 BS	57 SS	58 SS
59 SS	60 SS	61 BS	62 BS	63 SS	64 SS	65 BS	66 SS	67 BS	68 SS	69 SS	70 SS
71 BS	72 SS	73 BS	74 BS	75 SS	76 SS	77 BS	78 SS	79 BS	80 SS	81 SS	82 BS
83 SS	84 BS	85 BS	86 SS	87 SS	88 SS	89 BS	90 BS	91 BS	92 SS	93 SS	94 BS

Figure 4.6 Typing of the BSS interspecific backcross DNA panel for the trinucleotide repeat locus AGC5.

BB = homozygous C57BL/6J; SS = homozygous SPRET/Ei; BS = heterozygous.

Figure 4.7 BSS interspecific backcross map position for trinucleotide repeat locus AGC5



N_2 offspring	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50	51-55	56-60	61-65	66-70	71-75	76-80	81-85	85-90	91-94	Recom cM	Standard error
D6Ertd157e	BBSSB	BSBBS	BBBSS	BBBSS	SBBSB	BSSSS	BSBBS	BBBBB	SBBSB	BSSSS	BSBSB	BSSSS	BBSSB	SBSSS	BSBBS	SBSBS	SBSBB	SSSBB	BSSB		
					X															1.06	1.06
AGC5/D6Abb4	BBSSB	BSBBS	BBBSS	BBBSS	SBSSB	BSSSS	BSBBS	BBBBB	SBBSB	BSSSS	BSBSB	BSSSS	BBSSB	SBSSS	BSBBS	SBSBS	SBSBB	SSSBB	BSSB		
					X					X	X									3.19	1.81
Gata2	BBSSB	BSBBS	BBBSS	BBBSS	SBSBB	BSSSS	BSBBS	BBBBB	SBBSB	BSSBS	BSBSS	BSSSS	BBSSB	SBSSS	BSBBS	SBSBS	SBSBB	SSSBB	BSSB		

B= BS heterozygous C56BL/6J / SPRET/Ei S= SS homozygous SPRET/Ei

Recom = % recombination in cM Standard error = standard error of recombination between flanking loci on map

Figure 4.11 BSS interspecific backcross map position for trinucleotide repeat locus AGC5/*D6Abb4*

Mapped to chromosome 6, 35.2 cM distal from the most centromeric anchor locus.

Showing linkage relationship to nearest flanking recombinant markers *D6Ertd157e* and *Gata2*.

Figure 4.8 Linkage map positions of mouse trinucleotide repeat loci

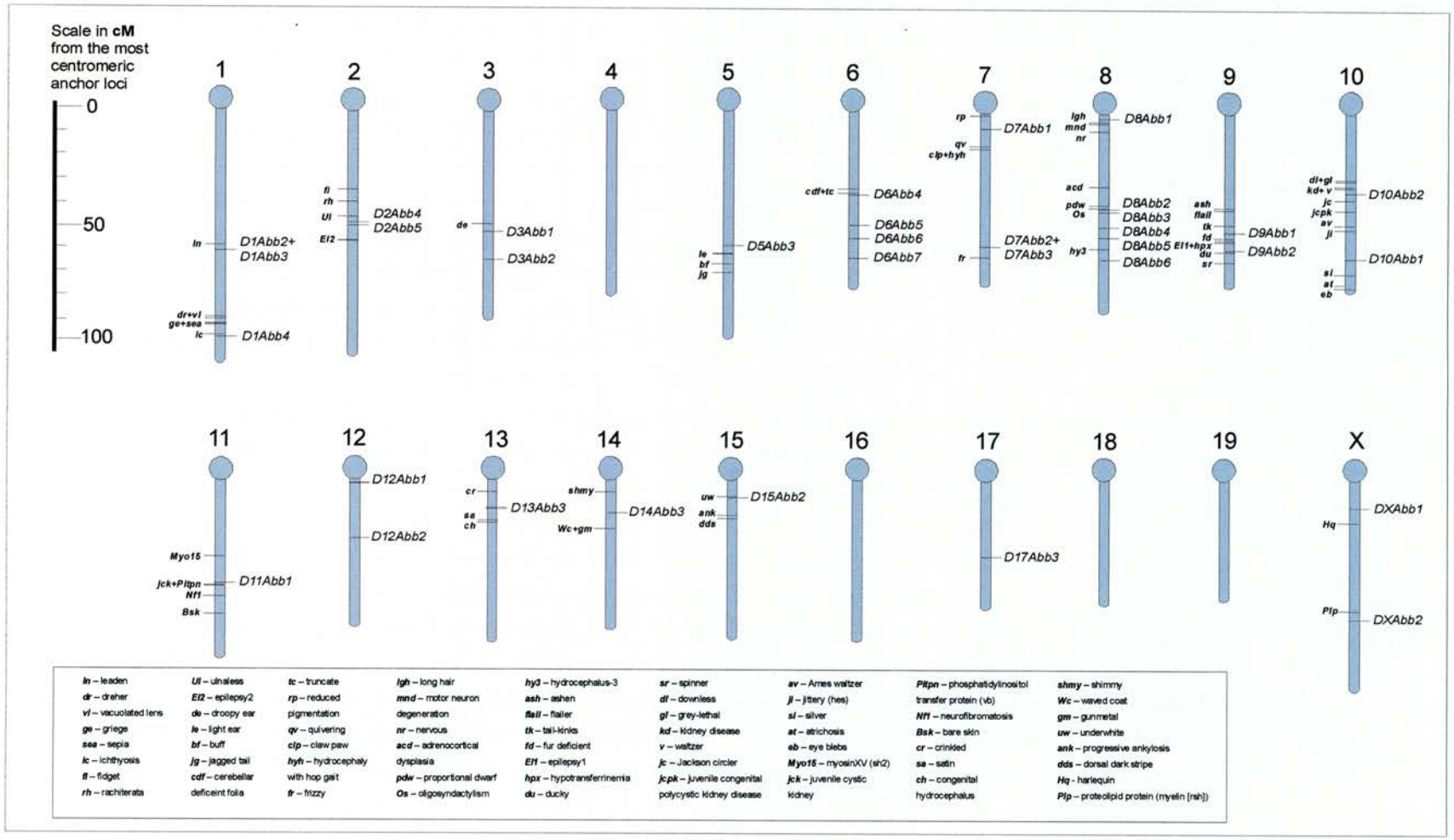


Table 4.4 Map positions of mouse trinucleotide repeat arrays

Chrom.	Locus symbol	Repeat name	Position cM distal to centromeric anchor locus	Flanking recombinant markers	% Recom. in N_2 offspring	Standard error ±
1	*D1Abb2*	AAC4	60.2 cM	*D1Mit440*	1.06 cM	1.06
				D1Ertd507e	2.13 cM	1.49
1	*D1Abb3*	AGG1	60.2 cM	*D1Mit440*	1.06 cM	1.06
				D1Ertd507e	2.13 cM	1.49
1	*D1Abb4*	AAC10	98 cM	*Hkp1*	1.06 cM	1.06
				D1Ertd75e	1.06 cM	1.06
2	*D2Abb4*	AAC13	46.5 cM	*D2Mit158*	1.06 cM	1.06
				Prg2	2.13 cM	1.49
2	*D2Abb5*	AAG6	49.5 cM	*D2Mit299*	2.13 cM	1.49
				Alx4	3.19 cM	1.81
3	*D3Abb1*	AAC15	52.6 cM	*Gabpb2*	2.13 cM	1.49
				Gja8	1.06 cM	1.06
3	*D3Abb2*	AGC1	64 cM	*D3Mit13*	1.06 cM	1.06
				Pitx2	1.06 cM	1.06
5	*D5Abb3*	ATC5	59 cM	*Vdr*	1.06 cM	1.06
				Pxn	1.06 cM	1.06
6	*D6Abb4*	AGC5	35.2 cM	*D6Ertd157e*	1.06 cM	1.06
				Gata2	3.19 cM	1.81

6	*D6Abb5*	ACC4	49 cM	*Lmcd1*	1.06 cM	1.06
				Mbd4	1.06 cM	1.06
6	*D6Abb6*	AAG9	55 cM	*D6Mit288*	1.06 cM	1.06
				Syb1	1.06 cM	1.06
6	*D6Abb7*	AAT3	64.4 cM	*Kcna1*	1.06 cM	1.06
				Emp1	1.06 cM	1.06
7	*D7Abb1*	ACC2	7.7 cM	*Ier4*	1.06 cM	1.06
				Aplp1	2.13 cM	1.49
7	*D7Abb2*	AAG1	59 cM	*D7Ertd629e*	1.06 cM	1.06
				D7Mit8	1.06 cM	1.06
7	*D7Abb3*	AGG11	59 cM	*D7Ertd629e*	1.06 cM	1.06
				D7Mit8	1.06 cM	1.06
8	*D8Abb1*	AGG2	4 cM	*D8Ertd387e*	1.06 cM	1.06
				D8Mit61	3.19 cM	1.81
8	*D8Abb2*	AAC2	45 cM	*Grid1-rs*	2.13 cM	1.49
				Mt2	1.06 cM	1.06
8	*D8Abb3*	AAC1	45 cM	*Grid1-rs*	2.13 cM	1.49
				Mt2	1.06 cM	1.06
8	*D8Abb4*	AAG12	50 cM	*D8Hun12*	1.06 cM	1.06
				Terf2	1.06 cM	1.06
8	*D8Abb5*	AAT4	56 cM	*D8Ertd107e*	1.06 cM	1.06
				Erdt107e	1.06 cM	1.06
8	*D8Abb6*	ATC4	66 cM	*Fkh14*	1.06 cM	1.06
				Acta1	1.06 cM	1.06

9	*D9Abb1*	AAT1	51 cM	*Ctsh*	3.19 cM	1.81
				Pk3	2.13 cM	1.49
9	*D9Abb2*	AGC10	60 cM	*Gpx1*	1.06 cM	1.06
				D9Ertd241e	1.06 cM	1.06
10	*D10Abb2*	AGC3	33 cM	*Denn*	1.06 cM	1.06
				Egr2	2.13 cM	1.49
10	*D10Abb1*	AGC7	63 cM	*D10Hun5*	1.06 cM	1.06
				D10Mit35	1.06 cM	1.06
11	*D11Abb1*	AGC2	44 cM	*D11Mit279*	4.25 cM	2.08
				Abr	1.06 cM	1.06
12	*D12Abb1*	AAG10	1 cM	*Apob*	0.00 cM	0.00
				Syndl	1.06 cM	1.06
12	*D12Abb2*	CCG11	24 cM	*Ifnl*	1.06 cM	1.06
				D12Mit54	4.25 cM	2.08
13	*D13Abb3*	ACC1	12 cM	*D13Mit3*	1.06 cM	1.06
				Gpld1	1.06 cM	1.06
14	*D14Abb3*	AAG5	13.5 cM	*D14Ertd817e*	1.06 cM	1.06
				Grid1	1.06 cM	1.06
15	*D15Abb2*	AGG5	6.8 cM	*Npr3*	0.00 cM	0.00
				D15Ertd600e	1.06 cM	1.06

17	*D17Abb3*	AAC5	32 cM	*D17Mit180* *D17Ertd479e*	2.13 cM 1.06 cM	1.49 1.06
X	*DXAbb1*	CCG10	12.4 cM	*DXTrk1* *DXMit48*	2.13 cM 2.13 cM	1.49 1.49
X	*DXAbb2*	AGG3	60 cM	*DXMit4* *Gucy2f*	2.13 cM 1.06 cM	1.49 1.06

KEY:

Chrom. = Chromosome

Recom. = % Recombination observed in N_2 offspring with nearest flanking marker, expressed in cM

4.2.6 Assesment of trinucleotide repeats as 'candidates' for characterised mouse mutant phenotypes

The mapped trinucleotide repeat arrays were assessed as 'candidate loci', for mapped mutant phenotypes. In other words, to determine whether the trinucleotide repeat arrays were equivalent to other mapped loci characterised only at the level of mutant phenotypes.

Obviously, the BSS mapping panel used to localise the trinucleotide repeat arrays was not used to provide simultaneous map information for the majority of phenotypically defined loci. Therefore it was necessary to compare map positions derived from different crosses, and those obtained by completely different mapping techniques.

The search for potentially equivalent mutationally defined loci began by ruling out loci considered too distant to be relevant. Database lists of all loci were scanned, and mutationally defined loci lying approximately 10 cM or less away on the map were considered worth assessing further. The database lists screened were the most up to date (1999) mouse chromosome committee reports at the time (www.informatics.jax.org/ccr/searches/index.cgi?year=1999). Descriptions of the phenotypes associated with the loci identified by this scan were also obtained at The Jackson Laboratory website (www.informatics.jax.org/searches/marker_form.shtml).

Once a list of loci had been compiled for further consideration, the likelihood of their having equivalent map positions to the trinucleotide repeat loci was noted. This was based on the confidence level of each map position derived from the type and resolution of mapping used. An attempt was then made to obtain DNA from all these mouse mutants. However, where DNA was not available for screening, the list was amended accordingly.

To assess whether the trinucleotide repeat (or an expanded version of the trinucleotide repeat) was likely to be responsible for each mutant phenotype, the size of the array in mutant mice and non mutant mice from the same strain were determined by PCR. This was carried out on DNA from all the mouse mutants listed in table 4.5. and figure 4.8.

Table 4.5 Mutant loci screened for link with trinucleotide repeats

Locus symbol	Repeat name	Position cM from centromeric anchor locus	Mouse mutant loci	Position cM from centromeric anchor locus	Confidence level of map position*
D1Abb2 *D1Abb3*	AAC4 AGG1	60.2 cM 60.2 cM	ln /leaden	59 cM	intermediate
D1Abb4	AAC10	98 cM	dr /dreher vl /vacuolated lens ge /griege sea /sepia ic /ichthyosis	88.5 cM 89.7 cM 92.1 cM 92.1 cM 97.2 cM	intermediate intermediate intermediate intermediate intermediate
D2Abb4 *D2Abb5*	AAC13 AAG6	46.5 cM 49.5 cM	fi /fidget rh /rachiterata Ul /ulnaless El2 /epilepsy2	34 cM 38 cM 45 cM 57 cM	intermediate intermediate intermediate low
D3Abb1	AAC15	52.6 cM	de /droopy ear	48.8 cM	high
D3Abb2	AGC1	64 cM	No mutants available	-	-
D5Abb3	ATC5	59 cM	le /light ear bf /buff jg /jagged tail	60 cM 64 cM 67 cM	low low low
D6Abb4	AGC5	35.2 cM	cdf /cerebellar deficient folia tc /truncate	33.5 cM 35.7 cM	intermediate low
D6Abb5 *D6Abb6* *D6Abb7*	ACC4 AAG9 AAT3	49 cM 55 cM 64.4 cM	No mutants available	- - -	- - -
D7Abb1	ACC2	7.7 cM	rp /reduced pigmentation qv /quivering clp /claw paw hyh /hydrocephaly with hop gait	2 cM 14.4 cM 15.2 cM 15.2 cM	intermediate intermediate intermediate intermediate
D7Abb2 *D7Abb3*	AAG1 AGG11	59 cM 59 cM	fr /frizzy	64 cM	intermediate

D8Abb1	AGG2	4 cM	lgh /long hair	5 cM	intermediate
			mnd /motor neuron degeneration	6 cM	intermediate
			nr /nervous	8 cM	intermediate
D8Abb2	AAC2	44 cM	acd /adrenocortical dysplasia	31 cM	intermediate
D8Abb3	AAC1	45 cM			
D8Abb4	AAG12	50 cM	pdw /proportional dwarf	39 cM	intermediate
			Os /oligosyndactylism	40 cM	intermediate
			hy3 /hydrocephalus-3	58 cM	intermediate
D8Abb5	AAT4	56 cM	hy3 /hydrocephalus-3	58 cM	intermediate
D8Abb6	ATC4	66 cM			
D9Abb1	AAT1	51 cM	ash /ashen	41 cM	intermediate
			flail /flailer	42 cM	high
			tk /tail-kinks	48 cM	high
			fd /fur deficient	54 cM	intermediate
			El1 /epilepsy1	55 cM	intermediate
			hpx /hypotransferrinemia	56 cM	intermediate
D9Abb2	AGC10	60 cM	El1 /epilepsy1	55 cM	intermediate
			hpx /hypotransferrinemia	56 cM	intermediate
			du /ducky	60 cM	intermediate
			sr /spinner	64 cM	intermediate
D10Abb2	AGC3	33 cM	dl /downless	29 cM	intermediate
			gl /grey-lethal	29 cM	intermediate
			kd /kidney disease	30 cM	intermediate
			v /waltzer	30.3 cM	high
			jc /Jackson circler	32 cM	intermediate
			jcpk /juvenile congenital polycystic kidney disease	38.6 cM	high
			av /Ames waltzer	40.2 cM	low
			ji /jittery (hesitant)	43 cM	intermediate
D10Abb1	AGC7	63 cM	si /silver	70 cM	high
			at /atrichosis	75 cM	intermediate
			eb /eye blebs	77 cM	intermediate

D11Abb1	AGC2	44 cM	Myo15 /myosinXV (shaker2)	33.9 cM	high
			jck /juvenile cystic kidney	44 cM	high
			Pitpn /phosphatidylinositol transfer protein (vibrator)	44.11 cM	high
			Nf1 /neurofibromatosis	46.06 cM	high
			Bsk /bare skin	58 cM	intermediate
D12Abb1	AAG10	1 cM	No mutants available	-	-
D12Abb2	CCG11	24 cM		-	-
D13Abb3	ACC1	12 cM	cr /crinkled	6 cM	low
			sa /satin	17 cM	high
			ch /congenital hydrocephalus	18 cM	low
D14Abb3	AAG5	13.5 cM	shmy /shimmy	6.5 cM	low
			Wc /waved coat	20 cM	intermediate
			gm /gunmetal	20.7 cM	intermediate
D15Abb2	AGG5	6.8 cM	uw /underwhite	6.7 cM	high
			ank /progressive ankylosis	14.4 cM	intermediate
			dds /dorsal dark stripe	15.9 cM	intermediate
D17Abb3	AAC5	32 cM	No mutants available	-	-
DXAbb1	CCG10	12.4 cM	Hq /harlequin	17 cM	intermediate
DXAbb2	AGG3	60 cM	Plp /proteolipid protein (myelin {rump shaker})	56 cM	high

KEY:

*The confidence level of the map position based on the type and resolution of mapping used

High = mapped to a high resolution in at least one cross with multiple other loci

Intermediate = mapped in one or few crosses relative to one or a few markers

Low = mapped by *in-situ* hybridisation only or in a quantitative trait loci (QTL) analysis

4.2.6.1 Two highly variable repeat arrays as candidates for the frizzy mutant phenotype

Two repeat loci (AAG1 and AGG11) exhibited significant size increases when assayed in frizzy (*fr*) mutant DNA (see figure 4.9). These repeats co-localise to 59 cM (from the centromeric anchor locus) on chromosome 7, within 5 cM of the frizzy mutation.

The frizzy mutation arose in a stock of mixed origin at The Jackson Laboratory in 1949 [Falconer and Snell, 1952]. The stock was derived from a cross between DBA and a BALB/c strain carrying the pink-eyed, chinchilla and shaker-1 (*sh-1*) mutations. The frizzy mutation is recessive and results in wavy or curly vibrissae (snout hairs) and coat.

The repeat containing alleles were approximately 150 bp (AAG1) and 40 bp (AGG11) larger in frizzy DNA than in normal BALB/c and DBA/2J mice. Both these loci were therefore considered further as possible candidates for causing the frizzy mutant phenotype.

4.2.6.2 Molecular basis for the increase in allele sizes

The basis for the size increases in the frizzy alleles of AAG1 and AGG11 was determined by sequencing the PCR products and comparing any differences from BALB/c and DBA/2J.

DNA	AAG1	PCR product size
BALB/c	$(AGG)_9$, $(AAG)_{35}$	329 bp
DBA/2J	$(AGG)_8$, $(AAG)_7$, (AAGG), $(AGG)_{10}$, (AG), $(AAG)_9$	293 bp
frizzy	$(AGG)_6$, $(AAGGAG)_{22}$, $(AAG)_{35}$	440 bp

The massive size increase in the AAG1 frizzy allele was the result of a large hexanucleotide repeat arising at the junction between the AGG and AAG repeat tracts.

DNA	AGG11	PCR product size
BALB/c	$(AGG)_4$, $(AGGG)_2$, (AG), $(AGG)_4$, (AGC), $(AGG)_5$	566 bp
DBA/2J	$(AGG)_4$, $(AGGG)_2$, (AG), $(AGG)_4$, (AGC), $(AGG)_5$	566 bp
frizzy	$(AGG)_7$, (GAG), $(AGG)_6$, (AGGGAG), $(AGG)_4$, (AGC), $(AGG)_{10}$	608 bp

The AGG11 frizzy allele was caused by widespread incremental size increases in several components of this complex compound repeat.

Figure 4.9 Size increases observed in repeat arrays AAG1 and AGG11 in frizzy DNA

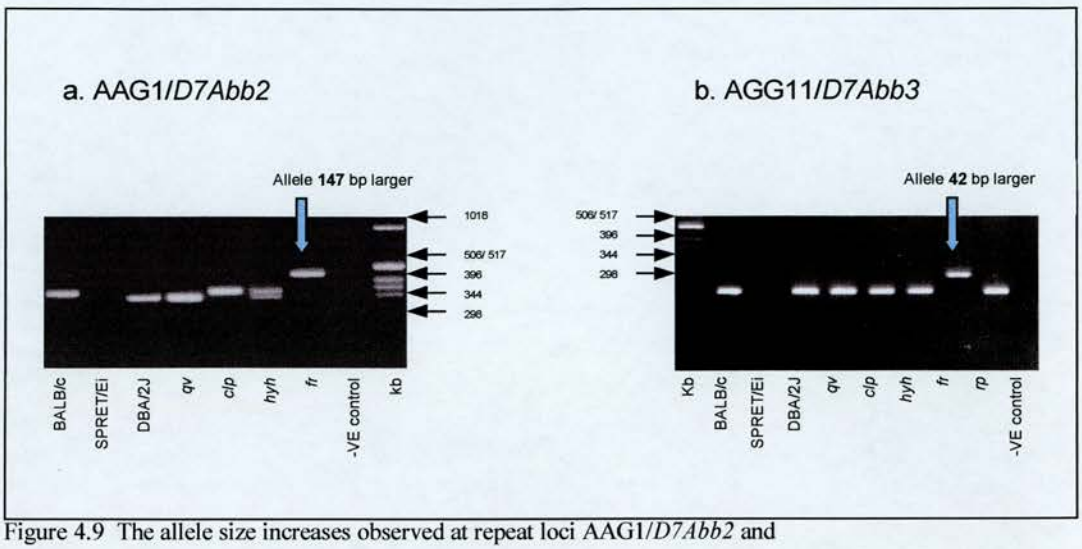


Figure 4.9 The allele size increases observed at repeat loci AAG1/*D7Abb2* and AGG11/*D7Abb3* in frizzy (*fr*) DNA, compared to normal BALB/c and DBA/2J DNA (the genetic background upon which the frizzy mutation arose).

Figure 4.10 The repeat changes observed at loci AAG1 and AGG11 are not present in other frizzy stocks

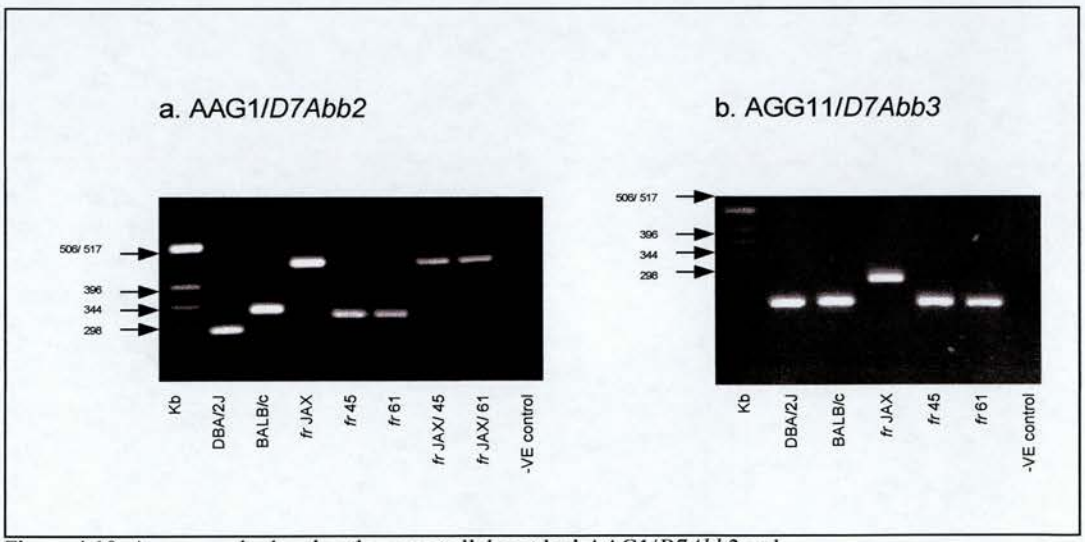


Figure 4.10 Agarose gels showing the repeat alleles at loci AAG1/*D7Abb2* and AGG11/*D7Abb3* in frizzy DNA from different stocks. *fr* JAX = frizzy DNA from The Jackson Laboratory. *fr* 45/ 61 = frizzy DNA from the Baylor College of Medicine, Houston, Texas. The repeats do not cause the frizzy phenotype.

To be sure that the two repeat loci (AAG1 and AGG11) were not simply allelic variants of the same locus (the CpG island library was constructed from a mixture of MF1 and 129 mouse DNA), their nucleotide sequence was compared for similarities using LALIGN at HGMP-RC (www.hgmp.mrc.ac.uk/Registered/Webapp/lalign/). No nucleotide sequence similarities between the two repeat containing loci were observed.

4.2.6.3 Analysis of DNA obtained from independently maintained frizzy stocks

To further assess whether changes at either repeat loci might cause the frizzy phenotype DNA was obtained from a living colony of frizzy mice. These mice were originally derived from The Jackson Laboratory strain, but had been maintained independently at Oak Ridge Laboratory, and then at the Baylor College of Medicine, Houston, Texas.

PCR analysis of DNA from two mouse lines (lines 45 and 61) from this independent colony of mice revealed that the repeat changes were unique to the Jackson Laboratory strain of frizzy mice, and therefore could not be the causative mutations (see figure 4.10).

4.2.6.4 Determination of the genetic background surrounding the AAG1, AGG11 and frizzy loci, using MIT markers

To determine the genetic background around the repeat loci AAG1/*D7Abb2*, AGG11/*D7Abb3* and the frizzy mutation, the region was typed as DBA or BALB/c utilising the known polymorphisms of MIT markers in the region. The typing was performed by PCR at markers *D7Mit7*, *D7Mit68*, *D7Mit206*, *D7Mit105*, *D7Mit12* and *D7Mit109*, with the following results:

LOCI	POSITION	TYPE OF DNA IN REGION
D7Mit7	54 cM	DBA
D7Abb2/AAG1	59 cM	
D7Abb3/AGG11	59 cM	
D7Mit68	60 cM	DBA
D7Mit206	60 cM	DBA
D7Mit105	63.5 cM	BALB/c
frizzy (*fr*)	64 cM	
D7Mit12	66 cM	BALB/c
D7Mit109	66 cM	BALB/c

The two repeat loci AAG1 and AGG11 are located in DNA of DBA origin, the frizzy mutation is assigned to a BALB/c genetic background.

4.2.6.5 Typing of frizzy DNA at all variable trinucleotide repeat loci

To establish whether the frizzy phenotype might cause widespread genomic instability, which might account for the size increases observed in the two chromosome 7 loci, the JAX frizzy DNA was checked at every variable repeat locus identified in this study. The frizzy DNA was not found to differ from BALB/c or DBA DNA at any of these repeat loci.

4.3 Discussion

4.3.1 Size variation of trinucleotide repeats

4.3.1.1 Effect of repeat array size

In this survey of mouse trinucleotide repeats, variability was generally observed to increase in line with the size of the repeat array. This relationship has previously been described as a key determinant in the length variability of human di- and trinucleotide repeats [Ashley and Warren, 1995; Riggins et al., 1992; Weber, 1990]. There were however notable exceptions to this trend, two relatively small repeats (AGC5 and AAG12) exhibiting the highest levels of variability, and a large repeat (AAG4) showing very little size variation. These anomalies indicate that at least one other modifying factor must be involved in the phenomenon of trinucleotide repeat size variability.

The majority of the trinucleotide repeats described in this study were small and stable, which could be a direct result of evolutionary pressure. Selection may have been exerted against larger trinucleotide repeats because of their propensity to expand resulting in deleterious phenotypic consequences.

4.3.1.2 Effect of CpG island status

Much of this project was based upon the observation that the disease causing trinucleotide repeat expansions in humans lie within or near CpG islands [Brock et al., 1999; Gourdon et al., 1997a]. From this known correlation we might extrapolate that something about the nature of CpG islands, either directly or indirectly affects trinucleotide repeats making them more susceptible to instability and expansion. For example the 'active' chromatin structure of CpG islands [Tazi and Bird, 1990], their association with promoters [Antequera and Bird, 1999], or origins of replication [Antequera and Bird, 1999; Delgado et al., 1998], could all potentially be the cause of this expansion capability. Certainly in the model organisms *Escherichia coli* and *Saccharomyces cerevisiae*, transcription and replication origins have been linked to the modification of trinucleotide repeat instability [Bowater et al., 1997; Kang et al., 1995; Miret et al., 1998].

This survey of variability in 51 mouse trinucleotide repeats revealed no correlation between high variability and CpG island location, as might have been expected. In fact 84 % of the most variable trinucleotide repeats (with variation scores of >100) were found in non CpG

island DNA regions. Moreover, the repeats in CpG island regions tended to be smaller (<10 tandem repeats) and inherently more stable (with variation scores of >25). This would seem to suggest that trinucleotide repeat instability is uncommon in mouse CpG islands.

The most likely explanation for these observations is that the mechanism underlying small polymorphic changes in trinucleotide repeat size (observed in this study) is different from that which results in large, dynamic expansions, and that it is the dynamic expansion mechanism which is associated with CpG islands. Alternatively the expansion of trinucleotide repeats might require location in or near a CpG island, but be much rarer events than normal polymorphic size changes. Evidence for this might not have been revealed by this study purely because of the relatively small number of trinucleotide repeats that were actually identified in CpG islands. Repeat size polymorphism and expansion capability might well be selected against in CpG islands, because of their association with genes [Bird, 1987]. For example, if a trinucleotide repeat were located in a coding region it could be governed by stricter evolutionary constraints, preventing overt variability, particularly if the consequences were potentially deleterious.

The link between human trinucleotide repeat expansions and CpG islands may eventually prove coincidental, revealing other putative phenomenon to screen in mice for association with elusive expansions.

4.3.1.3 Effect of trinucleotide repeat class

The data on trinucleotide repeat variability, size and class show that AGC and AAG repeats exhibit the widest range of variation scores, and that other classes of repeats (i.e. AGG) remain relatively stable irrespective of their size. The diverse range of variability scores associated with AGC and AAG repeats is not unsurprising, given that they are two of only three trinucleotide repeat classes known to undergo expansion. The AGC repeat is the most common progenitor of expanded repeat loci described in humans to date, accounting for approximately 70% of trinucleotide repeat expansions. It seems more than likely that AGC, AAG and CCG could have an increased propensity for polymorphism (compared to the other seven classes of trinucleotide repeats) in both humans and mice. However, AGC repeat expansions may not prove to be any more common than those of other repeats, their current abundance a result of ascertainment bias caused by the pathogenicity associated with expanded polyglutamine tracts.

4.3.1.4 Effect of flanking and compound repeats

The majority of repeats exhibiting variability well above and below the trend line, were accompanied by other repetitive elements in the library clones. This observation presents the conundrum that additional repetitive elements in the vicinity may serve to either stabilise, or de-stabilise other repeat tracts. These effects might be the result of secondary DNA structure formations interfering with DNA replication processes.

In total, 27% of the trinucleotide repeats assessed by PCR were compound (a repeat directly flanked by at least one other type or class of repeat) in nature. This figure is significantly higher ($p = 0.025$) than the 11% normally associated with dinucleotide repeats [Weber, 1990].

The size variability observed in compound repeat alleles was for simplicity's sake assumed to have occurred in the largest of the adjoining repeat elements. However, such repeats can vary in length at any, or all of the repeat arrays [Brinkmann et al., 1998; Bull et al., 1999; Garza and Freimer, 1996; Urquhart et al., 1994]. It was however unfeasible to sequence every allele of each compound repeat identified in this survey, to establish the exact molecular basis for every size variation.

The data assessed here did not elucidate any one strong modifier of trinucleotide repeat size variability. It seems most likely that a complex combination of factors orchestrate the phenomenon, possibly including: repeat array size; repeat class; and compound or flanking repeat elements. As this survey only revealed polymorphic repeat size variations, it is impossible to speculate how any of these factors might influence dynamic expansion mutations, if indeed they occur naturally in the mouse.

4.3.2 Interspecies variation of trinucleotide repeat arrays

In this study, 86% of trinucleotide repeats ($n \geq 4$) exhibited size polymorphisms in one or more of the nine classical *Mus musculus* inbred strains tested, 67% varied in size between C57BL/6J (*Mus musculus* subspecies) and *Mus spretus*, and 10% showed no size variation at all in the mouse DNA panel.

Previous studies of trinucleotide repeats (9 CAG repeats $n \geq 7$ [Abbott and Chambers, 1994], and 25 repeats $n \geq 5$ [Chambers and Abbott, 1996]) reported only 55% and 36% variation between inbred strains, significantly lower percentages than obtained here. This was despite these surveys having slightly larger starting points, for the size of repeat tracts assayed (7

and 5 respectively). The lower levels of size variation observed could be due to sampling error produced by the relatively small number of repeats identified in both these screens, or attributable to the different genomic regions sampled in the studies. The trinucleotide repeats from the two previous studies were selected from transcribed regions (cDNAs), whereas although the repeats identified in this study were obtained from a CpG island library, very few of them were actually associated with CpG islands. Trinucleotide repeats from transcribed regions could easily have been subject to greater evolutionary constraints, which restricted their variability.

The percentage of trinucleotide repeats exhibiting size variations between C57BL/6J (*Mus musculus* subspecies) and *Mus spretus* in the two previous screens (78% and 64% respectively) were not significantly different from that reported here (67%).

A large study of 300 dinucleotide repeats ($n \geq 15$ [Dietrich et al., 1992]) detected 50% polymorphism between inbred strains, a figure significantly lower than reported in this survey of trinucleotide repeats (86%). Conversely the level of variability between C57BL/6J and *Mus spretus* was much higher than was observed in any of the trinucleotide repeat studies. This would seem to indicate that trinucleotide repeats are more variable than dinucleotide repeats amongst mice derived from the same subspecies (in this case *Mus musculus* inbred strains), but that dinucleotide arrays are more variable between different species. However as the dinucleotide repeats ($n \geq 15$) assessed were considerably longer than the trinucleotide repeat arrays ($n \geq 4$), the data can not be reliably, directly compared.

The studies described here could potentially have underestimated the actual number of polymorphisms, because the detection methods employed (high percentage agarose gel separation) may not have resolved some small allele size differences.

4.3.3 Map locations of trinucleotide repeat arrays

The majority of the trinucleotide repeat arrays (approximately 50%) clustered on chromosomes 1, 6, 7 and 8 which are thought to be amongst the most gene rich (see Mouse Genome Database {MGD}: www.informatics.jax.org). Chromosomes 2, 4, 11 and 17 are also gene rich, but few or no trinucleotide repeats were found to map there. Obviously chromosomes 11 and 17 are relatively small, which could, in part explain this observation. The trinucleotide repeats were spread relatively evenly throughout the genome, and across the entire length of the chromosomes. They did not appear restricted to certain chromosomal regions, or localised in any specific pattern.

Two sets of repeats (AAC4/*D1Abb2* and AGG1/*D1Abb3*, AAG1/*D7Abb2* and AGG11/*D7Abb3*) were assigned to indistinguishable map regions. The phenomenon of microsatellite repeats clustering together and also with *Alu* elements has been previously described [Beckman and Weber, 1992]. This clustering suggests that in such instances the repetitive species may not be totally independent. Although it should be noted that although the repeats map to indistinguishable regions on the BSS map, they could potentially be separated by megabases of DNA.

Several of the trinucleotide repeats were informative when incorporated into the BSS backcross data and improved the overall resolution of the map. The mapped trinucleotide repeats will undoubtedly prove useful as genetic markers as they are highly variable and easily typed by PCR. They may also prove valuable in comparative mapping as trinucleotide repeats are often conserved between closely related species [Ricke et al., 1995], but show length variation between strains and species [Abbott and Chambers, 1994; Chambers and Abbott, 1996; Love et al., 1990]{this survey}.

4.3.4 Assessment of trinucleotide repeats as 'candidates' for characterised mouse mutant phenotypes

The assessment of trinucleotide repeats as candidates for mouse mutations mapped at the phenotypic level was unfortunately quite crude and incomplete. The crudity arose because although the repeats were well mapped themselves, they were compared to mutations positioned on different maps, by varying techniques with less stringent resolution capabilities. The assessment was incomplete firstly because several of the largest trinucleotide repeats identified in this survey could not be mapped, and secondly because the analysis relied on the availability of DNA from the appropriate mutant phenotypes. This second factor posed the severest limitations, because tracking down mutant mouse lines and obtaining DNA from researchers throughout the world proved difficult. The most accessible resource for mouse mutant DNA is probably The Jackson Laboratory, but in the majority of cases the mouse mutant phenotypes of interest had been described many years ago, had not been maintained at The Jackson Laboratory, nor had DNA been extracted from affected individuals, or sample tissues kept. The other major drawback of this survey was that a significant proportion of mouse mutations described to date have arisen through contrived mutagenesis programs. These studies have often utilised known mutagens such as ethylnitrosourea and X-irradiation, which most commonly produce point mutations and chromosomal rearrangements, respectively [Rinchik and Russell, 1990]. It seems unlikely

that either of these mutagenic processes could directly result in the expansion of naturally occurring trinucleotide repeats.

It should be considered that trinucleotide repeat expansions may not have been reported in laboratory mouse stocks, because of the essential prerequisite of certain haplotypes, absent from these inbred strains. Wild mice might therefore be more a more suitable reservoir of trinucleotide repeat variability and possibly expansion, but it would not be feasible to study large enough numbers of mice, to observe what would certainly be rare mutational events and phenotypes. Another obvious drawback of mice as a model species as far as trinucleotide repeat expansion related disorders are concerned, is that they mostly manifest as late onset phenotypes, which may not even have a chance to develop in such short lived organisms.

4.3.4.1 Two variable repeat arrays as candidates for the frizzy mutant phenotype

Of the two repeats which appeared larger in frizzy mice (AAG1 and AGG11), it was feasible that either could have been responsible for the mutant phenotype. It was more likely that the hexanucleotide repeat was involved because it was a large pure repeat tract, whereas the size increase in the AGG11 allele was the result of widespread incremental changes in several trinucleotide repeats, a phenomenon not known to be responsible for any of the repeat associated disorders in humans. At the time, the theory that a hexanucleotide repeat expansion (or insertion) might cause a phenotype was somewhat unusual. However, it has recently been discovered that a pentanucleotide repeat expansion is responsible for Spinocerebellar Ataxia type 10 [Matsuura et al., 2000] and that a dodecanucleotide repeat expansion causes Unverricht-Lundborg type Progressive Myoclonus Epilepsy (EPM1) [Lalioti et al., 1997]. This significantly broadens the potential scope for the size range of disease causing, repeat array expansions. The pentanucleotide repeat is particularly interesting as this clearly indicates that being a triplet, or a multiple of a triplet (dodecanucleotide) repeat array, is not necessary to result in a disease phenotype.

Further analysis revealed that neither repeat could be responsible for the frizzy phenotype because neither was present in the independently maintained frizzy stock. The increases in repeat size must have arisen in The Jackson Laboratory frizzy stock, at some point after the Baylor College mice had been separated from them. The fact that both these repeat size changes were homozygous in The Jackson Laboratory frizzy stock, implies that they were not the product of recent events. It is most likely that these size variations were the result of natural size polymorphisms, which arose through the stock being repeatedly bred and

maintained. However why such a large variation was observed in one stock, and non at all was found in the other, remains an enigma. It is possible that one stock has been passaged through significantly more generations than the other due to different breeding protocols, or temporary storage as frozen embryos or sperm. The genetic differences observed in The Jackson Laboratory and the Baylor College frizzy stocks, mean that these mice should now be considered substrains, and designated appropriately.

4.3.4.2 The genetic background surrounding the AAG1, AGG11 and frizzy loci

The typing of the genetic background surrounding the AAG1, AGG11 and frizzy loci, adds weight to the conclusion that the repeat and mutant loci are genetically distinct. The two repeat loci were flanked by a DBA type background, but the DNA region surrounding the location of the frizzy mutation was BALB/c in origin. However, this conclusion is not absolute as the frizzy mutation was not assigned a map location with as high a degree of certainty as the repeat loci were.

4.3.4.3 Analysis of frizzy DNA at all variable trinucleotide repeat loci.

As neither of the repeats themselves were the underlying factor responsible for the frizzy mutation, it was considered possible that the frizzy phenotype caused some type of genome-wide instability (for example a mismatch repair deficiency) that had resulted in these unusual occurrences. Analysis of the frizzy DNA at all other trinucleotide repeat loci described in this study revealed no other unusual instabilities, although this does not of course completely rule out a genome wide affect theory. To investigate this possibility further these repeats could be assessed for size variation and somatic instabilities in a much larger stock of frizzy mice. Other known polymorphic repeat loci could also be analysed in a similar manner.

The frizzy mutation could perhaps act in *cis* to cause the repeat size changes observed in this study. The underlying mutation causing the frizzy phenotype could affect repeats in the vicinity of the gene, by an as yet undescribed mechanism. To determine if this is a feasible assumption, the repeats themselves could be mapped relative to the frizzy mutation and to each other, to more accurately define the distances that lie between them.

The two repeat instabilities (AAG1 and AGG11) which map to indistinguishable regions of mouse chromosome 7, are separated by at least 50 base pairs of unique nucleotide sequence and are therefore not both part of the same larger, compound repeat. However the region as a whole could be a 'hotspot' for repeat length polymorphisms and natural variability.

CHAPTER 5

TRANSGENIC STUDY TO DETERMINE IF LOCATING A TRINUCLEOTIDE REPEAT WITHIN A CpG ISLAND WOULD EFFECT REPEAT INSTABILITY

5 Transgenic study to determine if locating a trinucleotide repeat within a CpG island would effect repeat instability

5.1 Introduction

The aim of this study was to determine whether locating a large trinucleotide repeat array within a well defined mouse CpG island would directly effect its stability on transmission. This was performed in an attempt to elucidate whether the observed correlation of expandable trinucleotide repeats in humans with CpG islands [Brock et al., 1999; Gourdon et al., 1997a] was coincidental, or directly linked to the expansion phenomenon.

5.1.1 A study on CpG island methylation in the mouse

The murine adenine phosphoribosyltransferase (*Aprt*) gene is a housekeeping gene with a well characterised CpG island, which extends across the promoter and includes the first two exons of the gene (see figure 5.1). In the native mouse *Aprt* gene the CpG island is completely free of methylation at the abundant CpG moieties, but the flanking DNA is methylated.

Previous studies revealed that transgenic constructs containing the native *Aprt* gene retained the endogenous CpG island conformation, including the protection from methylation. However the deletion or mutation of Sp1 binding sites located in the *Aprt* promoter proved sufficient to result in the *de novo* methylation of the transgenic CpG island [Brandeis et al., 1994; Macleod et al., 1994]. The native *Aprt* construct (pABS) used in this experiment, and the mutated gene construct (pAZM2) differed only in the nucleotide sequence of three GC boxes (the Sp1 binding sites), and the introduction of an *Xho*I site for identification purposes (see figure 5.2). These two constructs provided virtually identical vehicles into which large trinucleotide repeats could be cloned, and the putative effect of CpG island status on repeat instability studied.

Figure 5.1 Mouse adenine phosphoribosyltransferase gene

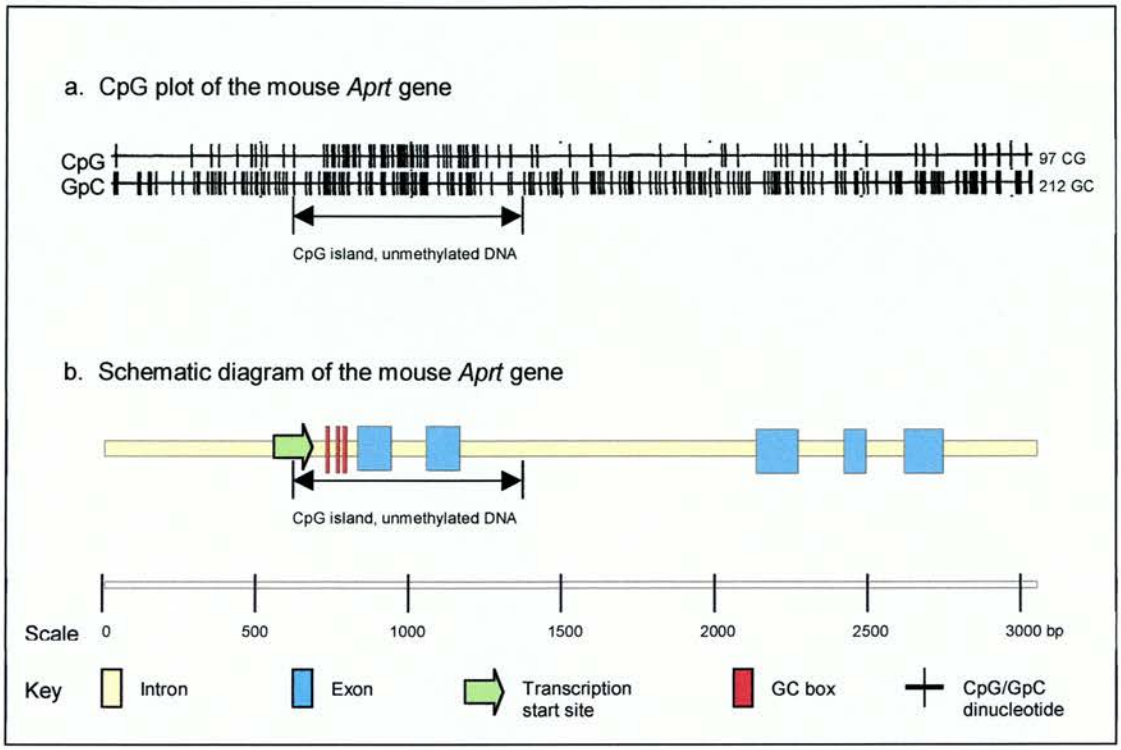
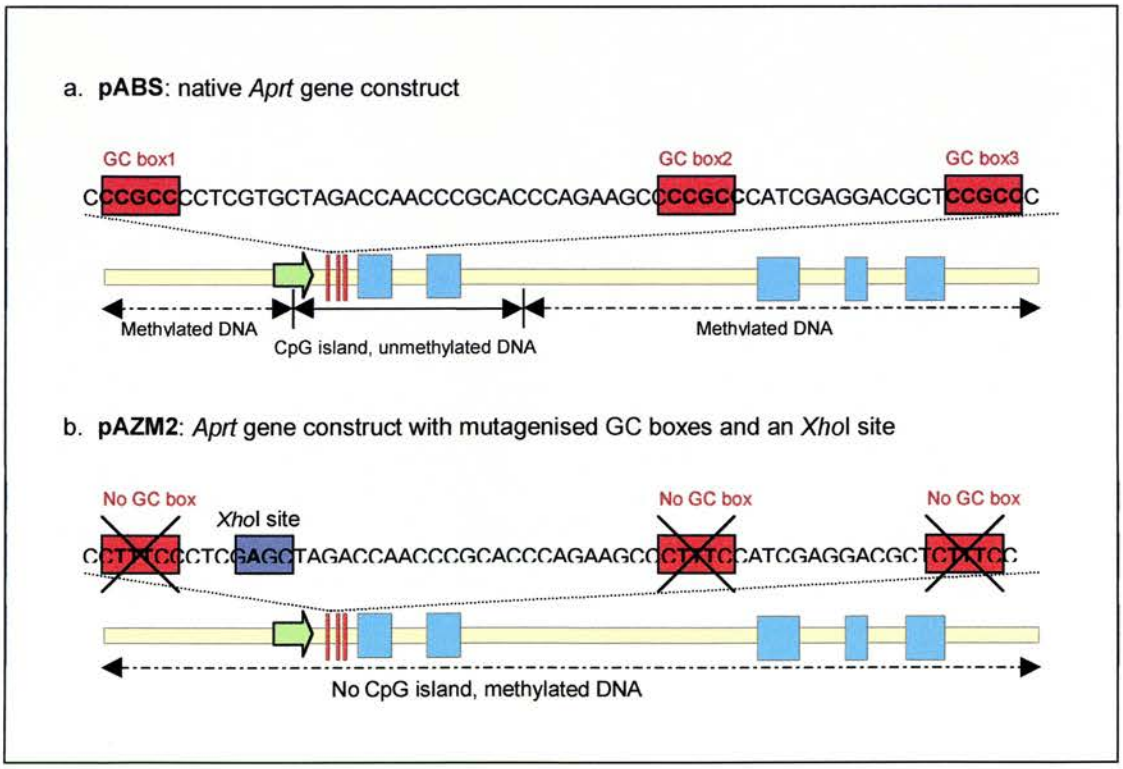


Figure 5.2 Transgenic constructs pABS and pAZM2



5.1.2 Collaborative transgenic study

This study was a collaborative project with Dr Donald Macleod and Professor Adrian Bird from the Institute of Cell and Molecular Biology (ICMB) at Edinburgh University. They provided the trangenic constructs pABS and pAZM2 into which large trinucleotide repeats were to be cloned. The trinucleotide repeats were to be cloned into a unique *Ngo*AIV site in the *Aprt* gene constructs, which was located in the middle of the CpG island, between the first two exons of the gene. The *Ngo*AIV site was chosen because it was the only unique restriction site located anywhere within the CpG island DNA.

5.2 Results

5.2.1 Cloning of a trinucleotide repeat into pABS and pAZM2.1

The original strategy for this project was to clone a trinucleotide repeat amplified by PCR from a human Huntington's disease patient, or from the mouse Metallothionein III gene (the repeat tract of which was expanded by propagation in *mutS*-deficient bacteria [Schmidt et al., 2000]), into the unique *Ngo*AIV sites of the constructs pABS and pAZM2. However although it proved possible to amplify these large repeats by PCR all attempts at cloning them into the constructs by adding *Ngo*AIV polylinkers to the PCR products or incorporating *Ngo*AIV sites into the PCR primers, or by blunt ended cloning were unsuccessful.

A database survey of all the expandable trinucleotide repeats identified at that point in time revealed that the human Myotonic Dystrophy causing repeat was closely flanked by two *Ngo*AIV sites. It was primarily for this reason that the Myotonic Dystrophy repeat was chosen to be incorporated into the mouse *Aprt* gene constructs pABS and pAZM2. Fortuitously the human Myotonic Dystrophy repeat subsequently proved the most unstable to date, when used to produce transgenic mice [Seznec et al., 2000].

5.2.1.1 Subcloning of pAZM2 into pUC18™ to produce pAZM2.1

The pABS construct was contained in the vector pBlueScribe™ and was suitable for a repeat to be cloned directly into the unique *Ngo*AIV site. The pAZM2 construct however had been cloned into a pTZ18™ vector which also contained an *Ngo*AIV site. Therefore this construct had to be subcloned into a suitable vector prior to the insertion of the trinucleotide repeat. The pAZM2 construct was released from the pTZ18™ vector by digestion with *Eco*RI and *Sph*I, and was subcloned into a pUC18™ vector (which had also been digested with *Eco*RI and *Sph*I) via the available sticky ends, to produce pAZM2.1 (see figure 5.3).

The construct pAZM2.1 was assayed by restriction digestion to check it was the desired cloning product. The construct was also amplified in segments by PCR, and partially sequenced to ensure that no changes had been introduced into the critical CpG island region.

The subcloning and construct analysis was performed by myself.

Figure 5.3 Transgenic constructs pABS, pAZM2, pAZM2.1 and their cloning vectors

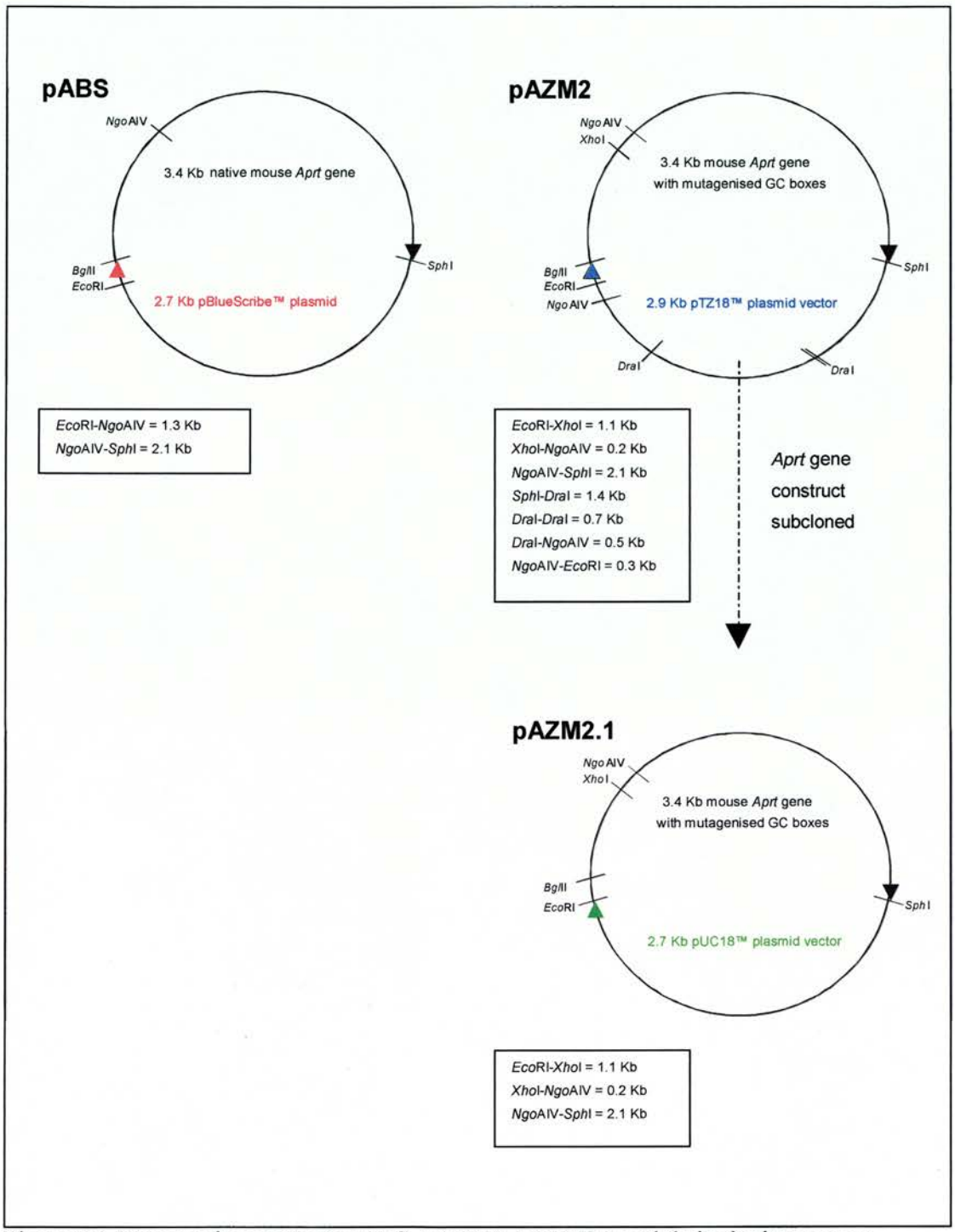


Figure 5.3 Transgenic constructs pABS, pAZM2, pAZM2.1 and their cloning vectors. pABS, the native mouse *Aprt* gene in pBlueScribe™. pAZM2 the mutagenised mouse *Aprt* gene in pTZ18™, and subcloned into pUC18™ to produce construct pAZM2.1.

5.2.1.2 PCR amplification of the trinucleotide repeat containing region from myotonic dystrophy patient DNA

PCR primers (DM) were designed to amplify the human myotonic dystrophy repeat containing region, including the flanking *Ngo*AIV restriction endonuclease sites (figure 5.4).

Primer name	Primer sequence	Reaction conditions	Other reagents	Product size*
DM forward	5'CCC AGC TCC AGT CCT GTG 3'	Anneal 58°C 60 ng primers	2 M Betaine	458 bp
DM reverse	5'ATC CAA ACC GCC GAA GCG G 3'			

*PCR product size when the patient DNA contains 11 trinucleotide repeats.

These primers were used to amplify the region from normal human DNA, and patient DNA containing expanded alleles with approximately 70 and 130 CTG trinucleotide repeats. The PCR was then optimised to enable the largest Myotonic Dystrophy allele to be amplified.

5.2.1.3 Examination and preparation of repeat containing DNA for cloning

An aliquot of each PCR product was digested with *Ngo*AIV, to release the repeat containing fragment for cloning and both digested and undigested PCR products were separated by agarose gel electrophoresis (see figure 5.5). The undigested expanded patient alleles were cut from the gel, purified with glass milk and sequenced (with the PCR primers) to establish that no mistakes had been incorporated into the repeat tract by the Taq polymerase during PCR amplification. The digested, expanded repeat containing alleles were also cut from the gel and purified with glass milk in preparation for cloning.

The amplification and preparation of trinucleotide repeat containing DNA was performed by myself.

5.2.1.4 Cloning of trinucleotide repeat region into pABS and pAZM2.1

The two constructs pABS and pAZM2.1 were also digested with *Ngo*AIV, and the trinucleotide repeat containing alleles were then ligated to both vectors via the available sticky ends. Only the smaller expanded allele containing 70 repeats (539 bp) was successfully cloned into either of these constructs.

This stage of construct preparation was performed by Dr Donald Macleod.

Figure 5.4 Human Myotonic Dystrophy PCR design

Excerpt of human Myotonic Dystrophy protein kinase nucleotide sequence:

Forward PCR primer

```
CACCGCCTGC CAGTTCACAA CCGCTCCGAG CGTGGGTCTC CGCCCAGCTC CAGTCCTGTG 13141
                                                        NgoAIV
TACCGGGCCC GCCCCCTAGC GGCCGGGGAG GGAGGGGCCG GGTCCGCGGC CGGCGAACGG 13201
                                  Normal trinucleotide repeat tract
GGCTCGAAGG GTCCTTGTAG CCGGGAATGC TGCTGCTGCT GCTGCTGCTG CTGCTGCTGC 13261

TGGGGGGATC ACAGACCATT TCTTTCTTTC GGCCAGGCTG AGGCCCTGAC GTGGATGGGC 13321

AAACTGCAGG CCTGGGAAGG CAGCAAGCCG GGCCGTCCGT GTTCCATCCT CCACGCACCC 13381

CCACCTATCG TTGGTTCGCA AAGTGCAAAG CTTTCTTGTG CATGACGCCC TGCTCTGGGG 13441

AGCGTCTGGC GCGATCTCTG CCTGCTTACT CGGGAAATTT GCTTTTGCCA AACCCGCTTT 13501
                                                        NgoAIV
TTCGGGGATC CCGCGCCCCC CTCCTCACTT GCGCTGCTCT CGGAGCCCCA GCCGGCTCCG 13561

CCCGCTTCGG CGGTTTGGAT ATTTATTGAC CTCGTCCTCC GACTCGCTGA CAGGCTACAG 13621
```

Reverse PCR primer

Forward primer: CCC AGC TCC AGT CCT GTG Reverse primer: ATC CAA ACC GCC GAA GCG G

PCR product = 458 bp when the trinucleotide tract contains 11 repeats

Figure 5.5 Human Myotonic Dystrophy PCR products

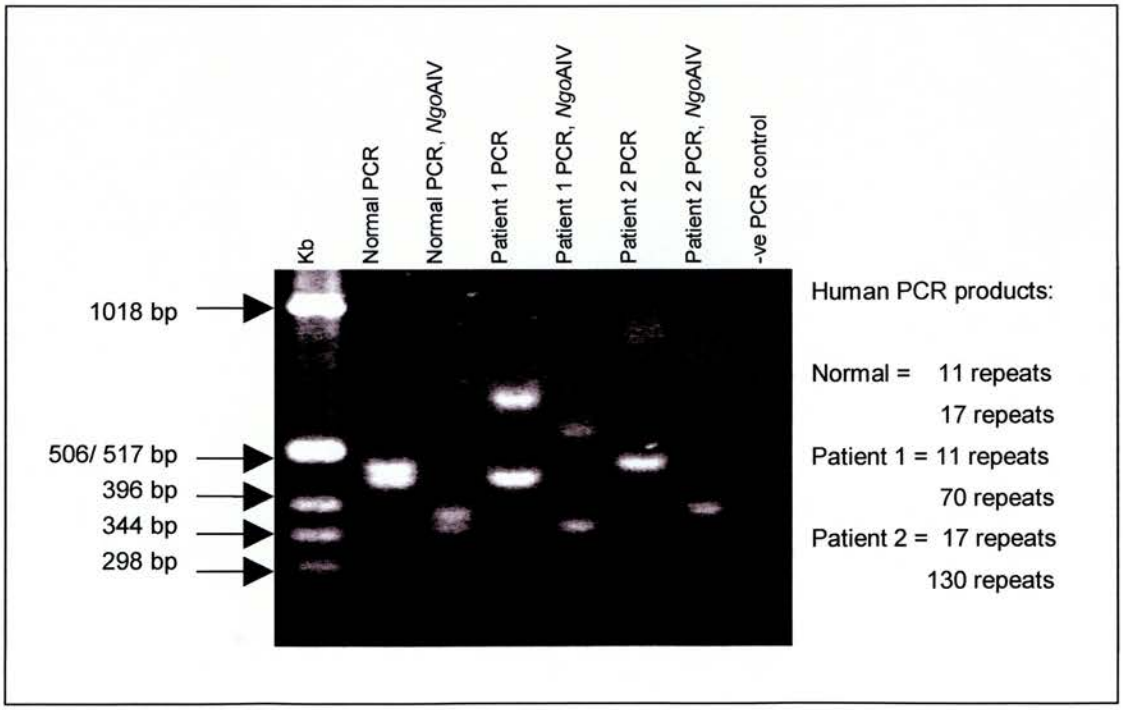


Figure 5.5 An agarose gel showing human Myotonic Dystrophy repeat containing PCR products.

5.2.1.5 Screening and analysis of constructs

After cloning, transformed *E. coli* colonies were screened to identify the desired constructs. As trinucleotide repeat orientation was known to have a profound effect on repeat stability in both *S. cerevisiae* [Freudenreich et al., 1997; Maurer et al., 1996; Miret et al., 1998] and *E. coli* [Hirst and White, 1998; Ji et al., 1996; Kang et al., 1995], pABS and pAZM2.1 constructs containing the trinucleotide repeat region in both possible orientations were isolated (pABSCTG, pAZM2.1CTG, pABSCAG and pAZM2.1CAG). These four constructs were identified by restriction digestion analysis (see figure 5.6) and were then partially sequenced (across the trinucleotide repeat and *Aprt* GC boxes) to ensure that no critical nucleotide sequence changes had been introduced during cloning and propagation. Two constructs were analysed by myself and the other two by Donald Macleod.

5.2.2 Production of transgenic mice.

5.2.2.1 Preparation of construct DNA for micro-injection.

The transgenic constructs were excised from their respective plasmid vectors by digestion with *Eco*RI and *Sph*I which removed all prokaryotic DNA sequence, except for a short region of polylinker containing a *Kpn*I restriction site. The plasmid and construct fragments were separated by agarose gel electrophoresis and the construct band cut from the gel and purified. The two constructs pABSCAG and pAZM2.1CAG were micro-injected into the pronuclei of mouse (F_1 CBA x C57BL/6J) embryos by myself, and the constructs pABSCTG and pAZM2.1CTG by Donald Macleod. The micro-injected embryos were cultured to the two cell stage and then transferred into the oviducts of CD1 mice to complete their gestation.

5.2.2.2 Identification of transgenic founder mice

The transgenic offspring (founder mice) were identified from amongst their littermates by PCR amplification of tail tip biopsy DNA, extracted at four weeks of age. The PCR primers (DMrepeat) were designed to amplify the human Myotonic Dystrophy trinucleotide repeat sequence which would only be present in transgenic individuals.

Primer name	Primer sequence	Reaction conditions	Other reagents	Product size*
DMrepeat forward (FAM)	5' CTC GAA GGG TCC TTG TAG CC 3'	Anneal 55°C	2 M Betaine	380 bp
DMrepeat reverse	5' CAC TTT GCG AAC CAA CGA TA 3'			

*PCR product size when the transgene contains 70x trinucleotide repeats.

Figure 5.6 The four constructs used to create transgenic mice

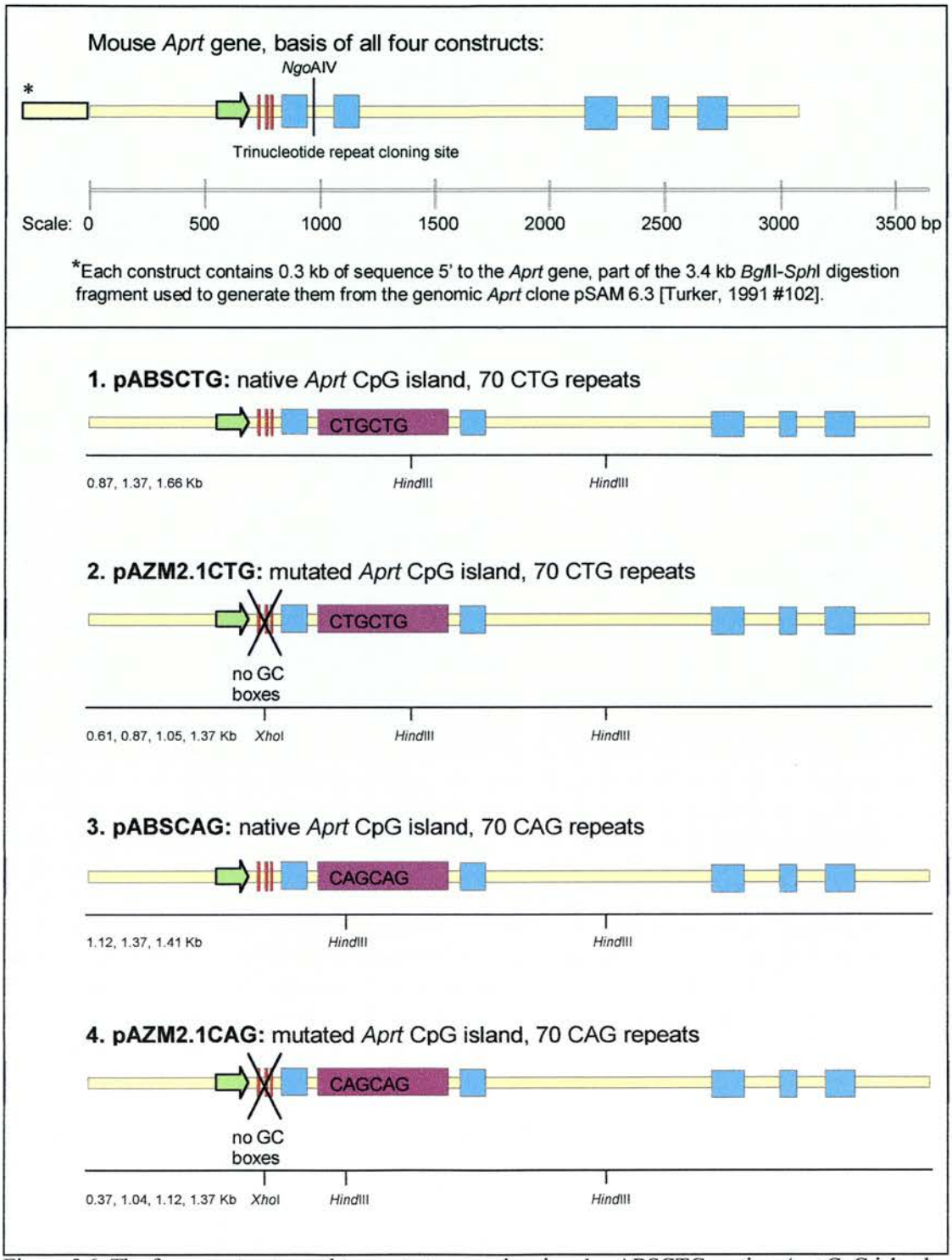


Figure 5.6 The four constructs used to create transgenic mice: 1. pABSCTG, native *Aprt* CpG island, 70 CTG repeats; 2. pAZM2.1CTG, mutated *Aprt* CpG island, 70 CTG repeats; 3. pABSCAG, native *Aprt* CpG island, 70 CAG repeats; 4. pAZM2.1CAG, mutated *Aprt* CpG island, 70 CAG repeats. These constructs were differentiated using *Hind*III and *Xho*I digestion.

5.2.3 Analysis of transgenic founder mice

The positive PCR products from each transgenic founder mouse were sequenced to check the orientation and size of the trinucleotide repeats within the integrated constructs. In the majority of transgenic mice (65%) the trinucleotide tract contained 70 repeats, the same number as in the construct. However the trinucleotide tract was seen to vary in size from 64 to 78 repeats in other transgenic mice.

The *Aprt* promoter region and GC boxes were sequenced to confirm that no changes had been introduced into the nucleotide sequence during transgene integration. The restriction digestion of large PCR products including APX (see below) was used to screen for any gross rearrangements of the incorporated transgenes.

The results of these analyses are detailed in table 5.1.

5.2.3.1 Determination of transgene insertion copy numbers

The copy number of the integrated construct in each transgenic founder was determined by the preparation of southern blots containing *Kpn*I digests of genomic tail tip DNA, and hybridisation to an *Aprt*-specific probe APX (see below). Transgenes were distinguished from the endogenous *Aprt* gene sequence by virtue of their different restriction fragment sizes (endogenous 5 kb, transgenic 2.2 kb {see figure 5.7}). The transgenic copy number was estimated by comparing the density of the transgene signal with that obtained from the endogenous mouse *Aprt* fragment. This fragment was known to originate from a single copy gene, and was therefore present as 2 copies in genomic DNA (see figure 5.8 and table 5.2).

The *Aprt*-specific probe APX was an 893 bp PCR product generated from the endogenous *Aprt* nucleotide sequence. The PCR primers AP7 and APX3 were used to produce the probe from bases 454-1347 of the native *Aprt* gene.

Primer name	Primer sequence	Reaction conditions	Other reagents	Product size
AP7 forward	5'GGC CCT TGT ACT ATG CGC G3'	Anneal 58°C	Standard	893 bp
APX3 reverse	5'GGG TGA ACA AGC GTC CAA GG3'			

These analyses of transgenic founder mice were carried out by the individual who generated them, i.e. pABSCAG and pAZM2.1CAG by myself, pABSCTG and pAZM2.1CTG by Donald Macleod. This data is summarised in table 5.1.

Table 5.1 Description of transgenic founder mice

Type of construct	Founder mouse/ name of line	Copy number	Number of repeats in transgene	Transgene transmitted	Line analysed further
pAZM2.1CAG	A196	2	64 GAC	Yes	Yes
	A197	4	70 GAC	No	No
	A197.1	1	78 GAC	Yes	Yes
pABSCAG	A198	≤1	70 GAC	Yes	Yes
	A199	10	70 GAC	Yes	Yes
	A200	?	70 GAC	?	No
	A201	3	70 GAC	Yes	Yes
	A201.1	≤1	71 GAC	No	No
pAZM2.1CTG	K60	1	70 CTG	Yes	Yes
	K65	20	70 CTG	Yes	Yes
	K68	2	71 CTG	Yes	Yes
	K71	1	70 CTG	Yes	Yes
	K79	2	69/70 CTG	Yes	No
pABSCTG	R13	1	66 CTG	Yes	No
	R35	1	70 CTG	Yes	Yes
	R72	6	70 CTG	Yes	Yes
	R203	3	70 CTG	Yes	Yes

Highlighted are lines that were not analysed beyond the founding transgenic mouse: mice A197 and A201.1 did not transmit the transgenes, mouse A200 died before breeding, the R13 transgene was rearranged and the K79 transgenes had different sized repeats which made further analysis difficult.

Figure 5.7 Maps of endogenous and transgenic *Aprt* sequence

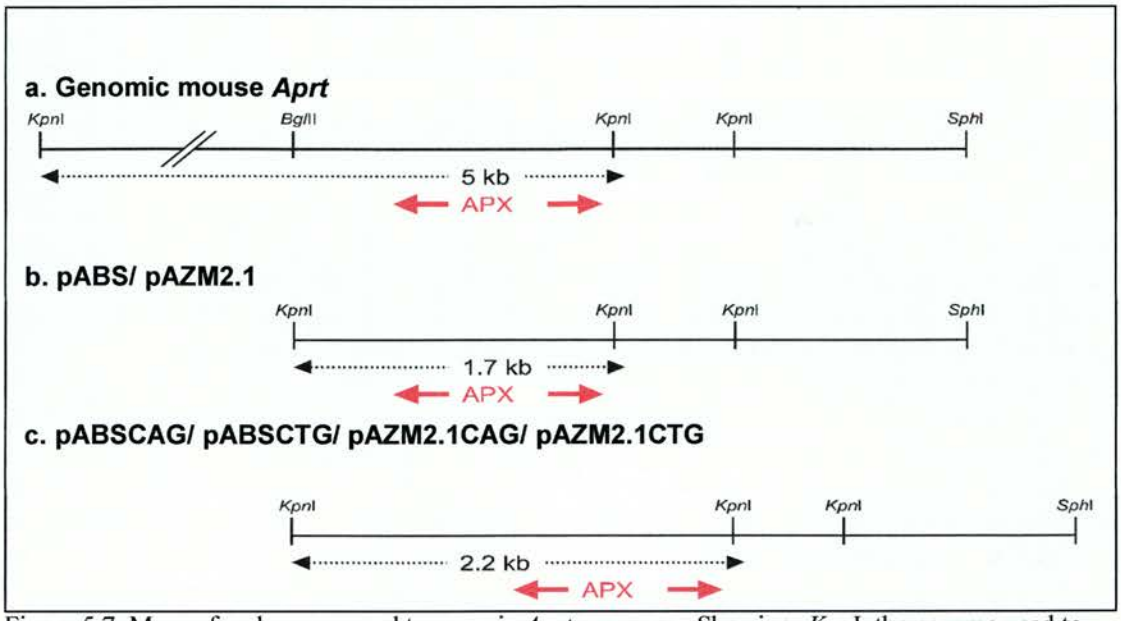


Figure 5.7 Maps of endogenous and transgenic *Aprt* sequence. Showing; *Kpn*I, the enzyme used to determine transgenic construct insertion copy number; *Bgl*II and *Sph*I, the enzymes used to clone the native *Aprt* fragment into pABS and pAZM2 [Macleod et al., 1994]; and APX the *Aprt*-specific probe.

Figure 5.8 Transgenic construct insertion copy number analysis

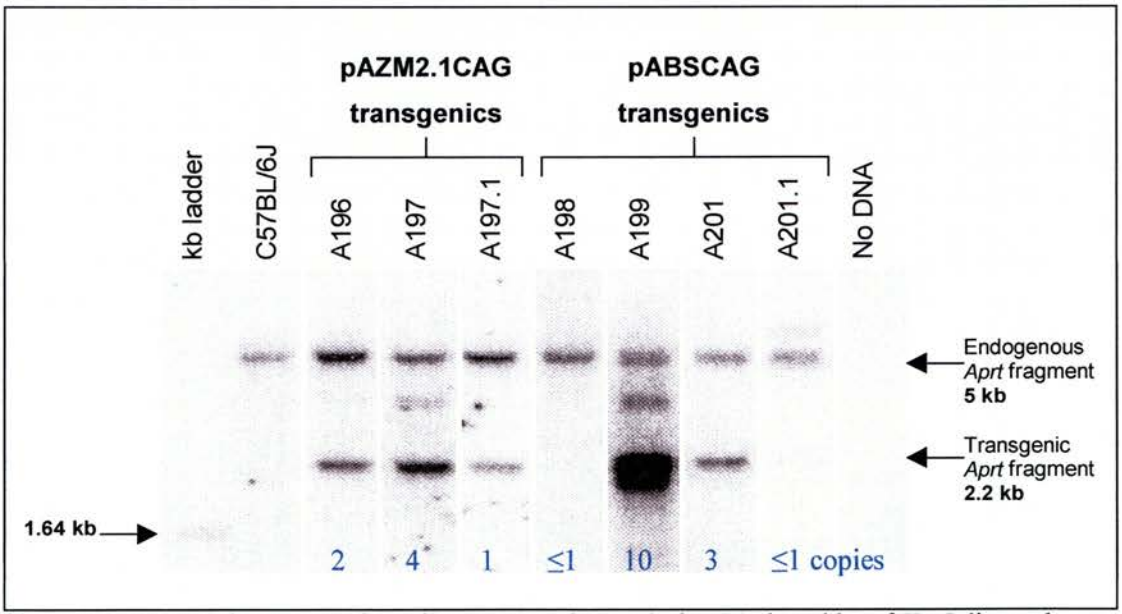


Figure 5.8 Transgenic construct insertion copy number analysis. Southern blot of *Kpn*I digested DNA from wild-type (C57BL/6J), and pABSCAG and pAZM2.1CAG transgenic mice, hybridised to a radio-labelled *Aprt*-specific probe (APX). Indicated by arrows are; the 1.64 kb fragment of the kb ladder; the 5 kb fragment of endogenous *Aprt*; and the 2.2 kb transgenic *Aprt* fragment.

5.2.3.2 Production of transgenic lines

The transgenic founder mice were crossed with F_1 CBA x C57BL/6J mice and offspring tail tip DNA analysed by two separate PCR assays to identify and confirm transgenic animals. The first PCR (DMrepeat see 5.2.2.2) detected transgenic individuals by amplifying the human myotonic dystrophy DNA across the trinucleotide repeat tract, this PCR product was also used subsequently to determine the copy number of each repeat (see 5.2.5). The second PCR (DMtrans) was designed to identify transgenic mice without amplifying across the trinucleotide repeat tract itself. This was in case the repeat tract had undergone expansion in some offspring to a size that could not be easily amplified by a standard PCR reaction, such individuals would potentially have been missed by the first assay alone. The primers for this PCR were complementary to the endogenous mouse DNA located 5' of the *Ngo*AIV site and to the human myotonic dystrophy sequence 5' of the cloned repeat. The second primer had to be different for the transgenes containing the repeat in the two opposite orientations.

Primer name	Primer sequence	Reaction conditions	Other reagents	Product size
DMtrans forward	5' CCC TCG TGC TAG ACC AAC C 3'	Anneal 57°C	-	
DMtrans reverse1 (CTG)	5' CAT TCC CGG CTA CAA GGA C 3'			442 bp
DMtrans reverse2 (CAG)	5' CCT CAC TTG CGC TGC TCT 3'			435 bp

If founder mice produced no transgenic offspring in their first litter, two further litters were generated and analysed before they were deemed not to be transmitting the transgene. The transgenic mice were bred on from second generation individuals through both the male and female germline.

5.2.4 Methylation analysis of transgenes

In this study the use of the two different constructs (pABS and pAZM2.1) was an attempt to produce similar transgenes containing trinucleotide repeats, but with different CpG island status and therefore methylation patterns. The transgenic methylation state was assessed at the predicted *Aprt* CpG islands using two different methods:

5.2.4.1 Methylation sensitive PCR

The methylation status of the transgenes was determined using a modified PCR assay [Singer-Sam et al., 1990] which utilised the methylation sensitivity of the restriction enzyme *Hpa*II. Methylation sensitive restriction endonucleases do not cleave DNA when their target restriction sites are methylated. The APRT PCR amplification region contains one *Hpa*II site in the endogenous *Aprt* sequence, and four sites in the transgenic *Aprt* sequence (figure 5.9).

Primer name	Primer sequence	Reaction conditions	Other reagents	Product size
APRT forward	5' CGT GCT GTT CAG GTG CGG T 3'	Anneal 54°C	2 M Betaine	Endogenous 58 bp
APRT reverse	5' GGG GAT GAG CGT ACA GCG 3'			Transgenic ~597 bp

When genomic DNA is digested with *Hpa*II prior to PCR, products are only obtained if all these restriction sites are methylated in the template DNA. Lack of methylation at just one of these sites should be sufficient to cleave the DNA, preventing PCR amplification. Test PCRs revealed that as expected the *Hpa*II site in the endogenous *Aprt* sequence was unmethylated, which meant it could be used as an internal control for successful *Hpa*II digestion. The restriction enzyme *Msp*I, an isoschizomer of *Hpa*II, does not have methylation sensitive activity. This enzyme was used as a control to show that the test DNA could be cleaved at these sites, and that obtaining a product from the *Hpa*II digested DNA was not the result of a digestion problem.

The test DNA was also amplified with a set of PCR primers which did not flank *Hpa*II, or *Msp*I restriction sites (AAT3, see table 4.2), to show that the DNA had not been subject to star activity degradation or contamination which might have effected the outcome of the assay. Examples of test PCRs are depicted in figure 5.10 and the results of the analysis are summarised in table 5.3.

Figure 5.9 Restriction map of endogenous and transgenic *Aprt* PCR sequence

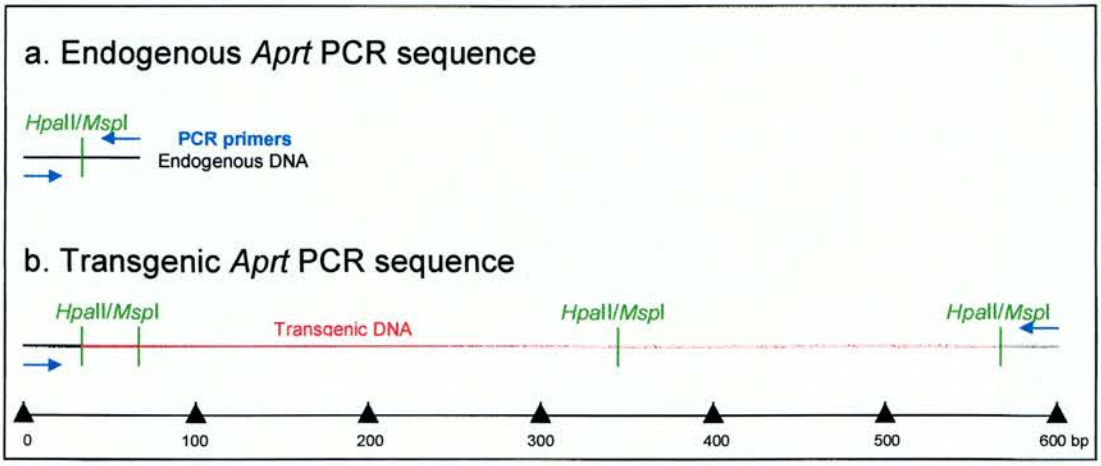


Figure 5.10 Analysis of transgene methylation status by PCR

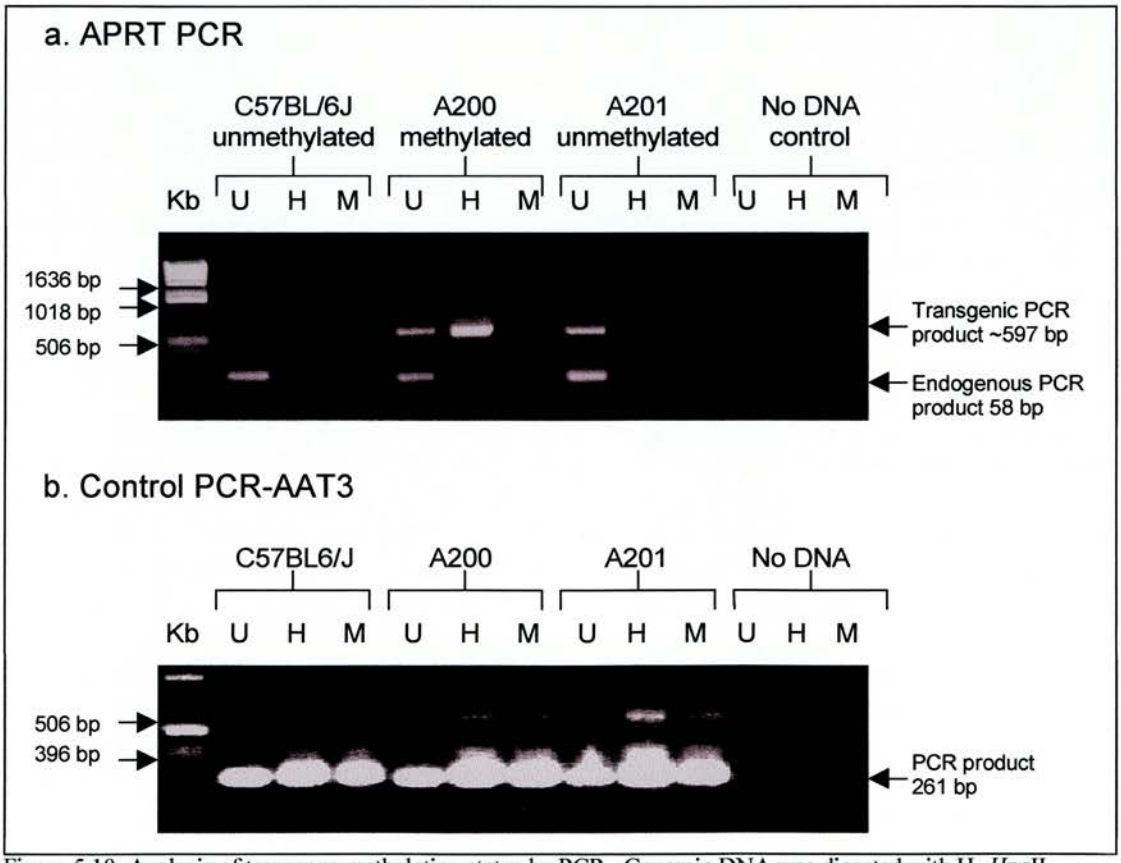


Figure 5.10 Analysis of transgene methylation status by PCR. Genomic DNA was digested with H- *Hpa*II, M-*Msp*I and U- undigested. Both endogenous (58 bp) and transgenic (597 bp) products should be obtained from uncut test DNA. When *Hpa*II digested DNA gave a product the target restriction sites were considered methylated (A199), no PCR product indicated that at least one restriction site was unmethylated at the CpG.

This analysis was also performed on offspring of both sexes which inherited the transgenes through both the female and male germline. The methylation patterns of the transgenes in these individuals were identical to those identified in the founder mice. Therefore the transgenic methylation status appears to remain stable on transmission.

All methylation PCR analyses were carried out by myself.

5.2.4.2 Methylation sensitive Southern blot

The second technique used here to determine transgene methylation status was a Southern blot assay. Tail tip DNA was digested with *Kpn*I and two methylation sensitive enzymes (*Sma*I and *Hpa*II), Southern blotted and then hybridised to an *Aprt*-specific probe APX (see 5.2.3). This method was previously used by Macleod et al [Macleod et al., 1994] to determine the methylation status of the pABS and pAZM2 transgenes. The detailed restriction maps of endogenous and transgenic *Aprt* used in this experiment are depicted in figure 5.11

This analysis was facilitated by the polylinker *Kpn*I site found at the 5' end of each construct. Transgenes were distinguished from the endogenous *Aprt* gene by virtue of their different fragment size when digested with *Kpn*I (endogenous fragment 5 kb, transgene 2.2 kb). The methylation status was assayed at several sites within the CpG island region using the methylation sensitive restriction enzymes *Sma*I and *Hpa*II. Double digestion of DNA with *Kpn*I and *Sma*I resulted in the reduction of intensity, or complete obliteration of the endogenous 5 kb (to 2.5 and 0.45 kb) and transgenic 2.2 kb (to 1.11 and 0.99 kb) fragments when the DNA was unmethylated. The 2.5 kb endogenous *Aprt* fragment served as an internal control for *Sma*I digestion, as the endogenous gene is unmethylated at these sites and should therefore always be completely eliminated. Double digestion of the DNA with *Kpn*I and *Hpa*II likewise resulted in the considerable reduction of the 5 kb (to 0.39 and 0.28 kb) and 2.2 kb (to 0.39 and 0.28 kb) fragments when the DNA was not methylated.

The Southern blots used to determine the methylation of pABSCAG and pAZM2.1GAC transgenes in founder mice are detailed in figure 5.12.

Figure 5.11 Restriction enzyme maps of endogenous and transgenic *Aprt*, used for methylation analysis

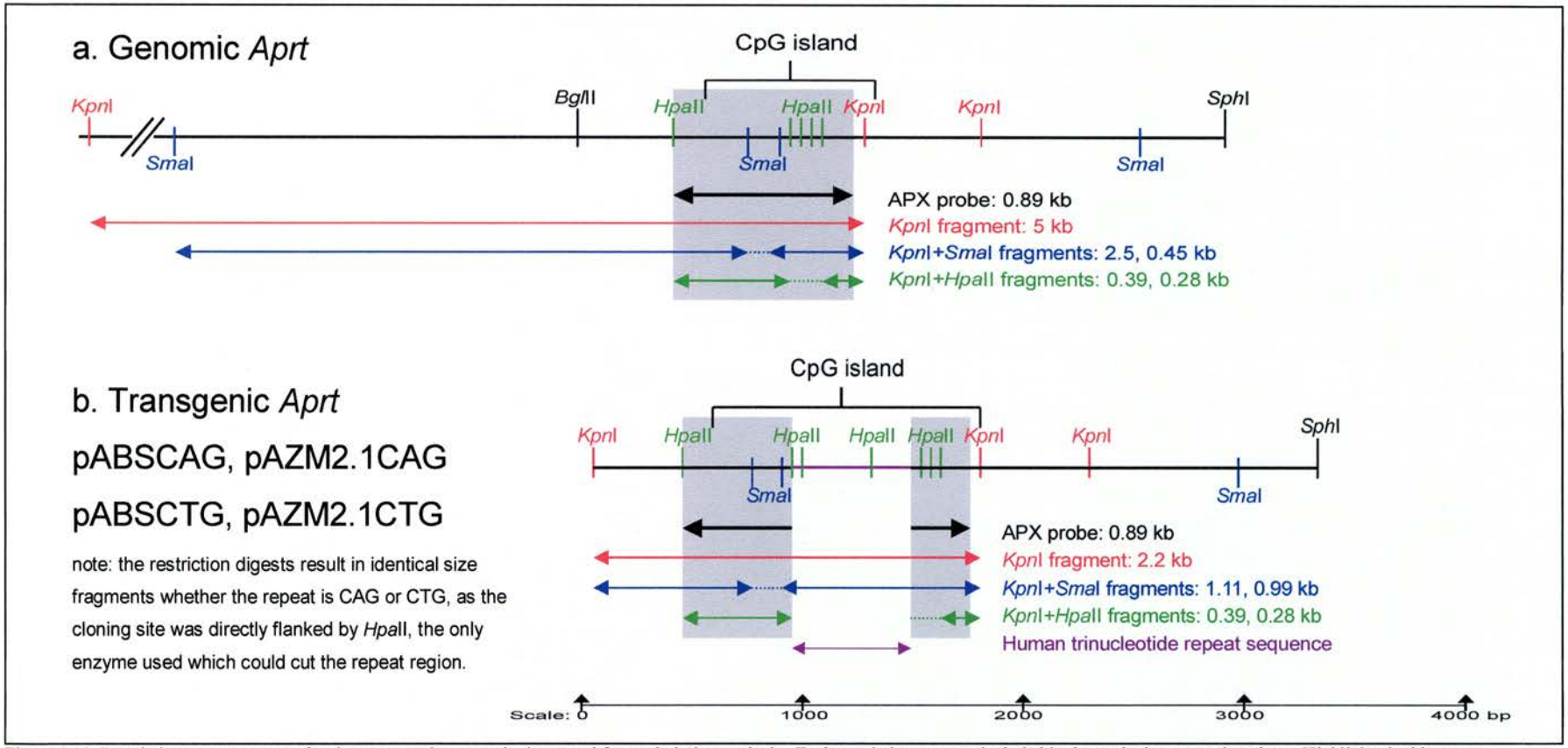


Figure 5.11 Restriction enzyme maps of endogenous and transgenic *Aprt*, used for methylation analysis. Each restriction enzyme included in the analysis appears in colour. Highlighted with a grey background are the regions of sequence complementary to the *Aprt*-specific probe APX. The *Kpn*I enzyme was used to produce clearly distinct size fragments from a. genomic (5 kb) and b. the transgenic (2.2 kb) *Aprt*. These fragments were digested further with two methylation-sensitive enzymes *Sma*I and *Hpa*II. The possible restriction fragments large enough to be visible on a southern blot, that would hybridise to the *Aprt*-specific probe are indicated by the coloured arrows.

Figure 5.12 Analysis of transgene methylation status by Southern blot

Figure 5.12 Analysis of transgene methylation status by Southern blot. DNA obtained from: a wild-type (C57BL/6J) mouse; a. pAZM2.1CAG transgenic mice; and b. pABSGAC transgenic mice, restricted with either K- *Kpn*I, S- *Kpn*I and *Sma*I, or H- *Kpn*I and *Hpa*II, and hybridised to an *Aprt*-specific probe APX (see figure 5.10). The endogenous *Aprt* fragment (5 kb) and transgenic *aprt* fragment (2.2 kb) produced by *Kpn*I digestion are indicated by arrows to the right side of the blots. The 1.64 kb fragment of the Kb ladder, and other digestion fragments of known sizes (5, 2.5, 0.45, 0.39 and 0.28 kb) are all indicated by arrows to the left of each blot. Annotated with coloured stars are the A201 digests, which show unmethylated *transgene and *endogenous *Aprt* sequences. If the DNA is unmethylated the *Kpn*I restriction fragments should not be present in either of the two double digests. Conversely if the *Kpn*I fragments are not reduced in intensity in the double digests, the sequence is highly methylated.

A summary of the transgene methylation status in mice from each line is shown in Table 5.2. It was difficult to interpret the methylation pattern of the A196 transgene from these Southern blots as the *Sma*I digest was incomplete, however subsequent Southern analysis revealed it to be partially methylated. Repeated Southern blots of founder mice, and where possible the offspring from these lines confirmed that these methylation observations were repeatable and inherited stably. The Southern blot analysis was performed by the individual who produced the transgenic lines.

5.2.4.3 Methylation status of transgenes

The two different methods of assessing transgene methylation produced predominantly concurring results, with three out of fifteen analyses mismatched (see table 5.2).

5.2.4.3.1 Modified transgenes are methylated differently to pABS and pAZM2

In the previous study [Macleod et al., 1994] the methylation status of the CpG islands within the transgenes pABS and pAZM2 was directly related to the functionality of the Sp1 binding sites (GC boxes) within these constructs. The southern blot methylation data obtained here was directly compared with that from the previous study in the two tables below. In the two instances where Southern blot data could not be obtained the PCR results were substituted as these were found to confirm the Southern blot data in the majority of lines:

Transgenic constructs with Sp1 binding sites	Lines methylated	Lines unmethylated
pABS	1	9
pABS + repeat	4	5

Transgenic constructs with mutated Sp1 sites	Lines methylated	Lines unmethylated
pAZM2	4	1
pAZM2.1+ repeat	4	4

It is clear from these tables that when the trinucleotide repeat was present in the constructs that the pABS transgenes were more frequently methylated, and the pAZM2 transgenes were more frequently unmethylated.

5.2.4.3.2 Influence of repeat orientation on transgene methylation

Another trend discernible from this data set is that the majority of CTG repeat containing transgenes (8/9) were unmethylated, and conversely most CAG repeat containing transgenes (6/8 or 7/8) were methylated to some degree.

Table 5.2 Methylation status of transgenes

Type of construct	Founder mouse	Methylation status of transgene determined by southern blot analysis	Methylation status of transgene determined by PCR analysis
pAZM2.1CAG	A196	Partially methylated	Methylated
	A197	Heavily methylated	Methylated
	A197.1	Partially methylated	Unmethylated
pABSCAG	A198	Could not be determined	Methylated
	A199	Partially methylated	Methylated
	A200	Could not be determined	Methylated
	A201	Unmethylated	Unmethylated
	A201.1	Partially methylated	Methylated
pAZM2.1CTG	K60	Unmethylated	Unmethylated
	K65	Heavily methylated	Methylated
	K68	Unmethylated	Unmethylated
	K71	Unmethylated	Methylated
	K79	Unmethylated	Unmethylated
pABSCTG	R13	Unmethylated	Unmethylated
	R35	Unmethylated	Unmethylated
	R72	Unmethylated	Methylated
	R203	Unmethylated	Unmethylated

Highlighted are transgenes which produced conflicting methylation patterns when determined by Southern blot and PCR analysis.

5.2.5 Analysis of trinucleotide repeat stability in transgenic lines

The size of the trinucleotide repeats in individuals from each transgenic line was determined by PCR. The positive transgenic PCR products (DMrepeat) contained a fluorescent FAM residue which meant they could be accurately sized on polyacrylamide gels analysed by an Automatic Laser Fluorescence (ALF™) system. The number of trinucleotide repeats in each PCR product was determined by comparisons with founder mouse PCR products (which had also been sequenced) run on the same gel, a 50 bp interval DNA ladder and internal size controls run with each product. An example of a PCR product size change observed in the transgenic line A196 is shown in figure 5.13. Also represented in this figure are human patient PCR products, the expanded allele of which gives a distinctive rounded peak of products due to the extreme somatic repeat instability. All the transgenic PCR products from this survey (which contained repeats of a similar size to the human allele) gave peaks which were more akin to the normal human PCR products, comprising three peaks of increasing intensity followed by a smooth drop off from the final peak. It was this last, largest peak which was considered to represent the true allele size. The data produced from the analysis of each transgenic line to date is summarised in the tables 5.3. The trinucleotide repeat size analysis of transgenes from all the transgenic lines was performed by myself.

A sample selection of PCR products containing different sized transgenic alleles were also sequenced to confirm that they were the product of trinucleotide repeat copy number variations.

5.2.5.1 Bias towards contraction in repeat size changes

Out of the 52 trinucleotide repeat size changes observed in this study, 94% were contractions. The majority of contracted tracts (65%) lost one repeat unit, less than half that number (26.5) lost two repeats, and only 2% contracted by three or four repeats.

5.2.5.2 Influence of parental sex on trinucleotide repeat instability

The other most striking observation from this data set is that 98% of the size changes observed on transgene transmission occurred through the female germline, only one size change (an expansion of one repeat) was seen through the male germline. This figure is highly statistically significant giving a Chi-square value of 79, and a chance probability (p) of ≤ 0.001.

Figure 5.13 ALF™ traces of transgene PCR products

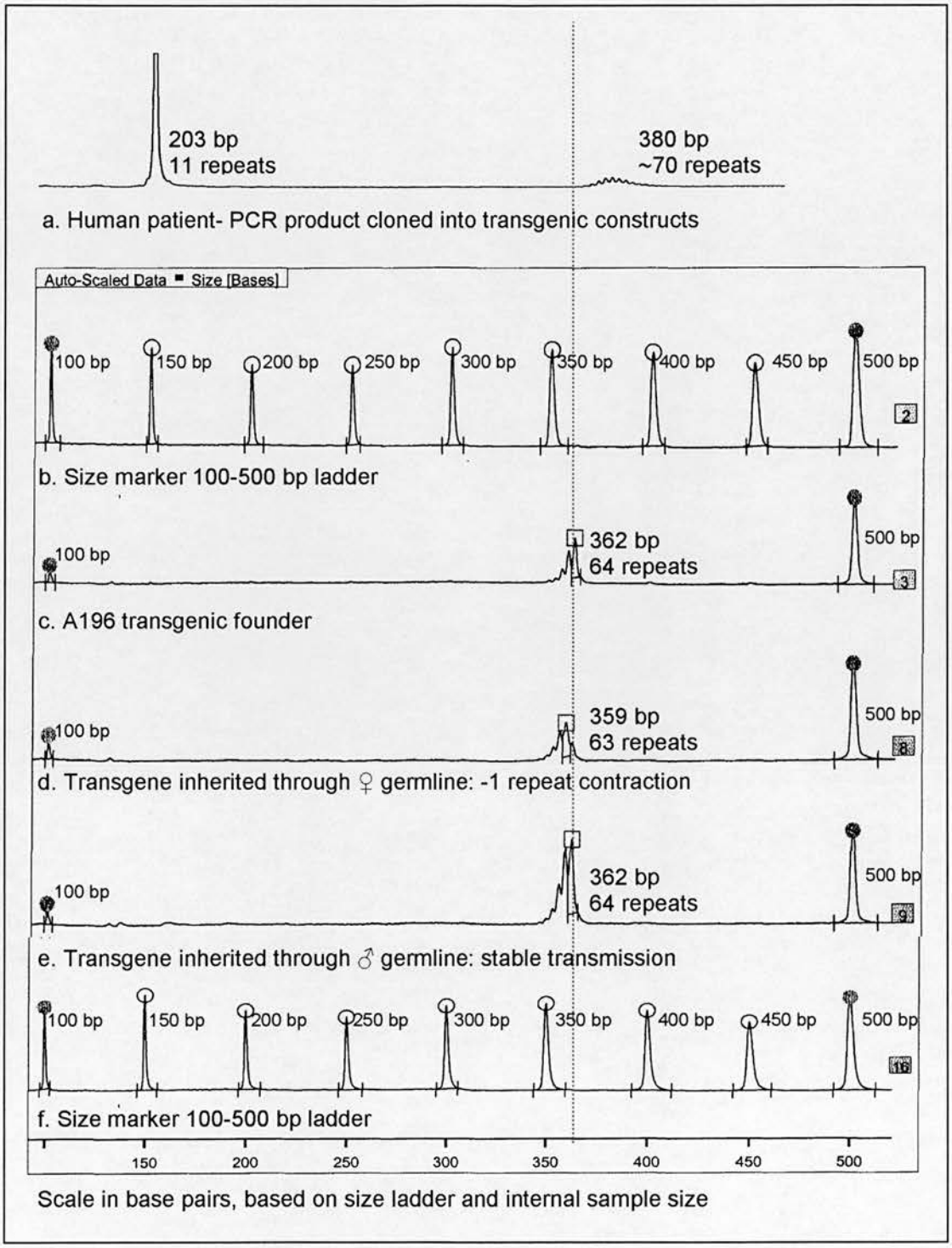


Figure 5.13 ALF™ traces of transgene PCR products. Trace: a. Human patient DNA with alleles containing 11 and ~70 repeats; b. and f. Size marker 100-500 bp ladder; c. Transgenic founder A196 transgene contains 64 repeats; d. Transgenic offspring of A196, transgene inherited through the female germline, repeat contracted on transmission; e. Transgenic offspring from A196 line, transgene stably inherited through the male germline.

Tables 5.3 Analysis of trinucleotide repeat stability in transgenic lines

Table 5.3.1 pAZM2.1CAG transgenic lines

Line	Repeat size on transmission from female parent		Repeat size on transmission from male parent		Number of transgenic mice analysed
	Female offspring	Male offspring	Female offspring	Male offspring	
A196	3 no change 3 –1 contractions **50% changes**	4 no change 1 –1 contraction 1 –2 contraction **33% changes**	11 no change **0% changes**	6 no change **0% changes**	29
A197.1	14 no change **0% changes**	9 no change **0% changes**	9 no change **0% changes**	1 +1 expansion 7 no change **12% changes**	40

Table 5.3.2 pABSCAG transgenic lines

Line	Repeat size on transmission from female parent		Repeat size on transmission from male parent		Number of transgenic mice analysed
	Female offspring	Male offspring	Female offspring	Male offspring	
A198	6 no change 3 –1 contractions **33% changes**	2 no change 5 –1 contractions 1 –4 contraction **75% changes**	5 no change **0% changes**	7 no change **0% changes**	29
A199	5 –2 contractions **100% changes**	1 –1 contraction 3 –2 contractions **100% changes**	15 no change **0% changes**	13 no change **0% changes**	37
A201	1 –1 contraction 1 –2 contraction **100% changes**	1 no change 3 –1 contractions 1 –4 contractions **80% changes**	16 no change **0% changes**	11 no change **0% changes**	34

Table 5.3.3 pAZM2.1CTG transgenic lines

Line	Repeat size on transmission from female parent		Repeat size on transmission from male parent		Number of transgenic mice analysed
	Female offspring	Male offspring	Female offspring	Male offspring	
K60	3 no change 3 –1 contractions **50% changes**	2 no change 1 –1 contraction **33% changes**	14 no change **0% changes**	3 no change **0% changes**	26
K65	2 no change **0% changes**	1 no change **0% changes**	- -	4 no change **0% changes**	7
K68	7 no change 1 –3 contraction **12% changes**	2 no change **0% changes**	- -	- -	10
K71	2 +1 expansions 3 no change 3 –1 contractions **62% changes**	8 no change 1 –1 contraction **11% changes**	3 no change **0% changes**	4 no change **0% changes**	24

Table 5.3.4 pABSCTG transgenic lines

Line	Repeat size on transmission from female parent		Repeat size on transmission from male parent		Number of transgenic mice analysed
	Female offspring	Male offspring	Female offspring	Male offspring	
R35	8 no change 2 –1 contractions **20% changes**	8 no change **0% changes**	5 no change **0% changes**	7 no change **0% changes**	30
R72	4 no change 2 –1 contractions 1 –3 contraction **43% changes**	2 no change 3 –1 contractions 3 –2 contractions **75% changes**	18 no change **0% changes**	18 no change **0% changes**	51
R203	- -	- -	8 no change **0% changes**	7 no change **0% changes**	15

The percentage of offspring exhibiting trinucleotide repeat size changes through female germline transmission varied considerably from line to line. This data was sorted into descending percentages of instability in table 5.4, alongside other relevant transgenic line details. From this table it was easier to discern any putative associations of trinucleotide repeat instability with other experimental factors.

5.2.5.3 Influence of offspring gender on trinucleotide repeat instability

In the majority of transgenic lines (6/11) the percentage of female progeny exhibiting repeat size changes was considerably higher than the percentage observed in male offspring. In three other lines the percentages of instability were approximately equal in both sexes, and in the remaining two lines this gender influence appeared to be reversed. Too few mice have yet been analysed to determine whether these distribution differences will prove statistically significant.

5.2.5.4 Influence of repeat orientation on trinucleotide repeat instability

There appeared to be a correlation between the trinucleotide repeat orientation within the transgene constructs and the overall degree of instability observed. Three out of four constructs exhibiting a high degree of variability (>50% of progeny had size changes) contained a CAG repeat, and five out of the seven more stably inherited constructs (<50% of progeny had size changes) contained a CTG repeat.

5.2.5.5 Influence of transgenic construct type on trinucleotide repeat instability

The functionality of the Sp1 sites within the transgenic constructs (pABS or pAZM2.1) may also have exerted some effect on repeat instability. If the cut off point of 50% instability is again considered, three out of four unstable lines were produced from constructs containing the native Sp1 binding sites (pABS), and six out of the seven more stable lines were produced from constructs with non functional Sp1 sites (pAZM2.1).

5.2.5.6 Influence of repeat size and transgene methylation on trinucleotide repeat instability

The founding trinucleotide repeat size and transgene methylation patterns did not appear to confer any clear effects on repeat instability.

Table 5.4 Summary table of trinucleotide repeat instability (through female germline transmission) and putative modifiers

Average % of offspring with transgene size changes	% of ♀ offspring with transgene size changes	% of ♂ offspring with transgene size changes	Number of mice analysed	Type of construct	Size of repeat tract in founder	Methylation status by Southern blot	Methylation status by PCR	Transgenic line
100%	100%	100%	n = 9	**pABSCAG**	70	Methylated	Methylated	A199
90%	100%	80%	n = 7	**pAZM2.1CAG**	70	Unmethylated	Unmethylated	A201
59%	43%	75%	n = 15	**pABSCTG**	70	Unmethylated	Methylated	R72
54%	33%	75%	n = 17	**pABSCAG**	70	?	Methylated	A198
41.5%	50%	33%	n = 12	**pAZM2.1CAG**	64	Methylated	Methylated	A196
41.5%	50%	33%	n = 9	**pAZM2.1CTG**	70	Unmethylated	Unmethylated	K60
36.5%	62%	11%	n = 17	**pAZM2.1CTG**	70	Unmethylated	Methylated	K71
10%	20%	0%	n = 18	**pABSCTG**	70	Unmethylated	Unmethylated	R35
6%	12%	0%	n = 10	**pAZM2.1CTG**	71	Unmethylated	Unmethylated	K68
0%	0%	0%	n = 23	**pAZM2.1CAG**	78	Methylated	Unmethylated	A197.1
0%	0%	0%	n = 3	**pAZM2.1CTG**	70	Methylated	Methylated	K65

5.3 Discussion

The transgenic study data presented here is unfortunately somewhat preliminary due to experimental time constraints relating to the generation of transgenic progeny. In some instances the number of transgenic lines generated for a specific construct and the numbers of progeny analysed are currently insufficient to draw firm conclusions from. As a consequence this project is ongoing in the laboratory phase.

5.3.1 Founding transgenic mice

5.3.1.1 Transgene trinucleotide repeat numbers

The trinucleotide repeat copy numbers in the transgenes of founding mice varied from 64-78 repeats. One founding mouse appeared to have one transgene containing 69 repeats and the other containing 70. The clones from which the constructs were derived carried 70 repeats, but it is possible that during their propagation small subsections of the population misreplicated the repeats, and that these were in a few cases the nucleotide sequences that were integrated into the mice. It is also possible that these size changes occurred during transgene integration.

5.3.1.2 'Mosaic transgenic mice

Of the eight transgenic founder mice produced by myself, one mouse failed to produce any transgenic offspring until its third litter, and two mice did not yield any transgenic progeny from the three litters screened. These mice were obviously mosaic for the transgenes, this percentage of 37.5% was slightly higher than that normally expected from transgenic production (10 - 30%) [2000] and was most likely a direct consequence of my being a novice to the techniques involved. This meant that it took longer for me to proceed to the injection stage than for a practised individual, presumably resulting in more embryos being nearer the two cell stage at the point of injection, increasing the likelihood of the transgenes integrating later in the developmental cycle.

5.3.2 Methylation of transgenes

The effects of CpG island status, most saliently methylation on trinucleotide repeat instability was the basis of this study. The previously tested constructs pABS and pAZM2 provided suitable vehicles in which to deliver these repeats into CpG island and non-island environments. The introduction of the human Myotonic Dystrophy repeats and 0.3 kb of flanking sequence did not theoretically disrupt the CpG island potential of the constructs (by NIX analysis).

5.3.2.1 Two methods for analysing transgene methylation

Two different assays were used to determine transgene methylation, the Southern blot to allow direct comparison with the results of the previous study, and PCR analysis to be assessed as a more rapid technique for screening large numbers of progeny.

For the most part the two techniques used to determine methylation patterns gave identical results. However in the three out of fifteen instances they did not concur. It is possible that both assays were correct if the transgenes were present in more than one copy, and were differentially methylated due to position effects. Of the two techniques the PCR was the more sensitive and therefore these results might be considered the most appropriate to rely on. Both these assays were reliant on complete enzymatic digestion of the DNA, but in both cases the endogenous *Aprt* digestion could be utilised as a reliable digestion control. Bisulphite or 'genomic' sequencing could also have been used to assess the methylation of the transgenes, but was considered too labour intensive for the number of samples to be handled in this study.

In retrospect the Southern blot analysis might have been clearer if a transgene-specific (human DMPK sequence) probe had been used instead of the *Aprt*-specific probe (APX). Some of the endogenous *Aprt* digested fragments were unfortunately of a similar size to the transgene fragments which complicated result interpretations.

5.3.2.2 Modified transgenes were methylated differently to pABS and pAZM2

In this experimental design we expected the functionality of the Sp1 binding sites to directly influence the methylation of the transgenes as it had in the previous study [Macleod et al., 1994]. However the cloning of the trinucleotide repeats into these constructs appeared to have altered this association. The modified pABS constructs were more frequently methylated than in the previous study, and the pAZM2.1 constructs were less frequently

methylated than expected. It is possible that this was caused by transgene position effects, but it seems more likely that the trinucleotide repeat or flanking DNA has directly resulted in altered transgene methylation patterns. It could be that the trinucleotide repeat or flanking sequence has effected Sp1 binding efficiency, or disrupted the CpG island sequence composition, thus altering the methylation. This is potentially feasible as the cloning site used to receive the repeat containing DNA was only a couple of hundred bases downstream of the Sp1 sites. This putative effect could be addressed by cloning the same trinucleotide repeat tract into another intronic region of the *Aprt* gene constructs, outside the CpG island.

It is also possible that the altered methylation was a result of the trinucleotide repeat tract being of a pathogenic human size. The altered methylation of DNA in repeat regions could well be involved in some of the pathogenic processes involved with (CAG/CTG) trinucleotide repeat expansions. For example large CTG repeats are already known to cause localised hypermethylation of the DMPK gene in severely affected DM patients [Steinbach et al., 1998]. The best way to explore this possibility would be to produce transgenic mice from similar constructs containing a normal sized trinucleotide repeat allele. This control is currently being constructed and will shortly be used to create transgenic animals.

5.3.2.3 Influence of repeat orientation on transgene methylation

The orientation of the trinucleotide repeat within the transgenes was not expected to effect their methylation pattern. However the majority of transgenes containing CAG repeats were methylated and most of those containing CTG repeats were unmethylated. This observation could of course be coincidental, a by product of position effects and other factors. However if that is not the case how could repeat orientation effect methylation?

The first possibility comes back to the putative effect of the trinucleotide repeat sequence on the Sp1 binding sites, perhaps the repeat or flanking sequence in the CAG orientation somehow inhibits the binding of SpI, resulting in the de novo methylation of the CpG island. Whereas in the opposite orientation this factor or influence is negated, or far enough away from the sites to permit Sp1 binding, protecting the transgene from methylation. Or it could be that the Sp1 binds normally, but that in the CAG orientation, the repeat sequence was not sufficiently island like in base composition resulting in the premature disruption of the island and protection from methylation. This putative effect could be clarified to by studying the expression levels of the *Aprt* genes with the repeat in different orientations. Unfortunately because the *Aprt* gene construct is a modified mouse gene, it would be practically impossible to determine the expression resulting only from the transgenes.

The other possibility is that as the different repeat orientation constructs were handled by different investigators, the person injecting the construct has somehow influenced their methylation. Perhaps the CAG repeat containing constructs being injected a few hours later in embryo development (due to my learning the technique) than the CTG containing constructs has subjected them to a pre-programmed timing, meiotic division effect, resulting in their methylation. It is unlikely that this methylation difference was caused by a genetic background effect as both investigators used the same type of embryos (F_1 CBA x C57BL/6J) for transgenic production.

5.3.3 Trinucleotide repeat stability

This study revealed moderate instability of trinucleotide repeats in ten out of the twelve transgenic lines produced. A strong bias towards repeat contraction and repeat instability on transmission through the female germline were observed. Other putative associations with repeat instability were tentative and more animals need to be analysed to support their existence.

5.3.3.1 Bias towards contraction in repeat size changes

Of the trinucleotide repeat size changes observed in transgenic offspring, 94% were repeat contractions. This phenomenon has repeatedly been noted in transgenic experiments, perhaps highlighting an organism specific difference in trinucleotide repeat behaviour. The only exception to this rule to date, was a study of transgenic constructs containing massive trinucleotide repeats (~300) embedded in 45 kb of its native human genomic *DMPK* sequence [Seznec et al., 2000]. It is therefore possible that the bias towards repeat contraction was negated in these mice due to the influence of the human flanking sequence, or the larger start size of the trinucleotide repeat array.

In humans, the tendency of trinucleotide repeats to contract or expand differs considerably between the various disease loci, suggesting that *cis* acting factors might be key determinants influencing the process [Brock et al., 1999].

5.3.3.2 Influence of parental sex on trinucleotide repeat instability

In this study 98% of the trinucleotide repeat instabilities occurred through female germline transmission. Only one transgenic line exhibited any instability through the male germline, and this repeat size change was a small expansion. The increased instability through female

germline transmission was also observed in a previous study where transgenic lines were created from constructs containing 3' UTR human DMPK sequence, and 162 CTG repeats [Monckton et al., 1997]. However Monckton *et al.* reported a higher level of instability through the female line than through the male, whereas in this study the instability occurred exclusively through a specific germline. The reproducible and exclusive sex of origin effects observed here were probably the result of a *cis* acting influence in the *Aprt* construct DNA, the few anomalous transgenic lines presumably the result of transgene position effects. It is interesting that the human trinucleotide repeat with such a small amount of flanking sequence (300 bp) was capable of producing such a strong sex of origin effect. The trinucleotide repeat was clearly out of its native genomic context, the only broad parallel being that the transgenic repeat tract was also located in a non-translated region of DNA.

The existence of sex-specific differences in trinucleotide repeat instabilities implies that a sex-specific *trans* acting factor must also be involved in the process. Differences in the number of mitotic divisions [Jansen et al., 1994], or levels of transcription during oogenesis and spermatogenesis might be responsible.

5.3.3.3 Influence of offspring gender on trinucleotide repeat instability

Although not statistically significant, clear differences in the number of trinucleotide repeat size changes observed in female and male offspring were seen in this study. Embryo gender has recently been reported to effect repeat instability in transgenic mice containing a Huntington's disease gene [Kovtun et al., 2000]. In the study reported here, six out of 11 transgenic lines exhibited increased numbers of size changes in female offspring, in three lines the number of size changes were approximately equal in both sexes, and in the remaining transgenic lines the offspring gender influence was reversed. It is difficult to imagine how transgene position effects could result in this phenomenon, but if these effects prove to be statistically significant this would seem the most likely cause. Local *cis* acting factors are unlikely to be responsible because the trend did not appear linked to transgenic construct type, repeat orientation or methylation patterns. As the effect appears reversed in several lines it is hard to envisage a sex biased *trans* acting factor causing the phenomenon.

5.3.3.4 Influence of repeat orientation on trinucleotide repeat instability

There may be a correlation between the trinucleotide repeat orientation within the transgenic constructs and the overall degree of instability. Three out of four constructs exhibiting the most instability (>50% progeny unstable) contained a CAG repeat, and five out of the seven

more stably inherited constructs (<50% of progeny unstable) contained a CTG repeat. This observation was not unexpected as trinucleotide repeat orientation relative to the origin of replication has a well documented effect on instability in *E. coli* and *S. cerevisiae* and was a considered factor in the experimental design.

5.3.3.5 Influence of construct type on trinucleotide repeat instability

The functionality of the Sp1 sites within the transgenic constructs (pABS or pAZM2.1) may also have exerted some effect on repeat instability. Three out of the four transgenic lines exhibiting the most instability were produced from constructs containing the native Sp1 binding sites (pABS), and six out of the seven more stable lines were produced from constructs with non functional Sp1 sites (pAZM2.1). This trend was what we might have expected if Sp1 functionality had directly influenced the transgene CpG island methylation and if this in turn directly influenced trinucleotide repeat instability. As the transgene methylation patterns did not appear strongly related to Sp1 functionality, it is difficult to account for this putative association.

5.3.3.6 Influence of trinucleotide repeat size and transgene methylation on trinucleotide repeat instability

The founding trinucleotide repeat size (64-78) varied very little, it was therefore not surprising to find that this factor had little effect on repeat instability. The methylation of the transgenes did not appear to greatly influence repeat instability either, which in itself suggests that the putative association between CpG islands and trinucleotide repeat expansions may be groundless. It is possible however that this association is indirect and that trinucleotide repeats are effected by another phenomenon, which is itself sometimes associated with CpG islands.

In an attempt to improve the propensity for trinucleotide repeat instability, such as that seen in a recent study [Seznec et al., 2000] trinucleotide repeats of a much larger magnitude are currently being cloned into these transgenic constructs. It is hoped that by also including a control construct containing a 'normal' sized trinucleotide repeat, and by studying more transgenic lines, that any effects of CpG island methylation on trinucleotide repeat instability will soon be elucidated.

CHAPTER 6

CONCLUDING REMARKS

6 Concluding remarks

6.1 Survey of trinucleotide repeats in mouse CpG islands

6.1.1 Aims achieved

The CpG island library screen for trinucleotide repeats described in this thesis, represents the most comprehensive study of all 10 classes of repeat in the mouse to date. Previous surveys have focused primarily on just one class of repeat CAG/CTG [Abbott and Chambers, 1994; Chambers and Abbott, 1996; Kim et al., 1997; King et al., 1998], but as an increasing number of repeat classes are being associated with expansion mutation, the relevance of screening for all possible motifs is of obvious importance.

6.1.2 Conclusions

The most striking observation gained from this survey is the apparent depletion of trinucleotide repeats in mouse, but not human CpG islands. If the putative association of repeat expansions with CpG islands holds true, this depletion could be a key factor in mediating why no endogenous expanded repeats have been identified in the mouse.

The most obvious limitations of the survey reported here were that: the library construction [Cross et al., 1997] included a PCR amplification step and that the presence of CpG islands was predicted rather than directly assayed by determining DNA methylation status. PCR amplification is known to struggle with long, especially GC rich repeat sequences which may have biased against their inclusion in library clones. Whether a region of DNA truly represents a CpG island is most likely reflected in the DNA methylation status rather than just the sequence base composition. However as these assay limitations were consistent in both the mouse and human surveys, the data are suitable for direct comparison and the depletion of repeats in mouse CpG islands may be a reasonable reflection of the genomic situation.

This survey revealed that large trinucleotide repeats, approaching human pathogenic sizes are present in the mouse genome. AAG1 contained 41 repeats in NZB/B1NJ DNA and AGC2 contained 24 repeats in *Rattus norvegicus* DNA.

6.1.3 Future work

At this point in time the sequencing of the entire human genome nears completion [Consortium, 2001] and that of the mouse is well under way. Therefore undertaking further genome screens for trinucleotide repeats may be of limited value at this juncture. Once the complete sequences are available unbiased comparisons of trinucleotide repeat frequencies in various genomic region and their relative sizes in both genomes will be easy to perform using bioinformatic techniques. This should rule out the limitations of sequence predictions, library construction and experimental inconsistencies. However the polymorphism and variability of repeat arrays will still need to be assessed experimentally until such time as complete sequence data is available from enough individuals of varied ethnic origins.

As a growing number of repeat types are being associated with expansion and disease, screens for large variable tetra-, penta-, hexanucleotides etc., may be valid exercises in identifying potential causes of both human and mouse pathogenic phenotypes.

As research to date has failed to identify any endogenous mouse trinucleotide repeat expansions, transgenic techniques may be the most suitable approach to explore their behaviour in this organism. Recent studies have revealed that under the right conditions many aspects of this mutational mechanism can be recreated in the mouse. It may well be that this disease mechanism does naturally exist in the mouse, but that the type of progressive disorders most commonly associated with it are not the easiest of phenotypes to identify in mouse stocks. Nor would repeat expansions be the type of mutations most likely to arise in contrived mutagenesis programs.

6.2 Size variation and mapping of mouse trinucleotide repeats

6.2.1 Aims achieved

The assessment of repeat variability in different mouse strains reported in this thesis represents the most in depth study of all trinucleotide repeat classes performed in the mouse to date. Although many highly variable trinucleotide repeats were identified in this survey no expansions were found, nor were any associated with a mouse phenotype. However a large hexanucleotide and a complex trinucleotide repeat expansion were isolated and may

yet prove to be linked with mouse mutant phenotypes. The mapped, variable trinucleotide repeat loci will undoubtedly prove useful as genetic markers in future mapping studies.

6.2.2 Conclusions

The analysis supports previous observations that repeat size is a strong modifier of repeat instability [Ashley and Warren, 1995; Riggins et al., 1992; Weber, 1990]. The study also revealed that repeat class may be associated with variability, AAG and AGC repeats displaying the widest range of variability in this mouse survey. It is interesting to note that these two classes of repeat are amongst those capable of expansion in humans. The location of mouse trinucleotide repeats within CpG islands did not appear to confer an increased rate of instability. However the phenomena of repeat instability and expansion mutation may be driven by different mechanisms and therefore modified by different factors.

6.2.3 Future work

Time limitations did not permit the mapping of every trinucleotide repeat identified in this survey. Several of the smaller repeats exhibited very small polymorphisms which could not be consistently resolved by agarose gel electrophoresis. The discovery that disease phenotypes such as PSACH and its allelic variant MED [Briggs et al., 1995; Briggs et al., 1998; Delot et al., 1999; Ikegawa et al., 1998], synpolydactyly [Akarsu et al., 1996; Johnson et al., 1998; Muragaki et al., 1996] and OPMD [Brais et al., 1998] are all caused by small duplications or deletions of repeats, makes mapping these smaller mouse repeats and examining them as candidates for mouse phenotypes a potentially worth while pursuit. These repeats could be easily mapped using fluorescent PCR primers and automated fragment detection which has a higher resolution capability than agarose gel electrophoresis. Likewise the large mapped repeats described in this study could be re-examined for small duplications and deletions in mouse mutant DNA, as in this study they were primarily assessed for large expansion mutations.

On preliminary analysis the two large expanded repeats identified in frizzy DNA do not appear to underlie the mutant phenotype. However a live stock of frizzy mice are being procured by our research group with the objective of studying the variability of these repeats on transmission through a pedigree. These mice may also be used to map the repeats more precisely in relation to each other and to the frizzy locus.

6.3 Transgenic study to determine if locating a trinucleotide repeat within a CpG island would effect repeat instability

6.3.1 Aims acieved

The object of this study was to determine whether locating a trinucleotide repeat within a CpG island would directly effect its transmissional instability. Unfortunately the inclusion of the trinucleotide repeat and 300 bp of flanking sequence into the transgenic constructs pABS and pAZM2, appears to have modified the direct relationship previously observed between Sp1 binding site functionality and CpG island methylation [Macleod et al., 1994]. However the transgenic lines which have been generated represent a new model for moderate trinucleotide repeat instability in the mouse. These mice manifest a strong sex of founder effect on repeat instability and may also display embryo gender effects.

6.3.2 Conclusions

How the addition of a trinucleotide repeat with a small amount of flanking sequence into the intron of a transgene has had such a profound effect on CpG island methylation is currently unclear. It may be that the repeat has in some way interfered with the normal binding of Sp1 to its target site located a couple of hundred base pairs upstream. Another possibility is that the repeat and flanking sequence have simply disrupted the CpG island sequence composition thus destroying the normal methylation pattern of the region. The sex of founder effect on repeat instability is most likely conferred by a *cis* acting element in the native *DMPK* sequence flanking the repeat, or coincidentally present in the *Aprt* gene sequence.

6.3.3 Future work

Mice from the transgenic lines produced in this study are still being generated and assessed. Increasing the number of progeny analysed should statistically strengthen the observations reported here. These mice may also display the founder age effects on instability reported in other transgenic studies.

The production of more transgenic lines from the same constructs may rule out the possibility that the modified methylation patterns observed here are the result of unfortunate position of integration effects. It is possible that the aberrant methylation of these transgenes

was due to the inclusion of an expanded trinucleotide repeat. To address this point control constructs are currently being prepared containing a similar, but nonexpanded trinucleotide repeat. Cloning the repeat into a downstream intron of the transgenic constructs could potentially resolve whether the repeat was directly interfering with normal Sp1 binding.

With hindsight the constructs we generated may not have been the most suitable to address the putative association between CpG islands and trinucleotide repeat instability. In the interim period since this project was started, it has been demonstrated that only extremely large repeats insulated by considerable contextual sequence show any significant degree of instability in the mouse [Seznec et al., 2000]. It has also been proposed that the association observed between repeat expansion and CpG islands may actually reflect a relationship between expansion and origins of replication. Any future transgenic experiment designed to study trinucleotide repeat instability would ideally incorporate or address the aforementioned points.

BIBLIOGRAPHY

Abbott, C., and Chambers, D. (1994). Analysis of CAG trinucleotide repeats from mouse cDNA sequences. Ann Hum Genet ***58***, 87-94.

Abdullah, A., Trifiro, M. A., Panet-Raymond, V., Alvarado, C., de Tourreil, S., Frankel, D., Schipper, H. M., and Pinsky, L. (1998). Spinobulbar muscular atrophy: polyglutamine-expanded androgen receptor is proteolytically resistant in vitro and processed abnormally in transfected cells. Hum Mol Genet *7*, 379-84.

Adamec, J., Rusnak, F., Owen, W. G., Naylor, S., Benson, L. M., Gacy, A. M., and Isaya, G. (2000). Iron-dependent self-assembly of recombinant yeast frataxin: implications for Friedreich ataxia. Am J Hum Genet ***67***, 549-62.

Akarsu, A. N., Stoilov, I., Yilmaz, E., Sayli, B. S., and Sarfarazi, M. (1996). Genomic structure of HOXD13 gene: a nine polyalanine duplication causes synpolydactyly in two unrelated families. Hum Mol Genet ***5***, 945-52.

Alwazzan, M., Newman, E., Hamshere, M. G., and Brook, J. D. (1999). Myotonic dystrophy is associated with a reduced level of RNA from the DMWD allele adjacent to the expanded repeat. Hum Mol Genet ***8***, 1491-7.

Amack, J. D., Paguio, A. P., and Mahadevan, M. S. (1999). Cis and trans effects of the myotonic dystrophy (DM) mutation in a cell culture model. Hum Mol Genet ***8***, 1975-84.

Amar, L. C., Dandolo, L., Hanauer, A., Cook, A. R., Arnaud, D., Mandel, J. L., and Avner, P. (1988). Conservation and reorganization of loci on the mammalian X chromosome: a molecular framework for the identification of homologous subchromosomal regions in man and mouse. Genomics ***2***, 220-30.

Ansari-Lari, M. A., Oeltjen, J. C., Schwartz, S., Zhang, Z., Muzny, D. M., Lu, J., Gorrell, J. H., Chinault, A. C., Belmont, J. W., Miller, W., and Gibbs, R. A. (1998). Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6. Genome Res ***8***, 29-40.

Antequera, F., and Bird, A. (1993). Number of CpG islands and genes in human and mouse. Proc Natl Acad Sci U S A ***90***, 11995-9.

Antequera, F., and Bird, A. (1999). CpG islands as genomic footprints of promoters that are associated with replication origins. Curr Biol ***9***, R661-7.

Ashley, C. T., Sutcliffe, J. S., Kunst, C. B., Leiner, H. A., Eichler, E. E., Nelson, D. L., and Warren, S. T. (1993). Human and murine FMR-1: alternative splicing and translational initiation downstream of the CGG-repeat. Nat Genet ***4***, 244-51.

Ashley, C. T., and Warren, S. T. (1995). Trinucleotide repeat expansion and human disease. Annu Rev Genet ***29***, 703-28.

Avner, P., Amar, L., Dandolo, L., and Guenet, J. L. (1988). Genetic analysis of the mouse using interspecific crosses. Trends Genet ***4***, 18-23.

Babcock, M., de Silva, D., Oaks, R., Davis-Kaplan, S., Jiralerspong, S., Montermini, L., Pandolfo, M., and Kaplan, J. (1997). Regulation of mitochondrial iron accumulation by Yfh1p, a putative homolog of frataxin. Science ***276***, 1709-12.

Balakumaran, B. S., Freudenreich, C. H., and Zakian, V. A. (2000). CGG/CCG repeats exhibit orientation-dependent instability and orientation-independent fragility in Saccharomyces cerevisiae. Hum Mol Genet ***9***, 93-100.

Banfi, S., Servadio, A., Chung, M. Y., Kwiatkowski, T. J., McCall, A. E., Duvick, L. A., Shen, Y., Roth, E. J., Orr, H. T., and Zoghbi, H. Y. (1994). Identification and characterization of the gene causing type 1 spinocerebellar ataxia. Nat Genet *7*, 513-20.

Banfi, S., Servadio, A., Chung, M., Capozzoli, F., Duvick, L. A., Elde, R., Zoghbi, H. Y., and Orr, H. T. (1996). Cloning and developmental expression analysis of the murine homolog of the spinocerebellar ataxia type 1 gene (Sca1). Hum Mol Genet ***5***, 33-40.

Bao, J., Sharp, A. H., Wagster, M. V., Becher, M., Schilling, G., Ross, C. A., Dawson, V. L., and Dawson, T. M. (1996). Expansion of polyglutamine repeat in huntingtin leads to abnormal protein interactions involving calmodulin. Proc Natl Acad Sci U S A ***93***, 5037-42.

Barker, D., Schafer, M., and White, R. (1984). Restriction sites containing CpG show a higher frequency of polymorphism in human DNA. Cell ***36***, 131-8.

Barnes, G. T., Duyao, M. P., Ambrose, C. M., McNeil, S., Persichetti, F., Srinidhi, J., Gusella, J. F., and MacDonald, M. E. (1994). Mouse Huntington's disease gene homolog (Hdh). Somat Cell Mol Genet ***20***, 87-97.

Beckman, J. S., and Weber, J. L. (1992). Survey of human and rat microsatellites. Genomics *12*, 627-31.

Beilin, J., Ball, E. M., Favaloro, J. M., and Zajac, J. D. (2000). Effect of the androgen receptor CAG repeat polymorphism on transcriptional activity: specificity in prostate and non-prostate cell lines. J Mol Endocrinol ***25***, 85-96.

Benders, A. A., Groenen, P. J., Oerlemans, F. T., Veerkamp, J. H., and Wieringa, B. (1997). Myotonic dystrophy protein kinase is involved in the modulation of the Ca2+ homeostasis in skeletal muscle cells. J Clin Invest ***100***, 1440-7.

Berul, C. I., Maguire, C. T., Aronovitz, M. J., Greenwood, J., Miller, C., Gehrmann, J., Housman, D., Mendelsohn, M. E., and Reddy, S. (1999). DMPK dosage alterations result in atrioventricular conduction abnormalities in a mouse myotonic dystrophy model. J Clin Invest ***103***, R1-7.

Bhagwati, S., Ghatpande, A., and Leung, B. (1996). Normal levels of DM RNA and myotonin protein kinase in skeletal muscle from adult myotonic dystrophy (DM) patients. Biochim Biophys Acta ***1317***, 155-7.

Biancalana, V., Serville, F., Pommier, J., Julien, J., Hanauer, A., and Mandel, J. L. (1992). Moderate instability of the trinucleotide repeat in spino bulbar muscular atrophy. Hum Mol Genet *1*, 255-8.

Bidichandani, S. I., Ashizawa, T., and Patel, P. I. (1998). The GAA triplet-repeat expansion in Friedreich ataxia interferes with transcription and may be associated with an unusual DNA structure. Am J Hum Genet ***62***, 111-21.

Bingaman, E. W., Baeckman, L. M., Yracheta, J. M., Handa, R. J., and Gray, T. S. (1994). Localization of androgen receptor within peptidergic neurons of the rat forebrain. Brain Res Bull ***35***, 379-82.

Bingham, P. M., Scott, M. O., Wang, S., McPhaul, M. J., Wilson, E. M., Garbern, J. Y., Merry, D. E., and Fischbeck, K. H. (1995). Stability of an expanded trinucleotide repeat in the androgen receptor gene in transgenic mice. Nat Genet ***9***, 191-6.

Bird, A. P. (1987). CpG islands as gene markers in the vertebrate nucleus. Trends Genet ***3***, 342-347.

Bonhomme, F., Benmehdi, F., Britton-Davidian, J., and Martin, S. (1979). Genetic analysis of interspecific crosses Mus musculus L. x Mus spretus Lataste: linkage of Adh-1 with Amy-1 on chromosome 3 and Es-14 with Mod-1 on chromosome 9. C R Seances Acad Sci D ***289***, 545-8.

Bonhomme, F., Guenet, J. L., and Catalan, J. (1982). Presence of a male sterility factor, Hst-2, segregating in interspecies crosses. M. musculus L. x M. spretus Lastaste and linked to Mod-1, and Mpi-1 on chromosome 9. C R Seances Acad Sci III ***294***, 691-3.

Bonhomme, F., Catalan, J., Britton-Davidian, J., Chapman, V. M., Moriwaki, K., Nevo, E., and Thaler, L. (1984). Biochemical diversity and evolution in the genus Mus. Biochem Genet ***22***, 275-303.

Bontekoe, C. J., de Graaff, E., Nieuwenhuizen, I. M., Willemsen, R., and Oostra, B. A. (1997). FMR1 premutation allele (CGG)81 is stable in mice. Eur J Hum Genet ***5***, 293-8.

Boucher, C. A., King, S. K., Carey, N., Krahe, R., Winchester, C. L., Rahman, S., Creavin, T., Meghji, P., Bailey, M. E., Chartier, F. L., and et al. (1995). A novel homeodomain-encoding gene is associated with a large CpG island interrupted by the myotonic dystrophy unstable (CTG)n repeat. Hum Mol Genet ***4***, 1919-25.

Boutell, J. M., Wood, J. D., Harper, P. S., and Jones, A. L. (1998). Huntingtin interacts with cystathionine beta-synthase. Hum Mol Genet *7*, 371-8.

Bowater, R. P., Jaworski, A., Larson, J. E., Parniewski, P., and Wells, R. D. (1997). Transcription increases the deletion frequency of long CTG.CAG triplet repeats from plasmids in Escherichia coli. Nucleic Acids Res ***25***, 2861-8.

Brais, B., Bouchard, J. P., Xie, Y. G., Rochefort, D. L., Chretien, N., Tome, F. M., Lafreniere, R. G., Rommens, J. M., Uyama, E., Nohira, O., Blumen, S., Korczyn, A. D., Heutink, P., Mathieu, J., Duranceau, A., Codere, F., Fardeau, M., Rouleau, G. A., and Korcyn, A. D. (1998). Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. Nat Genet ***18***, 164-7.

Branda, S. S., Yang, Z. Y., Chew, A., and Isaya, G. (1999). Mitochondrial intermediate peptidase and the yeast frataxin homolog together maintain mitochondrial iron homeostasis in Saccharomyces cerevisiae. Hum Mol Genet ***8***, 1099-110.

Brandeis, M., Frank, D., Keshet, I., Siegfried, 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***, 435-8.

Breschel, T. S., McInnis, M. G., Margolis, R. L., Sirugo, G., Corneliussen, B., Simpson, S. G., McMahon, F. J., MacKinnon, D. F., Xu, J. F., Pleasant, N., Huo, Y., Ashworth, R. G., Grundstrom, C., Grundstrom, T., Kidd, K. K., DePaulo, J. R., and Ross, C. A. (1997). A novel, heritable, expanding CTG repeat in an intron of the SEF2-1 gene on chromosome 18q21.1. Hum Mol Genet ***6***, 1855-63.

Briggs, M. D., Hoffman, S. M., King, L. M., Olsen, A. S., Mohrenweiser, H., Leroy, J. G., Mortier, G. R., Rimoin, D. L., Lachman, R. S., Gaines, E. S., and et al. (1995). Pseudoachondroplasia and multiple epiphyseal dysplasia due to mutations in the cartilage oligomeric matrix protein gene. Nat Genet ***10***, 330-6.

Briggs, M. D., Mortier, G. R., Cole, W. G., King, L. M., Golik, S. S., Bonaventure, J., Nuytinck, L., De Paepe, A., Leroy, J. G., Biesecker, L., Lipson, M., Wilcox, W. R., Lachman, R. S., Rimoin, D. L., Knowlton, R. G., and Cohn, D. H. (1998). Diverse mutations in the gene for cartilage oligomeric matrix protein in the pseudoachondroplasia-multiple epiphyseal dysplasia disease spectrum. Am J Hum Genet ***62***, 311-9.

Brinkmann, B., Klintschar, M., Neuhuber, F., Huhne, J., and Rolf, B. (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am J Hum Genet ***62***, 1408-15.

Brock, G. J., Anderson, N. H., and Monckton, D. G. (1999). Cis-acting modifiers of expanded CAG/CTG triplet repeat expandability: associations with flanking GC content and proximity to CpG islands. Hum Mol Genet ***8***, 1061-7.

Brook, J. D., McCurrach, M. E., Harley, H. G., Buckler, A. J., Church, D., Aburatani, H., Hunter, K., Stanton, V. P., Thirion, J. P., Hudson, T., and et al. (1992). Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. Cell ***69***, 385.

Broude, N. E., Chandra, A., and Smith, C. L. (1997). Differential display of genome subsets containing specific interspersed repeats. Proc Natl Acad Sci U S A ***94***, 4548-53.

Bull, L. N., Pabon-Pena, C. R., and Freimer, N. B. (1999). Compound microsatellite repeats: practical and theoretical features. Genome Res ***9***, 830-8.

Bulle, F., Chiannilkulchai, N., Pawlak, A., Weissenbach, J., Gyapay, G., and Guellaen, G. (1997). Identification and chromosomal localization of human genes containing CAG/CTG repeats expressed in testis and brain. Genome Res *7*, 705-15.

Burke, J. R., Wingfield, M. S., Lewis, K. E., Roses, A. D., Lee, J. E., Hulette, C., Pericak-Vance, M. A., and Vance, J. M. (1994). The Haw River syndrome: dentatorubropallidoluysian atrophy (DRPLA) in an African-American family. Nat Genet *7*, 521-4.

Burke, J. R., Enghild, J. J., Martin, M. E., Jou, Y. S., Myers, R. M., Roses, A. D., Vance, J. M., and Strittmatter, W. J. (1996). Huntingtin and DRPLA proteins selectively interact with the enzyme GAPDH. Nat Med *2*, 347-50.

Burright, E. N., Clark, H. B., Servadio, A., Matilla, T., Feddersen, R. M., Yunis, W. S., Duvick, L. A., Zoghbi, H. Y., and Orr, H. T. (1995). SCA1 transgenic mice: a model for neurodegeneration caused by an expanded CAG trinucleotide repeat. Cell ***82***, 937-48.

Campuzano, V., Montermini, L., Molto, M. D., Pianese, L., Cossee, M., Cavalcanti, F., Monros, E., Rodius, F., Duclos, F., Monticelli, A., and et al. (1996). Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. Science ***271***, 1423-7.

Cavanna, J. S., Greenfield, A. J., Johnson, K. J., Marks, A. R., Nadal-Ginard, B., and Brown, S. D. (1990). Establishment of the mouse chromosome 7 region with homology to the myotonic dystrophy region of human chromosome 19q. Genomics *7*, 12-8.

Chai, Y., Koppenhafer, S. L., Shoesmith, S. J., Perez, M. K., and Paulson, H. L. (1999). Evidence for proteasome involvement in polyglutamine disease: localization to nuclear inclusions in SCA3/MJD and suppression of polyglutamine aggregation in vitro. Hum Mol Genet ***8***, 673-82.

Chakrabarti, L., Bristulf, J., Foss, G. S., and Davies, K. E. (1998). Expression of the murine homologue of FMR2 in mouse brain and during development. Hum Mol Genet *7*, 441-8.

Chambers, D. M., and Abbott, C. M. (1996). Isolation and mapping of novel mouse brain cDNA clones containing trinucleotide repeats, and demonstration of novel alleles in recombinant inbred strains. Genome Res ***6***, 715-23.

Chen, X., Mariappan, S. V., Catasti, P., Ratliff, R., Moyzis, R. K., Laayoun, A., Smith, S. S., Bradbury, E. M., and Gupta, G. (1995). Hairpins are formed by the single DNA strands of the fragile X triplet repeats: structure and biological implications. Proc Natl Acad Sci U S A ***92***, 5199-203.

Chen, T. H., Brody, J. R., Romantsev, F. E., Yu, J. G., Kayler, A. E., Voneiff, E., Kuhajda, F. P., and Pasternack, G. R. (1996). Structure of pp32, an acidic nuclear protein which inhibits oncogene-induced formation of transformed foci. Mol Biol Cell *7*, 2045-56.

Chiurazzi, P., Pomponi, M. G., Willemsen, R., Oostra, B. A., and Neri, G. (1998). In vitro reactivation of the FMR1 gene involved in fragile X syndrome. Hum Mol Genet *7*, 109-13.

Clark, H. B., Burright, E. N., Yunis, W. S., Larson, S., Wilcox, C., Hartman, B., Matilla, A., Zoghbi, H. Y., and Orr, H. T. (1997). Purkinje cell expression of a mutant allele of SCA1 in transgenic mice leads to disparate effects on motor behaviors, followed by a progressive cerebellar dysfunction and histological alterations. J Neurosci ***17***, 7385-95.

Consortium, The Dutch-Belgian Fragile X. (1994). Fmr1 knockout mice: a model to study fragile X mental retardation. Cell ***78***, 23-33.

Consortium, International Human Genome Sequencing. (2001). Initial sequencing and analysis of the human genome. Nature ***409***, 860-921.

Cooper, D. N., and Krawczak, M. (1991). Mechanisms of insertional mutagenesis in human genes causing genetic disease. Hum Genet ***87***, 409-15.

Copeland, N. G., and Jenkins, N. A. (1991). Development and applications of a molecular genetic linkage map of the mouse genome. Trends Genet *7*, 113-8.

Cossee, M., Campuzano, V., Koutnikova, H., Fischbeck, K., Mandel, J. L., Koenig, M., Bidichandani, S. I., Patel, P. I., Molte, M. D., Canizares, J., De Frutos, R., Pianese, L., Cavalcanti, F., Monticelli, A., Cocozza, S., Montermini, L., and Pandolfo, M. (1997). Frataxin fracas. Nat Genet ***15***, 337-8.

Cossee, M., Puccio, H., Gansmuller, A., Koutnikova, H., Dierich, A., LeMeur, M., Fischbeck, K., Dolle, P., and Koenig, M. (2000). Inactivation of the Friedreich ataxia mouse gene leads to early embryonic lethality without iron accumulation. Hum Mol Genet ***9***, 1219-26.

Cox, R., and Mirkin, S. M. (1997). Characteristic enrichment of DNA repeats in different genomes. Proc Natl Acad Sci U S A ***94***, 5237-42.

Craig, J. M., and Bickmore, W. A. (1994). The distribution of CpG islands in mammalian chromosomes. Nat Genet *7*, 376-82.

Cross, S. H., Charlton, J. A., Nan, X., and Bird, A. P. (1994). Purification of CpG islands using a methylated DNA binding column. Nat Genet ***6***, 236-44.

Cross, S. H., Lee, M., Clark, V. H., Craig, J. M., Bird, A. P., and Bickmore, W. A. (1997). The chromosomal distribution of CpG islands in the mouse: evidence for genome scrambling in the rodent lineage. Genomics ***40***, 454-61.

Darlow, J. M., and Leach, D. R. (1995). The effects of trinucleotide repeats found in human inherited disorders on palindrome inviability in Escherichia coli suggest hairpin folding preferences in vivo. Genetics ***141***, 825-32.

Darlow, J. M., and Leach, D. R. (1998a). Evidence for two preferred hairpin folding patterns in d(CGG).d(CCG) repeat tracts in vivo. J Mol Biol ***275***, 17-23.

Darlow, J. M., and Leach, D. R. (1998b). Secondary structures in d(CGG) and d(CCG) repeat tracts. J Mol Biol ***275***, 3-16.

David, G., Abbas, N., Stevanin, G., Durr, A., Yvert, G., Cancel, G., Weber, C., Imbert, G., Saudou, F., Antoniou, E., Drabkin, H., Gemmill, R., Giunti, P., Benomar, A., Wood, N., Ruberg, M., Agid, Y., Mandel, J. L., and Brice, A. (1997). Cloning of the SCA7 gene reveals a highly unstable CAG repeat expansion. Nat Genet ***17***, 65-70.

Davies, S. W., Turmaine, M., Cozens, B. A., DiFiglia, M., Sharp, A. H., Ross, C. A., Scherzinger, E., Wanker, E. E., Mangiarini, L., and Bates, G. P. (1997). Formation of neuronal intranuclear inclusions underlies the neurological dysfunction in mice transgenic for the HD mutation. Cell ***90***, 537-48.

Davis, B. M., McCurrach, M. E., Taneja, K. L., Singer, R. H., and Housman, D. E. (1997). Expansion of a CUG trinucleotide repeat in the 3' untranslated region of myotonic dystrophy protein kinase transcripts results in nuclear retention of transcripts. Proc Natl Acad Sci U S A ***94***, 7388-93.

Day, J. W., Schut, L. J., Moseley, M. L., Durand, A. C., and Ranum, L. P. (2000). Spinocerebellar ataxia type 8: clinical features in a large family. Neurology ***55***, 649-57.

de la Monte, S. M., Vonsattel, J. P., and Richardson, E. P. (1988). Morphometric demonstration of atrophic changes in the cerebral cortex, white matter, and neostriatum in Huntington's disease. J Neuropathol Exp Neurol *47*, 516-25.

De Rooij, K. E., Dorsman, J. C., Smoor, M. A., Den Dunnen, J. T., and Van Ommen, G. J. (1996). Subcellular localization of the Huntington's disease gene product in cell lines by immunofluorescence and biochemical subcellular fractionation. Hum Mol Genet *5*, 1093-9.

Delgado, S., Gomez, M., Bird, A., and Antequera, F. (1998). Initiation of DNA replication at CpG islands in mammalian chromosomes. Embo J *17*, 2426-35.

Delot, E., King, L. M., Briggs, M. D., Wilcox, W. R., and Cohn, D. H. (1999). Trinucleotide expansion mutations in the cartilage oligomeric matrix protein (COMP) gene. Hum Mol Genet *8*, 123-8.

Dietrich, W., Katz, H., Lincoln, S. E., Shin, H. S., Friedman, J., Dracopoli, N. C., and Lander, E. S. (1992). A genetic map of the mouse suitable for typing intraspecific crosses. Genetics *131*, 423-47.

Doyu, M., Sobue, G., Mukai, E., Kachi, T., Yasuda, T., Mitsuma, T., and Takahashi, A. (1992). Severity of X-linked recessive bulbospinal neuronopathy correlates with size of the tandem CAG repeat in androgen receptor gene. Ann Neurol *32*, 707-10.

Dragatsis, I., Levine, M. S., and Zeitlin, S. (2000). Inactivation of Hdh in the brain and testis results in progressive neurodegeneration and sterility in mice. Nat Genet *26*, 300-6.

Dure, L. S., Landwehrmeyer, G. B., Golden, J., McNeil, S. M., Ge, P., Aizawa, H., Huang, Q., Ambrose, C. M., Duyao, M. P., Bird, E. D., and et al. (1994). IT15 gene expression in fetal human brain. Brain Res *659*, 33-41.

Duyao, M. P., Auerbach, A. B., Ryan, A., Persichetti, F., Barnes, G. T., McNeil, S. M., Ge, P., Vonsattel, J. P., Gusella, J. F., Joyner, A. L., and et al. (1995). Inactivation of the mouse Huntington's disease gene homolog Hdh. Science *269*, 407-10.

Ellerby, L. M., Andrusiak, R. L., Wellington, C. L., Hackam, A. S., Propp, S. S., Wood, J. D., Sharp, A. H., Margolis, R. L., Ross, C. A., Salvesen, G. S., Hayden, M. R., and Bredesen, D. E. (1999a). Cleavage of atrophin-1 at caspase site aspartic acid 109 modulates cytotoxicity. J Biol Chem *274*, 8730-6.

Ellerby, L. M., Hackam, A. S., Propp, S. S., Ellerby, H. M., Rabizadeh, S., Cashman, N. R., Trifiro, M. A., Pinsky, L., Wellington, C. L., Salvesen, G. S., Hayden, M. R., and Bredesen, D. E. (1999b). Kennedy's disease: caspase cleavage of the androgen receptor is a crucial event in cytotoxicity. J Neurochem *72*, 185-95.

Eriksson, M., Ansved, T., Edstrom, L., Anvret, M., and Carey, N. (1999). Simultaneous analysis of expression of the three myotonic dystrophy locus genes in adult skeletal muscle samples: the CTG expansion correlates inversely with DMPK and 59 expression levels, but not DMAHP levels. Hum Mol Genet *8*, 1053-60.

Evert, B. O., Wullner, U., Schulz, J. B., Weller, M., Groscurth, P., Trottier, Y., Brice, A., and Klockgether, T. (1999). High level expression of expanded full-length ataxin-3 in vitro causes cell death and formation of intranuclear inclusions in neuronal cells. Hum Mol Genet *8*, 1169-76.

Faber, P. W., Barnes, G. T., Srinidhi, J., Chen, J., Gusella, J. F., and MacDonald, M. E. (1998). Huntingtin interacts with a family of WW domain proteins. Hum Mol Genet *7*, 1463-74.

Falconer, D. S., and Snell, G. D. (1952). Two new hair mutants, rough and frizzy. J. Heredity *43*, 53-57.

Faust, C. J., Verkerk, A. J., Wilson, P. J., Morris, C. P., Hopwood, J. J., Oostra, B. A., and Herman, G. E. (1992). Genetic mapping on the mouse X chromosome of human cDNA clones for the fragile X and Hunter syndromes. Genomics *12*, 814-7.

Feinberg, A. P., and Vogelstein, B. (1983). A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal Biochem *132*, 6-13.

Feng, Y., Lakkis, L., Devys, D., and Warren, S. T. (1995). Quantitative comparison of FMR1 gene expression in normal and premutation alleles. Am J Hum Genet *56*, 106-13.

Feng, Y., Gutekunst, C. A., Eberhart, D. E., Yi, H., Warren, S. T., and Hersch, S. M. (1997). Fragile X mental retardation protein: nucleocytoplasmic shuttling and association with somatodendritic ribosomes. J Neurosci *17*, 1539-47.

Fortune, M. T., Vassilopoulos, C., Coolbaugh, M. I., Siciliano, M. J., and Monckton, D. G. (2000). Dramatic, expansion-biased, age-dependent, tissue-specific somatic mosaicism in a transgenic mouse model of triplet repeat instability. Hum Mol Genet *9*, 439-45.

Foury, F., and Cazzalini, O. (1997). Deletion of the yeast homologue of the human gene associated with Friedreich's ataxia elicits iron accumulation in mitochondria. FEBS Lett *411*, 373-7.

Foury, F. (1999). Low iron concentration and aconitase deficiency in a yeast frataxin homologue deficient strain. FEBS Lett *456*, 281-4.

Freudenreich, C. H., Stavenhagen, J. B., and Zakian, V. A. (1997). Stability of a CTG/CAG trinucleotide repeat in yeast is dependent on its orientation in the genome. Mol Cell Biol *17*, 2090-8.

Fu, Y. H., Kuhl, D. P., Pizzuti, A., Pieretti, M., Sutcliffe, J. S., Richards, S., Verkerk, A. J., Holden, J. J., Fenwick, R. G., Warren, S. T., and et al. (1991). Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. Cell *67*, 1047-58.

Fu, Y. H., Pizzuti, A., Fenwick, R. G., King, J., Rajnarayan, S., Dunne, P. W., Dubel, J., Nasser, G. A., Ashizawa, T., de Jong, P., and et al. (1992). An unstable triplet repeat in a gene related to myotonic muscular dystrophy. Science *255*, 1256-8.

Fu, Y. H., Friedman, D. L., Richards, S., Pearlman, J. A., Gibbs, R. A., Pizzuti, A., Ashizawa, T., Perryman, M. B., Scarlato, G., Fenwick, R. G., and et al. (1993). Decreased expression of myotonin-protein kinase messenger RNA and protein in adult form of myotonic dystrophy. Science *260*, 235-8.

Gacy, A. M., Goellner, G., Juranic, N., Macura, S., and McMurray, C. T. (1995). Trinucleotide repeats that expand in human disease form hairpin structures in vitro. Cell *81*, 533-40.

Gacy, A. M., Goellner, G. M., Spiro, C., Chen, X., Gupta, G., Bradbury, E. M., Dyer, R. B., Mikesell, M. J., Yao, J. Z., Johnson, A. J., Richter, A., Melancon, S. B., and McMurray, C. T. (1998). GAA instability in Friedreich's Ataxia shares a common, DNA-directed and intraallelic mechanism with other trinucleotide diseases. Mol Cell *1*, 583-93.

Gajdusek, D. C. (1991). The transmissible amyloidoses: genetical control of spontaneous generation of infectious amyloid proteins by nucleation of configurational change in host precursors: kuru-CJD-GSS-scrapie-BSE. Eur J Epidemiol *7*, 567-77.

Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. J Mol Biol ***196***, 261-82.

Garza, J. C., and Freimer, N. B. (1996). Homoplasy for size at microsatellite loci in humans and chimpanzees. Genome Res *6*, 211-7.

Gecz, J., Gedeon, A. K., Sutherland, G. R., and Mulley, J. C. (1996). Identification of the gene FMR2, associated with FRAXE mental retardation. Nat Genet ***13***, 105-8.

Gecz, J., Bielby, S., Sutherland, G. R., and Mulley, J. C. (1997). Gene structure and subcellular localization of FMR2, a member of a new family of putative transcription activators. Genomics ***44***, 201-13.

Gecz, J. (2000). FMR3 is a novel gene associated with FRAXE CpG island and transcriptionally silent in FRAXE full mutations. J Med Genet *37*, 782-4.

Gerber, H. P., Seipel, K., Georgiev, O., Hofferer, M., Hug, M., Rusconi, S., and Schaffner, W. (1994). Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science ***263***, 808-11.

Geschwind, D. H., Perlman, S., Figueroa, C. P., Treiman, L. J., and Pulst, S. M. (1997). The prevalence and wide clinical spectrum of the spinocerebellar ataxia type 2 trinucleotide repeat in patients with autosomal dominant cerebellar ataxia. Am J Hum Genet ***60***, 842-50.

Glaser, D., Wohrle, D., Salat, U., Vogel, W., and Steinbach, P. (1999). Mitotic behavior of expanded CGG repeats studied on cultured cells: further evidence for methylation-mediated triplet repeat stability in fragile X syndrome. Am J Med Genet ***84***, 226-8.

Godde, J. S., Kass, S. U., Hirst, M. C., and Wolffe, A. P. (1996). Nucleosome assembly on methylated CGG triplet repeats in the fragile X mental retardation gene 1 promoter. J Biol Chem *271*, 24325-8.

Goldberg, Y. P., Kalchman, M. A., Metzler, M., Nasir, J., Zeisler, J., Graham, R., Koide, H. B., O'Kusky, J., Sharp, A. H., Ross, C. A., Jirik, F., and Hayden, M. R. (1996a). Absence of disease phenotype and intergenerational stability of the CAG repeat in transgenic mice expressing the human Huntington disease transcript. Hum Mol Genet *5*, 177-85.

Goldberg, Y. P., Nicholson, D. W., Rasper, D. M., Kalchman, M. A., Koide, H. B., Graham, R. K., Bromm, M., Kazemi-Esfarjani, P., Thornberry, N. A., Vaillancourt, J. P., and Hayden, M. R. (1996b). Cleavage of huntingtin by apopain, a proapoptotic cysteine protease, is modulated by the polyglutamine tract. Nat Genet ***13***, 442-9.

Goodman, F. R., Mundlos, S., Muragaki, Y., Donnai, D., Giovannucci-Uzielli, M. L., Lapi, E., Majewski, F., McGaughran, J., McKeown, C., Reardon, W., Upton, J., Winter, R. M., Olsen, B. R., and Scambler, P. J. (1997). Synpolydactyly phenotypes correlate with size of expansions in HOXD13 polyalanine tract. Proc Natl Acad Sci U S A ***94***, 7458-63.

Gordenin, D. A., Kunkel, T. A., and Resnick, M. A. (1997). Repeat expansion--all in a flap? Nat Genet ***16***, 116-8.

Gostout, B., Liu, Q., and Sommer, S. S. (1993). Cryptic repeating triplets of purines and pyrimidines (cRRY) are frequent and polymorphic: analysis of coding cRRY in the proopiomelanocortin (POMC) and TATA-binding protein (TBP) genes. Am J Hum Genet ***52***, 1182-90.

Gourdon, G., Dessen, P., Lia, A. S., Junien, C., and Hofmann-Radvanyi, H. (1997a). Intriguing association between disease associated unstable trinucleotide repeat and CpG island. Ann Genet ***40***, 73-7.

Gourdon, G., Radvanyi, F., Lia, A. S., Duros, C., Blanche, M., Abitbol, M., Junien, C., and Hofmann-Radvanyi, H. (1997b). Moderate intergenerational and somatic instability of a 55-CTG repeat in transgenic mice. Nat Genet ***15***, 190-2.

Group, The Huntington's Disease Collaborative Research. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell ***72***, 971-83.

Gu, Y., Shen, Y., Gibbs, R. A., and Nelson, D. L. (1996). Identification of FMR2, a novel gene associated with the FRAXE CCG repeat and CpG island. Nat Genet ***13***, 109-13.

Guenet, J. L., Nagamine, C., Simon-Chazottes, D., Montagutelli, X., and Bonhomme, F. (1990). Hst-3: an X-linked hybrid sterility gene. Genet Res ***56***, 163-5.

Haaf, T., Sirugo, G., Kidd, K. K., and Ward, D. C. (1996). Chromosomal localization of long trinucleotide repeats in the human genome by fluorescence in situ hybridization. Nat Genet ***12***, 183-5.

Haldane, J. B. S. (1922). Sex ratio and unisexual sterility in hybrid animals. J Genet ***12***, 101-109.

Hancock, J. M. (1995). The contribution of slippage-like processes to genome evolution. J Mol Evol ***41***, 1038-47.

Hansen, R. S., Canfield, T. K., Lamb, M. M., Gartler, S. M., and Laird, C. D. (1993). Association of fragile X syndrome with delayed replication of the FMR1 gene. Cell ***73***, 1403-9.

Hansen, R. S., Canfield, T. K., Fjeld, A. D., Mumm, S., Laird, C. D., and Gartler, S. M. (1997). A variable domain of delayed replication in FRAXA fragile X chromosomes: X inactivation-like spread of late replication. Proc Natl Acad Sci U S A ***94***, 4587-92.

Harding, A. E., Thomas, P. K., Baraitser, M., Bradbury, P. G., Morgan-Hughes, J. A., and Ponsford, J. R. (1982). X-linked recessive bulbospinal neuronopathy: a report of ten cases. J Neurol Neurosurg Psychiatry ***45***, 1012-9.

Harley, H. G., Rundle, S. A., MacMillan, J. C., Myring, J., Brook, J. D., Crow, S., Reardon, W., Fenton, I., Shaw, D. J., and Harper, P. S. (1993). Size of the unstable CTG repeat sequence in relation to phenotype and parental transmission in myotonic dystrophy. Am J Hum Genet ***52***, 1164-74.

Harris, S., Moncrieff, C., and Johnson, K. (1996). Myotonic dystrophy: will the real gene please step forward! Hum Mol Genet ***5***, 1417-23.

Hayashi, Y., Kakita, A., Yamada, M., Koide, R., Igarashi, S., Takano, H., Ikeuchi, T., Wakabayashi, K., Egawa, S., Tsuji, S., and Takahashi, H. (1998). Hereditary dentatorubral-pallidoluysian atrophy: detection of widespread ubiquitinated neuronal and glial intranuclear inclusions in the brain. Acta Neuropathol (Berl) ***96***, 547-52.

Henke, W., Herdel, K., Jung, K., Schnorr, D., and Loening, S. A. (1997). Betaine improves the PCR amplification of GC-rich DNA sequences. Nucleic Acids Res ***25***, 3957-8.

Henricksen, L. A., Tom, S., Liu, Y., and Bambara, R. A. (2000). Inhibition of flap endonuclease 1 by flap secondary structure and relevance to repeat sequence expansion. J Biol Chem ***275***, 16420-7.

Hergersberg, M., Matsuo, K., Gassmann, M., Schaffner, W., Luscher, B., Rulicke, T., and Aguzzi, A. (1995). Tissue-specific expression of a FMR1/beta-galactosidase fusion gene in transgenic mice. Hum Mol Genet ***4***, 359-66.

Hirst, M. C., Barnicoat, A., Flynn, G., Wang, Q., Daker, M., Buckle, V. J., Davies, K. E., and Bobrow, M. (1993). The identification of a third fragile site, FRAXF, in Xq27--q28 distal to both FRAXA and FRAXE. Hum Mol Genet ***2***, 197-200.

Hirst, M. C., and White, P. J. (1998). Cloned human FMR1 trinucleotide repeats exhibit a length- and orientation-dependent instability suggestive of in vivo lagging strand secondary structure. Nucleic Acids Res ***26***, 2353-8.

Hodgson, J. G., Smith, D. J., McCutcheon, K., Koide, H. B., Nishiyama, K., Dinulos, M. B., Stevens, M. E., Bissada, N., Nasir, J., Kanazawa, I., Disteche, C. M., Rubin, E. M., and Hayden, M. R. (1996). Human huntingtin derived from YAC transgenes compensates for loss of murine huntingtin by rescue of the embryonic lethal phenotype. Hum Mol Genet ***5***, 1875-85.

Holmes, S. E., O'Hearn, E. E., McInnis, M. G., Gorelick-Feldman, D. A., Kleiderlein, J. J., Callahan, C., Kwak, N. G., Ingersoll-Ashworth, R. G., Sherr, M., Sumner, A. J., Sharp, A. H., Ananth, U., Seltzer, W. K., Boss, M. A., Vieria-Saecker, A. M., Epplen, J. T., Riess, O., Ross, C. A., and Margolis, R. L. (1999). Expansion of a novel CAG trinucleotide repeat in the 5' region of PPP2R2B is associated with SCA12. Nat Genet ***23***, 391-2.

Hoogeveen, A. T., Willemsen, R., Meyer, N., de Rooij, K. E., Roos, R. A., van Ommen, G. J., and Galjaard, H. (1993). Characterization and localization of the Huntington disease gene product. Hum Mol Genet ***2***, 2069-73.

Huang, X., and Harlan, R. E. (1994). Androgen receptor immunoreactivity in somatostatin neurons of the periventricular nucleus but not in the bed nucleus of the stria terminalis in male rats. Brain Res ***652***, 291-6.

Huynh, D. P., Del Bigio, M. R., Ho, D. H., and Pulst, S. M. (1999). Expression of ataxin-2 in brains from normal individuals and patients with Alzheimer's disease and spinocerebellar ataxia 2. Ann Neurol ***45***, 232-41.

Huynh, D. P., Figueroa, K., Hoang, N., and Pulst, S. M. (2000). Nuclear localization or inclusion body formation of ataxin-2 are not necessary for SCA2 pathogenesis in mouse or human. Nat Genet ***26***, 44-50.

Ikegawa, S., Ohashi, H., Nishimura, G., Kim, K. C., Sannohe, A., Kimizuka, M., Fukushima, Y., Nagai, T., and Nakamura, Y. (1998). Novel and recurrent COMP (cartilage oligomeric matrix protein) mutations in pseudoachondroplasia and multiple epiphyseal dysplasia. Hum Genet ***103***, 633-8.

Ikeuchi, T., Sanpei, K., Takano, H., Sasaki, H., Tashiro, K., Cancel, G., Brice, A., Bird, T. D., Schellenberg, G. D., Pericak-Vance, M. A., Welsh-Bohmer, K. A., Clark, L. N., Wilhelmsen, K., and Tsuji, S. (1998). A novel long and unstable CAG/CTG trinucleotide repeat on chromosome 17q. Genomics ***49***, 321-6.

Imbert, G., Trottier, Y., Beckmann, J., and Mandel, J. L. (1994). The gene for the TATA binding protein (TBP) that contains a highly polymorphic protein coding CAG repeat maps to 6q27. Genomics ***21***, 667-8.

Imbert, G., Saudou, F., Yvert, G., Devys, D., Trottier, Y., Garnier, J. M., Weber, C., Mandel, J. L., Cancel, G., Abbas, N., Durr, A., Didierjean, O., Stevanin, G., Agid, Y., and Brice, A. (1996). Cloning of the gene for spinocerebellar ataxia 2 reveals a locus with high sensitivity to expanded CAG/glutamine repeats [see comments]. Nat Genet ***14***, 285-91.

Ishikawa, K., Fujigasaki, H., Saegusa, H., Ohwada, K., Fujita, T., Iwamoto, H., Komatsuzaki, Y., Toru, S., Toriyama, H., Watanabe, M., Ohkoshi, N., Shoji, S., Kanazawa, I., Tanabe, T., and Mizusawa, H. (1999). Abundant expression and cytoplasmic aggregations of [alpha]1A voltage-dependent calcium channel protein associated with neurodegeneration in spinocerebellar ataxia type 6. Hum Mol Genet ***8***, 1185-93.

Iyer, R. R., Pluciennik, A., Rosche, W. A., Sinden, R. R., and Wells, R. D. (2000). DNA polymerase III proofreading mutants enhance the expansion and deletion of triplet repeat sequences in Escherichia coli. J Biol Chem ***275***, 2174-84.

Jackson, G. R., Salecker, I., Dong, X., Yao, X., Arnheim, N., Faber, P. W., MacDonald, M. E., and Zipursky, S. L. (1998). Polyglutamine-expanded human huntingtin transgenes induce degeneration of Drosophila photoreceptor neurons. Neuron ***21***, 633-42.

Jackson, I. J., and Abbott, C. M. (2000). Mouse Genetics and Transgenics, B. D. Hames, ed. (Oxford: Oxford University Press).

Jacobsen, P., Hauge, M., Henningsen, K., Hobolth, N., Mikkelsen, M., and Philip, J. (1973). An (11;21) translocation in four generations with chromosome 11 abnormalities in the offspring. A clinical, cytogenetical, and gene marker study. Hum Hered ***23***, 568-85.

Jankowski, C., Nasar, F., and Nag, D. K. (2000). Meiotic instability of CAG repeat tracts occurs by double-strand break repair in yeast. Proc Natl Acad Sci U S A ***97***, 2134-9.

Jansen, G.,.Willems, P., Coerwinkel, M., Nillesen, W., Smeets, H., Vits, L., Howeler, C., Brunner, H., and Wieringa, B. (1994). Gonosomal mosaicism in myotonic dystrophy patients: involvement of mitotic events in (CTG)n repeat variation and selection against extreme expansion in sperm. Am J Hum Genet ***54***, 575-85.

Jansen, G., Groenen, P. J., Bachner, D., Jap, P. H., Coerwinkel, M., Oerlemans, F., van den Broek, W., Gohlsch, B., Pette, D., Plomp, J. J., Molenaar, P. C., Nederhoff, M. G., van Echteld, C. J., Dekker, M., Berns, A., Hameister, H., and Wieringa, B. (1996). Abnormal myotonic dystrophy protein kinase levels produce only mild myopathy in mice. Nat Genet ***13***, 316-24.

Jaworski, A., Rosche, W. A., Gellibolian, R., Kang, S., Shimizu, M., Bowater, R. P., Sinden, R. R., and Wells, R. D. (1995). Mismatch repair in Escherichia coli enhances instability of (CTG)n triplet repeats from human hereditary diseases. Proc Natl Acad Sci U S A ***92***, 11019-23.

Jenster, G., van der Korput, H. A., van Vroonhoven, C., van der Kwast, T. H., Trapman, J., and Brinkmann, A. O. (1991). Domains of the human androgen receptor involved in steroid binding, transcriptional activation, and subcellular localization. Mol Endocrinol ***5***, 1396-404.

Ji, J., Clegg, N. J., Peterson, K. R., Jackson, A. L., Laird, C. D., and Loeb, L. A. (1996). In vitro expansion of GGC:GCC repeats: identification of the preferred strand of expansion. Nucleic Acids Res ***24***, 2835-40.

Jiang, J. X., Deprez, R. H., Zwarthoff, E. C., and Riegman, P. H. (1995). Characterization of four novel CAG repeat-containing cDNAs. Genomics ***30***, 91-3.

Jodice, C., Malaspina, P., Persichetti, F., Novelletto, A., Spadaro, M., Giunti, P., Morocutti, C., Terrenato, L., Harding, A. E., and Frontali, M. (1994). Effect of trinucleotide repeat length and parental sex on phenotypic variation in spinocerebellar ataxia I. Am J Hum Genet ***54***, 959-65.

Johnson, K. R., Sweet, H. O., Donahue, L. R., Ward-Bailey, P., Bronson, R. T., and Davisson, M. T. (1998). A new spontaneous mouse mutation of Hoxd13 with a polyalanine expansion and phenotype similar to human synpolydactyly. Hum Mol Genet *7*, 1033-8.

Jones, C., Penny, L., Mattina, T., Yu, S., Baker, E., Voullaire, L., Langdon, W. Y., Sutherland, G. R., Richards, R. I., and Tunnacliffe, A. (1995). Association of a chromosome deletion syndrome with a fragile site within the proto-oncogene CBL2. Nature ***376***, 145-9.

Kahlem, P., Green, H., and Djian, P. (1998). Transglutaminase action imitates Huntington's disease: selective polymerization of Huntingtin containing expanded polyglutamine. Mol Cell ***1***, 595-601.

Kalchman, M. A., Graham, R. K., Xia, G., Koide, H. B., Hodgson, J. G., Graham, K. C., Goldberg, Y. P., Gietz, R. D., Pickart, C. M., and Hayden, M. R. (1996). Huntingtin is ubiquitinated and interacts with a specific ubiquitin-conjugating enzyme. J Biol Chem ***271***, 19385-94.

Kalchman, M. A., Koide, H. B., McCutcheon, K., Graham, R. K., Nichol, K., Nishiyama, K., Kazemi-Esfarjani, P., Lynn, F. C., Wellington, C., Metzler, M., Goldberg, Y. P., Kanazawa, I., Gietz, R. D., and Hayden, M. R. (1997). HIP1, a human homologue of S. cerevisiae Sla2p, interacts with membrane-associated huntingtin in the brain. Nat Genet ***16***, 44-53.

Kang, S., Jaworski, A., Ohshima, K., and Wells, R. D. (1995). Expansion and deletion of CTG repeats from human disease genes are determined by the direction of replication in E. coli. Nat Genet ***10***, 213-8.

Kawaguchi, Y., Okamoto, T., Taniwaki, M., Aizawa, M., Inoue, M., Katayama, S., Kawakami, H., Nakamura, S., Nishimura, M., Akiguchi, I., and et al. (1994). CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nat Genet ***8***, 221-8.

Kawakami, H., Maruyama, H., Nakamura, S., Kawaguchi, Y., Kakizuka, A., Doyu, M., and Sobue, G. (1995). Unique features of the CAG repeats in Machado-Joseph disease. Nat Genet ***9***, 344-5.

Kaytor, M. D., Burright, E. N., Duvick, L. A., Zoghbi, H. Y., and Orr, H. T. (1997). Increased trinucleotide repeat instability with advanced maternal age. Hum Mol Genet ***6***, 2135-9.

Kaytor, M. D., Duvick, L. A., Skinner, P. J., Koob, M. D., Ranum, L. P., and Orr, H. T. (1999). Nuclear localization of the spinocerebellar ataxia type 7 protein, ataxin-7. Hum Mol Genet ***8***, 1657-64.

Kazantsev, A., Preisinger, E., Dranovsky, A., Goldgaber, D., and Housman, D. (1999). Insoluble detergent-resistant aggregates form between pathological and nonpathological lengths of polyglutamine in mammalian cells. Proc Natl Acad Sci U S A ***96***, 11404-9.

Kennedy, L., and Shelbourne, P. F. (2000). Dramatic mutation instability in HD mouse striatum: does polyglutamine load contribute to cell-specific vulnerability in Huntington's disease? Hum Mol Genet ***9***, 2539-44.

Kim, S. J., Shon, B. H., Kang, J. H., Hahm, K. S., Yoo, O. J., Park, Y. S., and Lee, K. K. (1997). Cloning of novel trinucleotide-repeat (CAG) containing genes in mouse brain. Biochem Biophys Res Commun ***240***, 239-43.

King, B. L., Sirugo, G., Nadeau, J. H., Hudson, T. J., Kidd, K. K., Kacinski, B. M., and Schalling, M. (1998). Long CAG/CTG repeats in mice. Mamm Genome ***9***, 392-3.

Klement, I. A., Skinner, P. J., Kaytor, M. D., Yi, H., Hersch, S. M., Clark, H. B., Zoghbi, H. Y., and Orr, H. T. (1998). Ataxin-1 nuclear localization and aggregation: role in polyglutamine-induced disease in SCA1 transgenic mice. Cell ***95***, 41-53.

Klesert, T. R., Otten, A. D., Bird, T. D., and Tapscott, S. J. (1997). Trinucleotide repeat expansion at the myotonic dystrophy locus reduces expression of DMAHP. Nat Genet ***16***, 402-6.

Klesert, T. R., Cho, D. H., Clark, J. I., Maylie, J., Adelman, J., Snider, L., Yuen, E. C., Soriano, P., and Tapscott, S. J. (2000). Mice deficient in Six5 develop cataracts: implications for myotonic dystrophy. Nat Genet ***25***, 105-9.

Knight, S. J., Voelckel, M. A., Hirst, M. C., Flannery, A. V., Moncla, A., and Davies, K. E. (1994). Triplet repeat expansion at the FRAXE locus and X-linked mild mental handicap. Am J Hum Genet ***55***, 81-6.

Koide, R., Ikeuchi, T., Onodera, O., Tanaka, H., Igarashi, S., Endo, K., Takahashi, H., Kondo, R., Ishikawa, A., Hayashi, T., and et al. (1994). Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). Nat Genet ***6***, 9-13.

Komure, O., Sano, A., Nishino, N., Yamauchi, N., Ueno, S., Kondoh, K., Sano, N., Takahashi, M., Murayama, N., Kondo, I., and et al. (1995). DNA analysis in hereditary dentatorubral-pallidoluysian atrophy: correlation between CAG repeat length and phenotypic variation and the molecular basis of anticipation. Neurology ***45***, 143-9.

Koob, M. D., Benzow, K. A., Bird, T. D., Day, J. W., Moseley, M. L., and Ranum, L. P. (1998). Rapid cloning of expanded trinucleotide repeat sequences from genomic DNA. Nat Genet ***18***, 72-5.

Koob, M. D., Moseley, M. L., Schut, L. J., Benzow, K. A., Bird, T. D., Day, J. W., and Ranum, L. P. (1999). An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). Nat Genet ***21***, 379-84.

Kooy, R. F., Reyniers, E., Verhoye, M., Sijbers, J., Bakker, C. E., Oostra, B. A., Willems, P. J., and Van Der Linden, A. (1999). Neuroanatomy of the fragile X knockout mouse brain studied using in vivo high resolution magnetic resonance imaging. Eur J Hum Genet *7*, 526-32.

Korade-Mirnics, Z., Tarleton, J., Servidei, S., Casey, R. R., Gennarelli, M., Pegoraro, E., Angelini, C., and Hoffman, E. P. (1999). Myotonic dystrophy: tissue-specific effect of somatic CTG expansions on allele-specific DMAHP/SIX5 expression. Hum Mol Genet ***8***, 1017-23.

Koshy, B., Matilla, T., Burright, E. N., Merry, D. E., Fischbeck, K. H., Orr, H. T., and Zoghbi, H. Y. (1996). Spinocerebellar ataxia type-1 and spinobulbar muscular atrophy gene products interact with glyceraldehyde-3-phosphate dehydrogenase. Hum Mol Genet *5*, 1311-8.

Koutnikova, H., Campuzano, V., Foury, F., Dolle, P., Cazzalini, O., and Koenig, M. (1997). Studies of human, mouse and yeast homologues indicate a mitochondrial function for frataxin. Nat Genet ***16***, 345-51.

Kovtun, I. V., Therneau, T. M., and McMurray, C. T. (2000). Gender of the embryo contributes to CAG instability in transgenic mice containing a Huntington's disease gene [In Process Citation]. Hum Mol Genet ***9***, 2767-75.

Koyano, S., Uchihara, T., Fujigasaki, H., Nakamura, A., Yagishita, S., and Iwabuchi, K. (1999). Neuronal intranuclear inclusions in spinocerebellar ataxia type 2: triple-labeling immunofluorescent study. Neurosci Lett ***273***, 117-20.

Krahe, R., Ashizawa, T., Abbruzzese, C., Roeder, E., Carango, P., Giacanelli, M., Funanage, V. L., and Siciliano, M. J. (1995). Effect of myotonic dystrophy trinucleotide repeat expansion on DMPK transcription and processing. Genomics ***28***, 1-14.

Kremer, E. J., Pritchard, M., Lynch, M., Yu, S., Holman, K., Baker, E., Warren, S. T., Schlessinger, D., Sutherland, G. R., and Richards, R. I. (1991). Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)n. Science ***252***, 1711-4.

La Spada, A. R., Wilson, E. M., Lubahn, D. B., Harding, A. E., and Fischbeck, K. H. (1991). Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature ***352***, 77-9.

La Spada, A. R., Peterson, K. R., Meadows, S. A., McClain, M. E., Jeng, G., Chmelar, R. S., Haugen, H. A., Chen, K., Singer, M. J., Moore, D., Trask, B. J., Fischbeck, K. H., Clegg, C. H., and McKnight, G. S. (1998). Androgen receptor YAC transgenic mice carrying CAG 45 alleles show trinucleotide repeat instability. Hum Mol Genet *7*, 959-67.

Lafreniere, R. G., Rochefort, D. L., Chretien, N., Rommens, J. M., Cochius, J. I., Kalviainen, R., Nousiainen, U., Patry, G., Farrell, K., Soderfeldt, B., Federico, A., Hale, B. R., Cossio, O. H., Sorensen, T., Pouliot, M. A., Kmiec, T., Uldall, P., Janszky, J., Pranzatelli, M. R., Andermann, F., Andermann, E., and Rouleau, G. A. (1997). Unstable insertion in the 5' flanking region of the cystatin B gene is the most common mutation in progressive myoclonus epilepsy type 1, EPM1. Nat Genet ***15***, 298-302.

Lalioti, M. D., Scott, H. S., Buresi, C., Rossier, C., Bottani, A., Morris, M. A., Malafosse, A., and Antonarakis, S. E. (1997). Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. Nature ***386***, 847-51.

Lamarche, J. B., Shapcott, D., Cote, M., and Lemieux, B. (1993). Handbook of Cerebellar Diseases, R. Lechtenberg, ed. (New York: Marcel Dekker), pp. 453-458.

Laval, S. H., Blair, H. J., Hirst, M. C., Davies, K. E., and Boyd, Y. (1992). Mapping of FMR1, the gene implicated in fragile X-linked mental retardation, on the mouse X chromosome. Genomics ***12***, 818-21.

Lavedan, C., Hofmann-Radvanyi, H., Shelbourne, P., Rabes, J. P., Duros, C., Savoy, D., Dehaupas, I., Luce, S., Johnson, K., and Junien, C. (1993). Myotonic dystrophy: size- and sex-dependent dynamics of CTG meiotic instability, and somatic mosaicism. Am J Hum Genet ***52***, 875-83.

Lavedan, C. N., Garrett, L., and Nussbaum, R. L. (1997). Trinucleotide repeats (CGG)22TGG(CGG)43TGG(CGG)21 from the fragile X gene remain stable in transgenic mice. Hum Genet ***100***, 407-14.

Lavedan, C., Grabczyk, E., Usdin, K., and Nussbaum, R. L. (1998). Long uninterrupted CGG repeats within the first exon of the human FMR1 gene are not intrinsically unstable in transgenic mice. Genomics ***50***, 229-40.

Lescure, A., Lutz, Y., Eberhard, D., Jacq, X., Krol, A., Grummt, I., Davidson, I., Chambon, P., and Tora, L. (1994). The N-terminal domain of the human TATA-binding protein plays a role in transcription from TATA-containing RNA polymerase II and III promoters. Embo J ***13***, 1166-75.

Lewis, H. A., Musunuru, K., Jensen, K. B., Edo, C., Chen, H., Darnell, R. B., and Burley, S. K. (2000). Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome. Cell ***100***, 323-32.

Li, S. H., McInnis, M. G., Margolis, R. L., Antonarakis, S. E., and Ross, C. A. (1993). Novel triplet repeat containing genes in human brain: cloning, expression, and length polymorphisms. Genomics ***16***, 572-9.

Li, X. J., Li, S. H., Sharp, A. H., Nucifora, F. C., Schilling, G., Lanahan, A., Worley, P., Snyder, S. H., and Ross, C. A. (1995). A huntingtin-associated protein enriched in brain with implications for pathology. Nature ***378***, 398-402.

Lia, A. S., Seznec, H., Hofmann-Radvanyi, H., Radvanyi, F., Duros, C., Saquet, C., Blanche, M., Junien, C., and Gourdon, G. (1998). Somatic instability of the CTG repeat in mice transgenic for the myotonic dystrophy region is age dependent but not correlated to the relative intertissue transcription levels and proliferative capacities. Hum Mol Genet *7*, 1285-91.

Lindblad, K., Savontaus, M. L., Stevanin, G., Holmberg, M., Digre, K., Zander, C., Ehrsson, H., David, G., Benomar, A., Nikoskelainen, E., Trottier, Y., Holmgren, G., Ptacek, L. J., Anttinen, A., Brice, A., and Schalling, M. (1996). An expanded CAG repeat sequence in spinocerebellar ataxia type 7. Genome Res ***6***, 965-71.

Lione, L. A., Carter, R. J., Hunt, M. J., Bates, G. P., Morton, A. J., and Dunnett, S. B. (1999). Selective discrimination learning impairments in mice expressing the human Huntington's disease mutation. J Neurosci ***19***, 10428-37.

Liu, Y. F., Deth, R. C., and Devys, D. (1997). SH3 domain-dependent association of huntingtin with epidermal growth factor receptor signaling complexes. J Biol Chem *272*, 8121-4.

Lodi, R., Cooper, J. M., Bradley, J. L., Manners, D., Styles, P., Taylor, D. J., and Schapira, A. H. (1999). Deficit of in vivo mitochondrial ATP production in patients with Friedreich ataxia. Proc Natl Acad Sci U S A ***96***, 11492-5.

Lorenzetti, D., Watase, K., Xu, B., Matzuk, M. M., Orr, H. T., and Zoghbi, H. Y. (2000). Repeat instability and motor incoordination in mice with a targeted expanded CAG repeat in the Sca1 locus. Hum Mol Genet ***9***, 779-85.

Love, J. M., Knight, A. M., McAleer, M. A., and Todd, J. A. (1990). Towards construction of a high resolution map of the mouse genome using PCR-analysed microsatellites. Nucleic Acids Res ***18***, 4123-30.

Lubahn, D. B., Joseph, D. R., Sullivan, P. M., Willard, H. F., French, F. S., and Wilson, E. M. (1988). Cloning of human androgen receptor complementary DNA and localization to the X chromosome. Science ***240***, 327-30.

Lyko, F., Ramsahoye, B. H., and Jaenisch, R. (2000). DNA methylation in Drosophila melanogaster. Nature ***408***, 538-40.

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***, 2282-92.

Mahadevan, M. S., Amemiya, C., Jansen, G., Sabourin, L., Baird, S., Neville, C. E., Wormskamp, N., Segers, B., Batzer, M., Lamerdin, J., and et al. (1993). Structure and genomic sequence of the myotonic dystrophy (DM kinase) gene. Hum Mol Genet *2*, 299-304.

Mangiarini, L., Sathasivam, K., Seller, M., Cozens, B., Harper, A., Hetherington, C., Lawton, M., Trottier, Y., Lehrach, H., Davies, S. W., and Bates, G. P. (1996). Exon 1 of the HD gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. Cell ***87***, 493-506.

Mangiarini, L., Sathasivam, K., Mahal, A., Mott, R., Seller, M., and Bates, G. P. (1997). Instability of highly expanded CAG repeats in mice transgenic for the Huntington's disease mutation. Nat Genet ***15***, 197-200.

Mankodi, A., Logigian, E., Callahan, L., McClain, C., White, R., Henderson, D., Krym, M., and Thornton, C. A. (2000). Myotonic dystrophy in transgenic mice expressing an expanded CUG repeat. Science ***289***, 1769-73.

Manley, K., Shirley, T. L., Flaherty, L., and Messer, A. (1999). Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. Nat Genet ***23***, 471-3.

Manly, K. F. (1993). A Macintosh program for storage and analysis of experimental genetic mapping data. Mamm Genome ***4***, 303-13.

Margolis, R. L., Breschel, T. S., Li, S. H., Kidwai, A. S., Antonarakis, S. E., McInnis, M. G., and Ross, C. A. (1995a). Characterization of cDNA clones containing CCA trinucleotide repeats derived from human brain. Somat Cell Mol Genet ***21***, 279-84.

Margolis, R. L., Breschel, T. S., Li, S. H., Kidwai, A. S., McInnis, M. G., and Ross, C. A. (1995b). Polymorphic (AAT) in trinucleotide repeats derived from a human brain cDNA library. Hum Genet ***96***, 495-6.

Margolis, R. L., Abraham, M. R., Gatchell, S. B., Li, S. H., Kidwai, A. S., Breschel, T. S., Stine, O. C., Callahan, C., McInnis, M. G., and Ross, C. A. (1997). cDNAs with long CAG trinucleotide repeats from human brain. Hum Genet ***100***, 114-22.

Marsh, J. L., Walker, H., Theisen, H., Zhu, Y. Z., Fielder, T., Purcell, J., and Thompson, L. M. (2000). Expanded polyglutamine peptides alone are intrinsically cytotoxic and cause neurodegeneration in Drosophila. Hum Mol Genet ***9***, 13-25.

Matilla, A., Koshy, B. T., Cummings, C. J., Isobe, T., Orr, H. T., and Zoghbi, H. Y. (1997). The cerebellar leucine-rich acidic nuclear protein interacts with ataxin-1. Nature ***389***, 974-8.

Matsuo, K., Clay, O., Takahashi, T., Silke, J., and Schaffner, W. (1993). Evidence for erosion of mouse CpG islands during mammalian evolution. Somat Cell Mol Genet ***19***, 543-55.

Matsuura, T., Yamagata, T., Burgess, D. L., Rasmussen, A., Grewal, R. P., Watase, K., Khajavi, M., McCall, A. E., Davis, C. F., Zu, L., Achari, M., Pulst, S. M., Alonso, E., Noebels, J. L., Nelson, D. L., Zoghbi, H. Y., and Ashizawa, T. (2000). Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. Nat Genet ***26***, 191-4.

Matsuyama, Z., Wakamori, M., Mori, Y., Kawakami, H., Nakamura, S., and Imoto, K. (1999). Direct alteration of the P/Q-type Ca2+ channel property by polyglutamine expansion in spinocerebellar ataxia 6. J Neurosci ***19***, RC14.

Maurer, D. J., O'Callaghan, B. L., and Livingston, D. M. (1996). Orientation dependence of trinucleotide CAG repeat instability in Saccharomyces cerevisiae. Mol Cell Biol ***16***, 6617-22.

McInnis, M. G. (1996). Anticipation: an old idea in new genes. Am J Hum Genet ***59***, 973-9.

McMurray, C. T. (1995). Mechanisms of DNA expansion. Chromosoma ***104***, 2-13.

Metzgar, D., Bytof, J., and Wills, C. (2000). Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res ***10***, 72-80.

Miller, W. J., Skinner, J. A., Foss, G. S., and Davies, K. E. (2000). Localization of the fragile X mental retardation 2 (FMR2) protein in mammalian brain. Eur J Neurosci ***12***, 381-4.

Miret, J. J., Pessoa-Brandao, L., and Lahue, R. S. (1998). Orientation-dependent and sequence-specific expansions of CTG/CAG trinucleotide repeats in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A ***95***, 12438-43.

Mitas, M., Yu, A., Dill, J., Kamp, T. J., Chambers, E. J., and Haworth, I. S. (1995). Hairpin properties of single-stranded DNA containing a GC-rich triplet repeat: (CTG)15. Nucleic Acids Res ***23***, 1050-9.

Monckton, D. G., Coolbaugh, M. I., Ashizawa, K. T., Siciliano, M. J., and Caskey, C. T. (1997). Hypermutable myotonic dystrophy CTG repeats in transgenic mice. Nat Genet ***15***, 193-6.

Monros, E., Molto, M. D., Martinez, F., Canizares, J., Blanca, J., Vilchez, J. J., Prieto, F., de Frutos, R., and Palau, F. (1997). Phenotype correlation and intergenerational dynamics of the Friedreich ataxia GAA trinucleotide repeat. Am J Hum Genet ***61***, 101-10.

Moore, H., Greenwell, P. W., Liu, C. P., Arnheim, N., and Petes, T. D. (1999). Triplet repeats form secondary structures that escape DNA repair in yeast. Proc Natl Acad Sci U S A ***96***, 1504-9.

Moseley, M. L., Schut, L. J., Bird, T. D., Koob, M. D., Day, J. W., and Ranum, L. P. (2000). SCA8 CTG repeat: en masse contractions in sperm and intergenerational sequence changes may play a role in reduced penetrance. Hum Mol Genet ***9***, 2125-30.

Mounsey, J. P., Mistry, D. J., Ai, C. W., Reddy, S., and Moorman, J. R. (2000). Skeletal muscle sodium channel gating in mice deficient in myotonic dystrophy protein kinase. Hum Mol Genet ***9***, 2313-20.

Muragaki, Y., Mundlos, S., Upton, J., and Olsen, B. R. (1996). Altered growth and branching patterns in synpolydactyly caused by mutations in HOXD13. Science ***272***, 548-51.

Nagafuchi, S., Yanagisawa, H., Sato, K., Shirayama, T., Ohsaki, E., Bundo, M., Takeda, T., Tadokoro, K., Kondo, I., Murayama, N., and et al. (1994). Dentatorubral and pallidoluysian atrophy expansion of an unstable CAG trinucleotide on chromosome 12p. Nat Genet ***6***, 14-8.

Naito, H., and Oyanagi, S. (1982). Familial myoclonus epilepsy and choreoathetosis: hereditary dentatorubral-pallidoluysian atrophy. Neurology ***32***, 798-807.

Nakamoto, M., Takebayashi, H., Kawaguchi, Y., Narumiya, S., Taniwaki, M., Nakamura, Y., Ishikawa, Y., Akiguchi, I., Kimura, J., and Kakizuka, A. (1997). A CAG/CTG expansion in the normal population. Nat Genet *17*, 385-6.

Nancarrow, J. K., Kremer, E., Holman, K., Eyre, H., Doggett, N. A., Le Paslier, D., Callen, D. F., Sutherland, G. R., and Richards, R. I. (1994). Implications of FRA16A structure for the mechanism of chromosomal fragile site genesis. Science *264*, 1938-41.

Nasir, J., Lin, B., Bucan, M., Koizumi, T., Nadeau, J. H., and Hayden, M. R. (1994). The murine homologues of the Huntington disease gene (Hdh) and the alpha-adducin gene (Add1) map to mouse chromosome 5 within a region of conserved synteny with human chromosome 4p16.3. Genomics *22*, 198-201.

Nasir, J., Floresco, S. B., O'Kusky, J. R., Diewert, V. M., Richman, J. M., Zeisler, J., Borowski, A., Marth, J. D., Phillips, A. G., and Hayden, M. R. (1995). Targeted disruption of the Huntington's disease gene results in embryonic lethality and behavioral and morphological changes in heterozygotes. Cell *81*, 811-23.

Nechiporuk, T., Huynh, D. P., Figueroa, K., Sahba, S., Nechiporuk, A., and Pulst, S. M. (1998). The mouse SCA2 gene: cDNA sequence, alternative splicing and protein expression. Hum Mol Genet *7*, 1301-9.

Nemes, J. P., Benzow, K. A., and Koob, M. D. (2000). The SCA8 transcript is an antisense RNA to a brain-specific transcript encoding a novel actin-binding protein (KLHL1). Hum Mol Genet *9*, 1543-51.

Neri, C., Albanese, V., Lebre, A. S., Holbert, S., Saada, C., Bougueleret, L., Meier-Ewert, S., Le Gall, I., Millasseau, P., Bui, H., Giudicelli, C., Massart, C., Guillou, S., Gervy, P., Poullier, E., Rigault, P., Weissenbach, J., Lennon, G., Chumakov, I., Dausset, J., Lehrach, H., Cohen, D., and Cann, H. M. (1996). Survey of CAG/CTG repeats in human cDNAs representing new genes: candidates for inherited neurological disorders. Hum Mol Genet *5*, 1001-9.

Novelli, G., Gennarelli, M., Zelano, G., Pizzuti, A., Fattorini, C., Caskey, C. T., and Dallapiccola, B. (1993). Failure in detecting mRNA transcripts from the mutated allele in myotonic dystrophy muscle. Biochem Mol Biol Int *29*, 291-7.

Oberle, I., Rousseau, F., Heitz, D., Kretz, C., Devys, D., Hanauer, A., Boue, J., Bertheas, M. F., and Mandel, J. L. (1991). Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. Science *252*, 1097-102.

Ohshima, K., Kang, S., Larson, J. E., and Wells, R. D. (1996a). Cloning, characterization, and properties of seven triplet repeat DNA sequences. J Biol Chem *271*, 16773-83.

Ohshima, K., Kang, S., and Wells, R. D. (1996b). CTG triplet repeats from human hereditary diseases are dominant genetic expansion products in Escherichia coli. J Biol Chem *271*, 1853-6.

Ohshima, K., Montermini, L., Wells, R. D., and Pandolfo, M. (1998). Inhibitory effects of expanded GAA.TTC triplet repeats from intron I of the Friedreich ataxia gene on transcription and replication in vivo. J Biol Chem *273*, 14588-95.

Okamura-Oho, Y., Miyashita, T., Ohmi, K., and Yamada, M. (1999). Dentatorubral-pallidoluysian atrophy protein interacts through a proline-rich region near polyglutamine with the SH3 domain of an insulin receptor tyrosine kinase substrate. Hum Mol Genet ***8***, 947-57.

Onodera, O., Roses, A. D., Tsuji, S., Vance, J. M., Strittmatter, W. J., and Burke, J. R. (1996). Toxicity of expanded polyglutamine-domain proteins in Escherichia coli. FEBS Lett ***399***, 135-9.

Orozco, G., Estrada, R., Perry, T. L., Arana, J., Fernandez, R., Gonzalez-Quevedo, A., Galarraga, J., and Hansen, S. (1989). Dominantly inherited olivopontocerebellar atrophy from eastern Cuba. Clinical, neuropathological, and biochemical findings. J Neurol Sci ***93***, 37-50.

Orr, H. T., Chung, M. Y., Banfi, S., Kwiatkowski, T. J., Jr., Servadio, A., Beaudet, A. L., McCall, A. E., Duvick, L. A., Ranum, L. P., and Zoghbi, H. Y. (1993). Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. Nat Genet ***4***, 221-6.

Otten, A. D., and Tapscott, S. J. (1995). Triplet repeat expansion in myotonic dystrophy alters the adjacent chromatin structure. Proc Natl Acad Sci U S A ***92***, 5465-9.

Oyake, M., Onodera, O., Shiroishi, T., Takano, H., Takahashi, Y., Kominami, R., Moriwaki, K., Ikeuchi, T., Igarashi, S., Tanaka, H., and Tsuji, S. (1997). Molecular cloning of murine homologue dentatorubral-pallidoluysian atrophy (DRPLA) cDNA: strong conservation of a polymorphic CAG repeat in the murine gene. Genomics ***40***, 205-7.

Pan, X., and Leach, D. R. (2000). The roles of mutS, sbcCD and recA in the propagation of TGG repeats in Escherichia coli. Nucleic Acids Res ***28***, 3178-84.

Papp, A. C., Snyder, P. J., Sedra, M., Guida, M., and Prior, T. W. (1996). Strategies for Amplification of Trinucleotide Repeats: Optimization of Fragile X and Androgen Receptor PCR. Mol Diagn ***1***, 59-64.

Parniewski, P., Jaworski, A., Wells, R. D., and Bowater, R. P. (2000). Length of CTG.CAG repeats determines the influence of mismatch repair on genetic instability. J Mol Biol ***299***, 865-74.

Parrish, J. E., Oostra, B. A., Verkerk, A. J., Richards, C. S., Reynolds, J., Spikes, A. S., Shaffer, L. G., and Nelson, D. L. (1994). Isolation of a GCC repeat showing expansion in FRAXF, a fragile site distal to FRAXA and FRAXE. Nat Genet ***8***, 229-35.

Paulson, H. L., Das, S. S., Crino, P. B., Perez, M. K., Patel, S. C., Gotsdiner, D., Fischbeck, K. H., and Pittman, R. N. (1997). Machado-Joseph disease gene product is a cytoplasmic protein widely expressed in brain. Ann Neurol ***41***, 453-62.

Peier, A. M., McIlwain, K. L., Kenneson, A., Warren, S. T., Paylor, R., and Nelson, D. L. (2000). (Over)correction of FMR1 deficiency with YAC transgenics: behavioral and physical features. Hum Mol Genet ***9***, 1145-59.

Pennacchio, L. A., Bouley, D. M., Higgins, K. M., Scott, M. P., Noebels, J. L., and Myers, R. M. (1998). Progressive ataxia, myoclonic epilepsy and cerebellar apoptosis in cystatin B-deficient mice. Nat Genet ***20***, 251-8.

Perez, M. K., Paulson, H. L., and Pittman, R. N. (1999). Ataxin-3 with an altered conformation that exposes the polyglutamine domain is associated with the nuclear matrix. Hum Mol Genet ***8***, 2377-85.

Petruska, J., Arnheim, N., and Goodman, M. F. (1996). Stability of intrastrand hairpin structures formed by the CAG/CTG class of DNA triplet repeats associated with neurological diseases. Nucleic Acids Res ***24***, 1992-8.

Philips, A. V., Timchenko, L. T., and Cooper, T. A. (1998). Disruption of splicing regulated by a CUG-binding protein in myotonic dystrophy. Science ***280***, 737-41.

Pianese, L., Cavalcanti, F., De Michele, G., Filla, A., Campanella, G., Calabrese, O., Castaldo, I., Monticelli, A., and Cocozza, S. (1997). The effect of parental gender on the GAA dynamic mutation in the FRDA gene. Am J Hum Genet ***60***, 460-3.

Prusiner, S. B. (1996). Molecular biology and genetics of prion diseases. Cold Spring Harb Symp Quant Biol ***61***, 473-93.

Pulst, S. M., Nechiporuk, A., Nechiporuk, T., Gispert, S., Chen, X. N., Lopes-Cendes, I., Pearlman, S., Starkman, S., Orozco-Diaz, G., Lunkes, A., DeJong, P., Rouleau, G. A., Auburger, G., Korenberg, J. R., Figueroa, C., and Sahba, S. (1996). Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. Nat Genet ***14***, 269-76.

Ranen, N. G., Stine, O. C., Abbott, M. H., Sherr, M., Codori, A. M., Franz, M. L., Chao, N. I., Chung, A. S., Pleasant, N., Callahan, C., and et al. (1995). Anticipation and instability of IT-15 (CAG)n repeats in parent-offspring pairs with Huntington disease. Am J Hum Genet ***57***, 593-602.

Reddy, S., Smith, D. B., Rich, M. M., Leferovich, J. M., Reilly, P., Davis, B. M., Tran, K., Rayburn, H., Bronson, R., Cros, D., Balice-Gordon, R. J., and Housman, D. (1996). Mice lacking the myotonic dystrophy protein kinase develop a late onset progressive myopathy. Nat Genet ***13***, 325-35.

Reddy, P. H., Williams, M., Charles, V., Garrett, L., Pike-Buchanan, L., Whetsell, W. O., Miller, G., and Tagle, D. A. (1998). Behavioural abnormalities and selective neuronal loss in HD transgenic mice expressing mutated full-length HD cDNA. Nat Genet ***20***, 198-202.

Richards, R. I., and Sutherland, G. R. (1992). Dynamic mutations: a new class of mutations causing human disease. Cell ***70***, 709-12.

Richards, R. I., and Sutherland, G. R. (1994). Simple repeat DNA is not replicated simply. Nat Genet ***6***, 114-6.

Richardson, C. C. (1965). Phosphorylation of nucleic acid by an enzyme from T4 bacteriophage- infected Escherichia coli. Proc Natl Acad Sci U S A ***54***, 158-65.

Ricke, D. O., Liu, Q., Gostout, B., and Sommer, S. S. (1995). Nonrandom patterns of simple and cryptic triplet repeats in coding and noncoding sequences. Genomics ***26***, 510-20.

Riess, O., Laccone, F. A., Gispert, S., Schols, L., Zuhlke, C., Vieira-Saecker, A. M., Herlt, S., Wessel, K., Epplen, J. T., Weber, B. H., Kreuz, F., Chahrokh-Zadeh, S., Meindl, A., Lunkes, A., Aguiar, J., Macek, M., Krebsova, A., Burk, K., Tinschert, S., Schreyer, I., Pulst, S. M., and Auburger, G. (1997). SCA2 trinucleotide expansion in German SCA patients. Neurogenetics *1*, 59-64.

Riggins, G. J., Lokey, L. K., Chastain, J. L., Leiner, H. A., Sherman, S. L., Wilkinson, K. D., and Warren, S. T. (1992). Human genes containing polymorphic trinucleotide repeats. Nat Genet *2*, 186-91.

Rinchik, E. M., and Russell, L. B. (1990). Germ-line deletion mutations in the mouse: Tools for intensive functional and physical mapping of regions of the mammalian genome. In Genetic and physical mapping, K. E. Davies, ed. (New York: Cold Spring Harbor Laboratory Press), pp. 121-158.

Robert, B., Barton, P., Minty, A., Daubas, P., Weydert, A., Bonhomme, F., Catalan, J., Chazottes, D., Guenet, J. L., and Buckingham, M. (1985). Investigation of genetic linkage between myosin and actin genes using an interspecific mouse back-cross. Nature ***314***, 181-3.

Robitaille, Y., Schut, L., and Kish, S. J. (1995). Structural and immunocytochemical features of olivopontocerebellar atrophy caused by the spinocerebellar ataxia type 1 (SCA-1) mutation define a unique phenotype. Acta Neuropathol ***90***, 572-81.

Rotig, A., de Lonlay, P., Chretien, D., Foury, F., Koenig, M., Sidi, D., Munnich, A., and Rustin, P. (1997). Aconitase and mitochondrial iron-sulphur protein deficiency in Friedreich ataxia. Nat Genet *17*, 215-7.

Rowe, L. B., Nadeau, J. H., Turner, R., Frankel, W. N., Letts, V. A., Eppig, J. T., Ko, M. S., Thurston, S. J., and Birkenmeier, E. H. (1994). Maps from two interspecific backcross DNA panels available as a community genetic mapping resource. Mamm Genome ***5***, 253-74.

Sakamoto, N., Chastain, P. D., Parniewski, P., Ohshima, K., Pandolfo, M., Griffith, J. D., and Wells, R. D. (1999). Sticky DNA: self-association properties of long GAA.TTC repeats in R.R.Y triplex structures from Friedreich's ataxia. Mol Cell *3*, 465-75.

Samadashwily, G. M., Raca, G., and Mirkin, S. M. (1997). Trinucleotide repeats affect DNA replication in vivo. Nat Genet *17*, 298-304.

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). Molecular cloning: a laboratory manual, 2nd Edition (Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory).

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A *74*, 5463-7.

Sanpei, K., Takano, H., Igarashi, S., Sato, T., Oyake, M., Sasaki, H., Wakisaka, A., Tashiro, K., Ishida, Y., Ikeuchi, T., Koide, R., Saito, M., Sato, A., Tanaka, T., Hanyu, S., Takiyama, Y., Nishizawa, M., Shimizu, N., Nomura, Y., Segawa, M., Iwabuchi, K., Eguchi, I., Tanaka, H., Takahashi, H., and Tsuji, S. (1996). Identification of the spinocerebellar ataxia type 2 gene using a direct identification of repeat expansion and cloning technique, DIRECT. Nat Genet *14*, 277-84.

Sarkar, P. S., Appukuttan, B., Han, J., Ito, Y., Ai, C., Tsai, W., Chai, Y., Stout, J. T., and Reddy, S. (2000). Heterozygous loss of Six5 in mice is sufficient to cause ocular cataracts. Nat Genet ***25***, 110-4.

Sato, A., Shimohata, T., Koide, R., Takano, H., Sato, T., Oyake, M., Igarashi, S., Tanaka, K., Inuzuka, T., Nawa, H., and Tsuji, S. (1999a). Adenovirus-mediated expression of mutant DRPLA proteins with expanded polyglutamine stretches in neuronally differentiated PC12 cells. Preferential intranuclear aggregate formation and apoptosis. Hum Mol Genet ***8***, 997-1006.

Sato, T., Oyake, M., Nakamura, K., Nakao, K., Fukusima, Y., Onodera, O., Igarashi, S., Takano, H., Kikugawa, K., Ishida, Y., Shimohata, T., Koide, R., Ikeuchi, T., Tanaka, H., Futamura, N., Matsumura, R., Takayanagi, T., Tanaka, F., Sobue, G., Komure, O., Takahashi, M., Sano, A., Ichikawa, Y., Goto, J., Kanazawa, I., and et al. (1999b). Transgenic mice harboring a full-length human mutant DRPLA gene exhibit age-dependent intergenerational and somatic instabilities of CAG repeats comparable with those in DRPLA patients. Hum Mol Genet ***8***, 99-106.

Saudou, F., Finkbeiner, S., Devys, D., and Greenberg, M. E. (1998). Huntingtin acts in the nucleus to induce apoptosis but death does not correlate with the formation of intranuclear inclusions. Cell ***95***, 55-66.

Schalling, M., Hudson, T. J., Buetow, K. H., and Housman, D. E. (1993). Direct detection of novel expanded trinucleotide repeats in the human genome. Nat Genet ***4***, 135-9.

Scherzinger, E., Lurz, R., Turmaine, M., Mangiarini, L., Hollenbach, B., Hasenbank, R., Bates, G. P., Davies, S. W., Lehrach, H., and Wanker, E. E. (1997). Huntingtin-encoded polyglutamine expansions form amyloid-like protein aggregates in vitro and in vivo. Cell ***90***, 549-58.

Schinzel, A., Auf der Maur, P., and Moser, H. (1977). Partial deletion of long arm of chromosome 11[del(11)(q23)]: Jacobsen syndrome. Two new cases and review of the clinical findings. J Med Genet ***14***, 438-44.

Schmidt, K. H., Abbott, C. M., and Leach, D. R. (2000). Two opposing effects of mismatch repair on CTG repeat instability in Escherichia coli. Mol Microbiol ***35***, 463-71.

Schmitt, I., Brattig, T., Gossen, M., and Riess, O. (1997). Characterization of the rat spinocerebellar ataxia type 3 gene. Neurogenetics ***1***, 103-12.

Schumacher, S., Fuchs, R. P., and Bichara, M. (1998). Expansion of CTG repeats from human disease genes is dependent upon replication mechanisms in Escherichia coli: the effect of long patch mismatch repair revisited. J Mol Biol ***279***, 1101-10.

Schweitzer, J. K., and Livingston, D. M. (1998). Expansions of CAG repeat tracts are frequent in a yeast mutant defective in Okazaki fragment maturation. Hum Mol Genet *7*, 69-74.

Servadio, A., Koshy, B., Armstrong, D., Antalffy, B., Orr, H. T., and Zoghbi, H. Y. (1995a). Expression analysis of the ataxin-1 protein in tissues from normal and spinocerebellar ataxia type 1 individuals. Nat Genet ***10***, 94-8.

Servadio, A., McCall, A., Zoghbi, H., and Eicher, E. M. (1995b). Mapping of the Sca1 and pcd genes on mouse chromosome 13 provides evidence that they are different genes. Genomics ***29***, 812-3.

Seznec, H., Lia-Baldini, A. S., Duros, C., Fouquet, C., Lacroix, C., Hofmann-Radvanyi, H., Junien, C., and Gourdon, G. (2000). Transgenic mice carrying large human genomic sequences with expanded CTG repeat mimic closely the DM CTG repeat intergenerational and somatic instability. Hum Mol Genet ***9***, 1185-94.

Shelbourne, P. F., Killeen, N., Hevner, R. F., Johnston, H. M., Tecott, L., Lewandoski, M., Ennis, M., Ramirez, L., Li, Z., Iannicola, C., Littman, D. R., and Myers, R. M. (1999). A Huntington's disease CAG expansion at the murine Hdh locus is unstable and associated with behavioural abnormalities in mice. Hum Mol Genet ***8***, 763-74.

Shimohata, T., Nakajima, T., Yamada, M., Uchida, C., Onodera, O., Naruse, S., Kimura, T., Koide, R., Nozaki, K., Sano, Y., Ishiguro, H., Sakoe, K., Ooshima, T., Sato, A., Ikeuchi, T., Oyake, M., Sato, T., Aoyagi, Y., Hozumi, I., Nagatsu, T., Takiyama, Y., Nishizawa, M., Goto, J., Kanazawa, I., Davidson, I., Tanese, N., Takahashi, H., and Tsuji, S. (2000). Expanded polyglutamine stretches interact with TAFII130, interfering with CREB-dependent transcription. Nat Genet ***26***, 29-36.

Silveira, I., Alonso, I., Guimaraes, L., Mendonca, P., Santos, C., Maciel, P., Fidalgo De Matos, J. M., Costa, M., Barbot, C., Tuna, A., Barros, J., Jardim, L., Coutinho, P., and Sequeiros, J. (2000). High germinal instability of the (CTG)n at the SCA8 locus of both expanded and normal alleles. Am J Hum Genet ***66***, 830-40.

Silver, L. M. (1995). Mouse genetics: concepts and applications (New York: Oxford University Press).

Singer-Sam, J., LeBon, J. M., Tanguay, R. L., and Riggs, A. D. (1990). A quantitative HpaII-PCR assay to measure methylation of DNA from a small number of cells. Nucleic Acids Res ***18***, 687.

Skinner, P. J., Koshy, B. T., Cummings, C. J., Klement, I. A., Helin, K., Servadio, A., Zoghbi, H. Y., and Orr, H. T. (1997). Ataxin-1 with an expanded glutamine tract alters nuclear matrix-associated structures. Nature ***389***, 971-4.

Slegtenhorst-Eegdeman, K. E., de Rooij, D. G., Verhoef-Post, M., van de Kant, H. J., Bakker, C. E., Oostra, B. A., Grootegoed, J. A., and Themmen, A. P. (1998). Macroorchidism in FMR1 knockout mice is caused by increased Sertoli cell proliferation during testicular development. Endocrinology ***139***, 156-62.

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol ***98***, 503-17.

Stallings, R. L. (1994). Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: implications for human genetic diseases. Genomics ***21***, 116-21.

Steinbach, P., Glaser, D., Vogel, W., Wolf, M., and Schwemmle, S. (1998). The DMPK gene of severely affected myotonic dystrophy patients is hypermethylated proximal to the largely expanded CTG repeat. Am J Hum Genet ***62***, 278-85.

Subramanian, P. S., Nelson, D. L., and Chinault, A. C. (1996). Large domains of apparent delayed replication timing associated with triplet repeat expansion at FRAXA and FRAXE. Am J Hum Genet ***59***, 407-16.

Suen, I. S., Rhodes, J. N., Christy, M., McEwen, B., Gray, D. M., and Mitas, M. (1999). Structural properties of Friedreich's ataxia d(GAA) repeats. Biochim Biophys Acta ***1444***, 14-24.

Sutherland, G. R., and Richards, R. I. (1995). Simple tandem DNA repeats and human genetic disease. Proc Natl Acad Sci U S A ***92***, 3636-41.

Sved, J., and Bird, A. (1990). The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc Natl Acad Sci U S A ***87***, 4692-6.

Takiyama, Y., Oyanagi, S., Kawashima, S., Sakamoto, H., Saito, K., Yoshida, M., Tsuji, S., Mizuno, Y., and Nishizawa, M. (1994). A clinical and pathologic study of a large Japanese family with Machado-Joseph disease tightly linked to the DNA markers on chromosome 14q. Neurology ***44***, 1302-8.

Takiyama, Y., Igarashi, S., Rogaeva, E. A., Endo, K., Rogaev, E. I., Tanaka, H., Sherrington, R., Sanpei, K., Liang, Y., Saito, M., and et al. (1995). Evidence for inter-generational instability in the CAG repeat in the MJD1 gene and for conserved haplotypes at flanking markers amongst Japanese and Caucasian subjects with Machado-Joseph disease. Hum Mol Genet ***4***, 1137-46.

Tan, J. A., Joseph, D. R., Quarmby, V. E., Lubahn, D. B., Sar, M., French, F. S., and Wilson, E. M. (1988). The rat androgen receptor: primary structure, autoregulation of its messenger ribonucleic acid, and immunocytochemical localization of the receptor protein. Mol Endocrinol ***2***, 1276-85.

Taneja, K. L., McCurrach, M., Schalling, M., Housman, D., and Singer, R. H. (1995). Foci of trinucleotide repeat transcripts in nuclei of myotonic dystrophy cells and tissues. J Cell Biol ***128***, 995-1002.

Tazi, J., and Bird, A. (1990). Alternative chromatin structure at CpG islands. Cell ***60***, 909-20.

Thornton, C. A., Wymer, J. P., Simmons, Z., McClain, C., and Moxley, R. T. (1997). Expansion of the myotonic dystrophy CTG repeat reduces expression of the flanking DMAHP gene. Nat Genet ***16***, 407-9.

Tome, F. M., and Fardeau, M. (1980). Nuclear inclusions in oculopharyngeal dystrophy. Acta Neuropathol ***49***, 85-7.

Torchia, B. S., Call, L. M., and Migeon, B. R. (1994). DNA replication analysis of FMR1, XIST, and factor 8C loci by FISH shows nontranscribed X-linked genes replicate late. Am J Hum Genet ***55***, 96-104.

Trinh, T. Q., and Sinden, R. R. (1993). The influence of primary and secondary DNA structure in deletion and duplication between direct repeats in Escherichia coli. Genetics ***134***, 409-22.

Trottier, Y., Biancalana, V., and Mandel, J. L. (1994). Instability of CAG repeats in Huntington's disease: relation to parental transmission and age of onset. J Med Genet ***31***, 377-82.

Trottier, Y., Lutz, Y., Stevanin, G., Imbert, G., Devys, D., Cancel, G., Saudou, F., Weber, C., David, G., Tora, L., and et al. (1995). Polyglutamine expansion as a pathological epitope in Huntington's disease and four dominant cerebellar ataxias. Nature ***378***, 403-6.

Trottier, Y., Cancel, G., An-Gourfinkel, I., Lutz, Y., Weber, C., Brice, A., Hirsch, E., and Mandel, J. L. (1998). Heterogeneous intracellular localization and expression of ataxin-3. Neurobiol Dis ***5***, 335-47.

Turmaine, M., Raza, A., Mahal, A., Mangiarini, L., Bates, G. P., and Davies, S. W. (2000). Nonapoptotic neurodegeneration in a transgenic mouse model of Huntington's disease. Proc Natl Acad Sci U S A ***97***, 8093-7.

Ulitzur, N., Rancano, C., and Pfeffer, S. R. (1997). Biochemical characterization of mapmodulin, a protein that binds microtubule-associated proteins. J Biol Chem ***272***, 30577-82.

Urieli-Shoval, S., Gruenbaum, Y., Sedat, J., and Razin, A. (1982). The absence of detectable methylated bases in Drosophila melanogaster DNA. FEBS Lett ***146***, 148-52.

Urquhart, A., Kimpton, C. P., Downes, T. J., and Gill, P. (1994). Variation in short tandem repeat sequences--a survey of twelve microsatellite loci for use as forensic identification markers. Int J Legal Med ***107***, 13-20.

Usuki, F., Ishiura, S., Saitoh, N., Sasagawa, N., Sorimachi, H., Kuzume, H., Maruyama, K., Terao, T., and Suzuki, K. (1997). Expanded CTG repeats in myotonin protein kinase suppresses myogenic differentiation. Neuroreport ***8***, 3749-53.

Usuki, F., and Ishiura, S. (1998). Expanded CTG repeats in myotonin protein kinase increase susceptibility to oxidative stress. Neuroreport ***9***, 2291-6.

Uyama, E., Nohira, O., Chateau, D., Tokunaga, M., Uchino, M., Okabe, T., Ando, M., and Tome, F. M. (1996). Oculopharyngeal muscular dystrophy in two unrelated Japanese families. Neurology ***46***, 773-8.

Vaesen, M., Barnikol-Watanabe, S., Gotz, H., Awni, L. A., Cole, T., Zimmermann, B., Kratzin, H. D., and Hilschmann, N. (1994). Purification and characterization of two putative HLA class II associated proteins: PHAPI and PHAPII. Biol Chem Hoppe Seyler ***375***, 113-26.

Verkerk, A. J., Pieretti, M., Sutcliffe, J. S., Fu, Y. H., Kuhl, D. P., Pizzuti, A., Reiner, O., Richards, S., Victoria, M. F., Zhang, F. P., and et al. (1991). Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell ***65***, 905-14.

Virtaneva, K., D'Amato, E., Miao, J., Koskiniemi, M., Norio, R., Avanzini, G., Franceschetti, S., Michelucci, R., Tassinari, C. A., Omer, S., Pennacchio, L. A., Myers, R. M., Dieguez-Lucena, J. L., Krahe, R., de la Chapelle, A., and Lehesjoki, A. E. (1997). Unstable minisatellite expansion causing recessively inherited myoclonus epilepsy, EPM1. Nat Genet ***15***, 393-6.

Vonsattel, J. P., Myers, R. H., Stevens, T. J., Ferrante, R. J., Bird, E. D., and Richardson, E. P. (1985). Neuropathological classification of Huntington's disease. J Neuropathol Exp Neurol ***44***, 559-77.

Wang, G., Sawai, N., Kotliarova, S., Kanazawa, I., and Nukina, N. (2000). Ataxin-3, the MJD1 gene product, interacts with the two human homologs of yeast DNA repair protein RAD23, HHR23A and HHR23B. Hum Mol Genet ***9***, 1795-803.

Wang, J., Pegoraro, E., Menegazzo, E., Gennarelli, M., Hoop, R. C., Angelini, C., and Hoffman, E. P. (1995). Myotonic dystrophy: evidence for a possible dominant-negative RNA mutation. Hum Mol Genet ***4***, 599-606.

Wang, Y. H., Amirhaeri, S., Kang, S., Wells, R. D., and Griffith, J. D. (1994). Preferential nucleosome assembly at DNA triplet repeats from the myotonic dystrophy gene. Science ***265***, 669-71.

Wang, Y. H., and Griffith, J. (1995). Expanded CTG triplet blocks from the myotonic dystrophy gene create the strongest known natural nucleosome positioning elements. Genomics ***25***, 570-3.

Wang, Y. H., and Griffith, J. (1996). Methylation of expanded CCG triplet repeat DNA from fragile X syndrome patients enhances nucleosome exclusion. J Biol Chem ***271***, 22937-40.

Wanker, E. E., Rovira, C., Scherzinger, E., Hasenbank, R., Walter, S., Tait, D., Colicelli, J., and Lehrach, H. (1997). HIP-I: a huntingtin interacting protein isolated by the yeast two-hybrid system. Hum Mol Genet ***6***, 487-95.

Waragai, M., Lammers, C. H., Takeuchi, S., Imafuku, I., Udagawa, Y., Kanazawa, I., Kawabata, M., Mouradian, M. M., and Okazawa, H. (1999). PQBP-1, a novel polyglutamine tract-binding protein, inhibits transcription activation by Brn-2 and affects cell survival. Hum Mol Genet ***8***, 977-87.

Warren, S. T. (1996). The expanding world of trinucleotide repeats. Science ***271***, 1374-5.

Warrick, J. M., Paulson, H. L., Gray-Board, G. L., Bui, Q. T., Fischbeck, K. H., Pittman, R. N., and Bonini, N. M. (1998). Expanded polyglutamine protein forms nuclear inclusions and causes neural degeneration in Drosophila. Cell ***93***, 939-49.

Webb, T. P., Bundey, S. E., Thake, A. I., and Todd, J. (1986). Population incidence and segregation ratios in the Martin-Bell syndrome. Am J Med Genet ***23***, 573-80.

Weber, J. L. (1990). Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms. Genomics ***7***, 524-30.

Wellington, C. L., Ellerby, L. M., Hackam, A. S., Margolis, R. L., Trifiro, M. A., Singaraja, R., McCutcheon, K., Salvesen, G. S., Propp, S. S., Bromm, M., Rowland, K. J., Zhang, T., Rasper, D., Roy, S., Thornberry, N., Pinsky, L., Kakizuka, A., Ross, C. A., Nicholson, D. W., Bredesen, D. E., and Hayden, M. R. (1998). Caspase cleavage of gene products associated with triplet expansion disorders generates truncated fragments containing the polyglutamine tract. J Biol Chem ***273***, 9158-67.

Wells, R. D., Parniewski, P., Pluciennik, A., Bacolla, A., Gellibolian, R., and Jaworski, A. (1998). Small slipped register genetic instabilities in Escherichia coli in triplet repeat sequences associated with hereditary neurological diseases. J Biol Chem ***273***, 19532-41.

Wells, R. D., and Warren, S. T. (1998). Genetic instabilities and hereditary neurological diseases (San Diego: Academic Press), pp. 4.
Wheeler, V. C., Auerbach, W., White, J. K., Srinidhi, J., Auerbach, A., Ryan, A., Duyao, M. P., Vrbanac, V., Weaver, M., Gusella, J. F., Joyner, A. L., and MacDonald, M. E. (1999). Length-dependent gametic CAG repeat instability in the Huntington's disease knock-in mouse. Hum Mol Genet ***8***, 115-22.

Williamson, P., Holt, S., Townsend, S., and Boyd, Y. (1995). A somatic cell hybrid panel for mouse gene mapping characterized by PCR and FISH. Mamm Genome ***6***, 429-32.

Wilson, R. B., and Roof, D. M. (1997). Respiratory deficiency due to loss of mitochondrial DNA in yeast lacking the frataxin homologue. Nat Genet ***16***, 352-7.

Wohrle, D., Hennig, I., Vogel, W., and Steinbach, P. (1993). Mitotic stability of fragile X mutations in differentiated cells indicates early post-conceptional trinucleotide repeat expansion. Nat Genet ***4***, 140-2.

Wohrle, D., Kennerknecht, I., Wolf, M., Enders, H., Schwemmle, S., and Steinbach, P. (1995). Heterogeneity of DM kinase repeat expansion in different fetal tissues and further expansion during cell proliferation in vitro: evidence for a casual involvement of methyl-directed DNA mismatch repair in triplet repeat stability. Hum Mol Genet ***4***, 1147-53.

Wood, J. D., Yuan, J., Margolis, R. L., Colomer, V., Duan, K., Kushi, J., Kaminsky, Z., Kleiderlein, J. J., Sharp, A. H., and Ross, C. A. (1998). Atrophin-1, the DRPLA gene product, interacts with two families of WW domain-containing proteins. Mol Cell Neurosci ***11***, 149-60.

Worth, P. F., Houlden, H., Giunti, P., Davis, M. B., and Wood, N. W. (2000). Large, expanded repeats in SCA8 are not confined to patients with cerebellar ataxia. Nat Genet ***24***, 214-5.

Yasuda, S., Inoue, K., Hirabayashi, M., Higashiyama, H., Yamamoto, Y., Fuyuhiro, H., Komure, O., Tanaka, F., Sobue, G., Tsuchiya, K., Hamada, K., Sasaki, H., Takeda, K., Ichijo, H., and Kakizuka, A. (1999). Triggering of neuronal cell death by accumulation of activated SEK1 on nuclear polyglutamine aggregations in PML bodies. Genes Cells ***4***, 743-56.

Yazawa, I., Nukina, N., Hashida, H., Goto, J., Yamada, M., and Kanazawa, I. (1995). Abnormal gene product identified in hereditary dentatorubral-pallidoluysian atrophy (DRPLA) brain. Nat Genet ***10***, 99-103.

Yu, S., Pritchard, M., Kremer, E., Lynch, M., Nancarrow, J., Baker, E., Holman, K., Mulley, J. C., Warren, S. T., Schlessinger, D., and et al. (1991). Fragile X genotype characterized by an unstable region of DNA. Science ***252***, 1179-81.

Yvert, G., Lindenberg, K. S., Picaud, S., Landwehrmeyer, G. B., Sahel, J. A., and Mandel, J. L. (2000). Expanded polyglutamines induce neurodegeneration and trans-neuronal alterations in cerebellum and retina of SCA7 transgenic mice. Hum Mol Genet ***9***, 2491-506.

Zeitlin, S., Liu, J. P., Chapman, D. L., Papaioannou, V. E., and Efstratiadis, A. (1995). Increased apoptosis and early embryonic lethality in mice nullizygous for the Huntington's disease gene homologue. Nat Genet ***11***, 155-63.

Zhuchenko, O., Bailey, J., Bonnen, P., Ashizawa, T., Stockton, D. W., Amos, C., Dobyns, W. B., Subramony, S. H., Zoghbi, H. Y., and Lee, C. C. (1997). Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the alpha 1A-voltage-dependent calcium channel. Nat Genet ***15***, 62-9.