



STUDIES OF FRAGILE SITES ON HUMAN CHROMOSOME 16

JULIE NANCARROW B.Sc

A thesis submitted for the Degree of Doctor of Philosophy

Department of Cytogenetics and Molecular Genetics,

Women's and Children's Hospital, North Adelaide, South Australia.

Department of Paediatrics, Faculty of Medicine, University of Adelaide, South Australia.

April, 1998

Abstract

A fragile site is defined as a specific point on a chromosome that is subject to the formation of non-random gaps or breaks when exposed to a specific chemical agent or condition of cell culture. Fragile sites are classified as rare or common, and also according to the chemistry of their induction. At the time this study commenced, the rare folate sensitive fragile site *FRAXA*, located on the X chromosome, had been characterised. The molecular basis of this fragile site was found to be expansion of a normally polymorphic CCG repeat tract, located adjacent to a CpG island that becomes methylated in fragile site expressing individuals. The project described in this thesis was performed in order to gain more knowledge regarding the characteristics of fragile sites at the molecular (or DNA) level, and enable comparisons with previously characterised fragile sites. This thesis mainly documents the molecular characterisation and study of certain aspects of the rare folate sensitive fragile site *FRA16A*, in band 16p13.11. Initial steps toward the cloning and characterisation of the band 16q23.1, common, aphidicolin inducible fragile site, *FRA16D* are also reported.

Fragile sites on human chromosome 16 were chosen for study by reason of the availability of a detailed somatic cell hybrid breakpoint map of this chromosome. All the known fragile sites on the chromosome formed part of the map, and were therefore localised to relatively small, accessible, regions between somatic cell hybrid translocation breakpoints.

The positional cloning of *FRA16A* commenced with long range pulsed field restriction mapping of uncloned genomic DNA in the region, utilising DNA markers previously localised to the same somatic cell hybrid breakpoint interval as the fragile site. This mapping identified co-localising *Not* I and *Bss*H II sites within the region of interest that appeared differentially cleaved on normal and *FRA16A* chromosomes. Yeast artificial

chromosomes (YACs) within the *FRA16A* region were identified and restriction mapped. One YAC, My769H1, was found to contain co-localising *Not* I and *Bss*H II sites adjacent to a CCG repeat hybridising sequence. My769HI was subcloned, and a clone containing the CCG repeat hybridising region isolated. Further analysis revealed that, as with *FRAXA*, the putative *FRA16A* clone contained a polymorphic CCG repeat tract that becomes expanded on chromosomes segregating with the fragile site. Further mapping definitively localised *FRA16A* to the cloned CCG repeat tract.

PCR analysis of normal *FRA16A* CCG repeat tract alleles in human DNA samples from three populations found significant differences in the allele frequencies between populations. In order to ascertain the reason for this, a subset of *FRA16A* alleles from each population was subject to DNA sequence analysis. The results of this analysis suggested that the longer and more variable alleles, found mainly in one population, were associated with the loss of an interruption within the CCG repeat tract. The implication of this finding (previously suggested by a number of authors) was that loss of a repeat interruption causes a greater predisposition to instability for a CCG repeat tract, thus leading to a greater variation of allele size, and ultimately to the genesis of a fragile site.

The CCG repeat tract associated with the fragile site *FRAXA* is located within the 5'UTR of the *FMRI* gene. CCG repeat tracts associated with two other folate sensitive fragile sites (*FRAXE* and *FRA11B*) are also located at the 5' ends of genes. The functional role of these repeat sequences is unknown.

In an attempt to obtain an insight into the function role of potentially unstable CCG repeat tracts in genes, the DNA region adjacent to *FRA16A* was subjected to a number of analyses that were designed to detect transcribed sequences. Fetal tissue Northern blot analysis detected a number of transcripts with homology to the region, therefore these results were inconclusive. Thirteen cDNA clones homologous to the region were isolated,

but no evidence of splicing could be found when compared to genomic DNA. Sequencing and computer analysis of the cDNA clones and *FRA16A* region genomic DNA predicted the presence of promoter regions and an exonic sequence. Searches of DNA and protein sequence databases with the exon sequence revealed limited levels of homology to distinct motifs present in certain extracellular proteins. Although the results obtained strongly suggest that a gene is associated with *FRA16A*, further studies are required to show this definitively.

Positional cloning of the fragile site *FRA16D* was initiated in order to obtain further information regarding DNA sequences associated common, aphidicolin inducible sites. That common fragile sites are expressed under conditions of folate stress (as well as by aphidicolin) suggests that the folate sensitive fragile sites (e.g. *FRA16A*) and the common fragile sites (e.g. *FRA16D*) have common properties that may be revealed by further analyses of fragile sites in both classes.

To identify YAC clones spanning the fragile site, DNA markers localised to the *FRA16D* region by somatic cell hybrid panel mapping were used for database searches to identify megaYACs in the region. Further database searches identified a contig of YACs spanning the fragile site. The results of FISH analysis indicated that three YACs within the contig span *FRA16D*. Due to time constraints, further steps toward cloning the fragile site were not included in the project.

Declaration

This work contains no material that has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give my consent to this thesis being made available for loan and photocopying if accepted for the award of PhD degree.

Julie Nancarrow

20th April 1998

Acknowledgments

The work presented in this thesis was carried out in the Department of Cytogenetics and Molecular Genetics, Women's and Children's Hospital, North Adelaide, South Australia.

My main supervisors for the project were Associate Professor Rob Richards and Professor Grant Sutherland, to whom I am most grateful for providing me with the opportunity to embark upon, and complete, this course of study. I would like to especially thank Rob for his support and guidance throughout my candidature, and for making himself available for frequent discussions about all aspects of the project. I would also like to thank Associate Professor David Callen for being my supervisor in the early stages of the project.

The laboratory work documented in this thesis would have seemed more arduous if not performed in an agreeable working environment. For providing such an environment, I would like to thank Kathie Friend, Marie Mangelsdorf, Lynne Hobson, Annette Osborne, Shirley Richardson and Sui Yu. I would like to thank Sui in particular for many conversations and helpful suggestions over the years.

With regard to their contributions to the project, I would like to thank Kathy Holman and Marie Mangelsdorf for typing *FRA16A* alleles in the CEPH pedigrees, and Elizabeth Baker, Helen Eyre, Jillian Nicholl, and Erica Woollatt for their efforts in the FISH studies necessary to the project.

On a personal level, I would like to thank Peter Rofe for helping me through some difficult times. Finally, and most importantly, I would like to acknowledge the love, help and support of all the members of my family, Rosemary (especial thanks for your speedy proof reading of this thesis), Roy, Jennifer, Fiona and Graham, during the course of this study.

Table of Contents

ABSTRACT	I
DECLARATION	IV
ACKNOWLEDGEMENTS	V
TABLE OF CONTENTS	VI
ABBREVIATIONS	VII
Chapter 1: Literature review	1
Chapter 2: Materials and methods	49
Chapter 3: PFGE long range restriction mapping within the interval containing FRA16A	76
Chapter 4: Molecular characterisation of the FRA16A fragile site	100
Chapter 5: A molecular basis for interpopulation variation of FRA16A repeat tract instability	124
Chapter 6: Identification and characterisation of a FRA16A associated transcript	142
Chapter 7: Construction of a megaYAC contig across the FRA16A fragile site	172
Chapter 8: Concluding remarks	185
References:	194
Appendix 1: FRA16A region DNA sequence and primers	223
Appendix 11: Publications	230

Abbreviations

DNAs	DNA samples
DTT	dithiothreitol
^{32}P	^{32} -Phosphorus
EDTA	ethylene diamine tetra acetic acid
FISH	fluorescence in situ hybridisation
kb	kilobase(s)
LANL	Los Alamos National Laboratory
L. medium	Luria Bertoni medium
Mb	megabase(s)
mM	millimolar
My	mega yeast artificial chromosome
ml	millilitres
PXE	Pseudoxanthoma elasticum
RT-PCR	reverse transcription polymerase chain reaction
SDS	sodium dodecyl sulphate
WIT	Whitehead Institute of Technology
y	yeast artificial chromosome
YAC	yeast artificial chromosome

Chapter 1

Literature Review

Chapter 1: Contents

Literature Review

1.1	Summary	5
1.2	Introduction	7
1.3	Chromosomal fragile sites	7
1.3.1	Definition.....	7
1.3.2	Classification.....	7
1.3.2	The induction and expression of fragile sites	8
1.3.3	Rare fragile sites	9
1.3.4	Fragile sites and neoplasia.....	9
1.3.5	Fragile sites located on human chromosome 16	10
1.3.6	Cytogenetic characterisation of the rare folate sensitive fragile site <i>FRA16A</i>	10
1.3.7	Cytogenetic expression of the common aphidicolin inducible fragile site <i>FRA16D</i>	11
1.3.8	Physical mapping of the <i>FRA16A</i> region	11
1.3.9	Duplication of DNA sequences within the <i>FRA16A</i> region.....	12
1.4	Fragile X syndrome	13
1.4.1	Clinical characteristics of the fragile X syndrome	13
1.4.2	Recognition of the fragile X syndrome	13
1.4.3	Cytogenetic expression of the marker X chromosome	14
1.4.4	The unique inheritance pattern of the fragile X syndrome.....	14
1.4.5	Departures from classical X linked inheritance	14
1.4.6	The dependence of fragile X syndrome penetrance upon position in pedigree	15

1.4.7	High mutation rate in sperm and absence of new mutations in ova at the fragile X locus.....	16
1.4.8	Models of fragile X syndrome inheritance.....	16
1.4.9	Correlations between X chromosome activation and the severity of fragile X syndrome penetrance in females	18
1.4.10	Laird's hypothesis	18
1.4.11	Additions to Laird's hypothesis.....	19
1.4.12	Relevance of Laird's hypothesis to the molecular characteristics of <i>FRAXA</i>	20
1.4.13	Genetic mapping of the fragile X syndrome locus.....	21
1.4.14	Physical mapping of the fragile X region.....	21
1.4.15	Identification of a hypermethylated CpG island in fragile X syndrome males.....	22
1.4.16	Cloning of DNA sequences across <i>FRAXA</i>	23
1.4.17	Pedigree analysis of fragile X families with probes specific for the mutated region.....	25
1.4.18	Classification of mutations at the fragile X locus	27
1.4.19	Sequence characterisation of a <i>Pst</i> I fragment detecting the fragile X syndrome mutation	27
1.4.20	Identification of a gene (<i>FMRI</i>) containing the <i>FRAXA</i> CCG repeat tract.....	28
1.4.21	Absence of <i>FMRI</i> gene expression in fragile X patients.....	30
1.4.22	Localisation of the fragile X syndrome heritable unstable element to a CCG repeat tract	31
1.4.23	Polymorphism of fragile X CCG repeat tract length in normal individuals.....	32
1.4.24	Studies of the <i>FMRI</i> CCG repeat tract on fragile X chromosomes.....	32
1.4.25	Number of CCG repeats is indicative of the level of intergenerational instability	33

1.4.26	Risk of expansion to the full mutation correlates with CCG repeat number: relationship to the Sherman paradox.....	33
1.4.27	Somatic variation at the fragile X locus	35
1.4.28	No evidence for new fragile X mutations	36
1.4.29	Evidence of founder chromosomes and linkage disequilibrium at the fragile X locus	36
1.4.30	Relationship between haplotype and fragile X CCG repeat copy number	38
1.4.31	Proposed mechanisms for founder mutations	38
1.4.32	Confirmation that the fragile X syndrome is a single gene disorder.....	39
1.4.33	<i>FMRI</i> gene transcription is repressed by methylation	40
1.4.34	Mechanism and timing of mutation events	40
1.4.35	Methylation at the fragile X locus.....	41
1.4.36	Laird's hypothesis and abnormal methylation of the <i>FMRI</i> CpG island.....	42
1.5	Dynamic mutations	44
1.5.1	Trinucleotide repeats in genes	45
1.5.2	Proposed mechanisms for repeat instability	45
1.5.3	Comparisons of autosomal and X-chromosomal folate sensitive fragile sites may provide insight into the nature of the mutational mechanism involved in their genesis	46
1.6	Project strategy	47

1.1 Summary

Fragile sites are defined as specific points on chromosomes that, when exposed to certain chemical agents or cell culture conditions, are liable to display non-random breaks or gaps that may be cytogenetically visible on stained metaphase chromosomes. They are classified on the basis of the conditions required to induce cytogenetic expression, and also according to whether they are considered to be rare or common (Sutherland and Richards, 1995b).

At the time of commencing the project described in this thesis, only one fragile site, the rare folate sensitive fragile site *FRAXA*, had been characterised at the molecular level. This fragile site has been the subject of considerable clinical and scientific interest for many years because of its positive association with the fragile X syndrome, a distinct form of X-linked mental retardation.

Segregation analysis of the fragile X syndrome revealed an unusual pattern of inheritance, including the phenomenon of increased penetrance in successive generations, termed the "Sherman Paradox" (for fragile X syndrome), or anticipation. Although many models were proposed to account for the apparently unique genetics of the disorder, none were successful in providing a full explanation, although the subsequent molecular characterisation of *FRAXA* indicated that some models were partially correct.

In order to characterise *FRAXA* at the molecular level, linkage mapping of the fragile X syndrome locus and *in situ* localisation of probes in the region was initially employed. These steps were followed by the identification of yeast artificial chromosome (YAC) clones (with human DNA inserts) spanning the fragile site. The YAC inserts were subcloned, and *FRAXA* localised to a CCG repeat tract within a CpG island that was found to be greatly expanded and unstable in fragile X syndrome affecteds (termed a full mutation) and unstable, but expanded to a lesser extent, in carriers (termed a premutation). Hypermethylation of the CCG repeat tract and CpG island was found to be associated with the full mutation repeat expansion. The fragile X CCG repeat tract was

localised to within the 5' untranslated region of a gene, *FMRI*, transcription of which was found to be suppressed by the hypermethylation associated with CCG repeat expansion.

Subsequent studies of the *FMRI* CCG repeat in fragile X families indicated that the risk of expansion to a full mutation from a premutation is directly correlated with the size (i.e. number of CCG repeats) of the premutation. This finding provided a complete resolution of the "Sherman Paradox", and provided a molecular basis for anticipation.

Analysis of alleles at microsatellite loci on fragile X associated and normal chromosomes found clear evidence of a founder effect for the fragile X mutation. It was therefore suggested that the initial change producing a founder chromosome must occur rarely and be carried silently by a proportion of the population in order to account for the high incidence of the fragile X syndrome.

The elucidation of the genetic mechanism of fragile X syndrome inheritance provided the first example of what has become known as dynamic mutation, a descriptive term for the mechanism of mutation in unstable heritable repeat disorders. Subsequently the genetic disorders spinobulbar muscular atrophy (SBMA) and myotonic dystrophy (DM) were found to be due to unstable CAG repeat tracts, thus demonstrating that this type of mutation had the potential for widespread impact in the field of human genetics.

1.2 Introduction

This chapter is a review of the literature on the subject of fragile sites and dynamic mutations that were primarily published prior to 1993, when the study described in this thesis commenced. However, in the interest of providing a more informative and complete discussion of the field, a small number of more recent papers are also included. A relatively detailed review of the literature documenting the molecular basis of the fragile X syndrome is presented, as the project of characterising *FRA16A* was undertaken in an attempt to address a number of existing questions in field of research relating to this disorder.

1.3 Chromosomal Fragile Sites

1.3.1 Definition

A fragile site is defined as a specific point on a chromosome that is liable to display a non-random gap or break when exposed to a specific chemical agent or condition of tissue culture. Fragile sites may be identified on stained chromosomes at the metaphase stage of cell division (Sutherland and Hecht, 1985).

1.3.2 Classification

Fragile sites may be classified as rare or common, and also according to the conditions of tissue culture under which they are expressed. The distinction between rare and common fragile sites is an arbitrary one, although common fragile sites are thought to be homozygous in all individuals (i.e. part of normal chromosome structure) and one fragile site classified as rare has been found in about 5% of individuals within a particular population (Sutherland and Richards, 1995b).

The rare fragile sites are divided into three classes, folate sensitive, distamycin A inducible and BrdU requiring. The common fragile sites are classified as aphidicolin, 5-azacytidine or BrdU inducible (Sutherland and Hecht, 1985).

1.3.2 The induction and expression of fragile sites

Although six fragile sites have now been characterised at the DNA level, what has not yet been made evident is the physical basis, or the molecular/conformational changes at the DNA and protein level, that are a requirement for expression of the chromosomal gaps visible under the microscope.

A variety of agents and/or conditions can induce the cytogenetic expression of fragile sites. A requirement for the expression of rare folate sensitive fragile sites in metaphase chromosome spreads is low levels of dTTP or dCTP (pyrimidine deoxyribonucleotides) in the tissue culture medium at the time of DNA synthesis. These conditions may occur in the absence of folic acid or thymidine. A large excess of thymidine inhibits the enzyme ribonucleotide reductase, thus leading to a decrease in the synthesis of the pyrimidine deoxyribonucleotides. Alternatively, the addition of inhibitors of folate metabolism (e.g. methotrexate), or of thymidylate synthesis (e.g. fluorodeoxyuridine) will achieve the same result. (Sutherland and Hecht, 1985).

Common fragile site expression requires the addition of aphidicolin to the culture medium. Aphidicolin is an inhibitor of DNA polymerase α , an enzyme largely responsible for chromosomal DNA replication. Expression of common aphidicolin fragile sites may occur without induction or under low folate conditions in a small proportion of metaphase spreads (Glover *et al.*, 1984; Glover and Stein, 1988). The addition of aphidicolin enhances expression of these sites in a concentration dependent manner, with high concentrations of aphidicolin leading to the disintegration of chromosomes.

Variability in expression has been observed for a given common fragile site between chromosomes in a cell, suggesting that polymorphic variation at the DNA sequence level may be present at these locations (E. Baker, personal communication). In addition, the

proportion of metaphases in which any common fragile site is expressed cytogenetically can vary between individuals (Sutherland and Richards, 1995b).

1.3.3 Rare fragile sites

As the cytological appearance of fragile sites is very variable, classification of a site as rare is dependent upon the observation of several essential properties: (1) there is a non-staining gap of varying width that usually involves both chromatids; (2) the fragile site is always at exactly the same point on the chromosome in an individual or kindred; (3) the fragile site is inherited in a Mendelian co-dominant manner; (4) the fragile site exhibits fragility under appropriate conditions of induction as evidenced by acentric fragments, deleted chromosomes, tri-radial figures and the like (Sutherland, 1979).

Of the fragile sites currently described, only the rare folate sensitive fragile sites *FRAXA* and *FRAXE*, on the human X chromosome, are considered to have direct phenotypic effects. As none of the autosomal folate sensitive fragile sites has been observed in a homozygous state, it is unknown if any phenotype would result from such a state.

1.3.4 Fragile sites and neoplasia

It has been proposed by a number of authors that fragile sites play a role in the pathogenesis of cancer (Daniel 1986; Le Beau 1986; Michels 1985; Shabtai *et al.* 1985), although there has been no neoplasm consistently associated with the presence of a specific fragile site. These proposals were based on the observation that a number of fragile sites are located at, or near, the breakpoints of some chromosomal rearrangements consistently found within tumour karyotypes (Le Beau 1986; Hecht and Glover 1984; Yunis and Soreng 1984; De Braekeleer *et al.* 1985; Hecht and Hecht 1986; Glover *et al.* 1986). It was therefore reasoned that an individual with a specific fragile site may be predisposed to the development of a neoplasm that is associated with a genetic alteration of the fragile site region (Craig-Holmes *et al.*, 1987). However, intensive mapping of a number of such fragile site locations has found no coincidence with cancer breakpoints (Simmers *et al.*, 1987).

1.3.5 Fragile sites located on human chromosome 16

Five fragile sites have been cytogenetically characterised on human chromosome 16. They are: *FRA16A*, a rare folate sensitive fragile site located in band 16p13.1; *FRA16B*, a rare distamycin A inducible fragile site located in band 16q22.1; *FRA16C*, a common aphidicolin inducible fragile site located in band 16q22.1; *FRA16D*, a common aphidicolin inducible fragile site located in band 16q23.2; and *FRA16E*, a rare, distamycin A inducible fragile site located within band 16p12.1 that has been recorded only in the Japanese population. Several of these fragile sites, in particular *FRA16A*, *FRA16B* and *FRA16D*, have facilitated the anchoring of the human chromosome 16 genetic and somatic cell hybrid breakpoint maps to the cytogenetic map of this chromosome. The fragile sites were therefore localised to defined intervals of chromosome 16 (between flanking hybrid cell line breakpoints), in relatively close proximity to DNA markers within the same interval. These mapped DNA markers provided an invaluable resource for the positional cloning of chromosome 16 fragile sites.

1.3.6 Cytogenetic characterisation of the rare folate sensitive fragile site *FRA16A*

The *FRA16A* fragile site was first reported by Day *et al.* (1967) as a fragmentation of the short arm of chromosome 16. This fragmentation was expressed in 20% of the cells of a man who was also carrying a translocation of chromosome 18. From within the South Australian population, four apparently unrelated *FRA16A* families have been ascertained. The individuals from these families who were initially referred for cytogenetic testing presented with a variety of phenotypic abnormalities, such as mental retardation, Laurence-Moon Biedl syndrome with a lymphoreticular malignancy, and an unusual facial appearance with seizures (Sutherland and Hecht, 1985). The diversity of abnormalities in the index cases, and the normality of other family members with the fragile site, suggested that *FRA16A* expression is not associated with an abnormal phenotype, at least in heterozygotes. A large extended *FRA16A* pedigree (Fig. 4.7), originally ascertained because the proband was affected with Laurence-Moon Biedl syndrome and a lymphoreticular malignancy, was used in linkage studies to map the

fragile site relative to genetic markers in the 16p13.11 region, thus forming a physical anchor for the genetic map (Simmers *et al.*, 1987; Callen *et al.*, 1989; Kozman *et al.*, 1992).

1.3.7 Cytogenetic expression of the common aphidicolin inducible fragile site *FRA16D*

FRA16D is the second most highly expressed aphidicolin inducible common fragile site after *FRA3B*, which is located on chromosome band 3p14. The third most highly expressed fragile site is *FRA16A*, which is located on band Xp22.31. These three fragile sites account for at least 30% of the breaks induced by aphidicolin in human chromosomes (Glover *et al.*, 1984; Craig-Holmes *et al.*, 1987).

1.3.8 Physical mapping of the *FRA16A* region

An early linkage and cytogenetic breakpoint map spanning the 16p13.11 region (Callen *et al.*, 1989) positioned *FRA16A* relative to DNA markers and somatic cell hybrid breakpoints in the region. A later physical map of human chromosome 16 (Callen *et al.*, 1992) contained many more markers and breakpoints, and thus enabled more detailed mapping of the fragile site region. The basis for this more detailed cytogenetically based physical map was a panel of 54 mouse/human somatic cell hybrids, with each hybrid containing a different translocated or rearranged section of human chromosome 16. The mapping of 235 DNA markers with respect to the hybrid breakpoints, four fragile sites, and the centromere led to the construction of a physical map of the chromosome, with an average of 1.6 Mb of DNA between breakpoints (Callen *et al.*, 1992).

Mapping of anonymous DNA markers with the hybrid panel localised the markers 16XE81 (D16S287), VK20 (D16S96) and 1.79 (D16S79 A and B) in the vicinity of

FRA16A (Callen *et al.*, 1992). These markers were utilised in the initial stage of the *FRA16A* project.

1.3.9 Duplication of DNA sequences within the *FRA16A* region

All markers that localised to within the interval containing *FRA16A* were subjected to fluorescent *in situ* hybridisation (FISH) analysis against chromosomes expressing the fragile site. Unusual FISH results were obtained with the probe 1.79, as signal was detected on both sides of the fragile site. Another anomalous finding was that 1.79 detected two separate *Taq* I polymorphisms in genomic DNA: locus A with 1.6 kb and 3.4 kb alleles; and locus B with 4.5 kb and 9.5 kb alleles. There were, however, no *Taq* I restriction sites found within the probe sequence to account for the existence of the two independent restriction fragment length polymorphisms found. Additionally, a cosmid spanning 1.79 sequences contained only a 4.5 kb locus B *Taq* I fragment, and no locus A fragments. When taken together with the FISH results, this finding was strongly suggestive of the existence of duplicated 1.79 sequences (corresponding to the A and B loci), with a copy located on each side of *FRA16A* (Callen *et al.*, 1992). Genetic linkage studies mapped locus D16S79A as distal to locus D16S79B (Kozman *et al.*, 1993). The presence of a duplication in the had to be included as a factor in the interpretation of experimental results obtained with all probes in the vicinity of *FRA16A*, as the extent of the duplicated region was unknown.

1.4 Fragile X Syndrome

The fragile X syndrome is an X-linked heritable disorder, and is the most common cause of familial mental retardation. The syndrome is associated with expression of the rare fragile site *FRAXA*, which is located in chromosome band Xq27.3. The prevalence of the fragile X syndrome is estimated to be 1:4000 or 2.4:10000 in males (Turner *et al.*, 1996). The disorder is the most prevalent cause of mental retardation after Down syndrome.

1.4.1 Clinical characteristics of the fragile X syndrome.

The fragile X syndrome has a clinical presentation that includes moderate to severe mental retardation, macroorchidism, large ears, prominent jaw, high pitched jocular speech, and behavioural problems. Males are generally more severely affected than females. About 30% of female carriers are mildly to moderately mentally retarded.

1.4.2 Recognition of the fragile X syndrome

Martin and Bell reported the existence of a pedigree demonstrating segregation of familial X-linked mental retardation in 1943. The segregation pattern of the inherited disorder in this family was subsequently recognised as identical to that of the fragile X syndrome. Therefore, the Martin-Bell syndrome is a synonym for the fragile X syndrome.

Initial observations of a marker X chromosome (i.e. *FRAXA* expression) in mentally retarded individuals were made by Lubs *et al.* (1969). His chromosome studies of mentally retarded members of an American family identified an unusual X chromosome with a non-staining gap near the end of the long arm. This observation was not repeated until 1976/1977, when a number of studies found the association between X-linked mental retardation and *FRAXA* expression to be relatively common (Giraud *et al.*, 1976; Harvey *et al.*, 1977; Sutherland, 1977). These studies definitively established the link between X-linked mental retardation and the marker X chromosome.

1.4.3 Cytogenetic expression of the marker X chromosome

The marker X chromosome (or *FRAXA*) was initially observed after culturing lymphocytes in a particular medium, titled TC199. Around the time the marker X was found in 1969, many cytogenetics laboratories switched to more modern culture media that was incapable of inducing folate sensitive fragile site expression, a major reason why the marker X chromosome was not observed again for five years.

Unlike modern culture media, which have been rigorously formulated, medium TC199 does not contain thymidine and has a very low concentration of folic acid. Expression of the fragile X and other autosomal folate sensitive fragile sites was subsequently found to be specifically due to these conditions in cell culture (Sutherland, 1977).

1.4.4 The unique inheritance pattern of the fragile X syndrome

The mode of inheritance of the fragile X syndrome family described by Lubs *et al.* (1969) was consistent with an X-linked recessive pattern. However, subsequent studies of the inheritance of mental retardation in fragile X syndrome families revealed a more complex picture.

A number of hypotheses were put forward to explain the departures from classical X-linked inheritance, but none satisfactorily accounted for all the observed discrepancies.

1.4.5 Departures from classical X-linked inheritance

A number of studies on fragile X syndrome pedigrees reported transmission of the disorder through males (Webb *et al.*, 1981a, Camerino, 1983; Neilsen *et al.*, 1981; Jacobs *et al.*, 1983; Froster-Iskenius *et al.*, 1984; Arinami *et al.*, 1986; Webb *et al.*, 1986; Voelckel *et al.*, 1988). Sherman *et al.* (1984, 1985) calculated the penetrance of the disorder to be 80%, or in other words, 20% of males inheriting the mutation did not express it either phenotypically or cytogenetically. About one third of females heterozygous for the fragile X syndrome mutation were reported to have the phenotype to some degree, with mental handicap ranging from borderline to severe (Brown *et al.*, 1978; Howard-Peebles *et al.*, 1979a, 1979b; Turner *et al.*, 1980; Webb *et al.*, 1982,

Turner and Jacobs, 1984; Sherman *et al.*, 1985; Sutherland and Hecht, 1985). These observations were inconsistent with an X-linked recessive model of inheritance for the disorder.

Parental sex dependent penetrance of the fragile X syndrome phenotype was reported by Lubs *et al.* (1984a, 1984b) and Sherman *et al.* (1984, 1985). The daughters of transmitting males were rarely mentally impaired, but phenotypic expression in the daughters of carrier mothers was common. Additionally, the penetrance of the fragile X syndrome phenotype in both male and female offspring was found to be related to their carrier mother's phenotype. Intellectually impaired carrier mothers were observed to have a significantly higher penetrance in their offspring than unimpaired carriers. (Sherman *et al.*, 1985)

As the variation of the phenotypic expression in males and females within the fragile X syndrome pedigrees was inconsistent with either X-linked recessive or dominant modes of inheritance, Mulley and Sutherland (1987) proposed an X-linked dominant model with incomplete penetrance to account for the observations.

1.4.6 Dependence of fragile X syndrome penetrance upon position in pedigree

Sherman *et al.* (1984, 1985) compared the penetrance of the fragile X syndrome phenotype between successive generations. This study found that there is a higher level of penetrance in the sons of the unaffected daughters of normal transmitting males (NTMs) (74%) than in the sons of the unaffected mothers of NTMs (18%). This observation could not be explained in terms of classical genetic theory, as all females in a family carrying the mutation would be assumed to have identical genotypes, and therefore a similar proportion of affected sons. The phenomenon of increased penetrance in successive generations was termed the "Sherman Paradox".

1.4.7 High mutation rate in sperm and absence of new mutations in ova at the fragile X locus

Segregation analysis of 206 fragile X syndrome pedigrees, assuming an X-linked dominant with incomplete penetrance model for the disorder, calculated that none of the affected males represented a new mutation. (Sherman *et al.*, 1984; 1985). This result was in contrast with the situation for X-linked recessive lethal genes, where one third of affected males are expected to be the result of a new mutation (Haldane, 1948)

Two models of mutation, containing one or two steps, were proposed by Sherman *et al.* (1985) in order to explain the apparent absence or rarity of sporadically affected males. The single step model assumed that the mutation events at the fragile X locus occur only in sperm. On the basis of this model, the mutation rate was estimated to be 7.2×10^{-4} per male gamete per generation. The two step model accounted for the absence of sporadically affected males by limiting the second mutation step to females. The first mutation step of the model was neutral and occurred with the same frequency in egg and sperm. These assumptions led to a calculated mutation rate of 2.4×10^{-4} per gamete per generation. The mutation rates calculated on the basis of the model assumptions were the highest for any known human genetic disease locus. The result of a further study (Sherman *et al.*, 1988) contradicted the previous finding, as it estimated the proportion of sporadic cases among affected males to be 24%.

Another theory proffered to explain the high prevalence of the disorder was that normal female carriers and transmitting males possess an increased level of fertility. On the basis of this proposal the mutation rate was calculated as between 1.10×10^{-4} and 2.05×10^{-5} per gamete per generation (Vogel, 1984).

1.4.8 Models of fragile X syndrome inheritance

In fragile X syndrome pedigrees there are three types of fragile X mutation carriers: transmitting males with daughters who are always mentally normal; affected males who do not reproduce; and unaffected or mildly affected females who may have affected children. These observations led to the suggestion by a number of authors that inheritance

of the fragile X syndrome is a two step process (Pembrey *et al.*, 1985, Jacobs *et al.*, 1985: Sherman *et al.*, 1985). The basis of this proposal was that the fragile X mutation consisted of two forms: a premutation state with no phenotypic effect, that is found in carrier females and NTMs, and a full mutation state found in mentally impaired individuals. Additionally, conversion of a premutation to a full mutation would only occur in an egg. Based on this proposal, the initial mutation rate was estimated to be 1.67×10^{-4} per gamete per generation (Winter, 1987).

Mechanisms for conversion of a premutation to a full mutation were proposed by a number of authors. Pembrey *et al.* (1985) and Sherman *et al.* (1985) suggested the existence of an inherited chromosome rearrangement that could generate a genetic imbalance through a recombination event at meiosis. Taking into account the culture conditions required for the induction of folate sensitive fragile site expression, Sutherland *et al.* (1985) proposed the existence of a repeating structure of polyd(AG)/polyd(TC) at the fragile X locus. This theory was extended by Nussbaum *et al.* (1986), who proposed that this repetitive sequence could be present on normal chromosomes, and become amplified through unequal crossover to produce a fragile X premutation. During meiosis in a female, recombination within the premutation could result in an elongated stretch of repetitive DNA representing the full mutation.

Nussbaum's hypothesis was refuted by the results of a subsequent linkage study by Winters and Pembrey (1986) using genetic markers flanking the fragile site. This study found (erroneously) a significant reduction in normal recombination rates across the fragile X locus in meioses giving rise to affected grandsons of normal transmitting males. An interpretation of this finding was that interference in a recombinatorial event might lead to the full fragile X mutation (Winter and Pembrey, 1986). However, later linkage studies clearly demonstrated that a reduction in recombination frequency at the fragile X locus was not a prerequisite for the production of a full mutation (Oberlé *et al.*, 1986; Goonewardena *et al.*, 1986; Mulley *et al.*, 1987; Suthers *et al.*, 1991a).

1.4.9 Correlations between non-random X chromosome inactivation and the severity of fragile X syndrome penetrance in females

A non-random X chromosome inactivation model was proposed by Jacobs *et al.* (1980) to account for the variation in severity of mental retardation in fragile X syndrome heterozygous females. According to this model, females with a greater proportion of active fragile X chromosomes would be more severely affected. Subsequent studies of the X-inactivation patterns in heterozygous females gave conflicting results or limited correlations with the model (Brown, 1990; Schmidt *et al.*, 1991; Jacobs *et al.*, 1986; Knoll *et al.*, 1984; Rosenberg *et al.*, 1991; Fryns *et al.*, 1984; Nielsen *et al.*, 1983).

1.4.10 Laird's hypothesis

An X-inactivation/imprinting model for the mechanism of inheritance and expression of the fragile X syndrome was proposed, in which the fragile X mutation acts as a block to the normal process of X-chromosome reactivation in the ovary (Laird, 1987). The following postulates formed the basis of the model: (1) the fragile X mutation is a potential cis-acting local block to the process of X chromosome reactivation that occurs in a female prior to oogenesis; (2) a cycle of X chromosome inactivation and incomplete reactivation in a female results in local heritable 'imprinting' of a fragile X chromosome; (3) the fragile X syndrome results from transcriptional inhibition that accompanies this chromosome imprinting; (4) transmitting males and some heterozygous females are unaffected because their fragile X chromosomes were not imprinted by a cycle of maternal inactivation and incomplete reactivation in a previous generation; (5) variable expression in females with an imprinted fragile X results from non-random inactivation of X chromosomes in somatic cells that affect the phenotype.

According to Laird's hypothesis, NTMs and carrier females carried an unimprinted fragile X mutation with the potential to block reactivation of a previously inactivated fragile site region and affected individuals carried an imprinted mutation that had been through a cycle of inactivation and incomplete reactivation. Penetrance values of 78% for male offspring and 38% for female offspring of heterozygous females were estimated on

the basis of the model. These values compared favourably with the penetrances calculated by Sherman *et al.* (1985), of 78% and 36%. Laird's hypothesis did not, however, suggest an explanation for the "Sherman paradox".

Late replication of the fragile X region had been found to occur on X chromosomes of affected individuals (Yu *et al.*, 1990; Hansen *et al.*, 1993). Laird (1987) proposed that this late replication of DNA at the fragile site was the basis for the local block to complete reactivation of the X chromosome. This incomplete reactivation was the cause of inappropriate methylation of DNA sequences at the locus, resulting in transcriptional inactivation of genes in the region. Conflicting with this proposal's predictions, expression studies of a number of genes in the region in fragile X patients detected no significant alteration in transcriptional activity. (Khalifa *et al.*, 1990, Steen *et al.*, 1991).

1.4.11 Additions to Laird's hypothesis

Laird *et al.* (1990), extended the X-inactivation/imprinting hypothesis for fragile X syndrome inheritance (Laird, 1987) in order to explain the observed differences in penetrance between brothers and grandsons of NTMs, previously reported as 18% and 74%, respectively (Sherman *et al.*, 1984, 1985). According to the extended model, fragile X chromosome imprinting occurred in about half the primary oocytes of a female, which was the expected frequency if X-inactivation was the initial step in fragile X mutation imprinting.

A biological explanation for the Sherman paradox was proposed after the analysis of fragile X syndrome family data according to the assumptions of Laird's extended model (1990). The pattern of inheritance was interpreted as being due the existence of a very limited number of oogonial progenitor cells, estimated to be two at the time of the initial event leading to chromosome imprinting. The estimated penetrance values predicted by

the model were 20% for brothers of transmitting males and 80% among grandsons of transmitting males, which was in line with the observed values.

The rare daughters of fragile X males are always phenotypically normal, and do not cytogenetically express the fragile site. In order to explain these features of fragile X inheritance, Laird (1991) proposed that erasure of the imprint occurs through male transmission, with re imprinting occurring in the daughter's primary oocytes.

1.4.12 Relevance of Laird's hypothesis to the molecular characteristics of *FRAXA*

When the fragile X syndrome disease locus was cloned and characterised, the X-inactivation model for fragile X syndrome inheritance (Laird, 1987) was modified to account for the molecular properties of the relevant DNA sequences.

Upon examining the stability of the presumed imprinted state in fragile X pedigrees, Follette and Laird (1992) concluded that the imprinting had a stability of about 96% through female meiosis, and also that 'erasure' may occur. The existence of this proposed imprint 'erasure' in females was demonstrated by fragile X pedigree analysis, utilising locus specific DNA probes to directly determine the different mutational states of fragile X alleles. Based on the results of this study, it was proposed that the deleterious state of the fragile X allele was not the result of any additional change in the DNA beyond the initial fragile X mutation. In other words, expression of the fragile X phenotype was due to hypermethylation alone, as predicted by the X-inactivation/imprinting model of fragile X inheritance (Laird, 1987).

The relevance of Laird's hypothesis to the molecular findings at the locus is further reviewed in section 1.4.34, in order to provide continuity with respect to publication dates.

1.4.13 Genetic mapping of the fragile X syndrome locus

Early linkage analysis of fragile X syndrome families with markers on the X chromosome found close linkage of the fragile X syndrome locus with the genes for G6PD deficiency, colorblindness and factor IX (Filippi *et al.*, 1983; Szabo *et al.*, 1984; Camerino *et al.*, 1983).

Between the years of 1984 and 1990, many DNA markers identifying RFLPs close to *FRAXA* were isolated and the corresponding loci ordered with respect to the fragile X locus and each other by linkage analysis (Boggs and Naussbaum, 1984; Drayna *et al.*, 1984; Drayna and White, 1985; Oberlé *et al.*, 1986; Buchanan *et al.*, 1987; Hofker *et al.*, 1987; Mulley *et al.*, 1987; Hyland *et al.*, 1989a, 1979b; Suthers *et al.*, 1989; Vincent *et al.*, 1989; Rousseau *et al.* 1990; Oostra *et al.*, 1990). These genetic markers were useful tools for carrier detection, confirmation of cytogenetic testing, and prenatal diagnosis in fragile X families.

Brown *et al.* (1988) performed multi-locus linkage analysis of 147 fragile X syndrome families. Linkage heterogeneity between loci flanking the fragile X region was demonstrated by this study, and also that of Thibodeau *et al.* (1988), leading to the proposal (suggested previously Winter and Pembrey, 1986) that the fragile site affects recombination rates in its vicinity. However, other multi-point linkage analysis studies found no evidence of linkage heterogeneity (Veenema *et al.*, 1987; Suthers *et al.*, 1991a), indicating that the existing genetic map of the *FRAXA* region was accurate and could be used for genetic counselling purposes and as a tool for positional cloning of the fragile X syndrome locus.

1.4.14 Physical mapping of the fragile X region

Poustka *et al.* (1991) described a physical map of rare cutter DNA restriction fragments covering the region of the X chromosome from Xq27.2 to the telomere, and thus spanning *FRAXA*. The map was assembled by determining the relative positions of overlapping DNA restriction fragments (usually hundreds of kilobases in size) produced by digestion of human genomic DNA (uncloned) with rare cutter restriction

endonucleases. Pulsed field gel electrophoresis and Southern blot hybridisation enabled the sizing and localisation of these fragments. The order and position of the DNA markers throughout the region was determined by their localisation to specific fragments.

The advent of YAC technology accelerated the physical mapping of the *FRAXA* region. Wada *et al.* (1990) and Schlessinger *et al.* (1991) established YAC contigs across the Xq24-28 region.

A panel of somatic cell hybrid cell lines containing translocated X chromosomes with breakpoints near *FRAXA* were of use in the rapid identification of DNA markers in close proximity to the fragile site (Suthers *et al.*, 1991b). DNA markers were mapped either proximal or distal to the fragile site by *in situ* hybridisation to metaphase chromosome spreads (Suthers *et al.*, 1991b), thereby providing points of reference from which chromosome 'walks' could be initiated.

Warren *et al.* (1987; 1990) demonstrated physical breakage at the fragile site under conditions of thymidine stress, with the production of somatic cell hybrids containing translocated fragile X chromosomes with thymidine stress induced breakpoints at or very near the fragile site. Two translocated chromosomes in these hybrids were subsequently found to break at the fragile site, and were therefore useful tools for cloning the fragile X (Yu *et al.*, 1991; Verkerk *et al.*, 1991).

1.4.15 Identification of a hypermethylated CpG island in fragile X syndrome males

An initial indication of the abnormal nature of DNA sequences at the fragile site was provided by comparative pulsed field gel electrophoresis studies of DNAs from normal and fragile X males. Vincent *et al.* (1991) reported that a DNA marker mapping near *FRAXA* hybridised to 620 kb *Bss*H II and *Sac* I fragments, and a 120 kb *Eag* I fragment in both normal males and normal transmitting males. These fragments appeared to be absent, or to have reduced intensity in mentally retarded fragile X syndrome males. The authors concluded that hypermethylation (of at least one site for each restriction enzyme) was the most likely explanation for these findings. A similar study by Bell *et al.* (1991)

produced similar results and conclusions. The evidence of abnormal methylation (or hypermethylation) at the fragile site was considered, at that time, to be consistent with Laird's X-inactivation imprinting hypothesis (Laird, 1987). The extensive pulsed field mapping of the region performed as part of the studies also demonstrated that the fragile X syndrome was probably not associated with large structural rearrangements of the DNA in band Xq23.7 (Bell *et al.* 1991).

1.4.16 Cloning of DNA sequences across *FRAXA*

YAC cloning technology, as well as accelerating the physical mapping of the *FRAXA* region, enabled more rapid positional cloning of disease loci than was possible with previous methods. It was with the use of this new technology that in 1991, 22 years after recognition of the association of *FRAXA* with mental retardation, the fragile X syndrome locus was independently cloned by three research groups within a short period of time.

Kremer *et al.* (1991a) reported the identification and restriction mapping of a 275 kb YAC (XTY-26) spanning *FRAXA*. The insert of this YAC was derived from the DNA of a fragile X patient. XTY-26 sequences appeared to span *FRAXA*, as direct fluorescent *in situ* hybridisation (FISH) of YAC DNA to metaphase chromosomes expressing the fragile site produced both proximal and distal signal. The localisation of two DNA markers previously mapped on opposite sides of the fragile site to within the YAC insert confirmed the initial FISH result. On the basis of these results YAC XTY-26 was considered a strong candidate to contain the *FRAXA* fragile site sequence, and was therefore subsequently mapped and subcloned in order to isolate the fragile X syndrome locus (Yu *et al.* 1991).

In order to accomplish the above goals, XTY-26 DNA was subcloned into a lambda vector. Following this step, a contig of the subclones spanning the fragile site was established. The location of each subclone (as proximal, central or distal) with respect to the fragile site was determined by FISH, and those appearing to span the fragile site were used as hybridisation probes against a panel of Southern blotted somatic cell hybrid DNAs. Two of the somatic cell hybrids, micro 21D and QIX, were thought to contain

translocated X chromosomes with breakpoints at the fragile site. These cell lines were derived from the fragile X parent cell line Y75-1B-M1 which had been treated before cell fusion to promote chromosomal breakage at the fragile site (Warren *et al.*, 1988). Hybridisation of one clone in the lambda contig, $\lambda 5$, to *Eco* RI digested micro 21D and QIX DNA found both to contain translocation breakpoints occurring within a 5 kb *Eco* RI restriction fragment detected by this probe. An *Eco* RI fragment of 5.9 kb was found in the parent Y75-1B-M1 cell line DNA (representing a fragile X chromosome), and a 5 kb fragment in hybrid cell line CY3 DNA (representing a normal X chromosome). Subsequently, the $\lambda 5$ subclone was used to probe *Eco* RI digested DNAs from normal and fragile X males. A 5 kb band was found for all normal individuals, whereas larger bands and smears varying in of sizes up to 7.5 kb were found for all fragile X males.

Within the 5 kb *Eco* RI fragment of $\lambda 5$, the region of altered DNA fragment mobility associated with the fragile X syndrome was localised to a 520 bp *Pst* I-*Nhe* I fragment containing repetitive human genomic sequences. Hybridisation of this fragment to *Pst* I digested DNAs from normal and unrelated fragile X individuals identified 1 kb bands in normal individuals, and from one to four bands varying from 1.5 to 3.5 kb in size or multiple bands appearing as a smear, in fragile X individuals.

Pedigree analysis of fragile X syndrome families with a locus specific probe (from $\lambda 5$) demonstrated segregation of altered mobility bands with the fragile X genotype (Yu *et al.*, 1991). Restriction fragments from the DNAs of non-penetrant transmitting males and carrier females, as well as affected males, were found to display altered mobility. It was observed that alterations in mobility occur within fragile X families, showing increase in size from generation to generation when females, but not males, transmitted the genotype. The restriction fragment sizes detected in DNA from affected individuals were, in general, found to be larger than those of carriers, but a clear relationship between the degree of increase in fragment size and phenotype was not apparent (Yu *et al.*, 1991).

On the basis of these findings, Yu *et al.* (1991) proposed that the degree of size increase in the unstable sequence segregating with the fragile X genotype may have a modulating effect on *FRAXA* expression and the associated syndrome. It was further speculated that

the molecular basis for the fragile X syndrome phenotype may be a result of the proximity of the unstable sequence to a CpG island, as the expression of the gene associated with the CpG island could be affected by an expansion.

Similar findings to those of Yu *et al.* (1991), as well as evidence of the occurrence of hypermethylation at the fragile X, were reported by Oberlé *et al.* (1991). In this study, the starting point for cloning the fragile X syndrome locus was the identification of two YACs (y141H5 and y209G4) mapping within the fragile X region, and containing the CpG island previously shown to be methylated specifically in fragile X males (Bell *et al.*, 1991; Vincent *et al.*, 1991).

YAC DNA sequences from the CpG island region were subcloned, and the subclones mapped and used to probe Southern blots of DNA from normal and fragile X individuals. One subclone (StB12.3) was found to detect abnormal fragments of increased size in *Bgl* II and *Bgl* II - *Eag* I digests of the majority of fragile X DNA samples. StB12.3 was also found to detect an *Eco* RI band of 5.2 kb in normal DNAs and abnormal single fragments of 5.35 kb and 5.6 kb in DNAs from normal transmitting males. Further mapping of the fragile X mutation region localised the unstable region to a 550 bp *Ban* I fragment containing the *Eag* I, *Bss*H II and *Sac* II sites that were known to be methylated in affected individuals. Sequencing of the normal (550 bp) *Ban* I fragment identified the repetitive methylatable sequence (GCC)₁₀TCC(GCC)₉ (Oberlé *et al.*, 1991).

1.4.17 Pedigree analysis of fragile X families with probes specific for the mutated region.

Forty-nine fragile X syndrome family pedigrees were subject to analysis with a probe specific for the unstable region (Oberlé *et al.*, 1991). The purpose of this study was to ascertain whether there were consistent patterns of mutational change related to the parental origin of the mutation, the position in the pedigree, and the phenotypic expression. The observations that emerged from the study were: (1) normal transmitting males and their daughters have band size increases in the 150 to 500 bp range; (2) the mutation is passed from a normal transmitting male to his daughters either unchanged, or

with a size increase of up to 200 bp; (3) in the next generation, about 80 % of males and females inheriting the fragile X chromosome have an average increase of 1.5 kb to 2.5 kb in the size of the abnormal fragment. Multiple discrete bands or a smear of bands were frequently observed, indicating fragment size heterogeneity.

The appearance of fragments with only a small size increase was associated with a lack of clinical expression and absence or low levels of fragile site expression. Exceptions to this rule were affected 'mosaic' individuals in whom a near normal fragment coexisted with larger ones of full mutation size.

It was concluded from the study that the detection of abnormally large fragments is a diagnostic indicator for the fragile X syndrome, as this feature, as well as the presence of abnormal methylation, is associated with phenotypic expression (Oberlé *et al.*, 1991).

Yu *et al.* (1992) determined the fragile X genotype, or the degree of trinucleotide repeat amplification, in individuals from 49 fragile X syndrome pedigrees. The aim of this study was to assess the instability of the trinucleotide repeat during transmission, as well as to determine a possible correlation of repeat expansion size and phenotype. The study determined that all affected males have abnormal restriction fragments, with an increase in size of more than 0.6 kb over normal. Males and females with fragment size increases of less than 0.6 kb were unaffected and did not usually have cytogenetic expression of the fragile site. Females with larger fragment size increases had frequent cytogenetic expression of the fragile site, and could be either mentally normal or affected. These observations implied that factors other than amplification might contribute to the phenotypic expression in female heterozygotes. The authors suggested that these factors could be attributed to normally occurring methylation differences (between chromosomes) within the *FRAXA* region of the X chromosome and/or non-random X inactivation in cells of target tissues (Yu *et al.* 1992).

1.4.18 Classification of mutations at the fragile X locus

Two classes of expanded fragments at the fragile X locus were distinguished at the molecular level. Normal transmitting males and some females, including all daughters of normal transmitting males, were found to have small expansions of 150 - 500 bp. In these individuals the adjacent CpG island was found to be unmethylated on the active X chromosome (as it is when no mutation is present), and methylated on inactive X chromosomes as in normal females. There was little or no clinical expression of the disorder and very low or zero fragile site expression in these individuals. In affected males or females with high fragile site expression, the expansion was found to be 1000 - 3000 bp or larger and often heterogeneous in size. In these individuals the CpG island was found to be completely methylated (Oberlé *et al.*, 1991).

The first class of fragment sizes were termed premutations, with a second class involving phenotypic expression being termed full mutations (Oberlé *et al.*, 1991, Rousseau *et al.*, 1991). These terms were originally used by Pembrey *et al.* (1985) and Nussbaum *et al.* (1986), who proposed that the inheritance of the fragile X syndrome was based on the existence of the two mutational states. The transition from a premutation to full mutation state was proposed to occur only after passage through a female, possibly as the result of a sequence amplification. This postulated event was subsequently found to occur at the fragile X locus, accounting for the usual inheritance patterns of the fragile X syndrome.

1.4.19 Sequence characterisation of a *Pst* I fragment containing the fragile X syndrome mutation

Kremer *et al.* (1991b) reported the full sequence (although this sequence was subsequently found to contain a deletion by Fu *et al.*, 1991) of a 1 kb *Pst* I fragment to which the fragile X unstable region associated with the fragile X syndrome had previously been localised (Yu *et al.* 1991). Within the sequence was found a CCG repeat tract and a CpG rich region with seven methylatable rare cutting restriction endonuclease sites, three of which were known to be methylated in fragile X males.

High resolution mapping by Southern blot hybridisation localised the instability to a region between two restriction sites immediately adjacent to the CCG repeat tract. FISH analysis with flanking probes confirmed this result, and localised the CCG repeat tract to the chromosomal gap (fragile site). In addition, the translocation breakpoints of micro 21D and QIX (Warren *et al.*, 1988) were precisely mapped to the CCG repeat tract.

The molecular characterisation of *FRAXA* did not fully clarify the relationship between the repeat tract expansion and cytological expression of the fragile site. However, it was proposed that the composition of the repeat sequence might render its replication sensitive to depletion of either dCTP or dGTP (as a consequence of high dTTP levels). This could result in the formation of single stranded DNA which might fail to package for mitosis and thus appear as a chromosomal gap or fragile site (Kremer *et al.*, 1991b).

1.4.20 Identification of a gene (*FMRI*) containing the *FRAXA* CCG repeat tract

The finding of a fragile X related CpG island that is hypermethylated in fragile X patients pointed to the existence of a gene, conceivably inactivated by methylation (thus producing the fragile X syndrome phenotype), in the genomic DNA adjacent to the CCG repeat tract and CpG island.

Verkerk *et al.* (1991) reported the isolation and characterisation of *FMRI*, a brain expressed gene. The *FMRI* transcript was found to include sequences homologous to those of the fragile X CpG island (including the CCG repeat tract) and the region 3' of this.

The initial step in the identification of *FMRI* was the isolation of a YAC (y209G4) spanning both the fragile site and a fragile X associated breakpoint cluster region (Warren *et al.*, 1988). The CpG island previously identified as being hypermethylated on fragile X chromosomes was mapped to within y209G4. The YAC DNA was subcloned into a cosmid vector, and from the resulting library a cosmid contig spanning the fragile X CpG island was constructed. A 5.1 kb *Eco* RI fragment subclone from a cosmid in the contig detected *Eco* RI fragments of increased size in all fragile X samples screened.

Cosmid subclones from the contig were used as hybridisation probes to screen a normal human fetal brain cDNA library. A positive cDNA clone was identified by this method, and used to screen for overlapping clones. The overlapping clones were again used to screen the cDNA library. Analysis of the resulting 3.8 kb cDNA contig identified a reading frame that remained open at the 5' end of the contig, indicating the 5' end of the mRNA was not represented. A repetitive sequence consisting of 28 CCG triplets interspersed with 2 AGG triplets was present near the 5' end of the cDNA contig sequence. A poly-A tract sequence was not found within the most 3' cDNA of the contig, therefore it could not be ascertained (at that time) whether the clones represented the total mRNA. A partial polypeptide of 657 amino acids, including a polyarginine tract encoded by the CCG repeat tract (translated as CGC), was predicted from an ORF sequence. A protein database search with the encoded protein sequence revealed no significant homologies to previously characterised genes.

Verkerk *et al.* (1991) assumed that the *FMRI* CCG repeat tract was translated. Mandel *et al.* (1992) stated that if this were the case, the functional role of the *FMRI* must be unaffected by the extremely variant number of arginine residues (~6 - 50) in the normal population. Subsequently, however, it was established that the *FMRI* CCG repeat was more likely to be 5' (upstream) to the true initiation AUG codon of the *FMRI* gene (Caskey *et al.*, 1992).

Northern blot analysis with one *FMRI* cDNA insert as a probe detected a 4.8 kb mRNA in human brain, placenta and lymphocytes, indicating the 3.8 kb cDNA contig did not represent the entire mRNA sequence. Hybridisation of the cDNA clone to DNAs from a number of different organisms detected bands for most species, indicating a high degree of sequence conservation at the DNA level.

Studies of the transcriptional orientation of the characterised gene found that the fragile X related CpG island was at the 5' end of the gene, within the putative promoter region. Further evidence of the involvement of *FMRI* in the fragile X syndrome was provided by the finding that the CCG repeat within the cDNA sequence is the fragile X related CCG

repeat. The gene was thus named *FMRI*, an abbreviation of fragile X mental retardation 1 (Verkerk *et al.*, 1991).

1.4.21 Absence of *FMRI* gene expression in fragile X patients

To further delineate the gene's role in the fragile X syndrome phenotype, Pieretti *et al.* (1991) studied *FMRI* expression in normal, carrier and affected individuals. In this study, leukocytes and lymphoblastoid cell lines of normal, carrier and fragile X individuals were used as sources of mRNA template for reverse transcription and PCR amplification (RT-PCR) of *FMRI* sequences with gene specific primers. Primers directed toward the *HPRT* gene were included as an internal standard in all RT-PCR reactions. Reduced levels or absence of *FMRI* mRNA were found in twenty fragile X males, with total methylation of the *Bss*H II site within the *FMRI* CpG island completely correlating with the total absence of expression found in 16 of these individuals. *FMRI* expression was found in normal males, normal members of fragile X pedigrees and heterozygous females.

A fragile X syndrome associated phenotype is found in 30% of heterozygous females, however the phenotypic severity did not correlate with the reduction in *FMRI* expression found in these females. To explain this finding, it was suggested that *FMRI* expression in affected heterozygous females is impaired in a relevant tissue (e.g. the brain) and/or at a particular stage of development by preferential inactivation of the normal X chromosome (Pieretti *et al.*, 1991).

In four of the twenty fragile X males previously mentioned, expression of *FMRI* by RT-PCR was found. In three of these cases a mosaic pattern of expanded size fragments, including some of normal and premutational size, were observed (with Southern blot hybridisation analysis). Partial methylation of the *Bss*H II site was associated with this mosaicism. The fourth individual did not have any normal or premutational size fragments (only the full mutation), but displayed only partial methylation of the *Bss*H II site. This finding indicated that the presence of an unmethylated *Bss*H II site (and therefore probably the entire CpG island) is sufficient for *FMRI* transcription, even in the presence of an expanded repeat.

Based on the results discussed above, Pieretti *et al.* (1991) concluded that there is a relationship between absence of *FMRI* expression and the presence of an amplified CCG repeat adjacent to a methylated *BssH* II site. Two alternative explanations for this were proposed: 1) *FMRI* mRNA expression is directly altered by the presence of the amplified CCG repeat region found in fragile X patients, resulting in transcriptional termination or production of an unstable mRNA; 2) the increased methylation of restriction sites within the *FMRI* CpG island is the result of the presence of a increased number of CCG repeats, and if the CpG island is a regulatory region for *FMRI*, the hypermethylation causes a reduction in transcriptional activity.

As previously stated, four of the fragile X males studied by Pieretti *et al.* (1991) were found to have near normal levels of *FMRI* mRNA (in leukocytes and lymphoblastoid cells). These findings suggested that lack of *FMRI* expression could not fully account for the fragile X phenotype. Explanations proposed to account for these results were: (1) the level of expression of *FMRI* in the patients did not reach the necessary threshold to provide normal levels of the protein in the relevant tissue(s) at the appropriate time in development; (2) the *FMRI* levels and DNA methylation found in leukocytes are not representative of other body tissues or other developmental stages; (3) detection of *FMRI* mRNA does not necessarily indicate the protein is normally expressed or has retained full activity.

The conclusion drawn from the study was that lack of *FMRI* expression is at least partly responsible for the fragile X syndrome phenotype, although the possibility that the regulation of other genes in the region is altered due to methylation could not be excluded (Pieretti *et al.*, 1991).

1.4.22 Localisation of the fragile X syndrome heritable unstable element to a CCG repeat tract

Individuals affected with fragile X syndrome were found to have abnormal variation in DNA restriction fragment size at the fragile site locus (Yu *et al.*, 1991; Oberlé *et al.*, 1991). The most likely cause was considered to be amplification of a polymorphic CCG

repeat present at the locus in all individuals (Kremer *et al.* 1991b). The studies of Fu *et al.* (1991) and Yu *et al.* (1992) utilised PCR methods to directly confirm that the instability and polymorphism were due to size variations of the CCG repeat tract.

1.4.23 Polymorphism of fragile X CCG repeat tract length in normal individuals

Southern blot analysis had previously found polymorphic variation of restriction fragment size at the fragile X locus among normal individuals (Kremer *et al.* 1991b). In order to obtain a higher degree of fragment size resolution, PCR was used to amplify sequences across the *FMRI* CCG repeat tract (Fu *et al.*, 1991). DNAs from unrelated normal individuals, representing 492 X chromosomes from the Caucasian, African-American, Hispanic and Asian races, were used as template for the reactions. PCR product sizes ranging from 239 bp to 383 bp and spaced at 3 bp intervals were amplified. These corresponded to variation of a triplet repeat from 6 to 54 units. The observed number of alleles in the overall population was 31, with a heterozygote frequency of 63%. A 29 CCG repeat allele was the most abundant, being present on 30% of chromosomes. The pattern of distribution of fragile X locus alleles was not significantly different between the races represented in the study (Fu *et al.*, 1991).

1.4.24 Studies of the *FMRI* CCG repeat tract on fragile X chromosomes

The development of a protocol for successful PCR amplification across the *FMRI* CCG repeat tract by Fu *et al.* (1991) allowed a more precise study of the premutation class of mutations in fragile X families. When PCR products of premutation size were successfully amplified from an affected male, Southern blot hybridisation analysis of DNA from the same individual always revealed mosaicism for fragment sizes, with both full and premutation fragments. Amplification of full mutation alleles produced a very low yield, only detectable by Southern blot analysis of the PCR product.

Fifty-six premutation alleles from individuals belonging to fragile X families were assayed. The PCR products amplified varied in size from 377 bp to 800 bp, representative

of 52 to 193 CCG repeats. A subsequent study found all normal transmitting males (who carry premutations) to have less than 240 copies of the CCG repeat (Yu *et al.*, 1992).

1.4.25 Number of CCG repeats is indicative of the level of intergenerational instability

Comparisons of normal and premutation range alleles found the size range of premutation alleles to exceed the size of all but one normal allele. The two largest normal alleles found were 46 and 54 CCG repeats in length, and the smallest premutation allele 52 CCG repeats in length. However, the normal 54 repeat allele was found to be unstable through all transmissions. The size ranges of normal and premutation alleles were therefore found to overlap.

Genotyping fragile X and normal families for alleles at the fragile X locus showed normal alleles of up to 46 repeats to be stable through the 75 transmissions scored. Alleles of 52 repeats and larger were unstable through all transmissions. Both increases and decreases in size were observed in the unstable alleles, with the former being far more common than the latter (Fu *et al.*, 1991).

1.4.26 Risk of expansion to the full mutation correlates with CCG repeat number: relationship to the Sherman paradox

Fu *et al.* (1991) studied the maternal transmission of premutation alleles ranging in size from 52 to 113 CCG repeats. Premutation alleles were found to expand to full mutations in 44 of 63 transmissions, with the frequency of expansion depending upon the size of the premutation allele. Alleles of 90 repeats or larger expanded to full mutations in all cases, and alleles of 59 repeats and below were not observed to increase in size. For alleles of between 59 and 90 repeats, the intermediate risks of expansion to a full mutation were found to vary from 17% in the 60 repeat range to 77% in the 70 to 86 repeat range. These findings therefore showed that the risk of expansion to a full mutation is directly correlated with the size of an allele.

Fu *et al.* (1991) considered that the results of the study provided resolution of the "Sherman paradox" at the molecular level. Sherman *et al.* (1984, 1985) had previously found the risk of phenotypic effects for an individual in a fragile X family to be dependent on the position of that individual in the pedigree. On the basis of empirical data from fragile X pedigrees, varying risks were assigned to members of each generation, with brothers of normal transmitting males given 9% risk, grandsons of normal transmitting males given 40% risk, and great-grandsons given 50% risk. These empirical risk determinations appeared to fit the reported expansion risks of Fu *et al.* (1991) for the different classes of alleles, as the 9% risk for brothers of NTMs would predict that mothers of NTMs are likely to have alleles in the 60 - 69 repeat range, with a 17% risk of expansion and therefore an 8.5% chance of an affected boy. Furthermore, the 40% risk for grandsons of NTMs would predict that daughters of NTMs would have alleles in the 70 - 89 repeat range, with a 77% risk of expansion and a 38.5% risk of having an affected boy. Once the number of repeats reaches 90 or above, the risk of expansion becomes 100%, and the consequent risk of having an affected boy is 50%. The expansion risks therefore provided an explanation for the empiric risk values of Sherman. They also predicted that alleles in the premutation range would undergo gradual increases in size, thereby accounting for the increasing risk in subsequent generations. The observations of variable risks for phenotypic expression within fragile X pedigrees were therefore completely accounted for by the finding that the risk of expansion to a full mutation is dependent upon the size of the carrier mother's premutation allele.

Studies of the transmission of the expanded unstable region through fragile X pedigrees by Yu *et al.*, (1992), found that marked amplification only occurs with maternal transmission of a mutation. Other observations from this study were that the risk of a carrier female having affected offspring increased in direct proportion to the length of her amplified sequence, whereas minimal, or no, amplification was associated with male transmission, whatever the size of amplified sequence carried. It was therefore concluded that the fragile X premutation is much less stable when transmitted by females than males.

Other studies established that: mutations in mothers and daughters of transmitting males are at different stages of progression (Yu *et al.*, 1992); and the length of the CCG repeat is correlated with the degree of meiotic instability (Fu *et al.*, 1991). In addition, the size of the amplified CCG repeat was observed to increase progressively through fragile X pedigrees (Yu *et al.*, 1992). These observations provided an explanation of why the daughters of transmitting males usually have larger amplifications than the mothers of transmitting males, and are therefore at greater risk of having affected sons. Yu *et al.* (1992) proposed that the pattern of transmission of the unstable element provided an explanation for the "Sherman paradox", in which the penetrance of the mutation, as measured by intellectual status, is greater in the offspring of the daughters of transmitting males than in the offspring of the mothers of transmitting males. This pattern of inheritance contradicted classical genetic theory, which dictated that the mothers and daughters of transmitting males should have an identical genotype at the fragile X locus, and therefore a similar ratio of affected and unaffected offspring.

1.4.27 Somatic variation at the fragile X locus

Somatic variation at the fragile X locus was studied by Southern blot analysis of fragile X male, carrier female and normal DNA from three different tissue sources: lymphocytes; lymphoblasts; and fibroblasts (Yu *et al.*, 1992). Somatic variation of restriction fragment size was found in the cultured lymphoblastoid cell lines of all 11 affected males tested, regardless of whether the lymphocyte sample from the same individual also displayed such variation. DNAs from eight normal individuals, and three carriers with small amplifications of 100 - 200 bp did not display somatic variation. However, clear evidence of somatic variation was found between the different tissues of a carrier with a larger amplification of 700 bp. Cytogenetic expression of the fragile X was evident in the cells of all individuals for whom somatic variation had been observed (Yu *et al.*, 1992).

DNA from 3 pairs of fragile X identical twins was analysed to establish the patterns of somatic variation. The band pattern of the fragile X mutation was strictly identical for

each pair of twins, indicating that this pattern must be established early in development, before zygote cleavage (Devys *et al.* 1992).

1.4.28 No evidence for new fragile X mutations

Rousseau *et al.* (1991), Smits *et al.* (1992) and Yu *et al.* (1992) performed independent fragile X family studies that found no evidence of new mutations at the locus (i.e. from normal allele ($n < 46$) to full mutation). To explain this result, Rousseau *et al.* (1991) suggested that this transition could not be made in one generation. Mandel and Heitz (1992), Morton and Macpherson (1992), and Yu *et al.* (1992) proposed that the generation of mutations may be continuous or multistep, rather than a simple one or two step process involving a premutation followed by a full mutation. An alternative proposal was that the full mutation causing the fragile X phenotype is derived from a predisposed state, which may consist of an initially minor amplification that becomes progressively larger in subsequent generations (Yu *et al.*, 1992).

1.4.29 Evidence of founder chromosomes and linkage disequilibrium at the fragile X locus

The fact that no new mutations were evident within fragile X pedigrees (Rousseau *et al.*, 1991; Smits *et al.*, 1992; Yu *et al.*, 1992) was inconsistent with the high mutation frequency that had been previously calculated from empirical data (Sherman *et al.*, 1984; 1985). In addition, founder effects were not expected to occur for X-linked diseases with a severely negative effect on male reproductive fitness (Imbert and Mandel, 1995). In order to resolve the contradiction between the observed and calculated mutation rates, normal and fragile X chromosome haplotype analyses were performed in order to determine whether the DNA segments containing normal and amplified CCG repeats were of divergent origin (Richards *et al.*, 1992).

Polymorphic AC repeat markers (FRAXAC1 and FRAXAC2) flanking, and within 10 kb of the fragile site, were used for the haplotype analysis. No recombination between the AC repeat or the fragile site loci had been detected and alleles at the two AC repeat loci

were found to be in linkage disequilibrium with each other. In addition, the frequencies of specific FRAXAC1 and FRAXAC2 alleles were found to be significantly different between normal X chromosomes and fragile X chromosomes.

The flanking markers were used to identify, as a haplotype, the intervening stretch of DNA encompassing the fragile X locus. Haplotypes were initially characterised in normal (CEPH) pedigrees by determining the co-inheritance of alleles at each locus. Subsequent analysis of fragile X DNAs found that distinct differences in the frequencies of certain haplotypes between fragile X and normal X chromosomes were evident. As an example, the three haplotypes AF, DD and DB together accounted for only 15% of the normal Caucasian population, yet were present in 58% of fragile X chromosomes examined. The results of the study were taken to indicate that only a few founder mutations are responsible for the majority of fragile X cases, and therefore that predisposing mutational events occur rarely (Richards *et al.* 1992). A similar study performed by Oudet *et al.* (1993a), with FRAXAC2 and a microsatellite marker 150 kb proximal to the CCG repeat, obtained comparable results, but indicated an increased degree of heterogeneity in fragile X associated haplotypes to that previously found.

Analysis of a more genetically homogeneous population group (Finnish), revealed a more striking linkage disequilibrium than had been reported for other populations studied (Oudet *et al.*, 1993b; Haataja *et al.*, 1994). Japanese population studies also found evidence of linkage disequilibrium at the fragile X locus, however the disease associated haplotypes were different from those of the Caucasian population (Arinami *et al.*, 1993; Richards *et al.*, 1994).

Results contradicting those of previous studies were reported by Macpherson *et al.* (1994), who found a notably different distribution of normal and fragile X haplotypes. This study found a greater diversity of haplotypes for fragile X chromosomes than for normal chromosomes, suggesting linkage disequilibrium/founder effects are not associated with the disorder.

1.4.30 Relationship between haplotype and fragile X CCG repeat copy number

PCR analysis of the fragile X locus alleles of normal individuals with the most common fragile X haplotypes found one of these haplotypes to be associated with *FMRI* CCG repeat alleles of greater than average copy number (Richards *et al.*, 1992). It had previously been established that the mutability of microsatellites is proportional to the copy number (Weber, 1990). It was therefore proposed that the higher incidence of fragile X syndrome associated with this specific haplotype could have a molecular basis in the comparatively frequent expansion (due to their greater mutability) of the associated normal, but high copy number *FMRI* CCG repeat alleles (Richards *et al.* 1992). More recent studies related to this hypothesis are discussed in Chapter 5.

1.4.31 Proposed mechanisms for founder mutations

Sequencing of normal fragile X alleles found the CCG repeat tract to be interrupted in many cases, and therefore to have only relatively short stretches of perfect repeat sequence (Hirst *et al.*, 1994; Kunst and Warren, 1994; Snow *et al.*, 1994). Weber (1990) had previously found that dinucleotide repeat loci are generally not as polymorphic when the repetitive sequence contains interruptions as when it is perfect and of equivalent length, implying that repeat tract interruptions predispose to a lower mutation rate. Based on this observation (before the sequencing of fragile X alleles was accomplished) Richards *et al.* (1992) proposed that fragile X founder chromosome mutations may eliminate interruptions in the repeat tract (e.g. a T to G transition), thus producing a more readily mutable length of sequence. Alternatively, while retaining the interruption, minor amplification of the perfect repeat tract section of an allele may occur, thereby increasing it to a length that is predisposed to mutate. Another possibility considered was that the predisposition to amplification of the repeat is conferred by specific flanking sequences that are associated with certain haplotypes, and is independent of repeat tract composition. With this mechanism, linkage disequilibrium would not be evident if the locus conferring predisposition to amplification was located elsewhere in the genome (Richards *et al.*, 1992; Morris *et al.*, 1995).

The greater diversity of fragile X haplotypes reported by Macpherson *et al.* (1994) led to the proposal that predisposing alleles are generated by two alternative mechanisms: the gradual increase in allele size previously suggested; or rare mutational events caused by generalised microsatellite instability, possibly due to a mutant DNA repair gene (Macpherson *et al.*, 1994; Morris *et al.*, 1995)

Richards *et al.* (1992) concluded that whatever the nature of the initial change producing a founder chromosome, it must occur rarely and be carried silently by a proportion of the population in order to account for the high incidence of the fragile X syndrome.

1.4.32 Confirmation that the fragile X syndrome is a single gene disorder

A patient was described by Gedeon *et al.* (1992) as having the typical clinical features of fragile X syndrome, but without cytogenetic expression of the fragile X or an amplified CCG trinucleotide repeat tract. Southern analysis of the patient's DNA using probes in the region found a previously uncharacterised submicroscopic deletion encompassing the CCG repeat tract, the entire *FMRI* gene, and about 2.5 Mb of flanking sequences. This finding indicated that expression of the fragile X phenotype is not dependent upon amplification of the CCG repeat or cytogenetic expression of the fragile site, and furthermore, that the disorder is genetically homogeneous and due entirely to loss of *FMRI* expression. This conclusion was supported by the work of Wöhrle *et al.* (1992), which characterised a 250 kb deletion encompassing the CCG repeat tract and part of the *FMRI* gene in a cytogenetically negative fragile X patient.

Subsequently, two fragile X patients were described as having *FMRI* intragenic loss of function mutations and lacking CCG repeat tract expansions (Lugenbeel *et al.*, 1995). In addition, De Boulle *et al.* (1993) reported the presence of a single missense point mutation in *FMRI* in a severely affected fragile X patient. These studies more precisely defined *FMRI* as the sole gene responsible for the disorder, as in these patients it was unlikely that nearby genes could be deleted or affected by hypermethylation.

1.4.33 *FMRI* transcription is repressed by methylation

The two changes at the molecular level that have been found in *FMRI* genes of fragile X individuals are: (1) the size increase of a CCG repeat tract within an exon and; (2) hypermethylation of a CpG island located 250 bp proximal to this repeat tract. It was established that repression of *FMRI* transcription is found in males affected with fragile X syndrome (Pieretti *et al.*, 1991), but further experimental data were required to determine the molecular change responsible for phenotypic expression, or whether both molecular changes contribute.

In order to address these issues, Sutcliffe *et al.* (1992) investigated differences in *FMRI* methylation and expression between a fetal tissue sample and corresponding chorionic villi sample of a fragile X syndrome fetus. The *FMRI* CpG island was found to be unmethylated in the chorionic villus sample and methylated in the fetal tissue sample. As analysis of the DNA extracted from both tissue sources found identical repeat expansion lengths, the methylation differences found between tissues was the only other apparent factor affecting *FMRI* gene expression.

The reverse transcription polymerase chain reaction (RT-PCR) technique was utilised in order to determine the transcriptional activity of *FMRI* in the tissue samples. In this study, the fluorescence intensity of dye labelled RT-PCR products from a primer pair downstream of the *FMRI* CCG repeat was compared to that of control hypoxanthine phosphoribosyl transferase (*HPRT*) derived RT-PCR products. By this method, it was found that *FMRI* transcription in the fragile X fetal tissue is approximately 3% of the transcription of the fragile X hypomethylated chorionic villi, relative to the *HPRT* transcription in the respective tissue samples.

1.4.34 Mechanism and timing of mutation events

Although the molecular basis of the fragile X syndrome had been established, factors that remained unknown were those dictating both the timing of the transition from premutation to full mutation during transmission, and the mechanism leading to this occurrence.

Laird's hypothesis (1987) predicted that the transition would occur on a previously inactive X chromosome where the premutation had locally blocked reactivation prior to oogenesis. As X inactivation causes methylation of CpG islands, it was proposed that hypermethylation could be the factor preceding or facilitating the size increase by interfering with replication (Laird, 1987)

Oberlé *et al.* (1991) and Yu *et al.* (1992) found Laird's predictions difficult to reconcile with the finding that some fragile X males are mosaics, carrying both the unmethylated premutation and the hypermethylated full mutation, as this implied that a secondary somatic mutation could result in a return to premutation status, and to loss of methylation. Both authors considered it more likely that the hypermethylation correlated with amplification size, and therefore that it is a consequence of amplification and not associated with maternal X-inactivation.

An alternative hypothesis proposed by Oberlé *et al.* (1991) was that the transition to full mutation occurs at random, with a frequency that is dependent on the exact structure of the premutation. In addition, it was suggested that the small size variations occurring during the transmission of a premutation are the result of the comparatively minor instability of that state, but the results of this process have an effect on the probability of subsequent amplification to a full mutation.

Later studies designed to clarify the mechanism(s) associated with the timing of repeat expansion are discussed in Chapter 4.

1.4.35 Methylation at the fragile X locus

DNA from the tissues of a fragile X male fetus was digested to determine the methylation status of a *Sac* II site in the fragile X CpG island. Varying degrees of methylation were observed in the fetal tissues, but none from a chorionic villi sample (Sutherland *et al.*, 1991a; Devys *et al.*, 1992). As all the tissues were derived from the same single cell zygote, it was concluded that the fragile X hypermethylation is most likely to be a consequence of the CCG repeat amplification. The occurrence of

methylation could therefore differ from one tissue to the next as it was not the result of a single event (Sutherland *et al.*, 1991a).

On the basis of these results, Sutherland *et al.* (1991a) considered that, since the amplified CCG repeat tract sequence presents many additional targets for methylation, the hypermethylation is a function of the amplification. It was further considered that although hypermethylation may affect phenotypic expression, whether it also has a role in the transmission of the disorder was not apparent from the results obtained.

1.4.36 Laird's hypothesis and abnormal methylation of the *FMRI* CpG island

The hypermethylation of the *FMRI* CpG island on fragile X chromosomes was interpreted by Follette and Laird (1992) as being due to persistence of methylation usually only present on the inactive X chromosome. Furthermore, it was proposed that the hypermethylation could silence the functioning of the *FMRI* gene as well as conferring instability on the repeated region. Expansion could then occur after imprinting, as a result of somatic instability (Webb, 1991).

Hansen *et al.* (1992) considered that a number of the predictions of Laird's X-inactivation imprinting model had been verified by the characterisation of the fragile X locus at the molecular level. The findings listed in support of this statement were: (1) abnormal DNA hypermethylation at the fragile X locus in affected individuals; (2) the abnormal pattern of hypermethylation resembles the methylation pattern of a normal allele at the fragile X locus on an inactivated X chromosome; (3) the *FMRI* gene is not transcribed in affected males; (4) cellular mosaicism for the hypermethylation at the fragile X locus is detectable in some affected individuals.

Hansen *et al.* (1992) studied CpG dinucleotide methylation patterns at the fragile X locus. This study revealed that the CpG dinucleotides within both the *FMRI* CCG repeat tract and 3' adjacent region form part of the *FMRI* CpG island, and as such become methylated during normal X chromosomal inactivation in females. Analysis of fragile X carriers and affecteds found that the *FMRI* CpG islands of fragile X carriers were unmethylated (as for normal active X chromosomes) but those of affected individuals

with full mutations had a methylation status similar to that of a normal inactive X allele. However, certain restriction sites within the *FMRI* CpG island were found to be methylated when a full mutation was present, but not when associated with the normal or premutational states on an inactive X chromosome (Hansen *et al.*, 1992). In addition, examination of the methylation status of the *FMRI* CpG island in cell lines containing either normal active X, normal inactive X and fragile X chromosomes established that alterations in methylation state did not occur during tissue culture (Hansen *et al.*, 1992).

Hansen and co-authors (1992) proposed that these observations were consistent with the hypothesis that the *FMRI* hypermethylation associated with fragile X full mutations and normal inactive X alleles has a common origin, as predicted by the X-inactivation/imprinting model (Laird, 1987). According to this model, the hypermethylation associated with full mutations is a consequence of a pattern originating in the X-inactivation process rather than the mutation itself, and therefore the primary CCG expansion of the fragile X mutation is neither necessary or sufficient for extensive methylation of the CpG island of the *FMRI* gene.

The experimental data of Hansen *et al.* (1992) did not conclusively show whether the hypermethylation is due to the presence of the expanded repeats, or if as predicted by Laird's hypothesis (1987), the abnormal persistence of hypermethylation leads to the secondary expansion of CCG repeats.

The proposal by Sutherland *et al.* (1991a) that the fragile X hypermethylation is due to the presence of an expanded repeat (and not vice-versa) was supported by the observation that methylase activity is markedly greater toward substrates with unusual DNA structures, such as those containing an unpaired or mismatched C in the CpG recognition sequence (Smith *et al.*, 1991). Another suggestion by Hansen *et al.* (1992) was that an oogonial specific monitoring system (Doerfler *et al.*, 1992) marks abnormal sequences for inactive chromatin formation, thereby causing hypermethylation of the region.

1.5 Dynamic Mutations

Anticipation is a term used to describe an aspect of inheritance that occurs when the penetrance and/or severity of a disorder increases, or age of onset decreases, in successive generations of affected pedigrees. Whether anticipation is a true genetic/biological phenomenon was for many years a subject of some debate. The cloning and characterisation of the fragile X syndrome locus in 1991 established that an increase in the copy number of an unstable repeat in successive generations of fragile X families could account for the genetic property of anticipation. For the fragile X syndrome this property (as previously mentioned) was termed the “Sherman paradox”, and strongly characterises the inheritance of this disorder.

Another disorder for which the inheritance patterns displayed evidence of anticipation was myotonic dystrophy (DM). It was therefore proposed that the genetic basis for the DM anticipation was similar to that of the fragile X syndrome (Sutherland *et al.* 1991b). Identification of the mutations responsible for myotonic dystrophy and the rare X-linked disorder spinal and bulbar muscular atrophy (SBMA) (Brook *et al.*, 1992; La Spada *et al.*, 1992) as amplifications of the trinucleotide AGC (or CAG) confirmed this prediction, and demonstrated that heritable unstable repeats were a more universal mutational mechanism, responsible for a variety of human genetic diseases.

The AGC repeat tracts at the DM and SBMA disease loci both displayed disease associated mutations consisting of increases in copy number. In common with the fragile X, the rate of the mutation at the loci was found to be related to the repeat copy number, with the product of a change in repeat copy number having a different mutation rate to that of its predecessor. It was for this reason that Richards and Sutherland (1992) used the term dynamic mutation to describe the mechanism of mutation in the unstable heritable repeat disorders.

1.5.1 Trinucleotide repeats in genes

Richards and Sutherland (1992) suggested that, because potentially unstable trinucleotide repeats in genes may have deleterious consequences, they must have a functional significance that outweighs this factor. This functional significance was evident for repeats within the coding sequence of a protein (with a function requiring a particular homopolymer), although with the degenerate genetic code repeat tracts within the coding sequence would not necessarily be required.

Searches of the Genbank database for sequences containing any of the ten possible trinucleotide repeat tracts identified a number of human genes with five or more repeat copies. Five was therefore suggested to be significant number of repeats with regard to the occurrence of instability, as it is the most common normal copy number of the unstable myotonic dystrophy repeat tract (Richards and Sutherland, 1992). The majority of CCG repeat tracts in genes were found to be located in the 5' untranslated region, suggesting a functional constraint other than coding capacity must maintain the presence of these repeats. It was suggested that the repeat tract may be a binding site for a transcription factor, and also that the characterisation of more genes with unstable repeat sequences, followed by analysis of to what degree (if any) these are related, might lead to clarification of this issue (Richards and Sutherland, 1992).

1.5.2 Proposed mechanisms for repeat instability

A number of models of molecular level events have been proposed in order to explain the ability of an unstable trinucleotide repeat to increase up to ten fold in length in one transmission. A model invoking a strand slippage mechanism was sufficient to explain small increases in size of short tandem repeat elements, but not the larger increases observed. Another model was based on the observation that DNA polymerases have difficulty in replicating GC rich sequences, therefore during replication of the repeats there may be a number of incomplete strands that are the result of premature termination and reinitiation events. Large increases in length could then be generated by strand switching between the incomplete strands. This model predicted that a longer allele

would have a greater frequency of conversion to an expanded allele (Kuhl *et al.*, 1993), which correlated with previous observations (Fu *et al.*, 1991).

Richards *et al.* (1992) suggested two theories to account for the molecular changes that constitute an unstable repeat founder mutation. One proposal was that the founder mutation causes the acquisition of a DNA binding site for an activity that destabilises the repeated sequence, resulting in different rates of mutation for the founder and normal chromosomes. Another proposal was that the acquisition of increased instability is due to a rare increase in perfect repeat copy number, but once a repeat tract is beyond a certain length, mutation events are frequent.

Based on the previous findings that interruptions in simple tandem repeat sequences are associated with reduced mutability (Weber, 1990), and that interruptions occur in fragile X (CCG)_n repeat alleles (Oberlé *et al.*, 1991; Kremer *et al.*, 1991b; Verkerk *et al.*, 1991), Richards *et al.* (1992) suggested that a founder mutation may occur if an interrupting base mutated back to the normal repeat motif, thus producing a perfect repeat of an unstable length.

1.5.3 Comparisons of autosomal and X-chromosomal folate sensitive fragile sites may provide insight into the mutational mechanism(s) involved in their genesis

Earlier sections of this chapter provided an overview of hypotheses put forward with regard to the mechanism of mutation at the fragile X locus, both prior to, and following molecular characterisation. The most notable hypothesis was that of Laird (1987), which implicated the process of X-chromosomal inactivation as performing a central role in the unusual inheritance pattern of the fragile X syndrome. When the fragile site was cloned, the methylation observed on fragile X chromosomes was initially taken to be an indication that the hypothesis may be correct (Hansen *et al.*, 1992). As this could be neither proved or refuted directly, it was considered that the most likely role of X-inactivation in the mechanism of fragile site genesis might therefore be defined by the characterisation of a fragile site located in a region not subject to X-inactivation. As autosomes are not subject

to methylation due to X-inactivation processes, the finding that an autosomal folate sensitive fragile site has the same features, i.e. it is an unstable heritable CCG repeat tract that becomes methylated when amplified, would provide strong evidence against the occurrence of X inactivation/imprinting events in fragile X inheritance.

The characterisation of an autosomal rare folate sensitive fragile site was also of interest because it was considered that it may result in a clearer picture of whether the propensity for the development of an amplified CCG repeat mutation is inherent only in the unstable repeat tract, or whether the instability is also influenced by features of the immediate flanking sequences and/or chromosomal environment. Such features could include inheritance, amplified size or stability of the mutated repeat, in addition to the DNA sequence and the variation in normal allele sizes at the fragile site loci.

1.6 Project Strategy

The initial aim of the study was to characterise fragile sites on human chromosome 16 (beginning with *FRA16A*) at the molecular level, with the degree of success in this endeavour obviously being limited by time constraints and occurrence of technical difficulties. Successful characterisation of *FRA16A* would then allow a number of interesting comparisons with *FRAXA* with regard to features such as methylation status and the inheritance of fragile site mutations.

Chromosome 16 fragile sites were chosen for study because of the availability of detailed genetic and physical maps of the chromosome. There was also a ready availability of the markers and somatic cell hybrid cell lines that had been used to construct the physical map of the chromosome.

Although there was no information available that could directly facilitate the cloning of other classes of fragile sites, the finding that *FRAXA* (which like *FRA16A*, is a rare folate sensitive fragile site) is a potentially unstable CCG repeat tract that is part of the *FMRI*

CpG island, had the potential to greatly decrease the amount of effort required to find other folate sensitive fragile sites. Therefore, in order to achieve the goal of cloning *FRA16A* whilst taking into account the features associated with *FRAXA*, the following strategies were chosen: 1) pulsed field long range restriction mapping of the *FRA16A* interval in DNAs from non-*FRA16A* and *FRA16A* individuals. This approach could potentially determine the size of the region of interest, the positions of DNA markers with respect to one another, and also the closest DNA markers to *FRA16A*. Additionally, anomalies in band size or intensity that may be due to hypermethylation at a fragile site locus could potentially be identified; 2) identification of megaYAC(s) with the DNA marker mapping closest to the fragile site. Pulsed field long-range restriction mapping of the YAC(s) could be performed in order to identify any restriction sites of interest found by the uncloned DNA mapping. Restriction mapping could also, by comparing maps of individual YACs, evaluate the degree of YAC insert rearrangement; 3) screening the Southern blotted DNAs of isolated YACs with a CCG repeat probe in order to identify those containing hybridising sequences, and therefore possibly a CCG repeat tract as identified at the fragile X locus.

Chapter 2

Materials and Methods

Chapter 2: Contents

Materials and Methods

2.1	Introduction	53
2.2	Materials.....	53
2.2.1	Enzymes and suppliers.....	53
2.2.2	Electrophoresis reagents and suppliers	54
2.2.3	Radio-chemicals.....	54
2.2.4	Buffers and solutions.....	54
2.2.5	Bacterial and yeast media.....	55
2.2.6	Bacterial strains.....	56
2.2.7	Vectors	56
2.2.8	Miscellaneous materials.....	57
2.2.9	Miscellaneous fine chemicals.....	57
2.3	Methods.....	58
2.3.1	DNA isolation.....	58
2.3.1.1	Large scale isolation of plasmid and cosmid DNA.....	58
2.3.1.2	Small-scale isolation of plasmid DNA: a modified mini alkaline lysis/PEG precipitation procedure.....	59
2.3.1.3	Rapid method for small scale isolation with BIO101 Inc. RPM™ kit.....	60
2.3.1.4	Isolation of peripheral lymphocyte DNA.....	60
2.3.2	General methods for purification of DNA	61
2.3.2.1	Phenol and chloroform extraction of DNA.....	61
2.3.2.2	Ethanol precipitation of DNA	61
2.3.2.3	Bio-Rad Prep-a-Gene kit purification of PCR products for sequencing and labelling.....	61

2.3.3	Subcloning human DNA fragments	62
2.3.3.1	Preparation of plasmid vector DNA and human DNA inserts.....	62
2.3.3.2	Dephosphorylation of plasmid vector DNA.....	62
2.3.3.3	Ligation reactions.....	62
2.3.3.4	Competent cells and transformation.....	63
2.3.4	Restriction endonuclease digestion, gel electrophoresis, and hybridisation analysis of DNA	63
2.3.4.1	Restriction endonuclease digestion of DNA	63
2.3.4.2	Gel electrophoresis.....	64
2.3.4.3	Molecular weight markers.....	64
2.3.4.4	Random primed ³² P labelling of DNA.....	64
2.3.4.5	Probe purification.....	64
2.3.4.6	Pre-reassociation of repetitive DNA	65
2.3.4.7	Transfer of DNA to nylon membranes.....	65
2.3.4.8	Southern blotting.....	65
2.3.4.9	Filter hybridisation and washing.....	66
2.3.4.10	Stripping nylon filters.....	66
2.3.5	Pulsed field gel electrophoresis	67
2.3.5.1	Encapsulation of cells in agarose beads for PFGE.....	67
2.3.5.2	Preparation of agarose blocks containing yeast DNA (lithium method).....	68
2.3.5.3	Preparation of mammalian cell line DNA in agarose blocks.....	69
2.3.5.4	Restriction digestion of DNA in agarose beads and blocks.....	69
2.3.5.5	Loading agarose beads into wells.....	70
2.3.5.6	Loading agarose blocks into wells	70
2.3.5.7	Switching intervals.....	70

2.3.5.8	DNA size markers for PFGE.....	71
2.3.5.9	Southern blotting and hybridisation of pulsed field gels.....	71
2.3.6	Polymerase chain reaction (PCR).....	71
2.3.6.1	PCR across CCG repeat tracts.....	72
2.3.6.2	PCR primers	72
2.3.7	DNA sequencing.....	73
2.3.7.1	Fluorescent automated DNA sequencing	73
2.3.7.2	Manual DNA sequencing	73
2.3.7.3	Electrophoresis of manual sequencing samples	75
2.3.8	Resolving minisatellite allele PCR products	75

2.1 Introduction

The majority of the methods used in this project are in routine use, and therefore well established in the laboratories of the Department of Cytogenetics and Molecular Genetics at the Women's and Children's Hospital, Adelaide (Australia).

This chapter will describe the general materials and methods used throughout the study. Included are the basic molecular genetic methods for DNA isolation, cloning DNA sequences, genomic DNA analysis by Southern blot hybridisation, the polymerase chain reaction and DNA sequencing. The materials and methods that were used for a specific section of the project are presented at the beginning of the relevant thesis chapter.

Restriction endonucleases and DNA modifying enzymes used in the project were obtained from commercial sources and used in accordance with the manufacturer's specifications. All solvents and chemicals are of analytical grade and were purchased from Ajax Chemicals (Sydney, Australia), unless stated otherwise.

2.2 Materials

2.2.1 Enzymes and suppliers

HK phosphatase	Epicentre Technologies
Calf intestinal phosphatase (CIP)	Boehringer Mannheim, Germany
<i>E.coli</i> DNA polymerase (Klenow fragment)	Amersham Australia, Pty. Ltd.
Proteinase K	Merck, Germany
RNase A	Boehringer Mannheim
T4 DNA ligase	Promega, Wisconsin, USA
<i>Taq</i> polymerase	Boehringer Mannheim
<i>Taq</i> polymerase (sequencing grade)	Perkin Elmer, California, USA

All restriction endonucleases were obtained from New England, Biolabs (Beverly, Massachusetts, USA) or Progen (Brisbane, Queensland, Australia)

2.2.2 Electrophoresis reagents and suppliers

Acrylamide	Bio-Rad Laboratories, California, USA
Agarose - nucleic acid grade	Pharmacia, Uppsala, Sweden
- low melting temp	FMC, Rockland, Maine
Ammonium persulphate	Bio-Rad Laboratories
Bromophenol blue	BDH Chemicals, Dorset, England
Ethidium bromide	Boehringer Mannheim
Molecular weight markers:	
- pUC19/ <i>Hpa</i> II	Bresatec, Adelaide, Australia
- Spp1/ <i>Eco</i> RI	Bresatec
- DRigest III	Pharmacia
- Lambda-PFGE	Pharmacia
- Yeast DNA-PFGE	Pharmacia
N, N, N1, N1-tetramethylethylene diamine (TEMED)	Bio-Rad Laboratories
Urea	Bio-Rad Laboratories
Xylene cyanol	Tokyo Kasei, Tokyo, Japan

2.2.3 Radio-chemicals

α - ³² P-dCTP, 3000Ci/mmol	Radiochemicals Centre, Amersham
--	---------------------------------

2.2.4 Buffers and solutions

Buffers and solutions routinely used in this study were:

Formamide loading buffer	92.5% (v/v) formamide 20 mM EDTA 0.1% (w/v) xylene cyanol 0.1% (w/v) bromophenol blue
--------------------------	--

10 x loading buffer	50% glycerol 1% (w/v) SDS 100 mM EDTA 0.1% (w/v) xylene cyanol 0.1% (w/v) bromophenol blue
10 x ligation buffer (Promega)	300 mM Tris-HCl, pH 7.8 100 mM MgCl ₂ 100 mM DDT 5 mM ATP
2 x PCR mix	33 mM (NH ₄) ₂ SO ₄ 133 mM Tris-HCl 2% (v/v) β-mercaptoethanol 13 mM EDTA 0.34 mg/ml BSA 20% (v/v) DMSO 3 mM dATP, dGTP, dTTP, dCTP
20 x SSC	3 M NaCl 0.3 M tri-sodium citrate
TBE	89 mM Tris-base 89 mM Boric acid 2.5 mM EDTA (pH 8.3)
TE	10 mM Tris-HCl (pH 7.5) 0.1 mM EDTA
TSB	10% (w/v) polyethylene glycol (PEG) MW3300 5% (v/v) DMSO 10 mM MgSO ₄ 10 mM MgCl ₄ L. Broth to required volume

2.2.5 Bacterial and yeast media

Liquid media

All liquid media were prepared with millipore water and autoclave sterilised.

The composition of the various media were as follows:

AHC medium (for <i>S. cerevisiae</i>)	0.67% (w/v) yeast nitrogen base without amino acids 1% (w/v) casein hydrolysate - acid 0.006% adenine 2% glucose
--	--

L. Broth (for <i>E. coli</i>)	1% (w/v) Bacto-tryptone 0.5% (w/v) Bacto-yeast extract 1% (w/v) NaCl pH to 7.5 with NaOH
--------------------------------	---

Solid Media

L. Agar	L. Broth 1% (w/v) Bacto-agar
L. Ampicillin agar	L. Broth 1% (w/v) Bacto-agar ampicillin (100 µg/ml)
L. Kanamycin Agar	L. Broth 1% (w/v) Bacto-agar kanamycin (50 µg/ml)

Antibiotics

The antibiotics used in the project are listed below with the suppliers:

Ampicillin	Sigma
Kanamycin	Boehringer Mannheim

2.2.6 Bacterial strains (*E. coli*)

XLI-Blue	genotype: rec A1, end A1, gyrA96, thi-1 hsd R17, sup E44, rel A1, lac [F ¹ proAB, lac I ^q ZAM15, tn10 (tet ^r)] supplier: Stratagene, California, USA
----------	---

2.2.7 Vectors

Plasmid and phagemid vectors

pUC19	amp ^r supplier: New England Biolabs (cat. # 304-1S) reference: Yanish-Peron <i>et al.</i> , (1985)
-------	--

pBluescript®

amp^r

supplier: Stratagene

(phagemid no longer commercially available)

2.2.8 Miscellaneous materials

Hybond N+™ Nylon Membrane

Amersham

Sephadex G-50 (fine)

Pharmacia

X-ray film

Kodak or Dupont

2.2.9 Miscellaneous fine chemicals5-bromo-4-chloro-3-indolyl-β-D
galactoside (X-Gal)

Boehringer Mannheim

Chemicals for oligonucleotide synthesis

Beckman, California, USA

Deoxynucleotide triphosphates
and Dideoxynucleotide triphosphates

Boehringer Mannheim

Dimethylsulphoxide (DMSO)

Sigma

Isopropyl thio-β-galactoside (IPTG)

Boehringer Mannheim

Random priming labelling kits

Amersham

Phenol

Wako, Japan

Salmon sperm DNA

Calbiochem

Sarkosyl

Ciba Geigy, Switzerland

Sodium dodecyl sulphate (SDS)

Sigma

Spermidine

Sigma

PRISM Ready Reaction Cycle
Sequencing Kits with
AmpliTaQ® DNA Polymerase, FSTaq Dye Deoxy™ Terminator
(cat. # 402080)

Dye Primer:

-21M13 primer (cat. # 402111)

M13 Rev primer (cat. # 402109)

2.3 Methods

2.3.1 DNA isolation

2.3.1.1 Large scale isolation of plasmid DNA and cosmid DNA (modification of Sambrook *et al.*, 1989)

Ten ml L. Broth containing ampicillin (50 µg/ml) was inoculated with a single fresh bacterial colony from a streaked plate. The culture was incubated at 37°C for 5 - 7 hours with vigorous shaking, and then transferred to 100 ml L. Broth with ampicillin (50 µg/ml).

After overnight incubation at 37°C with vigorous shaking, the 100 ml culture was transferred to two 50 ml Falconer tubes. The tubes were spun at 3000 rpm for 10 min in a Jouan CR3000 centrifuge at 4°C. The supernatant was discarded and the cell pellet resuspended in 300 µl of TE and glucose (50 mM Tris-HCl, 20 mM EDTA and 50 mM glucose). Sixty µl of 80 mg/ml lysozyme was added, and the cell suspension mixed gently and incubated at room temperature for 4 min, then placed on ice for 1 min. To the cell suspension was added 1.2 ml of 0.2 M NaOH/1% SDS, followed by gentle mixing and a 5 min incubation on ice. Nine hundred µl of ice cold 3 M potassium acetate (pH 4.3) was added to the suspension and mixed by inversion. The mixture was spun in a Beckman J2-21M/E centrifuge at 15K for 15 min and the supernatant was transferred to a fresh tube and spun as before. As much supernatant as possible was transferred to a fresh tube, mixed with 5.5 ml of ethanol, then incubated at room temperature for 5 min. Precipitated material (including plasmid DNA) was removed from the mixture by centrifugation in the Beckman centrifuge at 15000 rpm for 5 min, and decanting of the supernatant. The resulting pellet was washed in 2 ml of 70% ethanol, air dried, and resuspended in 100 µl TE. Ten µl of 1 mg/ml RNase (Boehringer) was added to the re-dissolved pellet and the mixture incubated for 15 min at 37°C. To eliminate proteins in the DNA preparation, 100 µl of 3 x proteinase K buffer (10 mM NaCl/10 mM EDTA), 10 µl of 10% SDS and 2 µl of 10 mg/ml proteinase K were added to the DNA solution. The mixture was then incubated at 37°C for 1 hour. For DNA extraction, an equal volume

of phenol (saturated with 10 mM Tris-HCl) was added and briefly mixed. The mixture was then spun in an eppendorf centrifuge at 10000 rpm for 10 min, with the top aqueous phase then transferred to a fresh tube and thoroughly mixed with an equal volume of phenol:isoamyl alcohol. After centrifugation, the top aqueous phase was transferred to a fresh tube and mixed with 1/3 volume 7.5 M ammonium acetate and two volumes ethanol, then left at 20°C overnight for DNA precipitation. After centrifugation, the DNA pellet was washed with 70% ethanol, desiccated, and dissolved in 200 µl of TE.

2.3.1.2 Small scale isolation of plasmid DNA: a modified mini alkaline lysis/PEG precipitation procedure

(modification of protocol supplied with Perkin Elmer *Taq* DyeDeoxy™ Terminator Cycle Sequencing Kit)

A single bacterial colony was used to inoculate 10 ml of L. Broth containing ampicillin (50 µg/ml). The culture was incubated at 37°C overnight with vigorous shaking, and bacterial cells were pelleted by centrifugation in a Jouan CR3000 centrifuge at 3000 rpm for 10 min. The pellet was resuspended in 200 µl of GTE buffer (50 mM glucose, 25 mM Tris-HCl pH 8.0, 10 mM EDTA pH 8.0). Three hundred µl of freshly prepared 0.2 M NaOH/1% SDS was added and mixed by inversion. The mixture was then incubated on ice for 5 min. Cellular debris was removed by centrifugation for 10 min at room temperature in an eppendorf centrifuge. The supernatant was transferred to a clean tube, RNaseA (DNase free) was added to a final concentration of 20 µg/ml and the tube was incubated at 37°C for 20 min. The supernatant was extracted twice with 400 µl chloroform, the layers being mixed by hand for 30 seconds each time. The tube was centrifuged for 1 min to separate the phases, and the aqueous phase removed to a clean tube. The DNA was precipitated with an equal volume of 100% isopropanol, and the tube immediately centrifuged for 10 min at room temperature. The DNA pellet was washed with 500 µl of 70% ethanol and dried under vacuum. The dried pellet was dissolved in 32 µl of deionised water, and the plasmid DNA precipitated by the addition of 8 µl of 4 M NaCl and 40 µl of autoclaved 13% PEG₈₀₀₀ solution. After thorough mixing the sample was incubated on ice for 20 min, and the plasmid DNA pelleted by centrifugation

in a Hettich Mikro Rapid/K centrifuge for 15 min at 4°C. The supernatant was removed and the pellet rinsed with 500 µl of 70% ethanol. The pellet was dried under vacuum and resuspended in 20 µl of sterile deionised water.

2.3.1.3 Rapid method for small scale isolation of plasmid DNA with BIO 101, Inc. (California, USA) RPM™ kit (cat. # 2070-406)

The protocol supplied with the kit was followed, except that 10 ml instead of 1.5 ml of bacterial culture was used for plasmid DNA extraction.

2.3.1.4 Isolation of peripheral lymphocyte DNA (modification of Wyman and White, 1980)

Blood samples were collected in 10 ml tubes containing EDTA and were allowed to cool to room temperature before being stored at -70°C. For the purpose of isolating lymphocyte DNA, the frozen blood sample was thawed and transferred to a 50 ml falconer tube. Cell lysis buffer (0.32 M sucrose, 10 mM Tris-HCl, 5 mM MgCl₂, 1% Triton X-100) was added to the tube until a total volume of 30 ml was reached. After mixing of the contents by inversion, the tube was left on ice for 30 min. The unlysed white cell suspension was spun in a Jouan centrifuge (4°C) at 3500 rpm for 15 min. The supernatant was vacuum aspirated down to 5 mls then cell lysis buffer again added to a volume of 30 ml and mixed. The centrifugation was repeated, and the supernatant carefully removed to leave a pellet of white cells. 3.25 ml of 1 x Proteinase K buffer, 0.5 ml of 10% SDS and 0.2 ml of Proteinase K (10 mg/ml) were added to the pellet and mixed. The tube was capped and sealed with laboratory plastic film, then incubated overnight at 37°C on a rotating wheel at 10 rpm. Next day, the solution was extracted twice with an equal volume of chloroform:isoamyl alcohol (24:1). To precipitate the DNA, 1/10 volume of 3 M sodium acetate (pH 4.6), and 2 volumes of ice-cold ethanol were added to the tube. The mixture was inverted several times until the precipitated DNA was visible. The precipitated DNA was transferred with a cut-off pipette tip to an eppendorf tube and washed twice with 70% ethanol. The DNA was desiccated and then

dissolved in 100 μ l of TE. The phenol and chloroform were handled in a fume hood and gloves were used throughout the procedure.

2.3.2 General methods for purification of DNA

2.3.2.1 Phenol and chloroform extraction of DNA (modification of Moore, 1993)

Solutions containing DNA were extracted with phenol/chloroform to remove proteins and other contaminants. An equal volume of TE saturated phenol was added to the DNA solution and mixed by vortexing for 1 min. The mixture was then centrifuged at 12000 rpm in an eppendorf centrifuge. The upper aqueous phase containing the DNA was removed leaving a white interface of denatured protein and the lower organic phase. To the aqueous phase was added an equal volume of chloroform. Vortexing, centrifugation and aqueous phase removal steps were repeated, and the DNA was ethanol precipitated.

2.3.2.2 Ethanol precipitation of DNA (modification of Moore, 1993)

A 3 M sodium acetate solution (pH 5.2) was added to the DNA sample for a final concentration of 300 mM. Two volumes of ice-cold ethanol were mixed with the sample, followed by incubation at -20°C for 1 hour. Precipitated DNA was pelleted at 14000 rpm for 10 min and washed once with 70% ethanol. After vacuum drying, DNA was redissolved in water or TE.

2.3.2.3 Bio-Rad Prep-a-Gene Kit (cat. # 732-6011) purification of PCR products for sequencing and labelling

Typically, in order to generate PCR amplified DNA for use as sequencing templates and probes, a 50 μ l PCR reaction was performed. A 5 μ l sample of PCR product was electrophoresed on an agarose gel stained with EtBr and visualised under UV light to ensure a band of the correct size was present. The remainder of the reaction was purified according to the protocol provided with the Prep-a-Gene kit. After purification the DNA was resuspended in 20 μ l elution buffer and quantitated by spectrophotometry.

2.3.3 Subcloning of human DNA fragments

2.3.3.1 Preparation of plasmid vector DNA and human DNA inserts (modification of Struhl, 1987)

Five hundred ng of vector DNA (pUC19) was digested with a polylinker specific restriction endonuclease, according to manufacturer's instructions. Human DNA (previously cloned in a cosmid, lambda or plasmid vector) was digested with the same restriction enzyme(s), and under the same conditions, that were used to cleave the vector. The digested DNA samples were electrophoresed on an agarose gel to ensure complete digestion had occurred, then phenol/chloroform extracted and ethanol precipitated.

2.3.3.2 Dephosphorylation of plasmid vector DNA (modification of Struhl, 1987)

In order to prevent self-ligation of the vector when digested with a single restriction enzyme, the 5' terminal group of the vector DNA molecule was removed by treatment with calf intestinal alkaline phosphatase (CIAP). The vector DNA was digested to completion with an appropriate restriction enzyme in a total volume of 50 μ l. Ten μ l of 10 x CIAP buffer (500 mM Tris-HCl pH 8.5, 10 mM $MgCl_2$, 10 mM $ZnCl_2$) and 5 μ l of 1 unit/ μ l, CIAP were added and the volume made up to 100 μ l with water. The reaction was incubated at 37°C for 1 hour and then phenol/chloroform extracted and ethanol precipitated. To test the efficiency of dephosphorylation, a ligation reaction was performed with 1 μ l of dephosphorylated DNA and transformed into *E. coli* strain XLI-Blue. If the 5' terminal phosphate group was removed, the vector could not recircularise. Therefore few colonies would be seen on a plate, as the transformation of DNA in a linear form is comparatively inefficient.

2.3.3.3 Ligation reactions (modification of Struhl, 1987)

Ligation reactions were carried out with a vector:insert molar ratio of approximately 1:3 to maximise intermolecular ligation rather than intramolecular ligation. Typically, a ligation reaction mixture consisted of 100 ng of linearised, phosphatased vector, ~200 ng of insert DNA, 2 μ l of 10 x ligation buffer and 1 - 2 units of T4 DNA ligase, with water

added to a total volume of 20 μ l. The reaction was incubated at 4°C overnight, or for 1 hour at room temperature.

2.3.3.4 Competent cells and transformation (Modification of Chung *et al.*, 1989)

Stationary phase *E. coli* strain XLI-Blue cells from an overnight culture were diluted 1:50 (v/v) into 50 ml LB. The cells were grown with constant shaking for 2.25 hours. The cells were pelleted in a Jouan centrifuge at 3000 rpm for 10 min and the cell pellet resuspended in 5 ml (1/10 culture volume) of fresh, ice-cold TSB. The cells were incubated on ice for 10 min and were then ready for use. An aliquot of the ligation reaction was added (typically 1 - 2 μ l) to the competent cells and the mixture incubated on ice 5 - 30 min. Nine hundred μ l of TSB and 20 μ l of 1 M glucose were added and the tube incubated at 37°C with constant shaking for 1 hour. After incubation, an aliquot of 100 μ l of the transformation mixture was spread plated onto 750 mm diameter L. Amp plates that had been previously spread with 100 μ l of 1 M IPTG and 6% X-Gal (dissolved in dimethyl-formamide). Plates were incubated overnight at 37°C. Recombinant plasmids were detected as white colonies with blue colonies containing only recircularised plasmid vector.

2.3.4 Restriction endonuclease digestion, gel electrophoresis and Southern blot analysis of DNA

2.3.4.1 Restriction endonuclease digestion of DNA.

Restriction endonuclease digestion of DNA was carried out in enzyme compatible buffers from the New England Biolabs or Promega commercial buffer systems. In general, 4 units of enzyme was added for each microgram of DNA to be digested, and the reaction mix incubated at the appropriate temperature for 2 - 4 hours for cosmid, plasmid and phage DNAs, and overnight for genomic DNAs. To ensure the enzymatic activity was not affected by glycerol, the volume of restriction enzyme added did not exceed 1/10 of the final volume of the reaction mix. Reactions were terminated by the addition of a 0.1 x volume of 10 x loading buffer.

2.3.4.2 Gel electrophoresis (Sambrook *et al.*, 1989)

Restriction endonuclease digested DNA samples for Southern blot hybridisation analysis were electrophoresed on 0.8 - 1.2% agarose gels in 1 x, or 0.5 x TBE. Gels were run in BRL horizontal tanks containing 0.5 or 1 x TBE buffer at 15 - 100 mA, until the bromophenol blue had migrated an appropriate distance for the required separation of the DNA fragments to have taken place (according to empirical observations).

Analytical agarose minigels (generally used for analysis of restriction digested and PCR product DNAs) were electrophoresed for ~1 hour at 100 V in a Bio-rad Mini-subTM DNA cell. In all cases, DNA in an agarose gel was visualised under UV light after staining in 0.02% EtBr/0.5 x TBE for 10 - 30 min.

2.3.4.3 Molecular weight markers

SPP1/*Eco* RI (Progen cat. # 500-0020), pUC19/*Hpa* II (Progen cat. # 500-0030), or DRigest (Pharmacia cat. # 27-4060-01) were used as molecular weight markers on agarose gels, the choice of marker depending on the size range of DNA fragments to be resolved.

2.3.4.4 Random primed ³²P labelling of DNA

³²P labelling of double stranded DNA was performed by the random primer method (Feinberg and Vogelstein, 1983) using a Amersham MegaprimeTM DNA labelling systems kit (cat. # RPN 1604). The protocol followed was supplied with the kit. In brief, a solution containing DNA template (25 - 50 ng) and random nonamer oligonucleotides was incubated at 100°C for 4 min and snap cooled on ice for 1 min. A buffer solution containing deoxyribonucleotides, $\alpha^{32}\text{P}$ -dCTP and Klenow enzyme was added. The mixture was incubated at 37°C for 1 - 2 hours.

2.3.4.5 Probe purification

Unincorporated radionucleotides were removed from the labelled probe by chromatography on a Sephadex[®] G-50 Fine column (Pharmacia cat. # 17-0573-01). Two-

drop fractions were collected from the column and monitored with a Geiger counter. The first of the 2 peaks eluted from the column contained the labelled probe, the second the unincorporated radionucleotides. The fractions corresponding to the first peak were pooled into 1 tube and incubated at 100°C for 10 min immediately prior to use.

2.3.4.6 Pre-reassociation of repetitive DNA (Sealy *et al.*, 1985)

³²P-labelled DNA probes containing repetitive sequences were pre-reassociated prior to Southern blot hybridisation in order to block high copy number repeats. To the unpurified labelled probe was added 1 µl of 0.5 M EDTA to inactivate the enzyme, 100 µl of 10 mg/ml or 2000 fold excess human placental DNA (previously boiled to produce DNA fragments of size range 300 - 700 bp), and 50 µl 20 x SSC. The sample was denatured in a 100°C heating block for 10 min, snap-cooled on ice for 1 min, then incubated in a 65°C waterbath for 1 - 2 hours. After incubation, the sample was added to the hybridisation bottle containing the hybridisation mix and Southern blot filter.

2.3.4.7 Transfer of DNA to nylon membranes

Plaque/colony lifting (Grunstein and Hogness 1975; Benton and Davis, 1977)

Bacteria containing a recombinant vector were plated out as described. After incubation overnight at 37°C, bacterial cells or phage were transferred by placing a nylon membrane disc (Amersham Hybond N+™ cat. # RPN82B or RPN137B) flat onto the plate surface and leaving 2 - 3 min. The filters and plates were marked so they could be re-aligned accurately, and filters were then transferred to a sheet of paper soaked in 0.5 M NaOH for 2 - 3 min to lyse host cells and denature DNA. Filters were then transferred to paper soaked in 1 M Tris-HCl (pH 8) for 2 - 3 min to neutralise, then rinsed in 2 x SSC and allowed to air dry.

2.3.4.8 Southern blotting (Reed and Mann, 1985)

Restriction endonuclease digested DNA or PCR products were separated on agarose gels, stained with EtBr, photographed, then transferred to Hybond N+ nylon membrane

using the alkaline transfer method (Reed and Mann, 1985). If DNA to be transferred was over 1 kb in size, an acid nicking step was included. This involved soaking the gel twice in 0.25 M HCl for 15 min with gentle shaking. The gel was then immersed in a NaOH/ NaCl solution for 30 min to denature the transferred DNA, and a Tris-HCl/NaCl solution to neutralise the DNA. Nylon membrane was cut to the size of the gel and rinsed in demineralised water and then in transfer solution (10 x SSPE). The neutralised gel was placed on the paper wick of a Southern blotting tray containing 10 x SSC as transfer solution. A nylon filter and blotting paper were placed on the gel and the DNA transferred by capillary action for 2 - 16 hours. DNA transferred to the filter was denatured in 0.5 M NaOH for 30 - 60 seconds, then neutralised in Tris/SSPE for 1 - 2 min and allowed to air dry.

2.3.4.9 Filter hybridisation and washing (modification of Brown, 1993)

Prior to the addition of labelled probe, nylon filters were wetted with 5 x SSPE solution, then prehybridised at 42°C for 30 - 60 min in a solution consisting of 50% (v/v) deionised formamide, 5 x SSPE, 2% SDS, 5 x Denhart's solution, 10% (w/v) Dextran Sulphate (Pharmacia), and 100 µg/ml salmon sperm DNA. Probe (1 - 10 ng/per ml hybridisation solution) was added directly to the hybridisation solution and filter, and incubated overnight at 42°C. After hybridisation, filters were subjected to a stringent wash. For this, filters were immersed in a solution of 2 x SSPE, 1% SDS for 30 min, and then in a solution of 0.1 x SSPE, 1% SDS for 30 min. Both washes were carried out at 65°C in a shaking waterbath.

2.3.4.10 Stripping nylon filters

Immersion in a denaturing 0.4 M NaOH solution stripped radioactive probes from Hybond N+ nylon filters. The denaturing solution was then replaced with a neutralising solution [0.2 M Tris-HCl (pH 7.5), 0.1 x SSC, 0.1 x SDS]. Both steps were carried out at 42°C for 30 min in a shaking waterbath. After stripping, the filters were air dried.

2.3.5 Pulsed field gel electrophoresis (PFGE) (Schwartz *et al.*, 1982; Chu *et al.*, 1986)

Long range restriction mapping of cloned YAC DNA and uncloned human genomic DNA within regions of interest required application of the pulsed field gel electrophoresis (PFGE) technique in order to resolve high molecular weight restriction fragments. For this purpose a Bio-Rad Chef Mapper™ apparatus was used. The gel was prepared by casting 100 ml of molten 1% agarose in 0.5 x TBE into a gel support tray (14 cm x 12.7 cm). Agarose beads (50 - 100 µl) containing YAC DNA, or agarose blocks (100 µl) containing lymphoblastoid cell line DNA, were loaded into wells. PFGE was performed in 0.5 x TBE at 14°C at the selected switching interval for the required time. The operation of the PFGE apparatus was according to the manufacturer's instructions.

2.3.5.1 Encapsulation of cells in agarose beads for PFGE.

In order to minimise shearing and preserve the integrity of large DNA molecules, it was necessary to isolate the DNA from intact cells after they had been embedded in agarose. The agarose matrix stabilises and immobilises the DNA molecules after the removal of cell membranes and proteins.

Preparation of agarose beads containing yeast DNA (Overhauser and Radic, 1989)

One to 4 hundred ml of AHC medium (1 litre of AHC medium consists of 6.7 mg of yeast nitrogen base, 10 mg of casein hydrolysate, 20 mg of adenine and 20 mg of glucose) was inoculated with a single yeast colony and incubated at 30°C with constant shaking for 2 - 3 days, or until the culture reached stationary phase. Yeast cells were pelleted by centrifugation at 3000 rpm for 10 min in a Beckman centrifuge. The supernatant was decanted and the yeast cells washed twice in 10 ml SE (75 mM NaCl, 25 mM Na₂EDTA pH 8) and resuspended in 4 ml SE. The cell suspension, as well as 1% low melting point agarose in SE, and paraffin oil, were equilibrated to 45°C. A beaker

containing 70 ml ice-cold SE and a magnetic stir bar was placed in an ice bucket on a stir plate set at medium speed. Four ml of the 1% low melting point agarose was added to the warmed cell suspension and mixed. Fifteen ml of pre-warmed paraffin oil was added to the suspension of cells in agarose and the mixture swirled vigorously to form a uniform emulsion. The emulsion was rapidly poured into the cold SE and stirred for 2 - 3 min. The mixture was transferred to 2 Falconer tubes and centrifuged in a Jouan centrifuge at 3000 rpm for 10 min. The majority of the paraffin oil and SE supernatant was removed leaving a layer of beads at the bottom of the tube. The beads were dispersed into the remaining supernatant and re-centrifuged. After centrifugation the excess SE and any unpelleted beads were removed and the pellets combined in a single tube. For removal of the yeast cell wall, 0.5 ml 2-mercaptoethanol, 5 mg zymolyase (Sigma) was added and the final volume adjusted to 10 ml with SE. The beads were mixed and incubated at 37°C for 2 hours. The beads were pelleted, resuspended in 20 ml (w/v) sarkosyl, 25 mM Na₂EDTA pH 8, 50 µg/ml proteinase K solution, and incubated overnight to degrade the yeast spheroblasts. The Proteinase K digested beads were pelleted and washed four times in 50 ml TE before storage at 4°C.

2.3.5.2 Preparation of agarose blocks containing yeast DNA (lithium method) (Chaplin and Brownstein, 1995)

Ten ml of AHC+Tet (50 µg/ml) medium was inoculated with a single yeast colony and incubated at 30°C overnight with shaking. Four ml of this culture was used to inoculate 400 ml AHC+Tet (50 µg/ml) medium before a 24 - 48 hour incubation at 30°C. After incubation, the yeast cells were pelleted (5000 rpm/5 min/4°C), resuspended in 50 mM EDTA solution, then pelleted again as before. The supernatant was decanted and the yeast cell pellet resuspended in 2 ml of a solution containing 1 M sorbitol, 20 mM EDTA, 14 mM 2-mercaptoethanol and 1 mg/ml lyticase. To the resuspended yeast cells was added 3 mls of molten 2% low melting point agarose in 1 M sorbitol, 20 mM EDTA and 2-mercaptoethanol at 45°C (agarose was dissolved by boiling, and then cooled to 45°C before the addition of 2-mercaptoethanol). The resulting solution was dispersed into plug

molds, which were set on ice. Plugs were extruded into 5 ml of 1 M sorbitol, 20 mM EDTA, 14 mM 2-mercaptoethanol, 10 mM Tris-HCl (pH 7.5), 1 mg/ml lyticase solution, then incubated at 37°C for 2 hours. After incubation the solution was removed and replaced with 5 mls of lithium lysis solution (1% lithium dodecyl sulphate, 100 mM EDTA, 10 mM Tris-HCl (pH 8.0), filtered). The resuspended plugs were incubated at 37°C for 1 hour, and the solution drawn off and replaced with fresh lithium lysis solution. After a 37°C overnight incubation, the lithium lysis buffer was removed, and the blocks washed 2 x 30 min in TE at 50°C, then 2 x 30 min in TE at room temperature. After washing, agarose blocks were stored in 0.5 M EDTA at 4°C.

2.3.5.3 Preparation of mammalian cell line DNA in agarose blocks (Kenrick *et al.*, 1987)

Cultured mammalian cells were washed twice in PBS (pellet cells at 3000 rpm, 5 mins at 4°C) and resuspended in PBS to a concentration of $\sim 2 \times 10^7$ cells/ml. The resuspended cells were warmed to 37°C and an equal volume of 1.4 % low melting point agarose in PBS at 50°C was added. The agarose/cell suspension mixture was maintained at 45°C until immediately before dispersion into plug molds, which when filled were placed on ice until the agarose was set. The solidified blocks were extruded into 2 volumes of 0.5 M EDTA, 1% SDS, 2 mg/ml proteinase K and incubated for 48 hours at 50°C. After incubation, blocks were washed 4 times in 10 volumes TE at 50°C (PMSF not required), then twice in 10 volumes TE at room temperature, with each washing step of 30 min duration. Washed blocks were stored in 0.5 M EDTA at 4°C.

2.3.5.4 Restriction digestion of DNA in agarose beads and blocks (modification of Finney, 1994)

For restriction endonuclease digestion of DNA contained within agarose beads, approximately 100 µl of bead suspension was placed in an eppendorf tube and washed twice with 1 ml of sterile distilled water. To the washed beads was added 20 µl of 10 x restriction buffer, and the volume made up to 200 µl. Restriction endonuclease was added

at a quantity of approximately 2 - 5 units/ μ g of DNA. The reaction was incubated at the appropriate temperature (specific for the restriction endonuclease) overnight. After incubation, the beads were centrifuged and the supernatant removed before addition of 0.1 volume of 10 x loading buffer.

2.3.5.5 Loading agarose beads into wells.

Agarose beads were loaded into wells prior to gel submersion in the buffer tank of the PFGE apparatus. The agarose bead sample was loaded into the well using a cut-off pipette tip, and the top of each well was sealed with 1% low melting point agarose before submersion in the electrophoresis chamber.

2.3.5.6 Loading agarose blocks into wells

Agarose blocks containing immobilised DNA were loaded into wells prior to gel submersion in the PFGE apparatus buffer tank. Each block was cut to the size of a well, then loaded using a small spatula, taking care that the interface between the block and the front of the well contained no bubbles. The top of each well was then sealed with 1% low melting point agarose before submersion.

2.3.5.7 Switching intervals

In order to achieve CHEF-PFGE separation of large DNA molecules it is generally necessary to empirically determine the appropriate switching intervals to achieve the desired result. The critical parameters used to calculate these values were described by Finney *et al.* (1994). However, this was unnecessary when utilising the Chef Mapper™ (Bio-rad Laboratories) used for this project, as this apparatus has an integrated software system whereby the parameters defining the required window of DNA separation are entered on the apparatus control panel, and an algorithm automatically selects the appropriate switching ratios to maximise resolution within that window. Alternatively, the software included barcode programs for the resolution of specific size ranges of DNA molecules.

The size range of DNA separation and switching intervals required for analysis of DNA in this project are described in the relevant figure captions.

2.3.5.8 DNA size markers for PFGE

A combination of commercially prepared λ DNA and yeast chromosomes were used as molecular weight markers for PFGE. For lower molecular weight DNA separations λ HindIII (Pharmacia cat. # 27-4048-01) provided markers spanning 2 - 23 kb. For higher molecular weight separations, λ DNA-PFGE (Pharmacia cat. # 27-4530-01), a preparation of λ DNA concatamers spanning 50 - 1000 kb, and Yeast DNA-PFGE (Pharmacia cat. # 27-4520-01) comprising whole *Saccharomyces cerevisiae* chromosomes spanning 200 - 2000 kb, were used as markers.

2.3.5.9 Southern blotting and hybridisation of pulsed field gels

Southern blotting and hybridisation of pulsed field gels was performed as described in section 2.3.4, except that immediately after EtBr staining a UV nicking step was added in order to reduce the large DNA fragments to a suitable size for Southern blot transfer. A Bio-Rad Gene Linker™ UV chamber on the 'nic' program setting (designed for UV nicking DNA in PFGE gels) was used for this additional step, as described in the instruction manual for the apparatus.

2.3.6 Polymerase Chain Reaction (PCR) (modification of Kogan *et al.*, 1987)

All PCR reactions were performed in a Perkin-Elmer Cetus 480 thermal cycler. Reactions typically comprised 10 μ l 2 x PCR mix, 1.5 - 3 mM $MgCl_2$ ($MgCl_2$ concentration was optimised for specific primer pairs, although in general, satisfactory results were obtained with a 1.5 mM concentration), 150 ng of each primer, template DNA, 1 unit of *Taq* DNA polymerase, with sterile water to 20 μ l. To generate radioactively labelled PCR products, which were usually required for microsatellite

analysis, 10 μ l PCR reactions were set up (containing the same proportions of reagents as the 20 μ l reactions) incorporating 0.5 μ l (5 μ Ci) of α - ^{32} P-dCTP.

After addition of all reagents, PCR reactions were mixed and overlaid with paraffin oil. The most frequently used thermal cycling program was: [94°C 1 min., 60°C 1.5 min., 72°C 1.5 min.] x 10 cycles, then [94°C 1 min., 55°C 1.5 min., 72°C 1.5 min.] x 25 cycles, then 72°C 10 min. If this regime was unsuccessful, 35 cycles of thermal cycling were performed with denaturation temperature 94°C, annealing temperatures in accordance with the calculated melting temperature (T_M) of the primers, and elongation temperature at 72°C (optimal for *Taq* polymerase). The elongation step time was assigned in accordance with the expected size of the PCR product.

2.3.6.1 PCR across CCG repeat tracts (Yu *et al.*, 1992)

PCR amplification across CCG repeat tracts was accomplished as described previously (section 2.3.6), except that dGTP was replaced with 7-deaza-dGTP in the reaction mix to compensate for the exceptionally high CG content of the region.

2.3.6.2 PCR primers (Saiki, 1989)

Oligonucleotides for PCR were designed to contain a similar proportion of purine and pyrimidine residues, no more than 4 consecutive residues of the same base, and no repeat sequence DNA. To inhibit the formation of primer-dimers, primers to be used as pairs did not have identical base residues at the 3' end. The length of oligonucleotides synthesised for the study varied from 20 to 25 basepairs. Oligonucleotide primers used in this project were synthesised in the laboratory by K. Holman and S. Richardson on an Applied Biosystems 391, or Beckman Oligo1000 DNA synthesisers. In the later stages of the project commercially synthesised oligonucleotide primers were obtained from Gibco BRL Life Technologies (Gaithersburg, Maryland, USA).

2.3.7 DNA Sequencing

2.3.7.1 Fluorescent automated sequencing

A Perkin Elmer (Applied Biosystems) 373A DNA sequencer was utilised for the fluorescent automated sequencing of plasmid and PCR amplified DNAs. Sequencing reactions were performed with reagents and protocols from Perkin Elmer PRISM ready reaction sequencing kits. Specifically, DyeDeoxy™ Terminator kits with custom primers were used for either plasmid and PCR product templates, and Dye Primer kits with dye labelled -21M13 and M13 Rev primers were used for plasmid templates only.

In general, for DyeDeoxy™ Terminator sequencing of plasmid inserts, ~500 - 1000 ng of a DNA preparation with 20 ng of vector or insert specific primer was used in each reaction. For Dye Primer sequencing, ~1000 ng of plasmid DNA was used in each reaction, and the fluorescently labelled vector specific primers were a component of the reaction mixtures provided with the kit. Sequencing of PCR generated templates was performed only by DyeDeoxy™ Terminator sequencing. For each reaction, ~100 - 200 ng of template DNA and 20 ng of template specific primer was used.

Fluorescently labelled sequencing reactions were run on denaturing polyacrylamide gels according to the protocols described in the Applied Biosystems 'Quick Reference Guide' for the 373A Sequencing System (Part Number 902347). The resulting DNA sequence data were software analysed and could be visualised as a 4-colour chromatogram.

2.3.7.2 Manual DNA sequencing

Sequencing of plasmid DNAs by a single extension chain termination method
(Peterson, 1988, Innis *et al.*, 1989)

Prior denaturation of the double stranded plasmids was required for this method. For template denaturation, 1 - 2 µg of plasmid DNA in 10 µl dH₂O was added to 2 µl of a 2 M NaOH/2 mM EDTA denaturing solution. The mixture was vortexed and incubated at room temperature for 10 min, then neutralised by the addition of 3 µl sodium acetate and 7 µl water. Plasmid DNA was precipitated by the addition of 60 µl ethanol and a 30 min incubation on ice. Precipitated DNA was pelleted with a 10 min at 14000 rpm

centrifugation, then rinsed with 70% ethanol and vacuum dried. The dried pellet was resuspended in 7 μ l H₂O in preparation for primer annealing (Chaplin and Brownstein, 1995).

For annealing of the sequencing primer, to the 7 μ l DNA template solution was added 2 μ l of a 5 x reaction buffer (250 mM Tris-HCl (pH8.8), 35 mM MgCl₂) and 1 μ l of sequencing primer (10 ng/ μ l). The mixture was incubated at 70°C for 3 min, then at 42°C for 10 min.

For the primer extension and labelling step, 2 μ l of a labelling mix (1.5 μ M dGTP, 1.5 μ M dATP, 1.5 μ M dTTP), 0.5 μ l α^{32} P-dCTP, and 2 μ l of 1 u/ μ l Perkin Elmer sequencing grade *Taq* DNA Polymerase (8 u/ μ l stock diluted in enzyme dilution buffer of 0.1 mM EDTA, 0.15% Tween 20®, 0.15% Nonidet P40®, 25 mM Tris-HCl, pH8.8) was added to the annealed template mixture. The resulting solution was mixed and incubated at 45°C for 5 min, then placed at room temperature. Termination was initiated by adding 4 μ l aliquots of a primer extension reaction to 4 separate tubes, each containing 4 μ l of either C, G, T, or A termination mix (each mix contained 20 μ M dGTP, 20 μ M dCTP, 20 μ M dTTP, 20 μ M dATP, and for termination at specific bases, 60 μ M ddGTP, 400 μ M ddCTP, 800 μ M ddTTP or 800 μ M ddATP). The contents of each tube were mixed, then incubated at 70°C for 5 min. Prior to the loading and electrophoresis of the sequencing samples on a 6% wedge polyacrylamide gel, 4 μ l of formamide loading buffer was added to each tube, and the DNA denatured at 95°C for 5 min.

Cyclist™ Exo-*Pfu* DNA sequencing

Typically, DNA sequencing with reagents from the Stratagene Cyclist™ Exo-*Pfu* DNA Sequencing Kit (cat. # 200326) was performed for the characterisation of DNA regions refractory to fluorescent automated sequencing methods. Sequencing reactions were performed according to the instruction manual provided with the kit, then loading dye provided with the kit) added. Immediately before loading, samples were heated at >80°C for 2 - 5 min.

2.3.7.3 Electrophoresis of manual sequencing samples (modification of Slatko and Albright, 1991)

In order to obtain the maximum amount of sequence data, double loading of sequencing samples was performed. Shark's teeth combs were used to form wells. Two to 4 μ l of each sample was loaded in a staggered fashion onto a 42 cm x 50 cm, 0.25 mm thick 6% (w/v) polyacrylamide/7 M urea wedge sequencing gel. Electrophoresis was performed at 2000 V, adjusting as necessary to maintain a constant temperature of 50°C. After first loading, samples were run until the xylene cyanol band from the loading dye reached the bottom of the gel. A duplicate loading was performed, and the gel run until the bromophenol blue from the loading dye reached the bottom of the gel. The gel was dried in an unfixed state, then autoradiographed at room temperature without intensifying screens (to reduce the scatter of radioactive signal).

2.3.8 Resolving minisatellite alleles (modification of Hudson *et al.*, 1994)

For typing minisatellite alleles, 42 cm x 50 cm, 0.4 mm thick 5% (w/v) polyacrylamide/7 M urea gels were used. Combs designed to form wells within the gel were used in preference to shark's teeth combs.

Before loading, 30 μ l of 2 x formamide loading buffer was added to each 10 μ l PCR reaction, and the mixture was heated for 5 min at 95°C. In general, gels were electrophoresed at 2000 V for 3 - 4 hours, depending on the size of the PCR product to be resolved. After electrophoresis, gels were dried unfixed and exposed to X-ray film at -70°C with intensifying screens.

Chapter 3

*Pulsed Field Long Range
Restriction Mapping
within the FRA16A
Interval*

Chapter 3: Contents

PFGE Long Range DNA Restriction Mapping within the FRA16A Interval

3.1	Summary	79
3.2	Introduction	80
3.3	Materials and Methods	81
3.3.1	DNA probes.....	81
3.3.2	YAC clones	81
3.3.3	Hybrid cell line panel	81
3.3.4	Fluorescent <i>in situ</i> hybridisation (FISH) analysis	81
3.3.5	DNA preparation and restriction digests for PFGE.....	82
3.3.6	Pulsed field gel electrophoresis.....	82
3.3.7	Southern blotting and hybridisation	82
3.3.8	Subcloning of cosmid and λ clones in order to isolate unique DNA fragments suitable for probing PFGE filters.....	82
3.3.9	Copies of the 1.79 probe sequence are present both proximal and distal to <i>FRA16A</i>	83
3.4	Results	84
3.4.1	FISH analysis with probes from within the <i>FRA16A</i> interval.....	84
3.4.2	Physical linkage of the DNA markers 1.79, 16XE81, VK20, and c305F6 on PFGE restriction fragments	84
3.4.3	Identification of anomalous <i>Not</i> I and <i>Bss</i> H II restriction fragments in DNA from <i>FRA16A</i> individuals.....	85
3.4.4	Ordering of loci with respect to cytogenetic breakpoints	87
3.4.5	Isolation and mapping of YAC clones containing the marker 16XE81	88
2.4.6	Identification of further cosmid clones mapping within the <i>FRA16A</i> interval	88

3.4.7	PFGE mapping of a somatic cell hybrid breakpoint.....	89
3.4.8	Identification of CEPH megaYACs with an STS from c37C6.....	89
3.4.9	Identification of a CCG repeat sequence within My769H1.....	90
3.5	Discussion	91
3.5.1	Strategy to localise <i>FRA16A</i>	91
3.5.2	Duplicated regions and chromosome 16 specific repeats in the <i>FRA16A</i> region.....	93
3.5.3	Rearrangements in megaYACs	94
3.5.4	Characterisation of <i>FRAXE</i> , <i>FRAXF</i> and <i>FRA11B</i>	95
3.5.5	The nature of the genomic DNA region surrounding <i>FRA16A</i>	98

3.1 Summary

Long-range pulsed field gel electrophoresis mapping of rare cutting restriction endonuclease sites within the somatic cell hybrid breakpoint interval containing the rare folate sensitive fragile site *FRA16A* was performed with *FRA16A* and non-*FRA16A* DNA samples. Within the interval, co-localising *Not* I and *Bss*H II sites were identified that appeared to be cleaved on all non-*FRA16A* (normal) chromosomes, but to remain uncleaved on *FRA16A* chromosomes. Given that hypermethylation at *FRAXA*, another rare folate sensitive fragile site, is associated with non-cleavage of rare cutter endonuclease sites, the differentially cleaved *Not* I and *Bss*H II sites in the *FRA16A* region were considered to potentially co-localise with the fragile site. To clone these restriction sites, YAC clones overlapping the closest flanking markers were identified and restriction mapped. The most distally mapping clone, My769H1, was found to contain the co-localising *Bss*H II and *Not* I sites of interest. Screening of the YACs with a CCG repeat probe identified a CCG repeat hybridising region within My769H1. Further restriction analysis revealed this region was localised immediately adjacent to the single *Not* I site within the YAC. Taken together, these findings indicated the CCG repeat hybridising region within My769H1 to be an excellent candidate for the *FRA16A* locus.

3.2 Introduction

Prior to the commencement of the project described in this thesis, many aspects of the inheritance, mutations and hypermethylation associated with the X-chromosomal rare folate sensitive fragile site *FRAXA* (the primary genetic feature of the fragile X syndrome) had been intensively studied (reviewed in section 1.3). Expression of this fragile site had been correlated with an expanded CCG repeat tract, as well as the hypermethylation of a CpG island of which the repeat tract forms part (the *FMR1* CpG island). As *FRA16A* was also classified as a rare folate sensitive fragile site (although autosomal), there was an expectation that it would have similar features, unless these were in some way related to the X-chromosomal location of *FRAXA*.

The hypermethylation of rare cutter restriction sites within the *FMR1* CpG island enabled the *FRAXA* fragile site to be accurately localised to the vicinity of certain rare cutting enzyme sites with the PFGE long range mapping technique (Bell *et al.*, 1991; Vincent *et al.*, 1991). In view of the success of this approach, a similar one was attempted for the localisation of *FRA16A*. An advantage of this strategy was that even if a *FRA16A* specific hypermethylated restriction site was not identified, other useful mapping data might be obtained (i.e. the identification of restriction fragments spanning the fragile site) thus determining the maximum distance to be bridged with YACs.

Physical mapping of many anonymous markers (cosmid and plasmid probes) on human chromosome 16 to intervals between hybrid cell line translocation breakpoints had identified four DNA markers (1.79, 16XE81, VK20, 305F6) within the interval containing *FRA16A*. With these probes as resources, the aims of this initial section of the *FRA16A* project were: 1) generation of an uncloned genomic DNA PFGE restriction map to order the probes with respect to each other and the surrounding hybrid breakpoints, and identification of any rare cutting restriction site(s) hypermethylated in *FRA16A* individuals; 2) screening of YAC libraries with a marker mapping in closest proximity to *FRA16A*, with the aim of identifying a YAC clone spanning the fragile site.

In this chapter the uncloned genomic and YAC PFGE restriction maps that facilitated the cloning of *FRA16A* will be presented and discussed.

3.3 Materials and Methods

3.3.1 DNA probes

Details of the probes used in this study are given in Table 3.1.

3.3.2 YAC clones

Details of the YAC clones used in this study are given in Table 3.2.

3.3.3 Hybrid cell line panel

DNAs from four somatic cell hybrids (CY185, CY183, CY163, CY11) containing either whole, or translocated sections of human chromosome 16 with breakpoints in the region of interest, were used in this study. A diagrammatic representation of the portion of human chromosome 16 present in each hybrid cell line is shown in Fig. 3.1.

Callen *et al.* (1986) described details of the construction of somatic cell hybrids and the human parental cell lines from which they were derived.

3.3.4 Fluorescent *in situ* hybridisation (FISH) analysis (performed by H. Eyre)

DNA probes 1.79, 16XE81, VK20 and c305F6 (not subclones) were used as probes for FISH analysis with metaphase chromosomes expressing *FRA16A*. Kremer *et al.* (1991b) described a protocol for the fluorescent *in situ* hybridisation (FISH) technique used in this study.

Table 3.1 DNA markers mapped between the breakpoints of the translocated chromosome 16s in the somatic cell hybrids CY11 and CY185, the interval containing *FRA16A*

locus	probe	insert size/vector	references
D16S79 A and B	1.79 or 36.1	1 kb/pKUN	Breuning <i>et al.</i> , 1989 Callen <i>et al.</i> , 1986 Gedeon <i>et al.</i> , 1989
D16S287	16XE81	15.1 kb/EMBL3	Phillips <i>et al.</i> , 1991
D16S96	VK20	14.9 kb/EMBL3	Hyland <i>et al.</i> , 1989a; 1989b
D16S1174	c305F6	27.14 kb/sCos-1	Stallings <i>et al.</i> , 1990; 1992
D16S1688	c37C6	29.167 kb/sCos-1	Stallings <i>et al.</i> , 1990; 1992

Table 3.2 CEPH YACs used for the *FRA16A* study. The YAC clones were derived from 2 total human DNA YAC libraries constructed by Dr. D. Le Paslier and colleagues at the Centre d'Etude du Polymorphisme Humain (CEPH), Paris, France. The method for YAC library construction was reported by Burke *et al.* (1987), and Abidi *et al.* (1990).

YAC	human DNA YAC insert size (kb)
y147C9	700
y168F4	485
y242E5	470
y326E8	340
y415A2	200
My769H1	830, 1670
My790A10	840

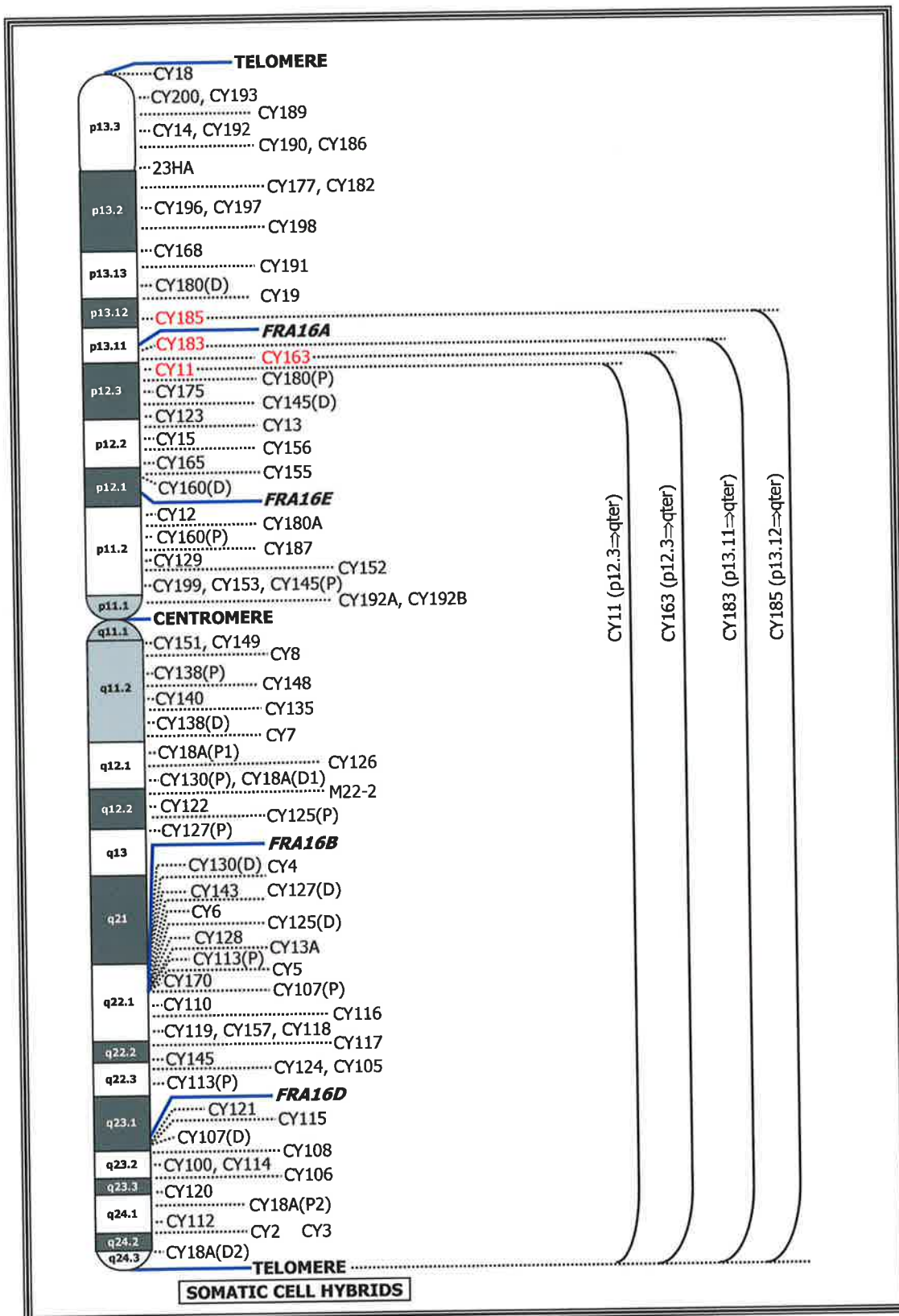


Figure 3.1 Map of somatic hybrid cell line breakpoints on human chromosome 16. The cytogenetic 'anchor points' of the map are indicated by blue lines. The broken lines indicate the position of hybrid breakpoints relative to each other. The cell lines used for mapping *FRA16A* are shown in red text. The large brackets define the portions of human chromosome 16 within the somatic cell hybrids CY11, CY163, CY183, CY185. Details of the construction and constitution of the somatic cell hybrids are given in Callen *et al.* (1986; 1989; 1990; 1992).

3.3.5 DNA preparation and restriction digests for PFGE (protocols described in sections 2.3.5.3 and 2.3.5.4)

Genomic DNAs embedded in agarose blocks were prepared from lymphoblastoid cell lines from *FRA16A* individuals, non-*FRA16A* individuals and hybrid cell lines. Agarose embedded DNAs were digested with the restriction endonucleases *Not* I, *Bss*H II and *Sal* I.

Yeast DNA containing YAC clones was prepared in agarose beads and digested with the restriction endonucleases *Not* I, *Bss*H II, *Sal* I and *Mlu* I.

3.3.6 Pulsed field gel electrophoresis

Pulsed field gel electrophoresis was performed on Bio-Rad Chef Mapper™ apparatus, according to manufacturer's instructions (further details are described in section 2.3.5). The pulsed field electrophoresis conditions used for optimal resolution of the DNA fragment(s) of interest are detailed in the figure captions.

3.3.7 Southern blotting and hybridisation

Southern blot filters of pulsed field gels were prepared according to the method described in section 2.3.4.8. The PFGE filters were probed and washed under stringent conditions (section 2.3.4.9). Between rounds of hybridisation, radiolabelled probes were stripped from filters (section 2.3.4.10).

3.3.8 Subcloning of cosmid and λ clones in order to isolate unique human DNA fragments suitable for probing PFGE filters

The λ clone 16XE81, and the cosmid clones c305F6 and c37C6 were known to contain human repetitive sequences and therefore to be unsuitable for use as probes on PFGE Southern blot filters (even when pre-reassociated). In order to generate the single copy probes for this application, it was therefore necessary to subclone a unique sequence restriction fragment from each of the 3 clones. This approach also enabled (technically) straightforward sequence analysis (ie. not directly sequencing from lambda or cosmid

clones), allowing subsequent design of STS primers for CEPH megaYAC library screening.

In brief, the procedure for subcloning unique fragments consisted of the following steps:

1) Lambda/cosmid clone DNA(s) were digested with *Eco* RI or *Pst* I and electrophoresed on 0.8% agarose gels for resolution and sizing of fragments.

2) Gels were Southern blotted, and filters probed with ^{32}P labelled total human DNA for identification of restriction fragments containing repetitive sequences.

3) Fragments containing no apparent repetitive sequences, and not derived from the vector, were subcloned (see section 2.3.3) and hybridised to PFGE southern blot filters.

Details of subclones used in the PFGE mapping are given in Table 3.3.

3.3.9 Copies of the 1.79 probe sequence are present both proximal and distal to *FRA16A*

Previous FISH analysis with probe 1.79 found hybridisation signals localising to the chromosomal regions immediately proximal and distal to *FRA16A*, at approximately equal frequency. This was initially taken to indicate that this probe might span the fragile site. However, complete sequencing of the insert did not reveal any repetitive tracts or other unusual sequences (A. Thompson, personal communication). In addition, the results of a study of two RFLPs detected by the probe were inconsistent with single copy status (see section 1.2.9), and when taken together with the FISH data, indicated a probable duplication of 1.79 sequences within the somatic cell hybrid interval containing *FRA16A* (Callen *et al.*, 1992).

Table 3.3 Plasmid subclones and primer pairs derived from Table 3.1 DNA clones

plasmid subclone	plasmid vector	human DNA insert size	insertion site(s)
p16XE81-1	pBR322	1.5 kb	<i>Hind</i> III
pVK20a*	pUC19	1.6 kb	<i>Sal</i> I, <i>Hind</i> III
p305F6-3	pUC19	0.8 kb	<i>Eco</i> RI
p37C6-2	pUC19	1.2 kb	<i>Eco</i> RI

primer name	primer sequence
16XE81F**	GCTTGTATTAGTCAGCATTCTCCAG
16XE81R**	TACAGACCATAGACTTGACAGTCTC
37C6F	CCTTAGGTGGAAGGAGTGGT
37C6R	CACTGGTGCATCAGCATTG

* subcloned by V.J. Hyland

** designed by H.A. Phillips

3.4 Results

3.4.1 FISH analysis with probes from within the *FRA16A* interval

FISH analysis with cosmid c305F6 detected signals both proximal and distal to the fragile site. Therefore this cosmid was assumed to also lie within a proximal/distal duplicated region (H. Eyre, personal communication). Probes pVK20a and p16XE81-1 hybridised only proximally to the fragile site, and were therefore assumed to lie within a unique DNA region. However, in order to prove these assumptions complete DNA sequencing of the region might be required.

3.4.2 Physical linkage of the DNA markers 1.79, 16XE81, VK20 and c305F6 on PFGE restriction fragments

Before the commencement of this study, a number of lines of evidence had led to the conclusion that D16S79 locus sequences are duplicated within the somatic cell hybrid interval containing *FRA16A* (Callen *et al.*, 1992). It was therefore assumed, for the purpose of interpreting the PFGE banding patterns produced by probes within the region, that the 1.79 probe could potentially hybridise to two separate sets of restriction fragments, representing both copies of the duplication (see section 1.2.9).

The results of successive hybridisations of 1.79, p16XE81-1, pVK20a and p305F6-3 to PFGE filters containing *Not* I, *Bss*H II and *Sal* I digested normal lymphoblastoid cell line DNAs provided strong evidence for the physical linkage of these probes. Probe 1.79 hybridised to two *Not* I bands of 1200 kb and 790 kb, four *Bss*H II bands of 1200 kb, 790 kb, 700 kb, 350 kb, and six *Sal* I bands of 1100 kb, 750 kb, 650 kb, 350 kb, 300 kb, and 100 kb (Fig. 3.2). Hybridisation of 1.79 and pVK20a to another filter (containing both non-*FRA16A* and *FRA16A* DNAs) found that pVK20a detected the 1.79 specific 790 kb *Not* I and *Bss*H II fragments (Fig. 3.3(1) A and B). Probe p16XE81-1 detected a set of fragments identical to those of pVK20a (data not shown). Probe pVK20a was also found to hybridise to the 1.79 specific 750 kb and 300 kb *Sal* I fragments (data not shown). The

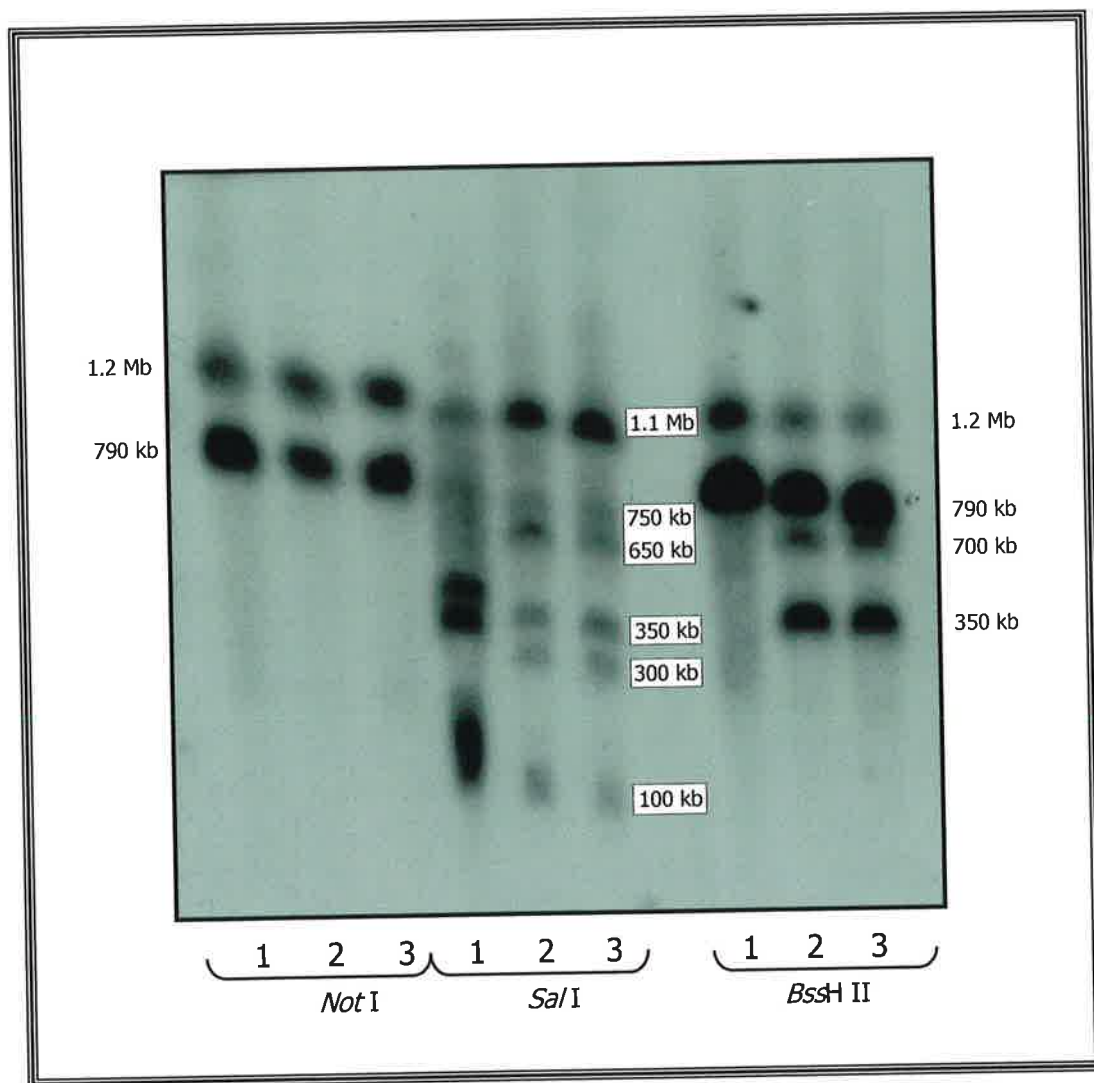


Figure 3.2 Hybridisation of the probe 1.79 to *Not* I, *Sal* I and *Bss*H II digested non-*FRA16A* DNAs. Samples: **1.** male (lymphoblastoid cell line DNA); **2.** female (lymphocyte DNA); **3.** male (lymphocyte DNA). Pulsed field gel electrophoresis performed on Bio-Rad Chef Mapper™ apparatus. Conditions for resolution of DNA fragments within the size range 200 - 2200 kb were: gel: 1% agarose in 0.5 x TBE, running temperature: 14°C, angle of electrodes: 120°, switch times: 60 - 110 sec, voltage gradient: 6V/cm, run time: 24 hours. Protocol obtained from Chef Mapper™ instruction manual (barcode 4).

Figure 3.3 (1) Hybridisation of probes to *Not* I and *Bss*H II digested lymphoblastoid DNAs from non-*FRA16A* and *FRA16A* individuals. Samples: 1) non-*FRA16A* male; 2) *FRA16A* male; 3) *FRA16A* male; 4) *FRA16A* male; 5) non-*FRA16A* female; 6) CY18 cell line; 7) non-*FRA16A* male. Restriction fragment localisations with respect to *FRA16A* : ∞ proximal and distal, ♠ proximal, ♠ distal.

(2) Proposed long range pulsed field restriction map of the *FRA16A* region. The predicted positions (in relation to the restriction map) of the *Not* I and *Bss*H II fragments detected by the probes are shown.

Pulsed field gel electrophoresis performed on Bio-Rad Chef Mapper™ apparatus. Conditions for resolution of DNA fragments within the size range 200 - 2200 kb were: gel: 1% agarose in 0.5 x TBE, running temperature: 14°C, angle of electrodes: 120°, switch times: 60 - 110 sec, voltage gradient: 6V/cm, run time: 24 hours. Protocol obtained from Chef Mapper™ instruction manual (barcode 4).

mapping data therefore indicated that the 16XE81, VK20 and proximal 1.79 (D16S79B) loci were located within 300 kb of each other.

The p305F6-3 probe (assumed to detect a duplicated sequence) was found to detect all the 1.79 restriction fragments not hybridised by p16XE81-1 and pVK20a, i.e. the 1200 kb *Not* I and 1200 kb, 350 kb *Bss*H II 1.79 bands (Fig. 3.3(1) A and C). As they were not detected by the proximal probes (p16XE81-1 and pVK20a), this subset of 1.79 restriction fragments was assumed to originate from the D16S79 locus (D16S79A), located distal to the fragile site. On the basis of this assumption the 305F6 and the 1.79 distal loci were mapped to within 350 kb of each other. The p305F6-3 probe also hybridised to 200 kb *Not* I and 300 kb *Bss*H II fragments that were not in common with those detected by 1.79 (or the other probes). Cosmid c305F6 sequences were thought to be duplicated, and were known to detect bands in common with the 1.79 distal band subset. Therefore, it was assumed for the purpose of constructing a restriction map that these bands represented restriction fragments associated with the proximal copy of the duplication (but not in common with p16XE81-1 or pVK20a). A proposed restriction map of the region, showing the locations of the *Not* I and *Bss*H II fragments discussed above, is shown in Fig. 3.3(2).

3.4.3 Identification of anomalous *Not* I and *Bss*H II restriction fragments in DNA from *FRA16A* individuals

Lymphoblastoid cell line DNAs from three *FRA16A* individuals were digested with the restriction endonucleases *Not* I and *Bss*H II, and the resulting fragments resolved by pulsed field gel electrophoresis. As controls, DNA samples from 3 lymphoblastoid cell lines (non-*FRA16A*) and 1 somatic cell hybrid were included. Direct comparisons of the 1.79 *Not* I bands revealed the 1200 kb and 790 kb bands from the *FRA16A* DNAs to be of a lower intensity than those from non-*FRA16A* DNAs. Furthermore, an unresolved *Not* I band of ~2 Mb was found for *FRA16A* DNAs that was not evident for the other samples on the Southern blot filter (Fig. 3.3(1)A).

Hybridisation of 1.79 to *Bss*H II digested *FRA16A* and non-*FRA16A* DNAs detected a more complex pattern of bands than found for *Not* I (Fig. 3.2). The 790 kb *Bss*H II band appeared identical in size to the smaller *Not* I fragment, indicating at least two *Bss*H II sites in the region co-localise with *Not* I sites. The *Bss*H II band of 790 kb appeared to be of lower intensity in all *FRA16A* DNAs. The 1200 kb *Bss*H II band also appeared to be reduced in intensity, or completely absent in *FRA16A* DNAs, as well as of reduced intensity in a subset of the non-*FRA16A* DNAs.

The reduced intensity of the 1200 kb *Bss*H II band found in non-*FRA16A* DNAs could be attributed to normally occurring features, such as partial digestion due to methylation, or an RFLP. The complete absence of the 1200 kb *Bss*H II band observed in some *FRA16A* DNAs could therefore come about as the result of the presence of the fragile site on one chromosome 16, and normal differences in methylation patterns or an RFLP (caused by the presence or absence of cleavage at a *Bss*H II site due to a base change), on the other.

The PFGE mapping data indicated that the two events (one normal and one abnormal) cause differential cleavage of *Bss*H II sites to occur at opposite ends of the *Bss*H II 1200 kb fragment. The appearance of a 350 kb *Bss*H II band in association with the absence or reduced intensity of the 1200 kb *Bss*H II band supported this explanation of the results. It could therefore be concluded that the 350 kb *Bss*H II band represents a complete digestion restriction fragment, and the 1200 kb *Bss*H II fragment a partial digestion fragment (containing the 350 kb fragment) that is present only when a certain *Bss*H II site is not cleaved.

A *Bss*H II fragment of more than 1200 kb in size (and variant in size) was detected in all *FRA16A* DNAs tested, but not in non-*FRA16A* DNAs. The variation in size of these larger fragments could be attributed to normally occurring differences (between chromosomes) in the methylation of the *Bss*H II sites occurring on the other side of the fragile site (Fig. 3.4). The assignment of the origin of 1.79 *Bss*H II fragments as either proximal or distal to the fragile site was determined by observations of hybridisation to common fragments by the proximal probes p16XE81-1 and pVK20a. It was subsequently

found that all 1.79 *Bss*H II fragments not detected by the proximal probes were common to fragments detected with p305F6-3.

Expression of the rare folate sensitive fragile site *FRAXA* is associated with methylation of adjacent rare cutting restriction sites, the locations of which were initially mapped by PFGE methods (Bell *et al.*, 1991; Vincent *et al.*, 1991). Therefore, for the purpose of constructing a PFGE restriction map of the *FRA16A* region consistent with the band patterns observed, it was assumed that, as for *FRAXA*, *FRA16A* is located adjacent to normally unmethylated rare cutting restriction fragments that may become methylated on chromosomes expressing the fragile site. The finding of non-cleavage at the *Not* I and *Bss*H II restriction sites located at the junction of the 1200 kb and 790 kb *Not* I/*Bss*H II fragments on *FRA16A* chromosomes was therefore attributed to the hypermethylation associated with folate sensitive fragile site expression. The PFGE and FISH results obtained further indicated that the location of *FRA16A* was between two duplicated regions of DNA. These factors have been accounted for in the proposed long-range restriction map of the region encompassing *FRA16A* that is presented in Fig. 3.3(2) and Fig. 3.4.

3.4.4 Ordering of loci with respect to cytogenetic breakpoints

Prior to commencement of the project, the 1.79, 16XE81, VK20 and c305F6 DNA markers had been mapped to an interval spanned by the breakpoints of hybrid cell lines CY11 and CY185 (Callen *et al.*, 1989; 1990). During the course of the project two more hybrid cell lines (CY163 and CY183) with translocation breakpoints in the interval became available (Callen *et al.*, 1992) and were used to further refine the ordering of markers with respect to the fragile site and each other. The localisations of markers with respect to hybrid cell line breakpoints in the vicinity of *FRA16A* are shown in Fig. 3.4.

Initial hybrid breakpoint mapping data from CY163 and CY185 (with available probes) identified 16XE81 as the closest known proximal marker to *FRA16A* because of its localisation distal to the breakpoint of the hybrid cell line CY163 (Fig. 3.4).

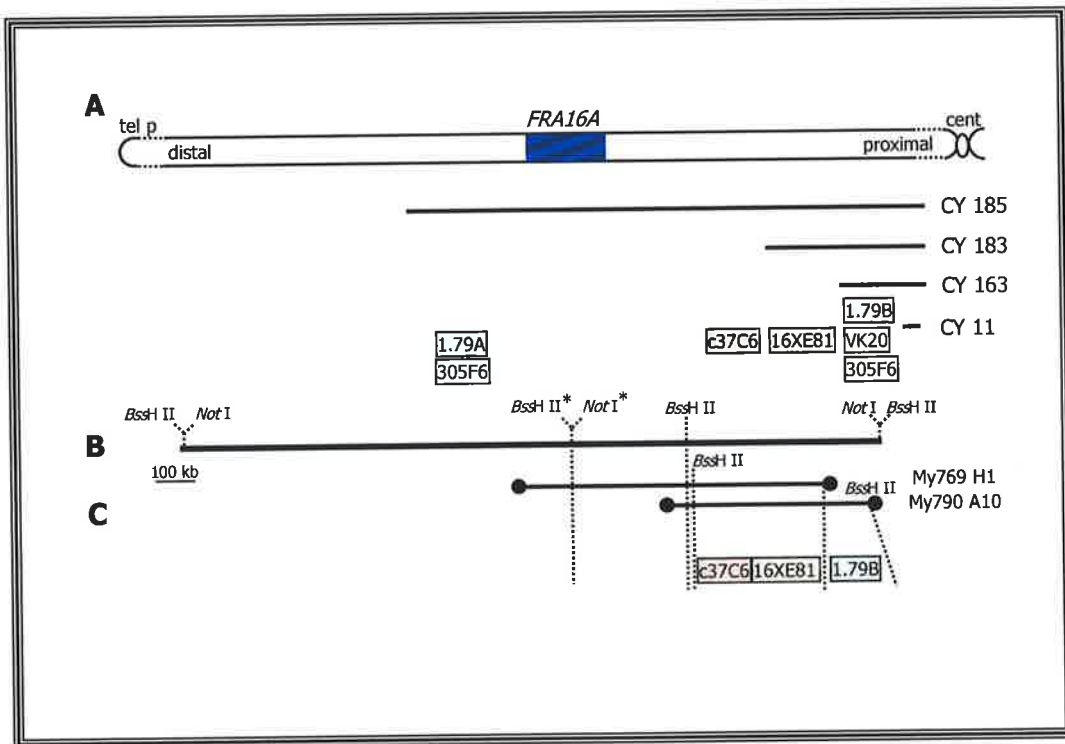


Figure 3.4 Mapping of *FRA16A*. **A.** Shows the relative positions of the somatic cell hybrid breakpoints on the short (p) arm of human chromosome 16 that were used to physically localise anonymous DNA probes (1.79, c37C6, 16XE81, c305F6 and VK20). Probe 1.79 hybridised to a duplicated DNA sequence (copies: 1.79A and 1.79B) present within two breakpoint intervals. **B.** The long range restriction map in the vicinity of *FRA16A* generated by pulsed field gel electrophoresis of *Not* I and *BssH* II digested normal and *FRA16A* DNAs probed with 1.79, c37C6 and 16XE81. The *Not* I and *BssH* II sites that exhibit differential methylation are indicated with an asterisk. **C.** Location and restriction map of My769H1 and My790A10 contig. y242E5 is not shown as the mapping information obtained from this YAC was redundant. *FRA16A* sequences were isolated from My769H1.

3.4.5 Isolation and mapping of YAC clones containing marker 16XE81 sequences (performed by H. Phillips and E. Kremer)

Primers designed for analysis of 16XE81 (AC)_n repeat polymorphism (16ACXE81F/16ACXE81R, Phillips *et al.*, 1991) were utilised for screening the CEPH YAC library (first generation). This screening identified the YACs y147C9, y168F4, y242E5, y326E8 and y415A2 as spanning the 16XE81 locus. Subsequent PFGE restriction mapping, however, did not identify co-localising *Bss*H II and *Not* I sites in any of these YACs. None of these YACs therefore appeared to span the region of interest, but could be utilised in subsequent steps toward cloning the fragile site.

3.4.6 Identification of further cosmid clones within the *FRA16A* interval (performed at Los Alamos National Laboratories (LANL) and by candidate)

DNA from YACs y168F4, y242E5, y326E8 and y415A2 was used as template for inter-Alu PCR, a method enabling the amplification of DNA sequences between Alu repeats within the human DNA inserts of YACs (Monaco *et al.*, 1991). The pre-reassociated inter-Alu PCR products were used to screen hybridisation membrane grids of chromosome 16 derived cosmid clone DNA representing ~1 x coverage of the chromosome. (Stallings *et al.*, 1990)

Cosmid clones corresponding to hybridisation signals on filters were mapped against the chromosome 16 somatic cell hybrid panel in order to localise them with respect to the existing mapped probes and breakpoints. Cosmids mapping in the vicinity of the fragile site were then used for FISH analysis against *FRA16A* chromosomes.

Based on this mapping data the cosmid c37C6 (a new marker in the *FRA16A* interval), which localised proximal to the fragile site by FISH and distal to the breakpoint of the somatic cell hybrid CY183, was identified as the closest proximal marker to the fragile site. This marker was therefore considered the most potentially useful one for subsequent approaches to cloning *FRA16A*. A unique sequence subclone of c37C6 (p37C6-2) was constructed for pulsed field mapping and sequence analysis (in order to generate STS primers for YAC library screening).

3.4.7 PFGE mapping of a somatic cell hybrid breakpoint

PFGE mapping of *FRA16A* region *Not* I and *Bss*H II sites in PK31.2 DNA (the parent cell line of somatic cell hybrid CY183) was performed in order to establish the maximum distance between the CY183 breakpoint (and therefore the c37C6 marker) and the *FRA16A* associated *Not* I or *Bss*H II sites located distal to the translocation breakpoint. In order to accomplish this, a *Bss*H II digest of PK31.2 DNA was probed with p37C6-2 and 1.79. The results of these hybridisations are shown in Fig. 3.5(A). A p37C6-2 *Bss*H II band of 350 kb was detected in PK31.2 that was not evident for any other cell line DNA digest (or detected by 1.79). The existence of this band was therefore attributed to the juxtaposition of a *Bss*H II site from another chromosome to a position 350 kb distal to the *FRA16A* *Bss*H II site, due to the CY183 (or PK31.2) translocation. On the basis of this assumption c37C6 was mapped to within 350 kb of *FRA16A*, a distance that could be spanned within the insert of one megaYAC. A restriction map of the region depicting the proposed localisation of the CY183 translocation breakpoint is shown in Fig. 3.5(B).

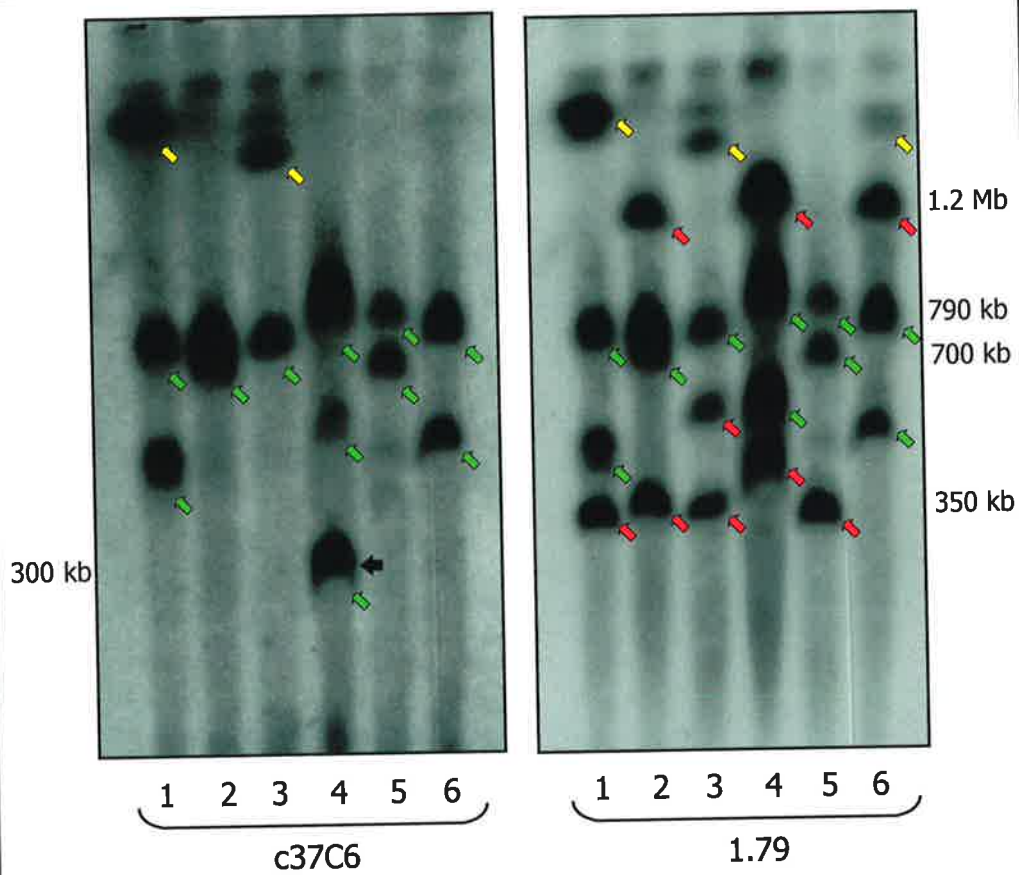
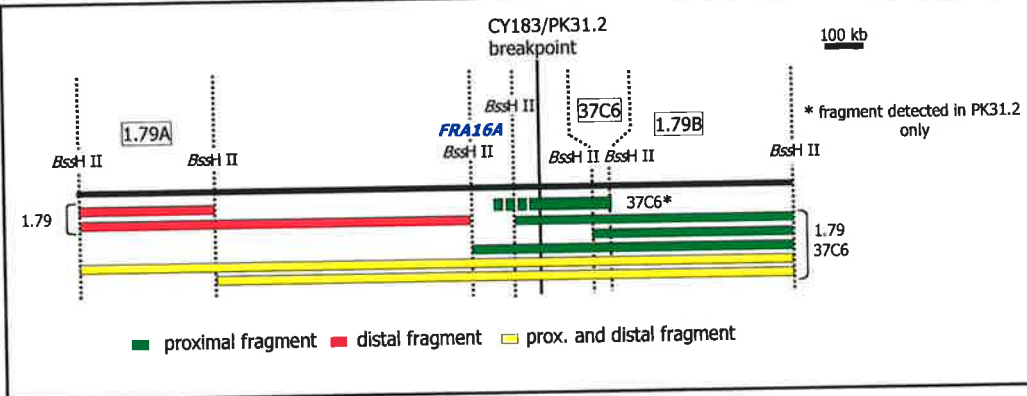
3.3.8 Identification of CEPH megaYACs by an STS from c37C6 (performed by CEPH and the candidate)

Greatly facilitating the next step in this project, the CEPH megaYAC library became available for PCR screening. The library contained YACs with insert sizes much larger than those of the previous CEPH YAC library (Evans, 1993). DNA pools of clones in the megaYAC library were PCR screened with the 37C6 STS primers. Screening results and the 37C6 STS primers were sent to CEPH, where further rounds of screening identified 11 megaYACs as potential positive clones.

DNA in agarose beads was prepared from each YAC and electrophoresed uncut on a pulsed field gel in order to resolve the yeast chromosomes and YACs. Southern blot hybridisation analysis with a 37C6 STS PCR product probe ascertained that only two megaYACs, My769H1 and My790A10, contained c37C6 sequences. The size of these megaYAC clones was estimated to be 800 kb and 650 kb, respectively. Screening (by

Figure 3.5 **A)** *Bss*H II digests of non-*FRA16A*, *FRA16A*, and CY183 parent cell line DNAs. Samples: 1) *FRA16A* male; 2) non-*FRA16A* male; 3) *FRA16A* male; 4) PK31.2 (parent cell line of CY183; 5) non-*FRA16A* male; 6) *FRA16A* male. A black arrow marks the position of the abnormal 300 kb fragment that has occurred due to the proximity of the 37C6 marker and the CY183 (PK31.2) breakpoint. The duplicated sequences detected by 1.79 do not detect this fragment because of *Bss*H II restriction site cleavage in the intervening region. Restriction fragment localisation with respect to *FRA16A*: ↖ proximal and distal; ↗ proximal; ↘ distal. **B)** Proposed pulsed field restriction map of the *FRA16A* region showing the predicted positions of *Bss*H II fragments detected by the probes. The position of the abnormal 300 kb fragment detected by 37C6 is shown as a shaded green bar.

Pulsed field gel electrophoresis performed on Bio-Rad Chef Mapper™ apparatus. Conditions for resolution of DNA fragments within the size range 200 - 2200 kb were: gel: 1% agarose in 0.5 x TBE, running temperature: 14°C, angle of electrodes: 120°, switch times: 60 - 110 sec, voltage gradient: 6V/cm, run time: 24 hours. Protocol obtained from Chef Mapper™ instruction manual (barcode 4).

A**B**

Southern blot hybridisation) of the YAC clones isolated from the previous CEPH YAC library found y242E5 to be the only clone containing c37C6 sequences.

PFGE filters were prepared for mapping *Not* I, *Bss*H II, *Mlu* I, *Nru* I and *Sfi* I restriction sites within YACs My769HI, My790A10, y242E5 and y147C9. Filters were probed with 1.79, p16XE81-1 and the 37C6F - 37C6R PCR product, in succession. The cloning vector pBR322 was used to identify the restriction fragments at both ends of each YAC, and the vector pUC19 was used to identify restriction fragments at one end of each YAC. An example of the banding patterns obtained is shown in Fig. 3.6.

The results obtained are summarised in the restriction map of the YACs presented in Fig. 3.7. The most promising finding from the mapping was the identification of co-localising *Not* I and *Bss*H II sites within My769HI. That these restriction sites mapped ~120 kb from one end of the megaYAC was indicated by the presence of identical *Not* I and *Bss*H II fragments of this size that were detectable with a pBR322 probe (Fig. 3.6 – red and yellow arrows). The orientation of these *Not* I/*Bss*H II sites with respect to the markers (1.79, 16XE81 and c37C6) localised on the YAC contig (in particular c37C6 which had been mapped to within 350 kb of the fragile site) strongly implied that *FRA16A* was located adjacent to the co-localising *Not* I and *Bss*H II sites within My769HI.

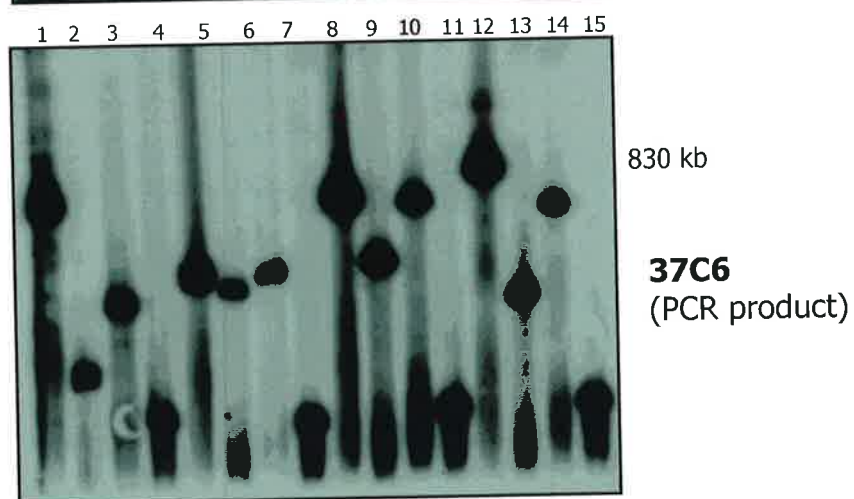
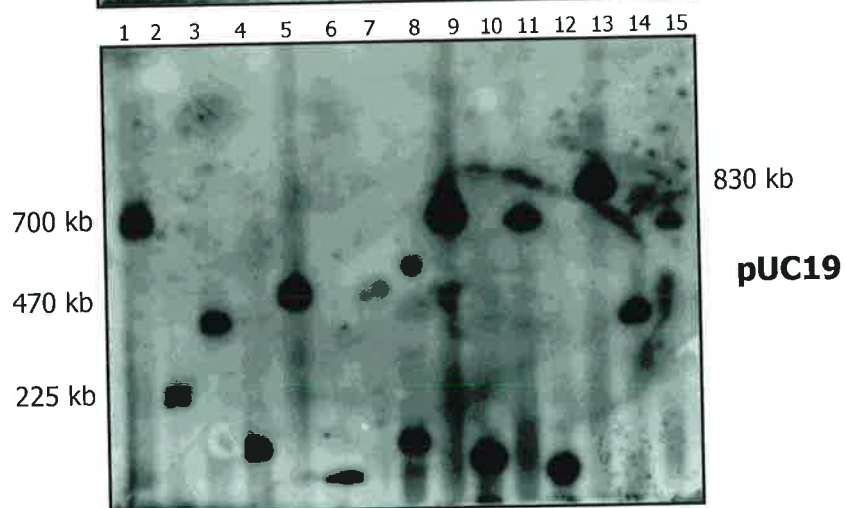
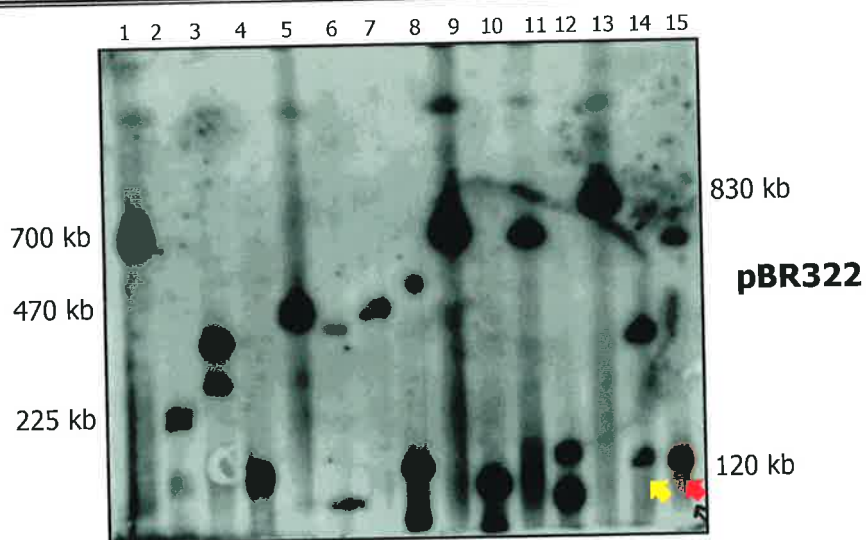
3.4.9 Identification of a CCG repeat tract sequence within My769HI

As normal *FRAXA* alleles were known to consist of comparatively short CCG repeat tracts, it was considered likely that normal *FRA16A* alleles were of a similar (if not identical) nature. In order to test this supposition, *Eco* RI and *Pst* I digested DNAs from *FRA16A* region YACs were Southern blotted and hybridised with a CCG repeat probe. A YAC containing the *FRAXA* CCG repeat tract was included as a positive control.

The CCG repeat probe hybridised to a fragment of 9 kb in *Eco* RI digested My769HI (data not shown). This band was not detected in *Eco* RI digested YAC DNAs not thought to span *FRA16A* (negative controls), and hence could not be attributed to the existence of cross hybridising yeast DNA sequences. Intensely hybridising bands of the expected sizes

Figure 3.6 Digests of YACs in the *FRA16A* region. Samples: 1) y147C9 uncut; 2) y147C9/*Bss*H II; 3) y147C9/*Not* I; 4) y147C9/*Sfi* I; 5) y242E5 uncut; 6) y242E5/*Bss*H II; 7) y242E5/*Not* I; 8) y242E5/*Sfi* I; 9) My790A10 uncut; 10) My790A10/*Bss*H II; 11) My790A10/*Not* I; 12) My790A10/*Sfi* I; 13) My769H1 uncut; 14) My769H1/*Bss*H II; 15) My769H1/*Not* I. On the pBR322 autoradiograph, a yellow arrow marks the ~120 kb My769H1 *Bss*H II band and a red arrow marks the ~120 kb My769H1 *Not* I band.

Pulsed field gel electrophoresis performed on Bio-Rad Chef Mapper™ apparatus. Conditions for resolution of DNA fragments within the size range 50 - 1000 kb were: gel: 1% agarose in 0.5 x TBE, running temperature: 14°C, angle of electrodes: 120°, switch times: 50 - 90 sec (linear ramp), voltage gradient: 6V/cm, run time: 22 hours. Protocol obtained from Chef Mapper™ instruction manual (barcode 3).



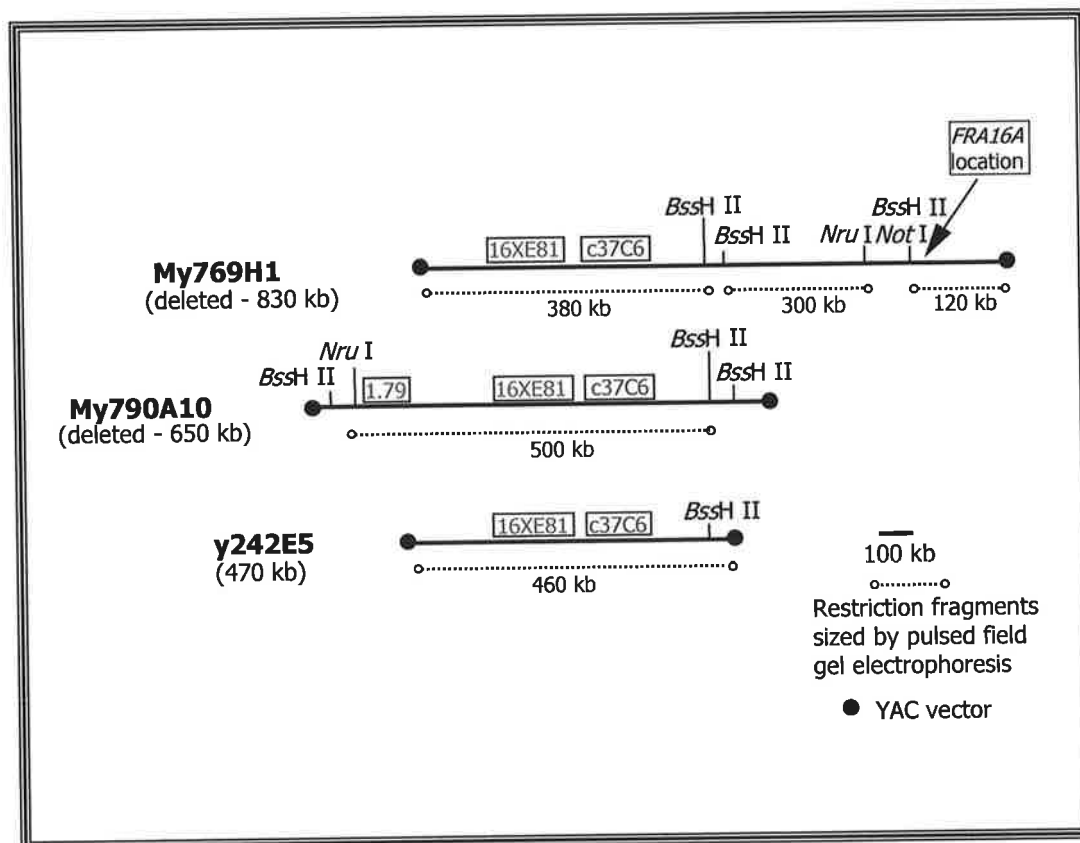


Figure 3.7 Restriction maps of YACs localised to the *FRA16A* region

were found for the *FRAXA* YAC, confirming the specificity of the probe's hybridisation to human CCG repeat tracts.

In order to determine whether the My769HI CCG repeat tract was located near the *FRA16A* associated *Not* I/*Bss*HI sites, YAC DNA digested with the restriction endonucleases *Eco* RI and *Not* I, and *Eco* RI alone was Southern blotted and hybridised with a CCG repeat probe (section 2.3.4). A 7 kb *Not* I/*Eco* RI fragment was detected by the probe, as well as the expected 9 kb *Eco* RI fragment. The change in fragment size that was observed upon the addition of *Not* I indicated that the putative *FRA16A* *Not* I site in My769HI was in close proximity to a CCG repeat tract sequence.

Taken together, the results presented in this chapter provided several lines of evidence that strongly indicated the CCG repeat tract within My769HI to be an excellent candidate sequence for the *FRA16A* locus. On this basis, further steps were subsequently undertaken in order to isolate the repeat and surrounding region of DNA. The cloning and characterisation of the CCG repeat sequence constituting the *FRA16A* locus will be presented in Chapter 4.

3.5 Discussion

3.5.1 Strategy to localise *FRA16A*

The construction of a long range PFGE restriction map of uncloned human genomic DNA spanning the region of localisation for *FRA16A* was the initial step toward characterising this autosomal folate sensitive fragile site at the DNA sequence level. Considering the location of duplicated genomic regions around the D16S79 (A, B) loci, and the characteristics of the other folate sensitive fragile sites sequenced, there were (particularly in hindsight) considerable advantages in this initial approach to the project over that of directly attempting to construct a YAC contig across *FRA16A* from the closest flanking markers, without prior restriction mapping of the region.

Before commencement of this project the rare folate sensitive fragile site *FRAXA* had been characterised (Fu *et al.*, 1991; Kremer *et al.*, 1991b) and found to consist of a CCG repeat tract adjacent to a CpG island that may become methylated when the repeat sequence is amplified. As CpG islands contain a high density of rare cutting restriction endonuclease sites (generally in an unmethylated state, Bird, 1986), they may be localised with respect to DNA markers located up to 2 Mb away by long-range PFGE restriction mapping methods that utilise these endonucleases. With the same methods, hypermethylation of rare cutter restriction sites adjacent to an expanded CCG repeat tract may be detected at a long range, as the absence or reduced intensity of a restriction fragment may be identified by a marker located up to 2 Mb away, but on the same restriction fragment. However, this mapping method is dependent upon the relative locations, within the chromosomal region to be characterised, of other rare cutter restriction sites, as fragments of a suitable size range for resolution on a pulsed field gel are necessary. The successful PFGE mapping of uncloned genomic DNA localised *FRAXA* prior to cloning (Bell *et al.*, 1991; Vincent *et al.*, 1991), demonstrating this approach to be applicable for the mapping of rare folate sensitive fragile sites in general.

Interestingly, the rare folate sensitive fragile site *FRAXE* was not initially localised by PFGE mapping of uncloned genomic DNA, but was mapped and cloned directly from YACs and cosmids found to span the fragile site by FISH analysis with *FRAXE* expressing chromosomes (Knight *et al.*, 1993). A *Not I* site that was subsequently found to be adjacent to *FRAXE* had been identified several years before in the course of preliminary PFGE restriction mapping of the *FRAXA* region (V. Hyland, personal communication). Comparative PFGE restriction mapping of DNA in this region may therefore have identified a hypermethylated restriction site in *FRAXE* individuals, and thus facilitated cloning of this fragile site locus.

3.5.2 Duplicated regions and chromosome 16 specific repeats in the *FRA16A* region.

The 1.79 probe identified two *Not* I fragments in human DNAs, therefore each fragment was assumed to contain one copy of the duplicated sequences located either side of *FRA16A* (Callen *et al.*, 1992). On the basis of FISH and PFGE mapping results, cosmid c305F6 also appeared to detect a duplicated region. The remaining markers on the map (c37C6, 16XE81, VK20) appeared to be unique, as FISH analysis found signal only proximal to *FRA16A*. In addition, their PFGE band patterns formed a subset of those detected with the duplicated probes. However, these results were not considered entirely conclusive, as it is possible that duplicated regions could occur within one restriction fragment, especially if the regions are close together.

The duplicated sequences near *FRA16A* were considered to have great potential to interfere with the FISH mapping of YACs in the region, as the YACs containing duplicated sequences could falsely appear to span the fragile site (i.e. give proximal and distal signals) when duplicated sequences on either side are present. In addition, the presence of the duplicated regions had to be always taken into account when interpreting the PFGE banding patterns from probes near *FRA16A*. Therefore, although FISH mapping has been generally the most useful tool for the identification of clones crossing cytogenetic breakpoints, additional approaches were required for the *FRA16A* project to gain useful mapping information.

Stallings *et al.* (1992) found a number of DNA regions containing chromosome 16-specific low abundance repetitive sequences during the assembly of a cosmid contig map of chromosome 16. Four main repeat rich regions were identified (by FISH), but of relevance to the *FRA16A* project was the localisation of one of these repeat regions to band 16p13.1. Hybridisation of cosmids containing the chromosome 16 low abundance repetitive sequences to chromosomes expressing *FRA16A* detected homologous regions both proximal and distal to the fragile site (E. Baker, personal communication). FISH analysis with a number of YACs near *FRA16A* detected hybridisation to the repeat-rich

regions, indicating the presence of homologous sequences within the YACs (data not shown).

The duplicated region, and chromosome 16 specific low abundance repetitive sequences, in the vicinity of *FRA16A* were initially the cause of some difficulties as they had to be taken into account when interpreting results of pulsed field, FISH and genetic mapping obtained with markers from the region (Kozman *et al.*, 1992).

Although YACs spanning *FRAXA* were successfully identified by direct *in situ* hybridisation screening against chromosomes expressing the fragile site (Kremer *et al.*, 1991a; Oberlé *et al.*, 1991), the same approach could not be applied with *FRA16A* because of the duplicated and/or low abundance repeat sequences in the region. The presence of these sequences within a given YAC could result in both proximal and distal FISH signal (when the YAC did not in fact span the fragile site), thus causing many YACs in the region to give the appearance of spanning *FRA16A*. The PFGE mapping of hypermethylated *Not* I and *Bss*H II sites on *FRA16A* chromosomes was therefore an essential initial step toward cloning this fragile site. Identification of *FRA16A* associated *Not* I and *Bss*H II sites enabled the rapid screening of YACs within the region by means of restriction analysis with these endonucleases. YACs containing these co-localising sites would be considered the most likely to contain *FRA16A*, and therefore suitable candidates for further analysis.

MegaYAC My769H1 was found to contain co-localising *Not* I and *Bss*H II sites, and was thus considered to be the most promising candidate clone. Therefore, subcloning of megaYAC 769HI was subsequently performed in order to enable isolation and characterisation of the CCG repeat tract mapped near the *Not* I and *Bss*H II sites.

3.5.3 Rearrangements in megaYACs

Large-scale genomic mapping of megaYACs has found that up to 60 % are chimaeric or contain deletions (Anderson, 1993). Such a high frequency of rearrangement has meant that, in order to be confident of the significance of the localisation of two markers within a YAC (i.e. that they both reside within a contiguous stretch of uncloned genomic DNA),

the identification of at least two YACs containing both markers is required. Alternatively, as was done for the *FRA16A* region, the accuracy of a YAC map may be validated by PFGE mapping of uncloned genomic DNA, which is obviously not subject to the YAC cloning associated rearrangements and instabilities.

The PFGE mapping of uncloned genomic DNA does not, however, make the region of interest more accessible to characterisation, therefore YAC identification mapping is also essential (although it may not be necessary to generate such comprehensive maps of YACs).

FISH mapping may identify a chimaeric YAC if a signal is apparent on two different chromosomes. However, deletion of a fragile site region may not be so obvious, as a deleted YAC may give proximal and distal signal, thus appearing to span the fragile site (even though deletion of the locus has taken place).

Comparison of the My769HI PFGE restriction map to that of *FRA16A* region uncloned genomic DNA found no clear discrepancies (although because of the normal methylation of the majority of rare cutter sites outside CpG islands, a direct comparison of restriction maps could not be done), suggesting the YAC insert DNA was not grossly rearranged. This observation enabled a higher level of confidence that the *FRA16A* associated *Not I/BssH II* sites initially identified were those on the YAC.

3.5.4 Characterisation of *FRAXE*, *FRAXF* and *FRA11B*

Three more rare folate sensitive fragile sites, *FRAXE*, *FRAXF* and *FRA11B*, were characterised after the *FRA16A* project commenced. These fragile sites were also found to consist of amplified CCG repeat tracts adjacent to methylatable CpG islands (Knight *et al.*, 1994; Parrish *et al.*, 1994; Ritchie *et al.*, 1994; Jones *et al.*, 1995). Comparison of the approaches used to isolate these fragile sites with that used for *FRA16A* illustrates that success in cloning the sites has depended on the existence of sufficient resources in the fragile site region (i.e DNA markers, known genes), as well as the nature of surrounding genomic DNA sequences (i.e the number of repeats present, both low and high copy

number). The approaches used in order to clone two of the folate sensitive fragile sites will therefore be discussed.

Rare folate sensitive fragile site *FRA11B* sequences were initially identified in the course of specific Genbank database searches for human genes containing CCG repeat tracts, which have been most often found in the 5' untranslated regions of genes. One of the genes identified, the *CBL2* proto-oncogene, mapped to the *FRA11B* chromosomal region. This was the only gene with such a repeat that mapped in the vicinity of a fragile site.

A YAC contig spanning *FRA11B* (by FISH) was found to contain the *CBL2* gene, a finding suggesting possible co-localisation. Cosmid cA0353, spanning the 5' *CBL2* CCG repeat tract, was used for FISH localisation on *FRA11B* expressing chromosomes. Signal proximal, central, and distal to the fragile site was found, indicating cA0353 also spanned *FRA11B*. The FISH results were therefore consistent with localisation of the fragile site to the *CBL2* CCG repeat tract.

For further evidence of co-localisation, Southern blots of DNAs from *FRA11B* families were probed with a 1.5 kb *Xba* I/*Sac* I fragment from cA0353. For all *FRA11B* DNAs, bands of increased size were observed, as had been found for other folate sensitive fragile sites. PCR analysis then further localised the *FRA11B* associated expansion to the *CBL2* CCG repeat tract. Corresponding to the findings at other folate sensitive fragile sites, restriction analysis of *CBL2* CpG island rare cutter sites in *FRA11B* and non-*FRA11B* DNAs found hypermethylation to be associated with fragile site expression.

The rare folate sensitive fragile site *FRAXF* had been mapped proximal to *FRAXE* and *FRAXA* in band Xq28, but unlike these fragile sites, *FRAXF* expression was not found to be associated with any phenotypic abnormality (Hirst *et al.*, 1993).

In contrast with *FRA16A*, construction of a long range uncloned genomic and/or YAC DNA restriction map of the Xq28 region was not a requirement for the cloning of *FRAXF*. This fragile site was found almost by chance when a cDNA clone (H10) containing a CCG repeat (isolated during in an effort to identify additional *FMR1* clones) was used to screen a flow sorted X chromosome library. The screening identified cosmids

containing homologous sequences that mapped to the genomic region in the vicinity of *FRAXF*. Sequence characterisation of the homologous region from the cosmids revealed the homology to be entirely due to the CCG repeat motifs present in both the cDNA and cosmid sequences. Corresponding with observations at the other fragile sites, Southern blot analysis of *FRAXF* family DNAs found an expanded repeat tract in association with fragile site expression.

Ritchie *et al.* (1994) also cloned *FRAXF*, but proceeded in a more deliberate manner. The initial step was the screening of megaYACs mapping near *FRAXF* for CCG repeat sequences. A megaYAC containing homologous sequences was subcloned into a λ vector, and the resulting λ library screened with total human DNA and CCG repeat probes in succession. A λ subclone hybridising to both probes was isolated and characterised, and an unstable CCG repeat tract that is expanded in *FRAXF* DNAs was identified.

The approach of Ritchie *et al.* (1994) to cloning *FRAXF* was similar to that employed for *FRA16A*, but did not include the initial detailed long range restriction mapping of uncloned genomic DNA that took place with *FRA16A*. This initial approach may not, however, be applicable to genomic regions with a high-density of, or few, rare cutting sites. Another drawback is that it may be time consuming and technically difficult. The more rapid approach of directly mapping/screening YACs is most applicable for cloning sequences of interest from a well characterised and finely mapped genomic region, as was the region of the X chromosome containing *FRAXF* (by virtue of its proximity to *FRAXA* and *FRAXE*).

The most unusual feature of the pathway to the cloning of *FRAXF* by Parrish *et al.* (1994) was that the initial step was accomplished in an indirect manner whilst attempting to identify *FMRI* sequence clones. The presence of a CCG repeat tract within the cDNA probe hybridising to the *FRAXF* cosmids was clearly the essential element of this approach to cloning *FRAXF*.

The isolation of *FRAXF* suggested that CCG repeat tracts are rare enough in the human genome to enable a rare folate sensitive fragile site associated unstable CCG repeat tract

to be readily identified by chromosome specific genomic library screening, without the necessity of mapping large numbers of positive clones to find those from the genomic region of interest.

Prior to commencement of this project, the approach of direct screening for CCG repeat tracts was attempted in an effort to identify a cosmid containing the *FRA16A* CCG repeat tract. To this end, LANL filters with cosmid DNAs representing a gridded 1 x coverage of human chromosome 16 (Stallings *et al.*, 1990) were hybridised with a CCG repeat probe. However, subsequent mapping with the chromosome 16 hybrid cell line panel excluded all CCG positive clones from the *FRA16A* interval (data not shown). The approach was therefore unsuccessful in this case. As it is within a C and G rich region, the *FRA16A* CCG repeat tract may have been under-represented within the cosmid library. Therefore clones containing fragile site sequences may have been successfully identified by screening filters containing more than a 1 x coverage of cosmids. In a later stages of the project, whilst attempting to identify a *FRA16A* associated gene, 10 x cosmid grid filters from the same LANL library, (which had not been previously available) were hybridised with a *FRA16A* specific clone (pf16A1). Eight positive signals (at least three of which were true positive) were found on the filters, indicating the *FRA16A* region was not under-represented in the source cosmid library. It could therefore be concluded that CCG repeat probe hybridisation to cosmid libraries, at least in densely mapped regions of the genome, may be a more efficient way to isolate folate sensitive fragile sites than the use of other positional cloning methods, such as YAC mapping.

3.5.5 The nature of the genomic region surrounding *FRA16A*

FRA16A region YACs (y242E5, My790A3, My769H1) were PFGE restriction mapped and arrayed into a contig encompassing ~990 kb of DNA (Figs. 3.4 and 3.7). The DNA loci mapping proximal to *FRA16A* (16XE81, 37C6) and the proximal locus of the duplicated marker (1.79) were spanned by this contig (Fig. 3.4C).

A *Not* I restriction site co-localising with a *Bss*H II site was mapped within My769H1. This was the only *Not* I site within the contig sequences. The location of the *Not* I/*Bss*H II

sites with respect to the DNA marker positions was consistent with that expected (from uncloned DNA PFGE mapping data, section 3.3.3) for the hypermethylated sites identified in *FRA16A* DNAs. Further evidence of the close proximity of the *Not* I/*Bss*H II sites to *FRA16A* was provided by the finding that the *Not* I site (in My769HI) was located within 7 kb of a CCG repeat tract. As My769HI was the only member of the contig to contain a CCG repeat tract, this sequence did not appear to be a common one in the vicinity of *FRA16A*.

Taken together, the results discussed in this chapter provided credible evidence that My769HI spanned the *FRA16A* *Not* I/*Bss*H II sites (although this could not be confirmed by FISH mapping), and additionally, that the fragile site may occur at an unstable CCG repeat tract. The pathway toward isolation of the putative *FRA16A* CCG repeat tract from the YAC therefore appeared to be much simplified, as a CCG repeat probe could be used to directly identify YAC subclones containing homologous sequences. Therefore, unlike the situation that existed for the characterisation of *FRAXA* and *FRAXE*, the construction of a contig of YAC subclones (λ or cosmid) spanning the fragile site was not a necessary step in the pathway toward characterising *FRA16A*. A detailed account of the subcloning of *FRA16A* sequences from My796HI is given in Chapter 4.

Chapter 4

*Molecular Characterisation
of the FRA16A
Fragile Site*

Chapter 4: Contents

Molecular Characterisation of the FRA16A Fragile Site

4.1	Summary	103
4.2	Introduction	104
4.3	Materials and Methods	105
4.3.1	DNA samples from <i>FRA16A</i> families	105
4.3.2	Identification of a CCG positive My769HI restriction fragment of a suitable size for subcloning into Lambda ZAP® II vector	105
4.3.3	Lambda ZAP® II library construction	105
4.3.4	Lambda ZAP® II library screening	106
4.3.5	Identification of a restriction fragment of intergenerationally unstable size on <i>FRA16A</i> chromosomes	107
4.3.6	Subcloning fragments of pfl6A1	107
4.3.7	Southern blot analysis of DNAs from <i>FRA16A</i> families	108
4.3.8	Sequencing of the <i>FRA16A</i> CCG repeat tract and adjacent region	108
4.3.9	PCR analyses of the <i>FRA16A</i> CCG repeat tract	108
4.3.10	SSCP analysis	108
4.4	Results	109
4.4.1	Characterisation of the <i>FRA16A</i> region	109
4.4.2	Detection of anomalous restriction fragments in <i>FRA16A</i> DNAs	109
4.4.3	Sequence of the <i>FRA16A</i> CCG repeat tract and adjacent region	109
4.4.4	PCR analysis of normal <i>FRA16A</i> alleles	110
4.4.5	PCR analysis of <i>FRA16A</i> families	111

4.4.6	Exclusion of sequences other than the cloned CCG repeat tract as the origin of <i>FRA16A</i> expansion.....	112
4.4.7	FISH analysis with probes immediately flanking <i>FRA16A</i>	114
4.4.8	<i>FRA16A</i> pedigree analyses.....	114
4.4.9	Methylation status of the <i>FRA16A</i> CCG repeat tract and CpG island.....	115
4.5	Discussion	116
4.5.1	Characteristics of rare folate sensitive fragile sites.....	116
4.5.2	Hypermethylation at rare folate sensitive fragile sites.....	117
4.5.3	Effect of a transcribed CCG repeat expansion upon subsequent translation	119
4.5.4	Parental biases in the transmission of expanded CCG repeat tracts	120
4.5.5	Somatic heterogeneity of the <i>FRA16A</i> CCG repeat tract expansion	122
4.5.6	Origin of the somatic heterogeneity in CCG repeat copy number.....	122

4.1 Summary

The YAC My769H1, containing a CCG repeat hybridising region considered to be an excellent candidate for the *FRA16A* locus, was subcloned into a lambda vector. The resulting library of subclones was screened with a CCG repeat probe. Several identical subclones containing the CCG repeat region were isolated, and used to probe Southern blots of DNAs from *FRA16A* and non-*FRA16A* individuals. Restriction fragments of abnormally large size, and displaying size variation between samples, were detected in the *FRA16A* DNAs. Analysis of *FRA16A* pedigrees demonstrated intergenerational instability in the size of these abnormal restriction fragments. These results were similar to those obtained for *FRAXA*, and thus provided evidence that *FRA16A* had been successfully cloned.

Sequence characterisation of the isolated CCG repeat region identified a CpG island containing an interrupted CCG repeat tract consisting of 20 CCG repeat units. PCR analysis of the *FRA16A* CCG repeat tract size in DNAs from non-*FRA16A* individuals found that it was highly polymorphic, with a heterozygosity of 91% in the Caucasian population. Subsequent pedigree analyses found that alleles at the *FRA16A* locus undergo Mendelian segregation.

Further evidence that the isolated CCG repeat tract represented the source of the *FRA16A* expansion was provided by the finding of CCG repeat locus null alleles segregating with *FRA16A*. Further sequencing, restriction mapping and SSCP analysis then excluded all regions of the My769H1 subclone as the origin of expansion, with the exception of the CCG repeat tract.

4.2 Introduction

The physical mapping of the *FRA16A* region described in the previous chapter indicated that the fragile site was in the vicinity of a CCG repeat tract and normally unmethylated *Not* I and *Bss*H II restriction sites (implying the presence of a CpG island). These findings were similar to those from the folate sensitive fragile site, *FRAXA*, located on the X chromosome.

Prior PFGE mapping of the *FRA16A* region eliminated the need to perform a chromosomal DNA walk originating from one or both sides of *FRA16A*, as a CCG repeat probe could be used for direct screening of a subcloned YAC DNA lambda library.

A major area of interest in the *FRA16A* project was to identify differential methylation patterns associated with expression of the fragile site. This was firstly achieved with the aid of PFGE mapping discussed in Chapter 3, but the ultimate aim was to perform investigations at a closer range with a DNA probe containing sequences immediately adjacent to the fragile site. It was considered that the results of such a study might determine the accuracy of the hypothesis that fragile X associated hypermethylation is a consequence of the X-inactivation process (Follette and Laird, 1992). A finding that hypermethylation occurs in association with a CCG repeat tract expansion at the *FRA16A* locus (which is not an apparent site of autosomal methylation imprinting), would imply that X-chromosomal inactivation is not necessary for the occurrence of abnormal methylation at folate sensitive fragile sites on the X chromosome.

In an attempt to understand more about the biological basis of fragile site expression and inheritance, it was considered to be of interest to compare and contrast *FRA16A* with other rare, folate sensitive fragile sites with regard to a features such as: (1) the sequence of the CCG repeat tract and immediate flanking regions; (2) the pattern of inheritance of the expanded repeat (determined by pedigree analysis); and (3) the polymorphism/repeat number variation of the CCG repeat tract.

4.3 Materials and Methods

4.3.1 DNA samples from *FRA16A* families

Lymphoblastoid cell line or peripheral lymphocyte DNAs from four *FRA16A* families were obtained from the DNA bank in the Department of Cytogenetics and Molecular Genetics at the Women's and Children's Hospital, Adelaide, South Australia. Metaphase spreads derived from individuals in families 1, 2, and 3 were subjected to cytogenetic scrutiny in order to determine their status with respect to *FRA16A* expression (Sutherland and Hecht, 1985; E. Baker, unpublished observations). Segregation of *FRA16A* expression in an extended pedigree of family 1 was reported by Callen *et al.* (1989).

4.3.2 Identification of a CCG positive My769HI restriction fragment of a suitable size for subcloning into Lambda ZAP®II vector

DNA from My769H1 was digested with the *Sac* I, *Xba* I, *Spe* I, *Eco* RI and *Xho* I restriction endonucleases, and Southern blotted. These particular endonucleases were chosen for their unique cleavage sites in the phage cloning vector Lambda ZAP®II polylinker (Stratagene, La Jolla, CA, USA cat. # 236201). The Lambda ZAP®II vector has a cloning capacity of up to 10 kb (Short *et al.*, 1988).

The Southern blot filter was probed with a CCG repeat probe, and washed under normally stringent conditions (section 2.3.4). Of the restriction fragments detected by hybridisation with the CCG repeat probe, a ~7 kb *Sac* I fragment was the only one of suitable size (<10 kb) for subcloning into Lambda ZAP®II.

4.3.3 Lambda ZAP® II library construction

Lambda ZAP®II vector was digested with *Sac* I and purified, in accordance with a protocol provided with the purchased vector. My769H1 DNA immobilised in agarose beads was *Sac* I digested and Prep-a-Gene purified (sections 2.3.5.4 and 2.3.2.3). Purified *Sac* I digested Lambda ZAP®II and My769H1 DNA was ligated (in accordance with the protocol provided with the vector) and the ligated DNA packaged, transfected and plated

in accordance with the protocol provided with a purchased Stratagene packaging kit (cat. # 200203).

4.3.4 Lambda ZAP®II library screening

The My769H1/Lambda ZAP®II library plates were plaque lifted (section 2.3.4.7), and the filters screened with a total human DNA probe in order to ascertain whether any of the subclones contained human DNA inserts (section 2.3.4). In terms of the goal of cloning a specific fragment, it was difficult to assess the significance of the number of hybridisation signals present, as only *Sac* I fragments of 10 kb or less were clonable (unless inserts were chimaeric). In addition, not all inserts would contain the high copy number repetitive sequences detectable by the total human DNA probe. Hence, it was not possible to determine beforehand, or by screening with a total human DNA probe, how many human DNA clones the library should contain in order for the fragment of interest to be represented.

My769H1/Lambda ZAP®II library plaque lift filters were stripped and reprobed with a CCG repeat probe. After normally stringent washing (section 2.3.4.9) 4 positive signals were detected on the filters. Lambda clones corresponding to 3 of the 4 positive signals were isolated after 3 rounds of screening. DNA from the 3 λ clones was purified and digested with *Sac* I in order to determine insert sizes. The 3 isolates were found to contain identical *Sac* I fragments of ~7 kb. The subcloned fragment within a pBluescript vector (plasmid pf16A1), was excised from the λ arms of one clone using the protocol and reagents supplied with the Lambda ZAP®II vector.

4.3.5 Identification of an intergenerationally unstable restriction fragment on *FRA16A* chromosomes

Prior to the comparison of restriction fragment sizes (to detect expansions) by means of Southern blot hybridisation analysis of non-*FRA16A* and *FRA16A* DNAs, it was considered that it may be of use to identify a CCG repeat containing fragment (of non-*FRA16A* origin) of a size such that any changes in mobility would be detectable. To this end, My769H1 DNA was digested with a panel of restriction endonucleases, Southern blotted, and probed with a CCG repeat probe (section 2.3.4). *Pst* I digestion, producing the smallest fragment of ~2 kb, was therefore used for the Southern blot hybridisation analysis of normal and *FRA16A* DNAs.

In order to detect any changes in mobility associated with *FRA16A* expression, a Southern blot filter containing *Pst* I digested DNAs from 3 *FRA16A* and 5 non-*FRA16A* individuals was prepared and hybridised with pre-reassociated pf16A1 (section 2.3.4).

4.3.6 Subcloning fragments of pf16A1

Restriction mapping of pf16A1 was performed in order to localise *Not* I, *Bss*H II and *Pst* I sites within the clone. A restriction map of the clone is presented in Fig. 4.1(D). Southern blot hybridisation analysis of the pf16A1 digests localised the CCG repeat tract to a 1.8 kb *Not* I - *Pst* I fragment. Therefore, in order to subclone this fragment, pf16A1 was digested with *Not* I and *Pst* I, and the resulting fragments ligated to a pBluescript vector (section 2.3.3). A subclone resulting from this (pf16A3.1) contained the 1.8 kb fragment of interest. An additional subclone (pf16A3.2), containing an adjacent 300 bp *Not* I - *Pst* I fragment, was also derived from the same ligation reaction.

In addition, *Rsa* I fragments of pf16A1 were ligated into pUC19, and a 2.7 kb clone (pf16A2) containing the CCG repeat was identified by bacterial colony lift screening (with CCG repeat probe). Clone pf16A2 was further subcloned, producing a 650 bp *Not* I - *Rsa* I/pUC19 clone containing the CCG repeat tract (pf16A2.1).

With the production pf16A2.1, the CCG repeat tract was isolated within a cloned DNA fragment of sufficiently small size for complete sequence analysis primed from the

pUC19 vector, assuming no inhibition of sequencing occurred due to the CCG repeat tract (primers pUC19/M13 forward and reverse, Promega cat. # Q5391 and Q5401).

4.3.7 Southern blot analysis of DNAs from *FRA16A* families

Genomic DNAs from *FRA16A* families were cleaved with *Pst* I according to the manufacturer's instructions. Digestions were Southern blotted and probed with pre-associated pf16A3.1 (section 2.3.4).

4.3.8 Sequencing of the *FRA16A* CCG repeat tract and adjacent region

Partial sequencing of pf16A3.1 and pf16A2.1 was accomplished with a combination of automated and manual sequencing methods (section 2.3.7).

4.3.9 PCR analyses of the *FRA16A* CCG repeat tract

PCR analysis of alleles at the *FRA16A* locus was performed with the primers *a* and *b*, the location and sequences of which are shown in Fig. 4.1(E). The PCR templates were DNAs from 42 unrelated individuals of European origin from the parental generation of the Centre d'Etude du Polymorphisme Humain (CEPH) pedigrees, and individuals from 3 (apparently) unrelated *FRA16A* families.

4.3.10 SSCP analysis (modification of Orita *et al.*, 1989)

In brief, 10 µl of 2 x loading buffer was added to each 10 µl radiolabelling PCR reaction. After being heated at 100°C for 3 min, 3 µl of each sample was loaded onto a non-denaturing, 5% (w/v) bis/acrylamide (1:60), 5% (v/v) glycerol gel in 0.5 x TBE (38 cm x 50 cm x 0.4 mm), and run at 600 V/room temperature for ~60 hours. Gels were dried and autoradiographed at -70°C overnight with intensifying screens.

4.4 Results

4.4.1 Characterisation of the *FRA16A* region

The steps toward the positional cloning and sequence analysis of the *FRA16A* fragile site locus are illustrated in Fig. 4.1.

4.4.2 Detection of anomalous restriction fragments in *FRA16A* DNAs

Further evidence of the localisation of *FRA16A* to pf16A1 was provided by the results of Southern blot hybridisation analyses of DNAs from 3 *FRA16A* and 5 non-*FRA16A* individuals with this probe. In *FRA16A* DNA lanes, expanded fragments of variable size were detected, in addition to the 2 kb and 4 kb *Pst* I fragments found for non-*FRA16A* DNAs. At this stage, the 2 kb *Pst* I fragment was assumed to be the source of the expansion, as all abnormal *FRA16A* fragments were observed to be less than, or equivalent in size, to the 4 kb fragment (a fragment size increase was expected as this had been found at the *FRAXA* locus). In addition, the size of the 2 kb *Pst* I fragment corresponded with that of the My769HI CCG repeat hybridising fragment size.

4.4.3 Sequence of the *FRA16A* CCG repeat tract and adjacent region

Sequencing the region of cloned DNA adjacent to the *Not* I site within clones pf16A3.1 and pf16A2.1 proved to be problematic. The difficulties encountered may have been due to the DNA secondary structures that are often present in C and G rich regions. The partial DNA sequence of pf16A3.1, immediately adjacent to the *Not* I site, was obtained with fluorescent automated sequencing. From this sequence PCR primer *a* was designed (Fig. 4.1 E). Five CCG repeats within pf16A3.1 were discernible on the sequence chromatogram, but the total length of the repeat tract could not be determined from this result (Fig. 4.2).

The use of a manual single extension sequencing method (section 2.3.7.2) was necessary in order to determine the full repeat tract and flanking region sequence. The sequencing gel autoradiograph is presented in Fig. 4.3, and the sequence of the repeat and flanking

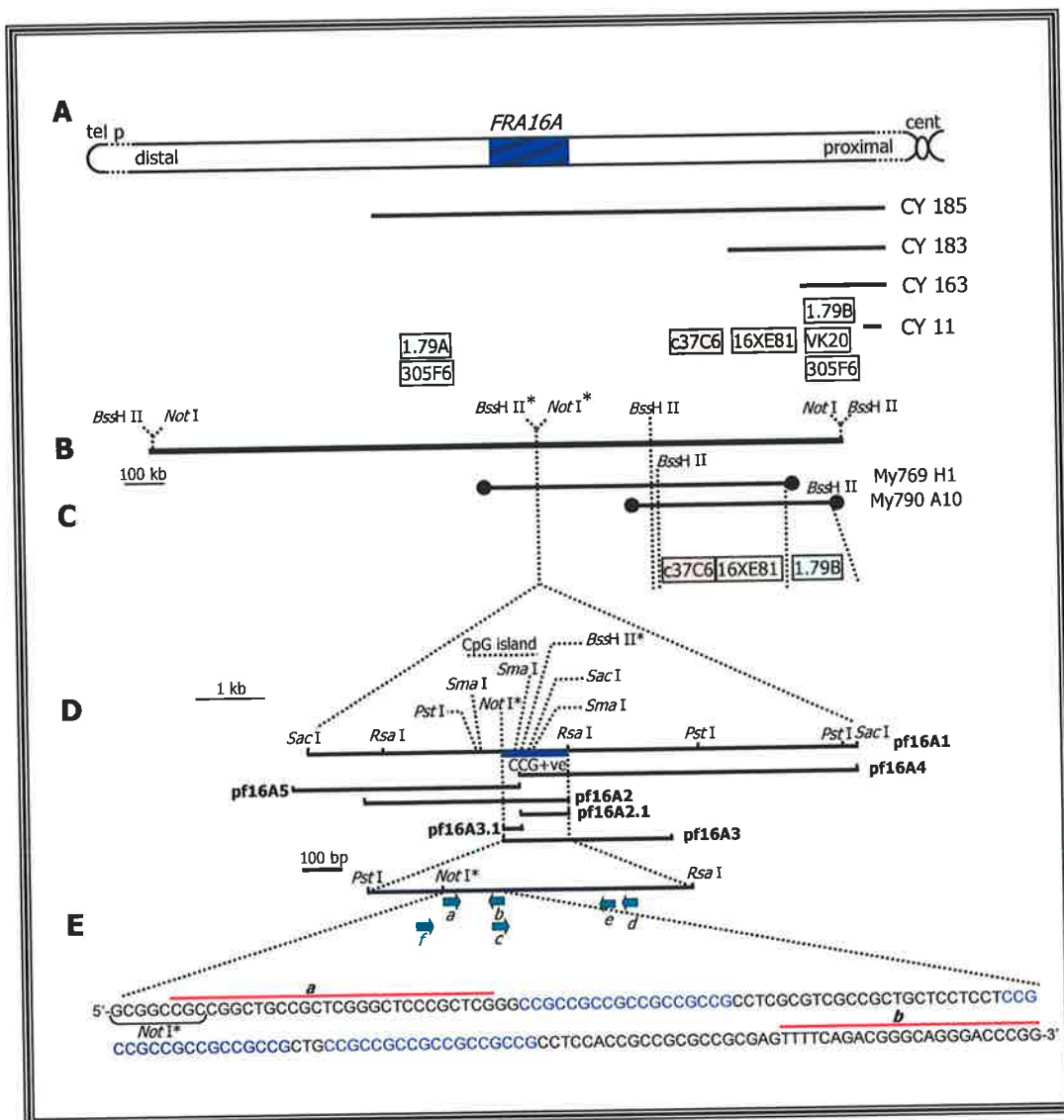


Figure 4.1 Positional cloning and sequence analysis of *FRA16A*. Sections A, B and C are discussed in Chapter 3. D. Map of the *FRA16A* subclones pf16A1, pf16A2, pf16A2.1, pf16A3, pf16A3.1, pf16A4, pf16A5, some of which contain the differentially methylated *Not* I site adjacent to a CCG repeat tract. The extent of the associated CpG island (dotted line) and the *Not* I - *Rsa* I fragment to which the instability was localised (blue shaded line) are indicated. E. Location of the PCR primers (green arrows) utilised to localise the region of *FRA16A* instability to a CCG repeat tract, and for sequence analysis of the region. The positions of repeat tract flanking primers *a* and *b* are indicated by red overlines.

PCR primer sequences

a 5' -GCCGGCTGCCGCTCGGGCTCCCGCT- 3'
b 5' -CCGGGTCCCTGCCCCGTCTGAAAA- 3'
c 5' -TTTTCAGACGGGCAGGGACCCG- 3'
d 5' -ATCCTACTCCCACTACGTCCTGAGG- 3'
e 5' -GCCTTCCCCATCACCCTCCCCTCCA- 3'
f 5' -CACGCTCCGGGCCGCGCCGCGCTC- 3'

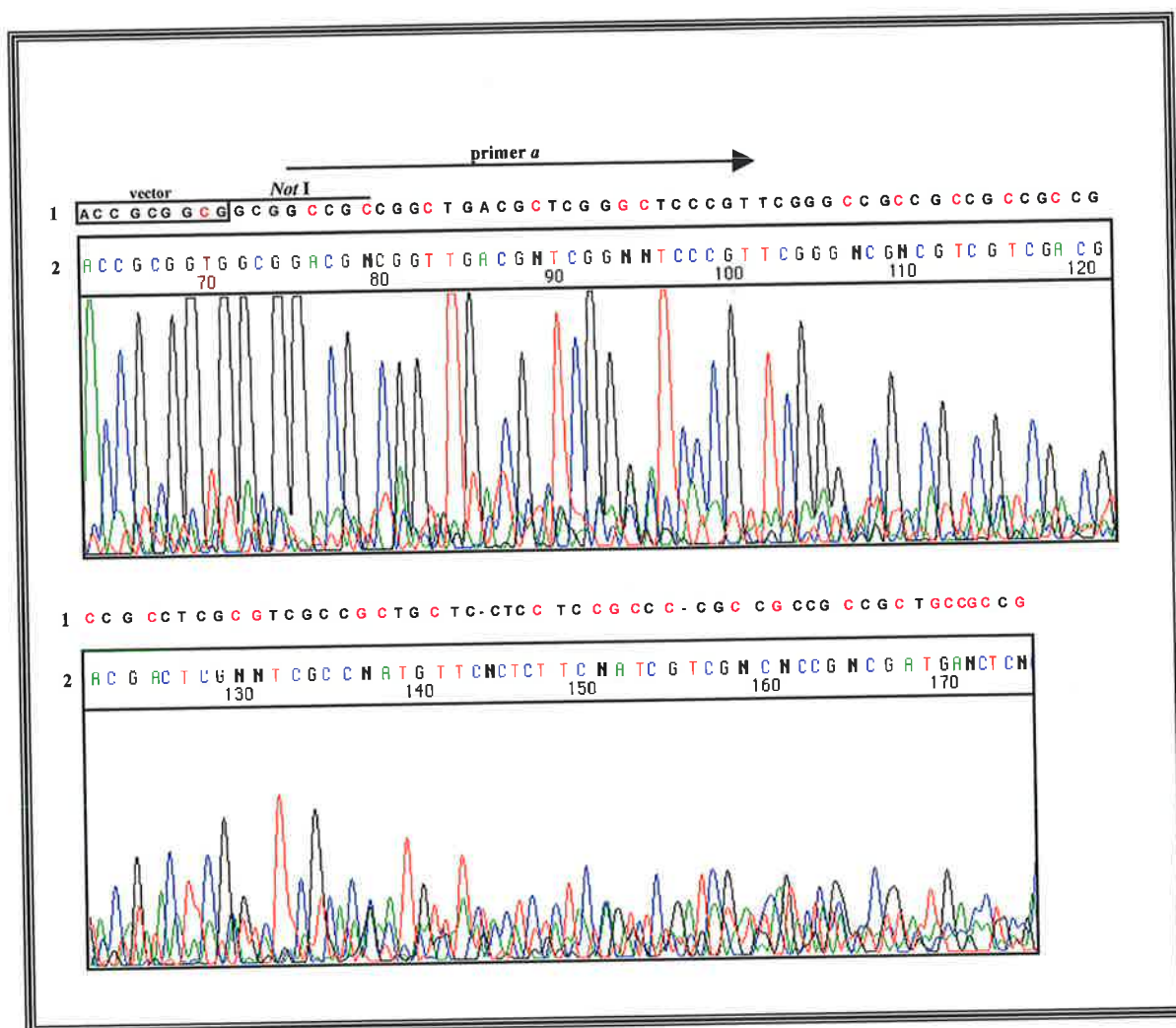


Figure 4.2 Attempted automated sequencing of the *FRA16A* CCG repeat tract. **1.** The DNA sequence as determined by a combination of manual and fluorescent automated methods - red bases are those which differ from the automated sequence assignments. **2.** Fluorescent automated sequence with base assignments from Perkin Elmer ABI sequence analysis software.

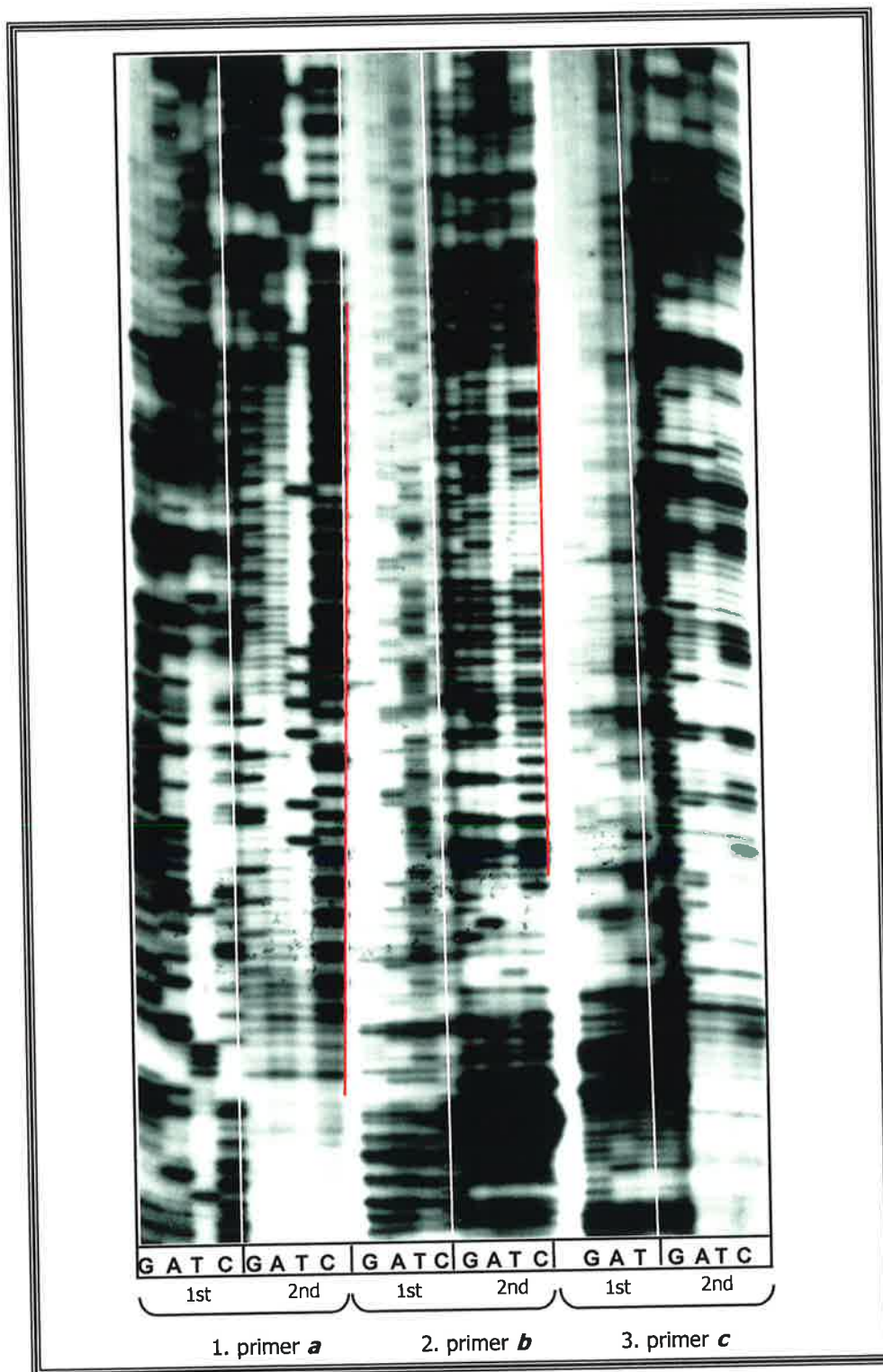


Figure 4.3 Manual sequencing analysis of pf16A2 (2.0 kb *Not* I - *Pst* I fragment) with primers **a**, **b**, and **c**, demonstrating the extensive cross banding due to DNA secondary structures in the region. The position of the *FRA16A* CCG repeat tract within the sequences is marked by red lines. For each set of sequencing reactions (1, 2 and 3), lanes were loaded in the order **GATC**. The 1st and 2nd labels refer to the first and second loadings of the sequencing reactions, respectively. For the 1st primer **c** loading, only three lanes are visible due to distortions during electrophoresis.

regions is presented in Fig. 4.1(E). From the Fig. 4.3 (sample 1) sequence data, a second flanking primer (*b*) was designed. The sequence and position of primer *b* is shown in Fig. 4.1(E).

Sequence characterisation of the CCG repeat tract isolated from My769HI identified a total of twenty copies of a CCG repeat motif. However, stretches of non-repetitive sequence or base substitutions (variant repeats) interrupted the repeat tract at intervals of six repeats or less. Further sequencing of pf16A3.1 revealed that the trinucleotide repeat sequence is bordered by a stretch of very C and G rich sequence that meets the criteria for classification as a CpG island (Bird *et al.*, 1986).

Complete sequencing of the pf16A2.1 insert was attempted by designing sequencing primers at the most 3' ends of the existing characterised sequence, at each end of the clone. Primer *c* (antisense to primer *b*) was used in a manual sequencing reaction in order to obtain further pf16A2.1 DNA sequence immediately adjacent to the repeat (Fig. 4.3, sample 3). Primers *d* and *e* were used to sequence pf16A2.1 from the opposite direction by the fluorescent automated sequencing method.

Manual sequencing of the CCG repeat tract and surrounding regions was hindered by extensive cross banding (Fig. 4.3, samples 1, 2, 3), probably due to the CG richness and repetitive nature of the sequence. The full sequence of pf16A2.1 (650 bp *Rsa* I/*Not* I fragment insert) was therefore very difficult to obtain, but was eventually completed while attempting to identify transcribed sequences in the region (see Chapter 6).

The DNA sequence of the CCG repeat tract and immediate flanking region is given in Fig. 4.1(E). The locations of primers *a* and *b* (*c* and *d* are the reverse complement) are underlined within the sequence.

4.4.4 PCR analysis of normal *FRA16A* alleles

A high degree of polymorphism at the *FRA16A* locus was apparent from the polyacrylamide gel analysis of *a* - *b* PCR products (Fig. 4.4). PCR amplification of the *FRA16A* CCG repeat tract in 46 CEPH parent DNAs identified twenty-one different

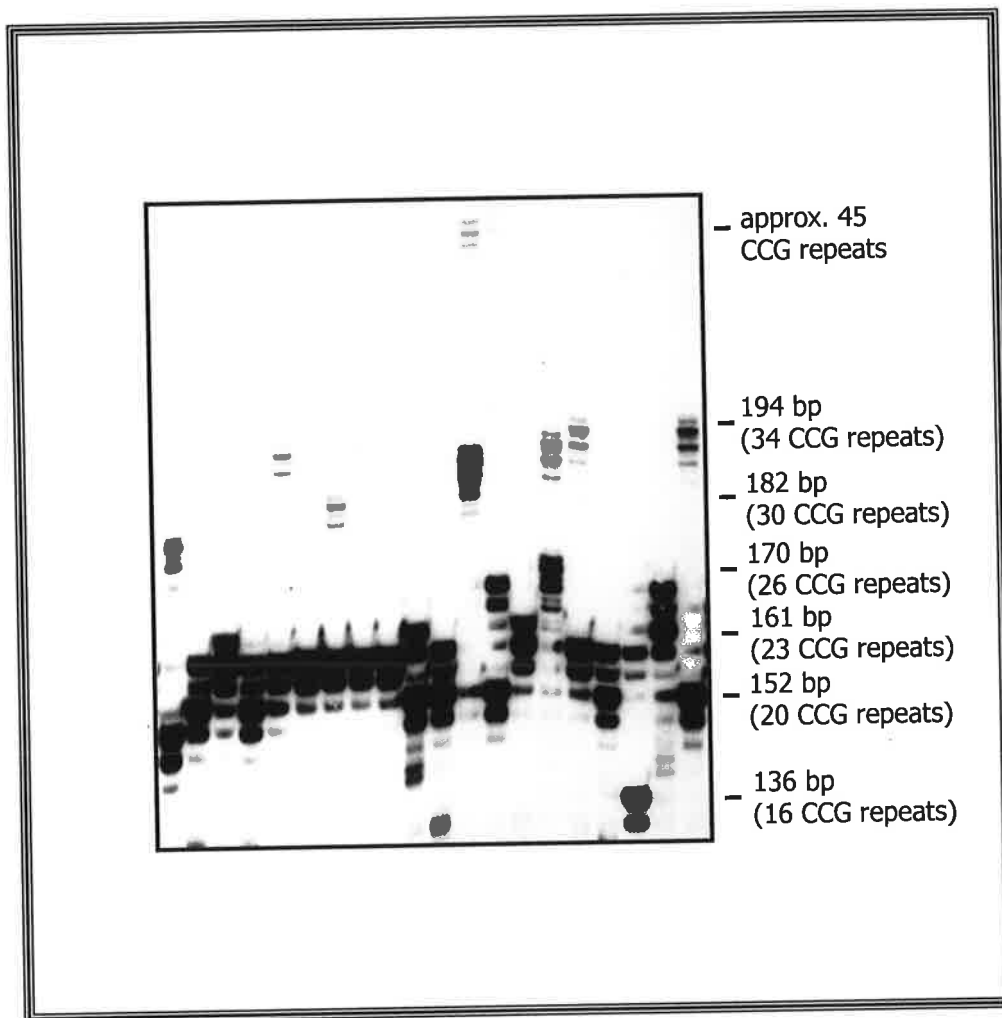


Figure 4.4 PCR analysis revealed a high degree of polymorphism at the *FRA16A* locus. *FRA16A* allele PCR products were amplified from primers *a* and *b*. The PCR templates were DNAs from twenty unrelated Caucasian individuals (CEPH parents).

allele types, with a heterozygosity of 91%. Assuming the variation in *FRA16A* allele size was entirely due to differences in CCG repeat number, copy numbers appeared to range from 16 to 49 repeats, with a mode of 22. Further details on the size and frequency of *FRA16A* alleles in the CEPH parents are given in Chapter 5.

Two of the CEPH parent DNAs typed were found have three *FRA16A* alleles each (two alleles from each sample were in the upper range). These results might be indicative of somatic instability of the CCG repeat tract *in vitro*, occurring during tissue culture of CEPH lymphoblastoid cell lines, or less likely, *in vivo* instability (i.e the individual from whom the cell line was derived was mosaic at the *FRA16A* locus). Pedigree analysis of the families of these CEPH parents, as well as those of a number of others with upper range *FRA16A* allele sizes, revealed normal Mendelian segregation of alleles. When taken together, these results indicated that *FRA16A* normal length alleles are not often intergenerationally unstable, but larger *FRA16A* alleles appear to be frequently mitotically unstable, at least when the cells have undergone many generations of cell culture, as have the CEPH parent cell lines.

The PCR analysis results indicated a single copy status for the *FRA16A* CCG repeat tract region, as a maximum of two *FRA16A* alleles were amplified from the majority of DNAs. This finding did not correspond with results obtained from the PCR analysis of a AC repeat tract at the duplicated D16S79 (1.79) locus, which identified up to four alleles per genome, suggesting a high level of conservation of the duplicated sequence in the vicinity of *FRA16A* (Phillips *et al.*, 1991).

4.4.5 PCR analysis of *FRA16A* families

The PCR analysis of the *FRA16A* CCG repeat tract in DNAs from three *FRA16A* families with primers *a* and *b* demonstrated the segregation of a probable null allele with

the fragile site (Fig. 4.5). As there was no evidence of null alleles in the typed CEPH pedigrees (data not shown), they appeared to be uncommon within the normal population. CCG repeat tracts are more refractory to PCR amplification than many sequences, probably due to a high degree of secondary structure (R. Richards, personal communication). It was therefore proposed that the null alleles were due to an expansion of the CCG repeat tract, causing the distance between the two primers to become too large for efficient PCR amplification of the exceptionally GC rich intervening region. However, finding *FRA16A* associated null alleles was not considered sufficient proof that the characterised CCG repeat tract was the origin of the abnormal expansions detectable by Southern blot hybridisation analysis.

4.4.6 Exclusion of DNA sequences other than the cloned CCG repeat tract as the origin of the *FRA16A* expansion

Further PCR analysis of the 650 bp DNA region, including the CCG repeat tract (spanned by the insert of pf16A2.1), was performed in order to exclude the possibility that the *FRA16A* associated null alleles were the result of base substitutions in the *a* or *b* primer binding sites in linkage disequilibrium with *FRA16A*. An additional primer (*f*) was designed from the existing sequence for this analysis. The position and sequence of primer *f* is shown in Fig. 4.1 (E).

PCR amplification of CEPH and *FRA16A* family DNAs with the primers *e* and *f* (flanking the *a* and *b* primer set) produced a set of larger PCR products containing the CCG repeat tract. These larger PCR products corresponded in banding pattern (on a polyacrylamide gel) to a set of smaller products from the same DNA templates, but amplified from the internal *a* and *b* primers. The null alleles segregating with *FRA16A* were still observed with the *e* and *f* primers (data not shown), and therefore could not be attributed to base changes in the annealing sites of primers *a* or *b*.

As the entire DNA sequence of pf16A2.1 (650 bp) had not been determined, analysis of the cloned DNA region (other than the CCG repeat) by another method was necessary in

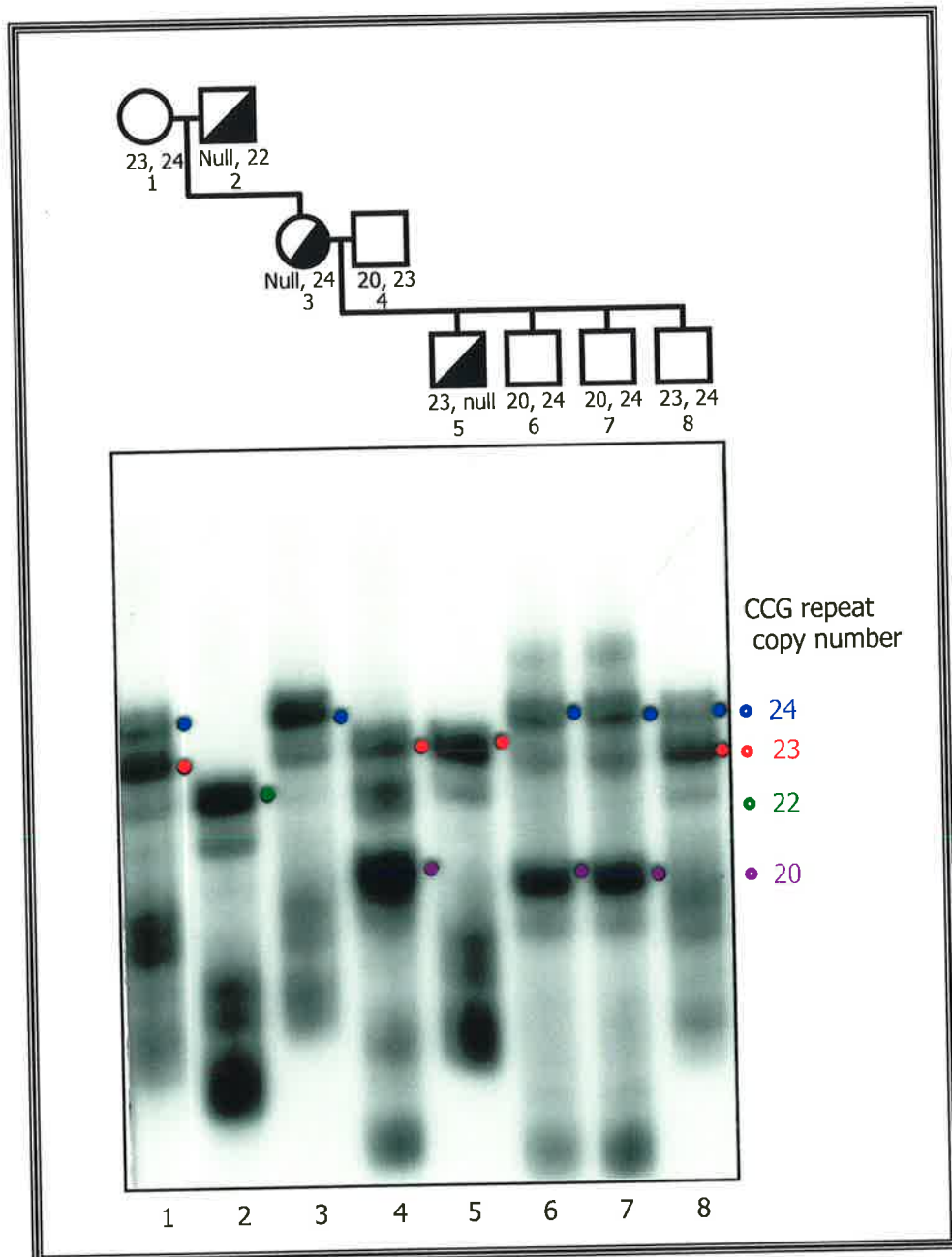


Figure 4.5 Segregation of a null allele with the fragile site in a *FRA16A* pedigree. Lanes 1 to 8 correspond to individuals numbered 1, 2, 6, 7, 9, 8, 10, 11 in Fig. 4.7. PCR products were generated with primers *a* and *b*. Each allele is represented by major and minor bands, typical of most simple tandem repeat PCR products. The major band scored for each allele is indicated by a dot of a specific colour for each size. Specifically, blue corresponds with 24 CCG repeat alleles, red with 23 repeat alleles, green with 22 repeat alleles, and violet with 20 repeat alleles. The variable, more rapidly migrating bands present in each lane are double-stranded hetero- and homo-duplexes of the scored single-stranded alleles.

order to exclude the possibility that the *FRA16A* expansion may have arisen from another uncharacterised repeat motif.

Single strand conformational polymorphism (SSCP) analysis (section 4.3.10) of PCR products from primers *c* and *d* with *FRA16A* DNA template enabled the exclusion of the region of DNA between the primers as the source of the expansion. The sequence and location of primers *c* and *d* are shown in Fig. 4.1(E). In order to achieve this exclusion, 2 SSCP alleles (i.e. 1 or 2 bands per sample were observed for individuals homozygous or heterozygous respectively at this SSCP locus) of the *c* - *d* PCR product were initially identified by the screening of CEPH parent DNAs. Mendelian inheritance of the *c* - *d* SSCP alleles was then demonstrated in a CEPH pedigree (Fig. 4.6 A). SSCP analysis of *FRA16A* DNAs found several individuals who had *a* - *b* CCG repeat null alleles to be heterozygous for the *c* - *d* SSCP (Fig. 4.6 B). This finding indicated that, since *c* - *d* PCR products of normal size are produced from both the normal and *FRA16A* chromosomes, the region between the *c* and *d* primer binding sites could be excluded as the origin of the fragile site expansion.

The DNA sequence between the *Rsa* I site and primer *d* annealing site (Fig. 4.1 E) appeared to be completely deficient in repeat motifs that could be taken as possible sites of amplification (assuming that no insertion or deletion events had taken place). Complete sequencing of clone pf16A3.1 (a 300 bp *Not* I - *Pst* I fragment insert) revealed a C and G rich region deficient in repeat motifs with the exception of a short stretch of 4 CAG repeats. This region was also excluded on the basis of Southern blot hybridisation results from probe pf163.1 to a *Pst* I - *Not* I digest of *FRA16A* carrier DNA (Fig. 4.8, Family 2, Individual 1). An expansion of 500 bp (premutational size), with an unmethylated adjacent *Not* I site was found in this individual (data not shown). The premutation was detectable with probe pf16A3.1, which does not contain 300 bp *Not* I - *Pst* I fragment sequences, thereby localising the *FRA16A* expansion to the 1.8 kb *Not* I - *Pst* I (pf16A3.1) fragment.

The results obtained from the Southern blot hybridisations and SSCP analyses precisely mapped the *FRA16A* expansion origin to the sequenced CCG repeat tract.

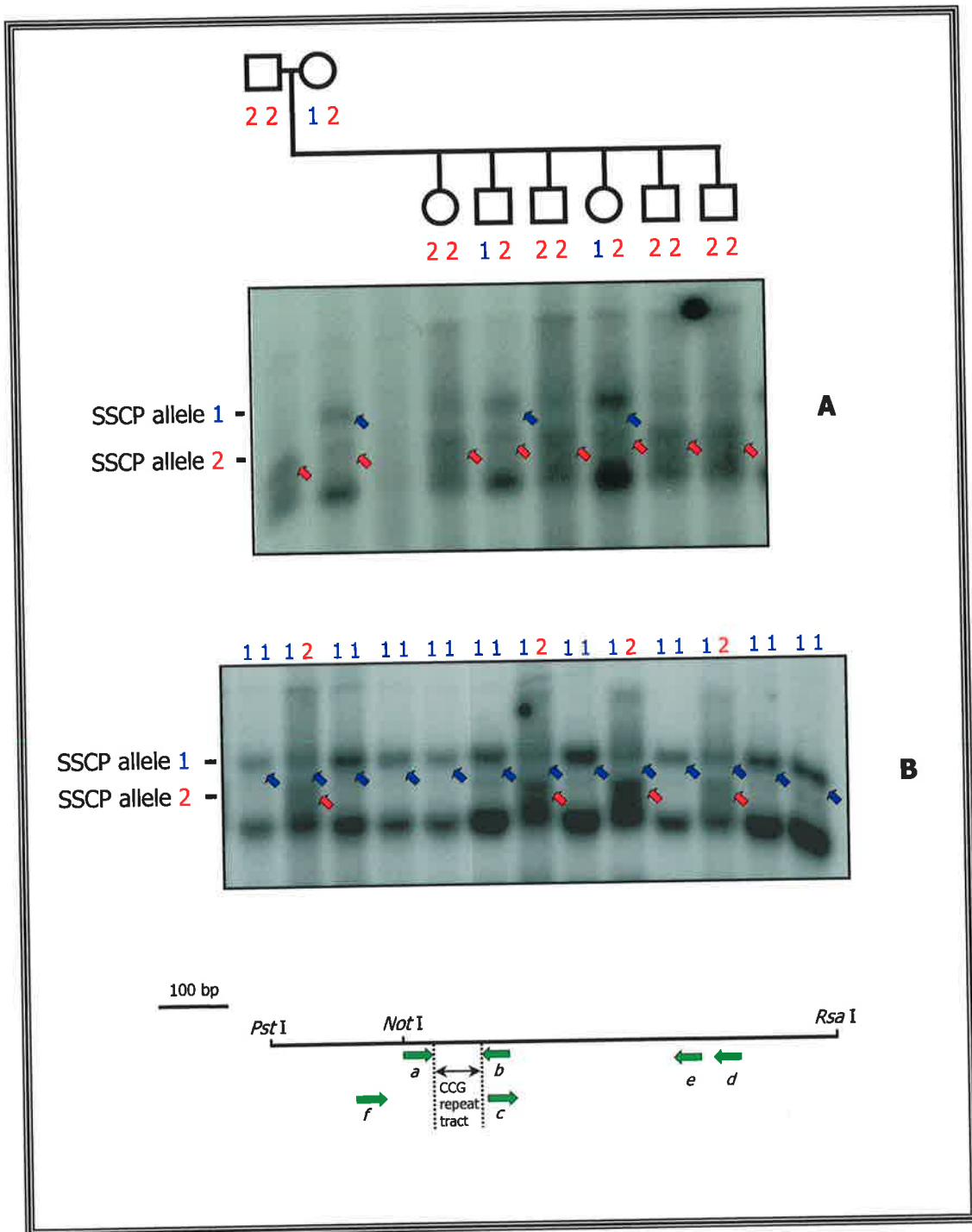


Figure 4.6 Exclusion of a DNA region immediately adjacent to the CCG repeat tract (between the binding sites of primer *c* and *e* - shown in the above line diagram) as the source of *FRA16A* expansion. **A.** Mendelian segregation of *c* - *e* SSCP alleles (1 and 2) in a CEPH family. **B.** SSCP analysis of *FRA16A* DNAs showing 4 individuals to be heterozygous at this SSCP locus, a result indicating the absence of an expansion between the binding sites of primers *c* and *e*.

4.4.7 FISH analysis with DNA clones immediately flanking *FRA16A*

FISH analysis was performed in order to provide additional confirmation of the localisation of *FRA16A* to within the insert of pf16A1. Subclones of pf16A1 (pf16A4 and pf16A5) were generated for this purpose (Fig. 4.1 D). Clone pf16A4 (containing the CCG repeat at one end of the insert) was localised distal to *FRA16A*, with scores of 12 distal signals, 1 central signal, and 2 proximal signals. Clone pf16A5 localised proximal to *FRA16A*, with scores of 7 proximal signals, 1 central signal, and 1 distal signal. These results provided further evidence for the localisation of *FRA16A* (on the cytogenetic level, as opposed to the molecular level) to within clone pf16A1.

4.4.8 *FRA16A* pedigree analyses

DNAs from individuals of 3 unrelated *FRA16A* families were analysed by Southern blot hybridisation in order to ascertain the size and degree of meiotic instability of the *FRA16A* expansion. The results are detailed as follows:

Family 1 - The Southern blot hybridisation results and pedigree are shown in Fig. 4.7. This large pedigree comprises 28 individuals over three generations.

The results of this pedigree analysis indicated that expression of *FRA16A* is associated with the presence of a *Pst* I fragment of expanded size (normal size 2.2 kb), which appears meiotically unstable only upon female transmission. There was no apparent decrease or increase in expanded fragment size upon three male transmissions (from one individual). The increase in the size of the 2.2 kb *Pst* I fragment related to the *FRA16A* expansion in this family was found to be from ~3.0 kb to ~5.7 kb (i.e. final size of *Pst* I fragment ~5.2 kb to ~7.9 kb).

The *FRA16A* expansion of individual 20 appeared as an indistinct smear of bands larger than 5.7 kb. For the other *FRA16A* individuals within the pedigree, an abnormally large discrete band representing the *FRA16A* expansion was present, indicating a high degree of somatic stability for the *FRA16A* full mutation.

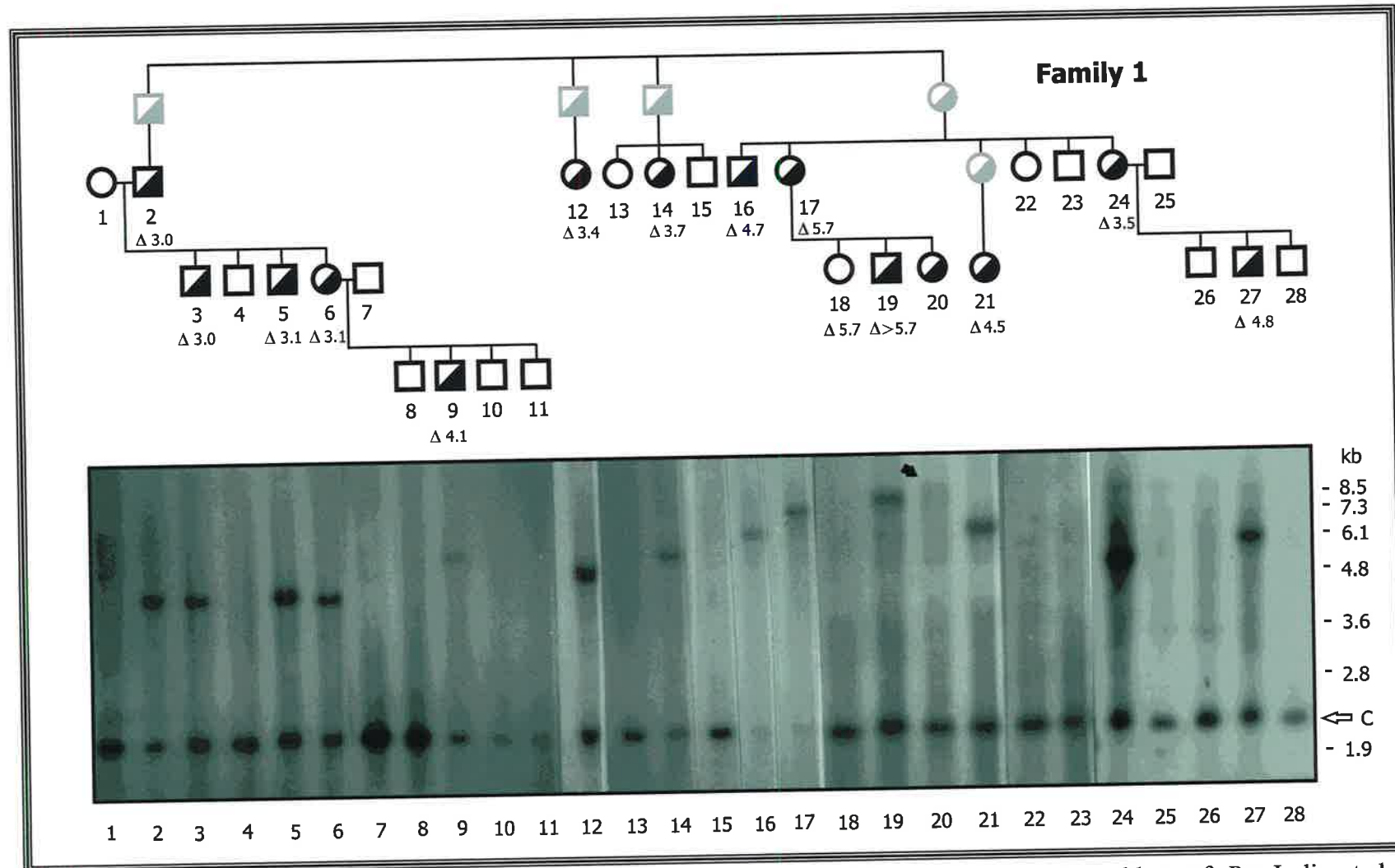


Figure 4.7 Detection of a heritable unstable element in a large *FRA16A* pedigree. Southern blots of *Pst* I digested chromosomal DNAs from members of a large *FRA16A* pedigree (**Family 1**) were probed with pf16A3. Only those members represented by black symbols were analysed to determine their *FRA16A* cytogenetic and molecular status. Cross hatched symbols represent *FRA16A* individuals or obligate carriers. The Δ value is the increase in kilobase pairs above the normal 2.2 kb fragment size. An arrow marks the position of a smear of bands, indicative of somatic heterogeneity.

Family 2 - The Southern blot hybridisation results and pedigree are shown in Fig. 4.8. This pedigree comprises three individuals over two generations. Previous cytogenetic analysis had shown that individual 1 did not express *FRA16A* (E. Baker, personal communication), however Southern blot analysis detected a ~500 bp expansion (a *FRA16A*), indicating the positive carrier status of this individual. Individual 3, the son of individual 1, was found to have an expansion of 2 kb associated with expression of the fragile site, thus demonstrating expansion of a *FRA16A* premutation to a full mutation state. A clear secondary expanded fragment was detected for individual 3, indicating somatic variation of the expansion.

Family 3 - The Southern blot hybridisation results and pedigree are shown in Fig. 4.8. The pedigree comprises 4 individuals over 2 generations. An expansion associated with *FRA16A* expression was found to have increased only marginally, by ~200 bp in size, after a female meiotic event. As for the majority of family 1, the expansion appeared somatically stable in this pedigree.

4.4.9 Methylation status of the *FRA16A* CCG repeat tract and CpG island

The methylation status of the *FRA16A* CCG repeat tract in 2 non-*FRA16A* and 3 *FRA16A* DNAs was assessed by means of digestion with the methylation sensitive restriction endonuclease *Fnu* 4HI. The recognition site of this enzyme is GCNGC, which is encoded by the CCG repeat motif. If the CCG repeat tract and surrounding *Fnu* 4HI restriction sites (N=G or C) are methylated, cleavage will not occur. The size of a restriction fragment spanning only the *FRA16A* CpG island (detectable by Southern blot hybridisation) may thus be unaffected, or only marginally affected by the addition of *Fnu* 4HI to a digestion reaction. The results of this methylation assay (presented in Fig. 4.9) indicated the existence of complete methylation of the *Fnu* 4HI site CpG dinucleotides in the expanded *FRA16A* CCG repeat tract (and adjacent region), and non-methylation of the normal alleles.

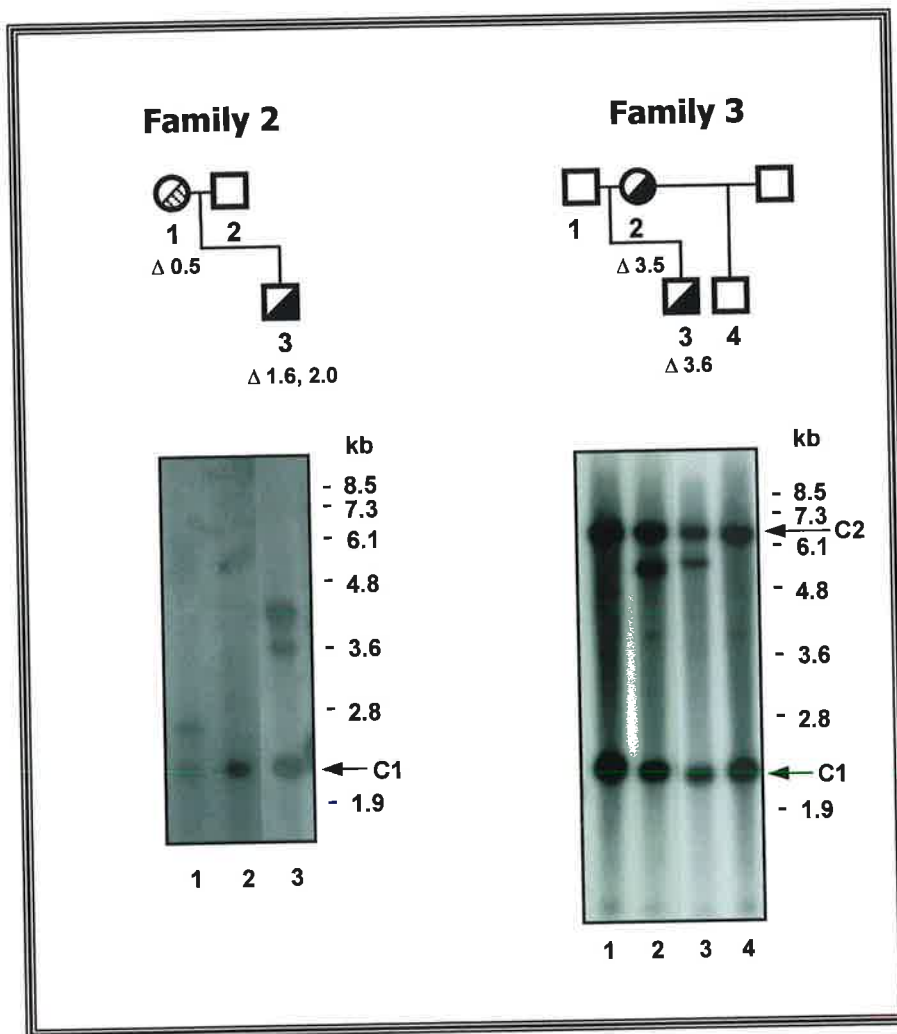


Figure 4.8 Detection of a heritable unstable element in two small *FRA16A* pedigrees. **Family 2:** Southern blot of *Pst* I digested DNAs probed with pf16A3, showing expansion of a *FRA16A* premutation (lane 1) to a somatically heterogeneous full mutation (lane 3). **Family 3:** Southern blot of *Pst* I DNAs probed with pf16A1, (displaying an additional constant 6.5 kb band detected by this probe). Cross hatched solid symbols represent those members of the pedigrees with cytogenetic expression of *FRA16A*. The shaded symbol represents a non-cytogenetically expressing *FRA16A* carrier. **C1** indicates the normal sized *Pst* I fragment (2.2 kb) that is the origin of the instability. **C2** indicates a normal constant band. The Δ value is the increase in kilobase pairs above the 2.2 kb normal fragment size.

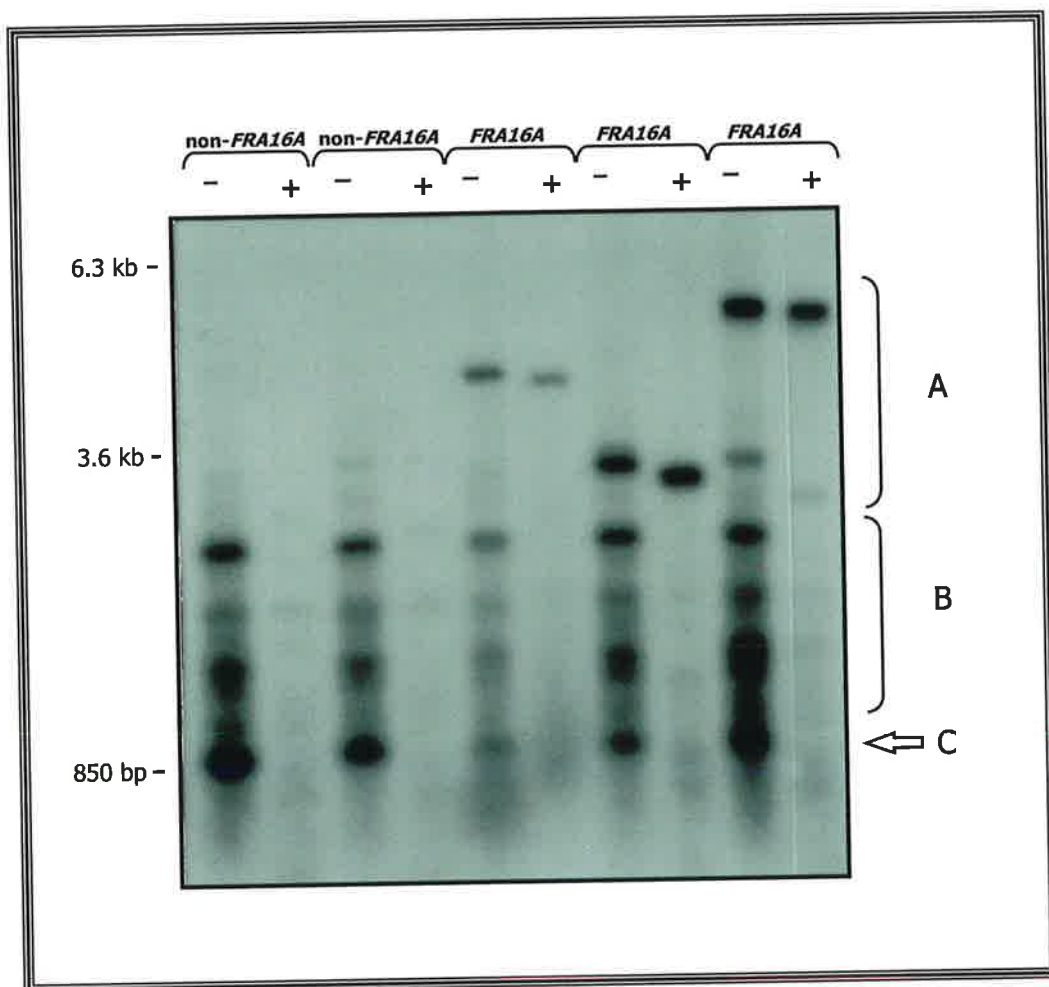


Figure 4.9 Methylation status of the *FRA16A* CCG repeat tract assessed by *Fnu* 4HI digestion. *Pst* I - *Rsa* I double digests of chromosomal DNAs from non-*FRA16A* (lanes 1 and 2) and *FRA16A* individuals (lanes 3 to 5) were treated with *Fnu* 4HI (+) or not (-), then subjected to Southern blot analysis with the 650 bp *Not* I - *Rsa* I fragment from pf16A3 as a probe. The probe detected an 850 bp constant band (C) in the non-*FRA16A* DNAs, several cross hybridising fragments from elsewhere in the genome (B), and expanded unstable fragments in *FRA16A* DNAs (A). Digestion with *Fnu* 4HI eliminated the normal (C) fragment, and the cross hybridising bands (B); however, only a minor alteration was afforded to the expanded *FRA16A* fragments (as not all *Fnu* 4HI site contain a methylatable CpG), thus indicating the resistance to digestion of the majority of the *Pst* I - *Rsa* I fragment containing the CCG repeat tract, due to hypermethylation.

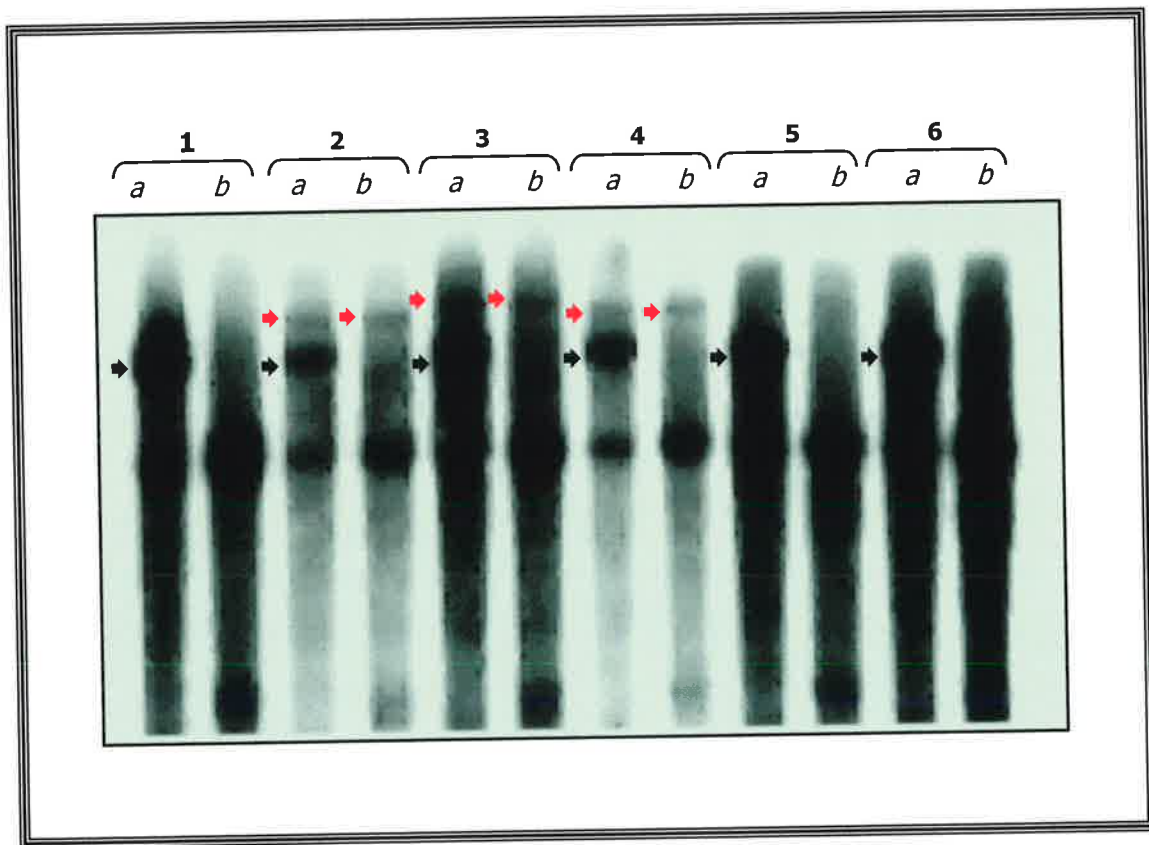


Figure 4.10 Methylation analysis of the *FRA16A* *Not* I site in DNAs from *FRA16A* and non-*FRA16A* individuals. Samples in lanes: **1.** non-*FRA16A* female (lymphoblastoid DNA); **2.** *FRA16A* male (lymphoblastoid DNA); **3.** *FRA16A* male (lymphoblastoid DNA); **4.** *FRA16A* male (lymphocyte DNA); **5.** non-*FRA16A* male (lymphocyte DNA); **6.** non-*FRA16A* individual (lymphocyte DNA). Lanes marked *a* contain *Eco* RI digested DNA only; those marked *b* contain DNA digested with both *Eco* RI and *Not* I. The red arrows denote the position of the *Eco* RI fragments detected by pf16A3.1 which remain uncut by *Not* I in *FRA16A* DNAs. The black arrows denote the position of the normal *Eco* RI fragment that increases in size as a result of *FRA16A* CCG repeat tract amplification.

The *Not* I restriction site adjacent to the *FRA16A* CCG repeat had been ascertained, by long range PFGE restriction mapping, to be always methylated (uncut) on *FRA16A* chromosomes, and unmethylated (cut) on normal chromosomes. For confirmation of this result, the methylation status of the *Not* I site was examined at closer range by hybridisation of the pf16A1 probe to *Eco* RI and *Not* I digested *FRA16A* and non-*FRA16A* DNAs. An uncut *Eco* RI band of variable size was found in the three *FRA16A*, but not in the non-*FRA16A* DNAs (Fig. 4.10). This result clearly established that the hypermethylated *Not* I site, initially identified by PFGE mapping, was within the sequence of clone pf16A1.

4.5 Discussion

4.5.1 Characteristics of rare folate sensitive fragile sites

In addition to *FRA16A*, the four other rare folate sensitive fragile sites characterised at the molecular level are *FRAXA*, *FRAXE*, *FRAXF* and *FRA11B*. Cloning of *FRAXF* and *FRA11B* was achieved after *FRA16A*, and *FRAXE* was characterised after commencement of this project (Knight *et al.*, 1993; Parrish *et al.*, 1994; Ritchie *et al.*, 1994; Jones *et al.*, 1995). Fragile site expression was found, in all cases, to be associated with expansion of a normally polymorphic CCG (or GGC) repeat tract located adjacent to a CpG island. Furthermore, both repeats and CpG islands were found to become hypermethylated when expansions were present, and the corresponding fragile site to be cytogenetically detectable. Molecular characterisation of the rare folate sensitive fragile site *FRA16A* ascertained that this fragile site has all the features mentioned above. Comparisons of allele sizes and repeat expansions of the cloned folate sensitive fragile sites are given in Fig. 4.11, but it is difficult to draw any further conclusions on the nature of fragile sites in general from this comparison. The precise molecular events leading to

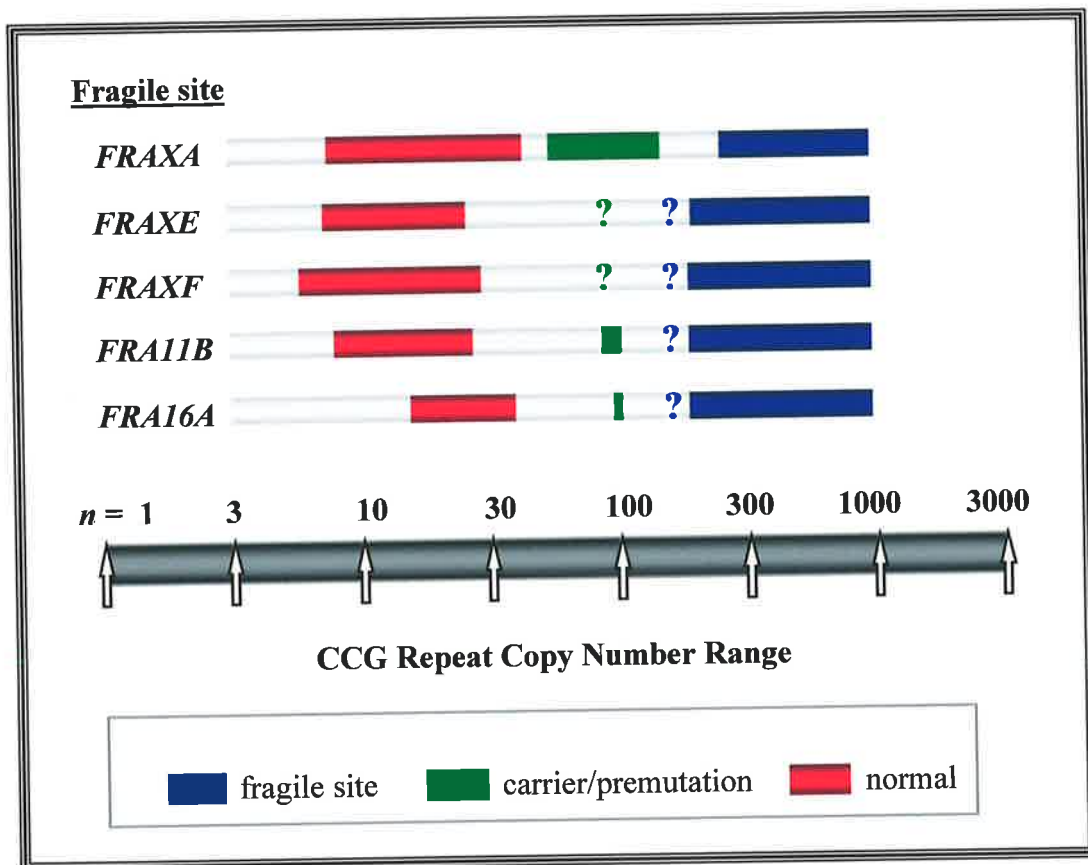


Figure 4.11 Comparison of CCG repeat unit number in the normal, carrier/premutation, and fragile site expression states at the *FRAXA*, *FRAXE*, *FRAXF*, *FRA11B* and *FRA16A* loci. Modification of a figure from Richards and Sutherland, (1997).

the changes in chromosome structure which can be viewed under the microscope remain unknown.

4.5.2 Hypermethylation at rare folate sensitive fragile sites

In response to the molecular findings at *FRAXA*, Laird (1991) proposed that hypermethylation at the fragile X locus was a consequence of failure to erase the X chromosome inactivation imprinting signal (of which methylation is a normal component) from the *FMRI* CCG repeat and CpG island. An opposing view was that hypermethylation occurs after and as a result of CCG repeat expansion, and therefore cannot be attributed to X inactivation (Oberlé *et al.*, 1991; Sutcliffe *et al.*, 1992).

The cloning of *FRAXE* (Knight *et al.*, 1993) did not provide further insight into the factor(s) contributing to fragile X hypermethylation, as this rare folate sensitive fragile site is also located on the X chromosome and therefore also subject to X-inactivation processes. Resolution of the conflicting views regarding the molecular events (or at least an exclusion of hypotheses) associated with fragile site hypermethylation could therefore be at least partially achieved by the characterisation and methylation analysis of an autosomal folate sensitive fragile site. If located within a CpG island (which is generally unmethylated), it would be expected that the sequences at such a fragile site would be unmethylated in the normal course of events, and would certainly not be subject to X-inactivation processes.

As expected, methylation of the *FRA16A* CpG island (including the CCG repeat tract) was not evident on normal chromosomes. Therefore, this region did not appear to be a normal site of imprinting, at least by methylation status. Based on this observation, it could be concluded that failure to erase a methylation imprint is not a necessary condition for rare folate sensitive fragile site genesis.

Other mechanisms (not invoking X inactivation/imprinting) have been put forward to account for the unique features of the folate sensitive fragile sites. It has been proposed that the hypermethylation at rare folate sensitive fragile sites the result of increased methylase activity at the fragile site CpG islands (Imbert and Mandel, 1995). Adding

weight to this proposal is the fact that methylase activity is considerably greater toward DNA sequences with the capacity to form unusual structures, such as those containing an unpaired or mismatched C in the CpG recognition sequence (Smith *et al.*, 1991). It has also been found that (CCG)₁₅ and (CGG)₁₅ oligonucleotides adopt unimolecular foldback conformations, which are substrates for methyltransferase (Smith *et al.*, 1994). The function of the resulting methylation is thought to be the stabilisation and marking of unusual structures for repair (Fry *et al.*, 1994).

If the events described above were the cause of fragile site hypermethylation, the process would be dependent only upon the presence of expanded CCG repeat sequences, and therefore entirely independent of normal parental imprinting, whether X or autosome. However, direct experimental evidence of such a methylation process may be difficult to obtain.

In addition to a number of studies using methylation sensitive restriction endonucleases, a direct genomic sequencing method was utilised in order to determine the methylation status of cytosine residues of the *FMRI* CpG island region. This approach enabled ascertainment of the methylation status of individual CpG dinucleotides within DNA regions containing a high density of potential restriction endonuclease cleavage sites (Hornstra *et al.*, 1993). The method was not used for methylation analysis of the *FRA16A* CpG island analysis, as reagents that are difficult to obtain and utilise are required.

The direct genomic sequencing found all CpG dinucleotides within the *FMRI* CpG island (including the CCG repeat tract) to be unmethylated on normal and transmitting male X chromosomes, and methylated on affected male (from lymphoblastoid cell lines) and normal inactive X chromosomes (Hornstra *et al.*, 1993). Considering the *Fnu* 4HI methylation assay results from *FRAXA* and *FRA16A* studies are similar, it appears likely that complete methylation of the *FRA16A* CpG island also occurs.

Hornstra *et al.* (1993) suggested that the pattern of fragile X hypermethylation could be related to X inactivation, as had been proposed by Laird (1991). However, it was also argued that for X-inactivation to be a contributing factor, the X chromosome carrying the fragile X mutation would have to be selectively inactivated in the female germ line

during gametogenesis, as all premutations above 90 repeat units always expand to a full mutation during oogenesis (Hornstra *et al.*, 1993).

The findings from two phenotypically normal brothers with fragile X full mutations and cytogenetic expression of *FRAXA* (Smeets *et al.*, 1995) provided further insights into the nature of this fragile site, that are probably also relevant to the other folate sensitive fragile sites. On the brothers' X chromosomes the *FMRI* CpG island was found to be unmethylated. The *FMRI* gene was transcribed at normal levels, but there appeared to be a low translation efficiency, as lower than normal FMRP levels were found. These findings also clearly indicated that it is the *FMRI* gene inactivation due to hypermethylation, and not a direct effect of the repeat tract expansion, that is responsible for the fragile X phenotype.

Another two conclusions, which may be relevant to all folate sensitive fragile sites, were drawn from this study. These conclusions were: (1) CCG repeat expansion does not necessarily induce methylation, and (2) methylation is not an absolute requirement for the induction of fragile sites (Smeets *et al.*, 1995).

The first conclusion could also be drawn from a study by Sutcliffe *et al.* (1992), that showed both *FMRI* transcription and an absence of methylation in the chorionic villi of a fragile X full mutation fetus. In all other tissues of this fetus hypermethylation and absence of *FMRI* expression was evident.

4.5.3 Effect of a transcribed CCG repeat tract expansion upon translation

Feng *et al.* (1995) studied the effect of expanded repeats upon the translation of *FMRI* transcripts. The results obtained from this study indicated that, even with normal *FMRI* mRNA levels, translation of transcripts with more than 200 CCG repeats is greatly reduced, probably leading to phenotypic expression of the mutation. Electron microscopy observations indicated that there was no migration of the ribosome along the repeat beyond this range, thus explaining the reduction in translation. These results however, appeared directly at odds with the findings of Smeets *et al.* (1995), that suggested hypermethylation alone was responsible for the fragile X phenotype.

4.5.4 Parental biases in the transmission of expanded CCG repeat tracts

Of the rare folate sensitive fragile sites characterised, *FRAXA* has been by far the most intensively studied in terms of pedigree analysis. This has occurred because of the clinically significant, disabling, fragile X syndrome phenotype, which is a marker for the fragile site, and also because of the high frequency of *FRAXA* compared with the other fragile sites (although this may be in part because there are no apparent phenotypic markers for the other fragile sites). Consequently, the patterns of transmission and amplification of the *FRAXA* CCG repeat tract have been extensively characterised, and are discussed in the Chapter 1.

A more recent study of the mutational events at the *FRAXA* locus has reported that reversion of the fragile X full mutation to a premutational state occurs in the male germ line, as only premutations could be found in the sperm of non-mosaic fragile X males (Reyniers *et al.*, 1993). This finding led to the proposal that *FMRI* CCG repeat tract expansion from premutation to full mutation is a post-zygotic event that occurs in somatic cells after separation from the germ cell lineage, thus sparing the germ line (Reyniers *et al.*, 1993). However, in order to account for the observation that expansions of premutations to full mutations occur only on the maternally derived X chromosomes, some form of imprinting had to be invoked.

It was also initially suggested that there could be selection within the testes against full mutation sperm. However, this hypothesis was disproved by the finding that *Fmr1* knockout mice have normal fertility (The Dutch-Belgian Fragile X Consortium, 1994).

Pedigree analysis of a number of families segregating *FRA11B* or *FRAXE* full and premutations found that the parental biases for expansion of the CCG repeat tract upon transmission appeared identical to those of *FRAXA*. In the case of *FRAXE*, expansions of maternal premutations to full mutations were evident, and paternal transmissions of full mutations resulted in daughters with no fragile site expression and small, premutation type expansions (Knight *et al.*, 1994; Mulley *et al.*, 1995). Paternal transmission of a 100 CCG repeat *FRA11B* premutation allele to 2 offspring resulted in a reduction of the allele size to 85 repeats in both instances (Jones *et al.*, 1995).

Although differing in that it was for the male transmission of a pre, instead of full, mutation, the above findings for the autosomal fragile site *FRA11B*, did not correlate with the fact that no apparent expansion size change had occurred for the three paternal transmissions of a *FRA16A* full mutation observed. However, in the same pedigree, this full mutation was always unstable upon maternal transmission.

In one of the three *FRAXF* families studied, a small decrease in the size of a full mutation allele upon maternal transmission was found (Ritchie *et al.*, 1994), indicating that a reduction in expansion size may be common for this fragile site. Such size decreases have been observed infrequently at the *FRAXA* locus, however the *FRA16A* family studies did not identify any such events during transmissions of pre, or full mutations.

From comparisons of *FRAXA*, *FRA11B*, *FRAXE* and *FRAXF* pedigree data, it was evident that the paternal inheritance pattern of the *FRA16A* full mutation is markedly different from that of the other cloned fragile sites. It would therefore be of interest to determine the biological basis of this at a future time

As with the other folate sensitive fragile sites studied at the molecular level, it is only upon maternal transmission of the *FRA16A* expanded CCG repeat tract that a size increase is observed. However, the paternal transmission pattern of the *FRA16A* expansion was found to differ from that of the X-linked fragile sites, as contraction to premutation size was not observed to occur in three paternal transmissions (from one individual in Family 1). The factor(s) that may cause the difference in stability upon male transmission are currently unknown. One possibility is that they are due to the different genomic locations of the fragile sites, i.e. autosomal versus X, as a number of other characteristics (such as the *FRA16A* repeat tract composition, methylation and expansion) appear identical to those of the fragile X.

One carrier of a *FRA16A* premutation was identified in this study (Fig. 4.8, Family 2), and for this individual both the expansion size and methylation status of the *Not I* site adjacent to the *FRA16A* repeat tract were determined. The small expansion size (500 bp),

and the lack of associated methylation were features consistent with unstable premutations that have been found at other fragile site loci.

4.5.5 Somatic heterogeneity of the *FRA16A* CCG repeat tract expansion

From comparison of CCG repeat tract expansion sizes and levels of somatic heterogeneity (i.e. secondary bands or smears present) of folate sensitive fragile site mutations at all cloned loci, it was evident that *FRA16A* CCG repeat tract expansions are both the largest and most stable. Of the seventeen *FRA16A* full mutations identified, three of the expanded repeat tracts showed visible evidence of somatic heterogeneity (~18% of *FRA16A* chromosomes). By comparison, full fragile X mutations have a somatic heterogeneity frequency of 41% (Nolin *et al.*, 1994).

For two *FRA16A* full mutations, somatic instability was indicated by the presence of a single secondary band of lower intensity than the primary band. Another *FRA16A* full mutation was represented as a diffuse smear of bands (expansions of more than 5.7 kb), Clear bands representing amplifications of over 5.7 kb were not evident for any of the DNA samples tested.

4.5.6 Origin of the somatic heterogeneity in expanded CCG repeat tract size

Wöhrle *et al.* (1993) reported findings from a study performed to establish the origin of mosaic (or somatically heterogeneous) fragile X CCG repeat expansions. For this study, dilutions of cells from mosaic individuals were made in order to establish clonal populations. Analysis of the cloned progeny found that the length of a particular expansion (from one parental cell) was maintained. It was therefore concluded that the somatic variation of CCG repeat length is based on the existence of a mosaic population of cells, containing a number of different but stable *FMRI* alleles, and thus permanent mitotic instability does not occur. Based on this conclusion, it was proposed that the mosaicism and expansion to full mutation are generated post-conceptionally by the same molecular mechanism, within a particular window of early development (Wöhrle *et al.*, 1993).

A subsequent study involving detailed analysis of the intact ovaries of full mutation fetuses provided sufficient evidence to refute the proposal of Wöhrle *et al.* (1993) regarding repeat expansion timing (Malter *et al.*, 1997). This study revealed that only full expansion alleles, in an unmethylated state, are detectable in fetal oocytes. It was therefore postulated that full expansions already exist within the DNA of fetal oocytes (i.e. during germline development in the transmitting mother). An alternative suggestion was that postzygotic expansions occur early in fetal development prior to germline segregation (Malter *et al.*, 1997).

Studies of the fragile X mutations in DNAs from the testes of 13-week and 17-week full mutation fetuses found no evidence of premutations for the 13-week fetus (i. e. only full mutations), but that some germ cells of the 17-week fetus exhibited premutation status. These findings indicated that contraction of a full mutation to a premutation state occurs within the immature testes. It was suggested that the contraction might be due to a subtle selective advantage for cells expressing FMRP. Collectively, the findings from the analysis of fragile X fetal germ cell served to discount the hypothesis of germline protection from full expansion (Malter *et al.*, 1997).

In conclusion, it has been of interest to compare features of full fragile X and *FRA16A* mutation male transmissions. The contrasting findings for the fragile sites has made it evident that, whatever the mechanism leading to the fragile X full mutation repeat contraction, it is not an intrinsic property of an expanded CCG repeat (located anywhere in the genome) undergoing the process of male gametogenesis.

Chapter 5

*A Molecular Basis for
Interpopulation Variation
of FRA16A Repeat
Tract Instability*

Chapter 5: Contents

Identification and Characterisation of a FRA16A Associated Transcript

5.1	Summary	126
5.2	Introduction	127
5.3	Materials and Methods	128
5.3.1	DNA samples.....	128
5.3.2	Genotyping the <i>FRA16A</i> and <i>FRAXE</i> CCG repeat tract polymorphisms	129
5.3.3	Sequence analysis of <i>FRA16A</i> alleles	129
5.4	Results.....	130
5.4.1	Frequencies of <i>FRA16A</i> alleles in different populations.....	130
5.4.2	Results of <i>FRA16A</i> alleles sequence analyses.....	131
5.5	Discussion	132
5.5.1	Interpopulation variation of normal <i>FRA16A</i> allele size.....	132
5.5.2	Variation of <i>FMRI</i> CCG repeat tract size	133
5.5.3	Molecular basis for instability at the <i>FRA16A</i> locus.....	133
5.5.4	Stabilising effect of interruptions in simple tandem repeat tracts.....	135
5.5.5	Models explaining the stabilisation of simple repetitive tracts by interruptions.....	136
5.5.6	Polarity of mutation events.....	136
5.5.7	Influence of adjacent sequences on repeat tract instability	137
5.5.8	Transitions between configurations.....	138
5.5.9	Frequency of <i>FRA16A</i> in the populations studied.....	138
5.5.10	Correlations between dynamic mutation loci allele size and disease frequency in human populations.....	139
5.5.11	Comparison of <i>FRA16A</i> with other dynamic mutation loci	140
5.5.12	Degree of polymorphism appears directly related to perfect repeat copy number	141

5.1 Summary

PCR analysis of polymorphic *FRA16A* CCG repeat tract in the European, Japanese, and Indian population groups revealed significant differences between these ethnic populations with regard to the number and frequency of the alleles at this locus. This result implied that certain groups of alleles, either absent, or present at a low frequency in some populations, have a greater predisposition to mutate (thus generating size variation) than others.

In order to determine a molecular basis for the normal *FRA16A* allele interpopulation variation observed, sequence analysis of *FRA16A* alleles was performed. The results of this analysis revealed the degree of polymorphism at the *FRA16A* locus is the combined result of repeat number variation within four separate regions of the complex repeat tract. Sequence analysis also revealed that the larger *FRA16A* alleles do not contain a CCG repeat tract interruption that is present in all smaller alleles, a finding suggesting that a perfect repeat configuration predisposes to an increased level of repeat tract instability. At the *FRA16A* locus, such an increased level of instability may lead to the genesis of the rare folate sensitive fragile site *FRA16A*.

5.2 Introduction

The origin of the rare, folate sensitive fragile site *FRA16A*, in common with at least four other fragile sites of the same class, is a CCG trinucleotide repeat tract that is of variable length in the normal population. Other trinucleotide repeat tracts of differing sequence, with repeating units CAG and GAA, have been found to exhibit similar levels of instability (Campuzano *et al.*, 1996; Warren *et al.*, 1996). It has been proposed that a common mechanism, termed dynamic mutation, is responsible for the trinucleotide repeat tract instability associated with fragile sites, as well as an ever increasing number of genetic disorders (Sutherland and Richards, 1995a).

An indication as to the nature of the mutational mechanism(s) or pathway involved with trinucleotide repeat tract instability was revealed by findings of specific haplotypes associated with the inheritance of a number of disease causing expanded repeat tracts (e.g. the fragile X, myotonic dystrophy and Huntington's disease associated repeat tracts) more often than would be expected by chance alone (Harley *et al.*, 1991; MacDonald *et al.*, 1992; Richards *et al.*, 1992). In addition, fragile X disease associated haplotypes were found to occur most commonly with *FMR1* alleles at the upper end of the normal size range (Richards *et al.*, 1992), indicating these alleles may be the ancestors of fragile X chromosomes.

Sequencing studies of both dinucleotide repeat loci alleles and normal/disease alleles at dynamic mutation loci have identified a sequence feature, often present within repeat tracts, that appears to strongly influence the outcome of the mutational process. It has been observed, for example, that relatively long and uninterrupted AC repeat tracts are more frequently highly polymorphic (or variable in size) than those containing repeat imperfections (Weber, 1990), implying that a higher level of stability is conferred by the presence of an interruption. An apparently direct association between the presence/absence of an interruption and repeat tract stability has been found at the spinocerebellar ataxia 1 (SCA1) locus (Chung *et al.*, 1993). Ninety-eight percent of SCA1 repeat tracts in unaffected individuals were found to be interrupted. However, analysis of disease

associated unstable repeat tracts invariably found them to be uninterrupted. These results were taken to indicate that loss of the interruption causes a predisposition to repeat expansion at the SCA1 locus (Chung *et al.*, 1993). Sequencing of *FMRI* alleles provided additional supportive evidence for this general mechanism, as the results implied that the loss of the interruption in longer repeat tracts is the founder mutation for the fragile X syndrome (Eichler *et al.*, 1994; 1995b; 1996; Hirst *et al.*, 1994; Snow *et al.*, 1994).

Sequencing of one *FRA16A* allele (from My769HI) revealed the CCG repeat tract to be highly interrupted, yet it was also found to be highly polymorphic in the European population (Chapter 4). The purpose of this study was therefore to find an explanation for these apparently conflicting features, as well as to determine the similarities and differences between alleles at the *FRAXA* and *FRA16A* loci, in order to gain further insight into the molecular basis of trinucleotide repeat instability.

5.3 Materials and Methods

5.3.1 DNA samples

(Indian samples were provided by M. Denton, Dept. of Biochemistry, University of Otago, Otago, New Zealand, and Japanese samples were provided by T. Hori, Division of Genetics, National Institute of Radiological Sciences, Chiba, Japan)

The Indian DNA samples were collected from individuals living in the state of Tamil Nadu, located in a region south of Bombay. There was no known European ancestry for any of these individuals. The Japanese DNA samples were collected from individuals living on three of the main islands of Japan, and the Okinawa islands. Richards *et al.* (1994) previously used the DNAs as normal controls for *FRAXA* studies. European DNA samples constituted the parental generation of the Centre d'Etude Polymorphisme Humain (CEPH) pedigrees.

5.3.2 Genotyping the *FRA16A* and *FRAXE* CCG repeat tract polymorphisms

Genotyping of the CCG repeat tract copy number of normal *FRA16A* alleles was undertaken using PCR primers *a* and *b* (section 4.3.3), and the PCR conditions of Yu *et al.* (1992), which are described in section 2.3.6.

Genotyping of the CCG repeat copy number of normal *FRAXE* alleles was undertaken using the flanking primers described by Knight *et al.* (1993), with the PCR conditions of Yu *et al.* (1992).

The size of normal *FRA16A* and *FRAXE* alleles was determined by electrophoresis of PCR products on denaturing 5% polyacrylamide gels (section 2.3.8). As *FRA16A* alleles have complex repeat structures, all alleles were scored on the basis of length rather than repeat copy number in order to facilitate a more direct comparison of the size variation between loci.

5.3.3 Sequence analysis of *FRA16A* alleles

Where DNAs from heterozygotes were chosen as the templates for PCR amplification of *FRA16A* alleles to be sequenced, they were selected on the basis of the degree of size difference between the two alleles amplified from the sample. It was a requirement that the PCR products (from 2 allele products in a single DNA sample) were sufficiently different in size to enable the clean isolation (without contamination by the other allele) by excision from a dried polyacrylamide gel. Alleles from homozygous individuals were preferentially chosen for the sequencing of more common alleles, in order to overcome problems of contamination with another allele in the same lane.

Initially, *FRA16A* CCG repeat tract alleles of ~500 bp were PCR amplified (with *Taq* polymerase in a 7-deaza-dGTP reaction mixture, Yu *et al.* 1992) with primers *a* and *d* which flank the repeat (Fig. 5.1 - diagram of primer positions). The resulting PCR products were separated on 5% denaturing polyacrylamide gels. Gels were dried unfixed, then autoradiographed in order to locate the band(s) of interest. Sections of dried polyacrylamide gel containing *FRA16A* allele PCR products were excised, and the DNA eluted overnight in 500 µl of TE. One µl of eluted PCR product was used as template for

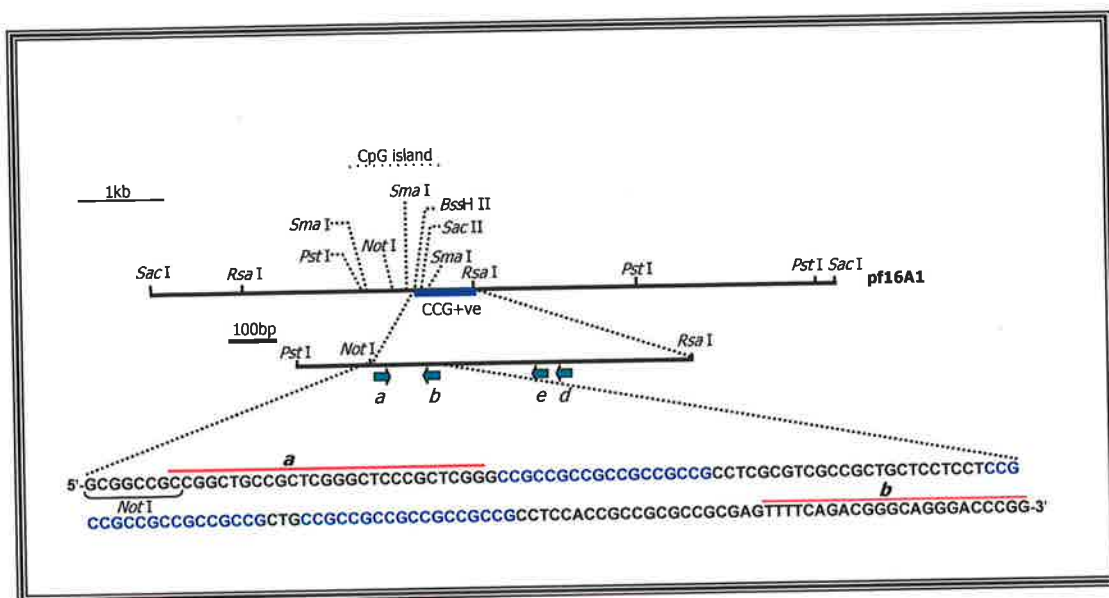


Figure 5.1 Location of the primers used for amplification and sequence analysis of *FRA16A* alleles. Diagram illustrates the position of the primers relative to the CCG repeat tract (green arrows). Within the DNA sequence the position of primers *a* and *b* (flanking the repeat tract) are indicated by red overlines.

PCR primer sequences

<i>a</i>	5' -GCCGGCTGCCGCTCGGGCTCCCGCT- 3'
<i>b</i>	5' -CCGGGTCCCTGCCCCGTCTGAAAA- 3'
<i>d</i>	5' -ATCCTACTCCCACTACGTCCTGAGG- 3'
<i>e</i>	5' -GCCTTCCCCATCACCTCCCCTCCA- 3'

a second round of PCR amplification with primer *a* and the nested primer *e* (50 µl PCR reaction using native *Pfu* polymerase and supplied buffer, Stratagene cat. # 600135). PCR products were directly purified using a Prep-a-Gene kit (section 2.3.2.3) and manually sequenced with Exo(-) *Pfu* Cyclist DNA Sequencing Kit reagents (Stratagene cat. # 200326), in accordance with the instructions supplied.

5.4 Results

5.4.1 Frequencies of *FRA16A* alleles in different populations

An initial study determined *FRA16A* allele sizes in Caucasian/European CEPH parent samples. This analysis revealed a broader distribution of CCG repeat alleles than had been evident for the Caucasian population group (in such a relatively small sample set) at other folate sensitive fragile site loci.

In order to assess whether this unusually broad distribution was due to an inherent property or instability of the *FRA16A* CCG repeat tract, DNAs from Japanese and Indian populations were similarly analysed for *FRA16A* CCG repeat allele distribution and size. Fig. 5.2 shows the results of this study in a graphical form.

The population heterozygote frequencies calculated from the allele typing results were: 91% for European; 66% for Japanese; and 59% for Indian. The number of *FRA16A* alleles (based on allele size, not configuration), and the allele size distribution (or range) found for each population group were: 21 alleles, ranging from 141 - 216 bp for Europeans; 4 alleles ranging from 153 - 168 bp for Japanese; 7 alleles ranging from 141 - 198 bp for Indians. These results clearly showed that, of the twenty-one alleles evident in the European population, 17 were absent from the Japanese population samples and 14 were absent from the Indian population samples.

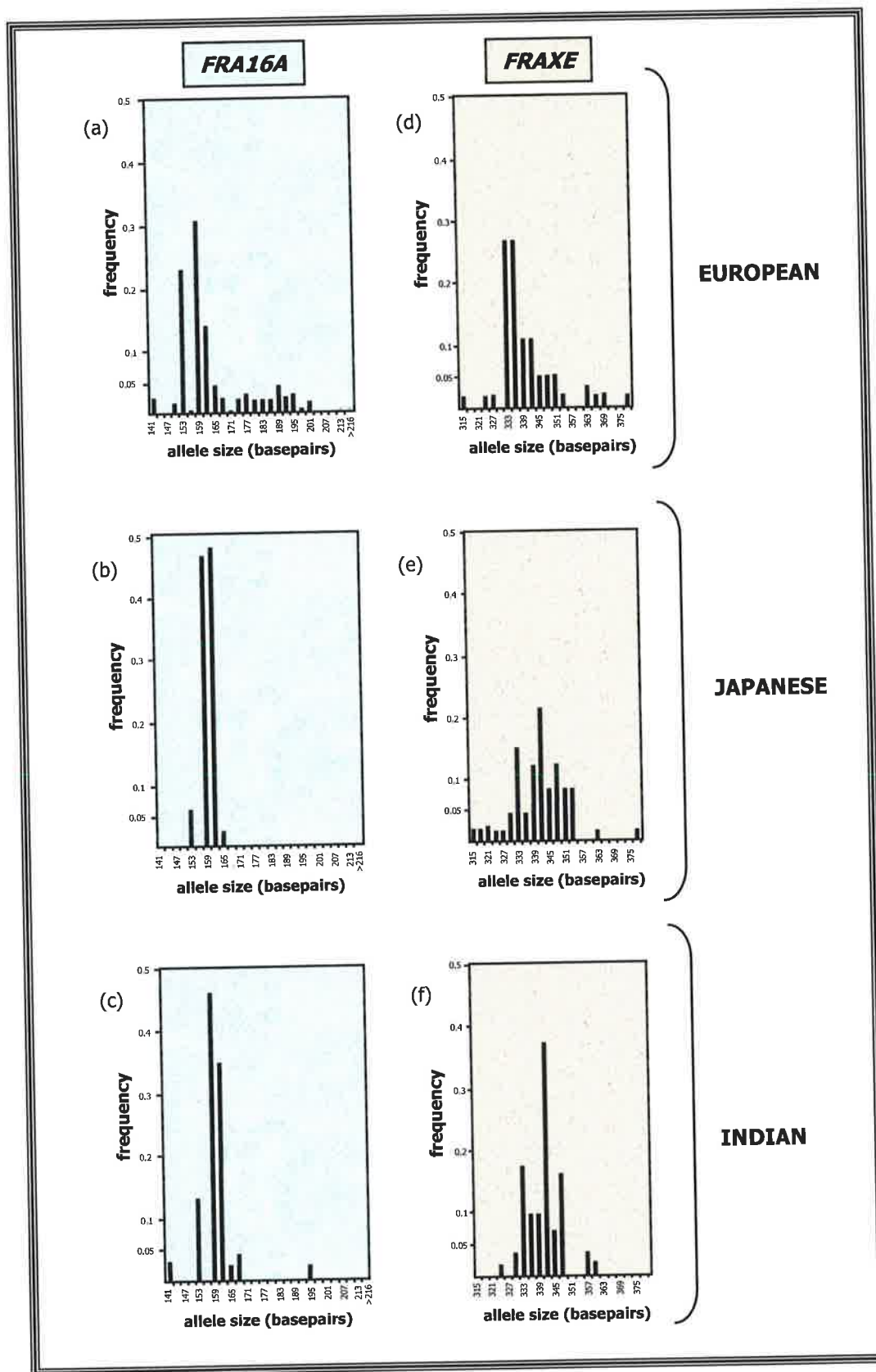


Figure 5.2 Allele frequency distributions of normal *FRA16A* (a - c) and *FRAXE* (d - f) CCG repeat tracts in the European (a, d), Japanese (b, e) and Indian (c, f) populations. Number of chromosomes represented on each graph were: a - 160; b - 114; c - 64; d - 68; e - 76; f - 78. Due to the complex nature of the repeat tract, all alleles are shown as the length of the corresponding PCR product in basepairs.

As a control for the population sampling, *FRAXE* CCG repeat alleles were typed in all three populations. The results of this study found a markedly greater similarity of allele distributions and frequencies between the three populations at the *FRAXE* locus than existed at the *FRA16A* locus (Fig. 5.2). It could therefore be concluded that the distribution and frequency differences are specific for the *FRA16A* locus and not due to biases that may occur from sampling within isolated or inbred Japanese and Indian population groups.

In order to assess the statistical significance of the *FRA16A* results, an arbitrary cut-off allele size of 168 basepairs was chosen. The frequency of alleles above and below this size was calculated to be significantly different within the three populations, as chi-squared analysis with one degree of freedom (Yate's correction), gave a value of 103.714 ($P < 0.0001$). A similar analysis between all three populations also found significant differences, with a chi-squared value of 49.82 ($P < 0.0001$, two degrees of freedom).

The higher level of variation and broader distribution of *FRA16A* alleles evident in the European population indicated the existence of an inherent instability within the larger alleles found almost exclusively in this population. In order to assess whether intergenerational instability of *FRA16A* alleles was of frequent occurrence, six CEPH pedigrees segregating 10 large (>168 bp) *FRA16A* alleles were typed. No intergenerational changes of allele size, and thus evidence of repeat tract instability, was found for the 47 transmissions within the pedigrees. Examples of the segregation of large *FRA16A* alleles within 2 CEPH pedigrees are shown in Fig. 5.3.

5.4.2 Results of *FRA16A* allele sequence analyses

Seventy-five normal *FRA16A* alleles, of which 46 were European, 16 Japanese and 13 Indian, were subjected to sequence analysis. Examples of the sequencing results obtained are shown in Figs. 5.4 and 5.5. The analysis revealed that the *FRA16A* repeat tract has a complex composition within the normal population, as four distinct simple tandem repeat regions (I - IV) contributing variously to the degree of polymorphism were identified (Figs. 5.4, 5.5 and 5.6). Four repeat tract configurations (A - D) were distinguishable by

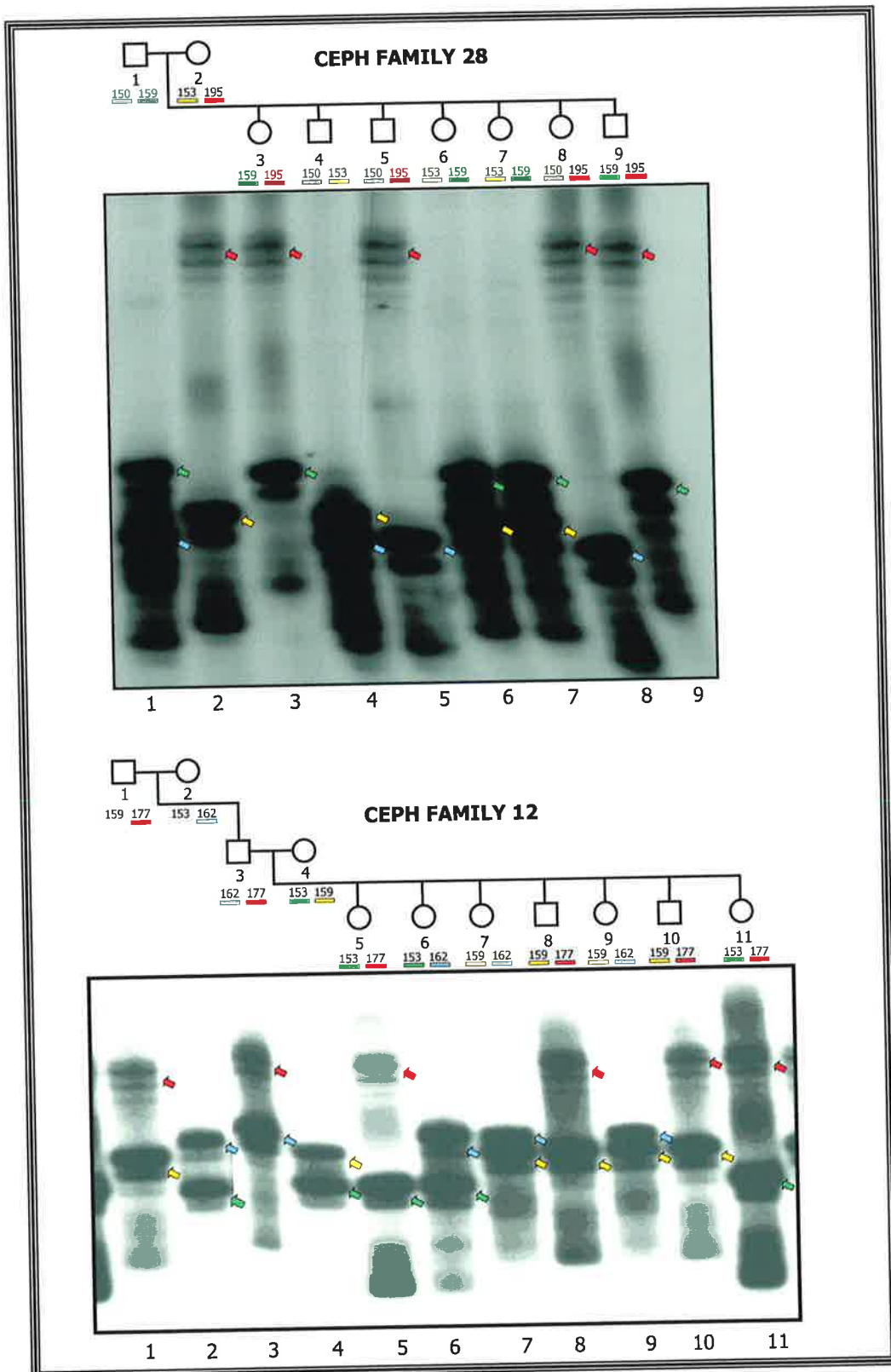


Figure 5.3 Segregation of large normal *FRA16A* alleles in two CEPH pedigrees. Allele sizes in basepairs are given under the symbol marking each individual's position in the pedigree. The colour of the rectangle under each number corresponds to that of the coloured arrow marking the position of the allele band. Changes in allele size, indicative of instability, were not observed for any transmission within these pedigrees.

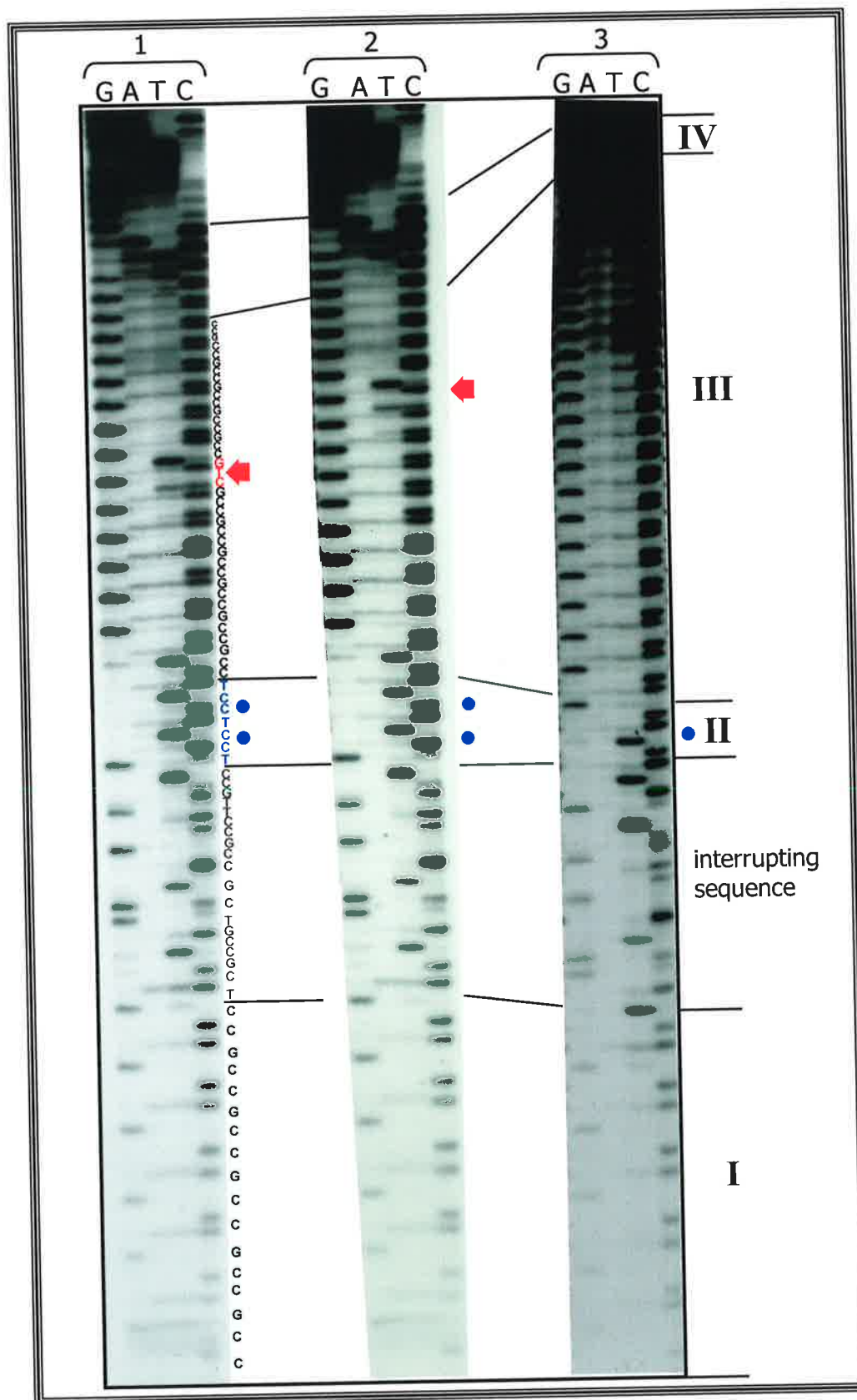


Figure 5.4 Sequence analysis of three *FRA16A* repeat tract alleles from primer *a*. Sequencing ladders are as follows: 1) 159 bp allele (configuration B); 2) 162 bp allele (configuration B); 3) 171 bp allele (configuration C). The four polymorphic regions within the *FRA16A* repeat tract are indicated. The number of CCT repeat units in each sequence is indicated by blue dots, and the CTG repeat interruptions are arrowed in red.

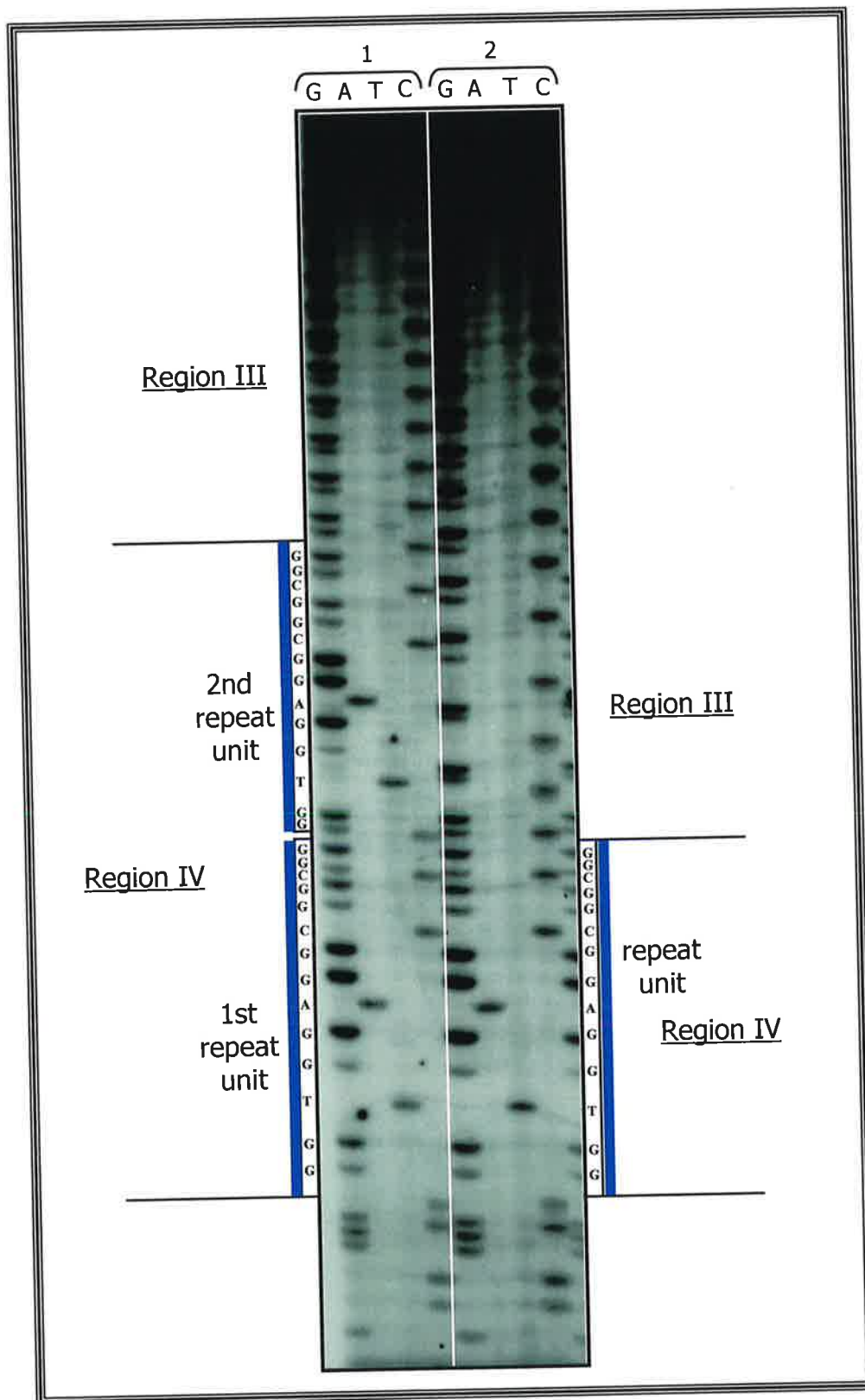


Figure 5.5 Sequence analysis of two normal *FRA16A* repeat tract alleles from primer *b*. Sequencing ladders are: 1) configuration C' allele; 2) configuration C allele. The positions of the repeat unit GGTGGAGGCGGCGG within the sequences are indicated by blue lines. The III and IV polymorphic regions within the CCG repeat tract sequence are indicated.

Table 5.1 Sequenced *FRA16A* CCG repeat tract configurations

Population	Repeat tract configuration*	Allele size (PCR product in basepairs)											
		141	150	153	159	162	165	168	171	174	177	180	>183
European	A	2											
	B		10	6	3	2	1						
	C	1						1	1	2	1	1	14
	D	1											
Japanese	A												
	B				8	8							
	C												
	D												
Indian	A												
	B			1	5	5							
	C												
	D	1											1

* Repeat configurations

A - (CCG)₃ --- (CCT)₂ (CCG)₆ CTG (CCG)₆ (CCGCCGCCTCCACC)₁--

B - (CCG)₆ --- (CCT)₂ (CCG)₆₋₉ CTG (CCG)₆₋₉ (CCGCCGCCTCCACC)₁--

C - (CCG)₆ --- (CCT)₁ (CCG)_{10, 19-41} (CCGCCGCCTCCACC)_{1, 2}--

D - (CCG)₆ --- (CCT)₂ (CCG)₉ (CCGCCGCCTCCACC)₁--

the presence of sequence differences (other than the number of CCG repeats) within the various repetitive regions. The frequency of these configurations was found to vary markedly between the European and other populations (Fig 5.6). In addition, the sequencing identified repeat region III, a CCG repeat tract interrupted by a single CTG trinucleotide in configurations A and B, as the most frequent and greatest source of variation within the complex repeat.

The sequence analysis also ascertained that the most common alleles of 153 bp, 159 bp and 162 bp are of identical configuration (B) in all populations (Table 5.1). The region III (CCG)_mCTG(CCG)_n values determined for these common alleles were: m=6, n=4 for the 153 bp allele; m=6, n=6 for the 159 bp allele; m=9, n=4 for the 162 bp allele.

The less common alleles of more than 168 bp in size were found to contain uninterrupted region III sequences of 19 or more CCG repeats, and to be of configuration C. One of the larger alleles (>168 bp), containing an uninterrupted tract of 25 CCG repeats, was found to have an unusual direct duplication of the region IV CCGCCGCCTCCACC sequence (Figs. 5.5 and 5.6).

The majority of alleles (of configuration C) with an uninterrupted region III CCG repeat also exhibited only one copy of the region II CCT repeat unit, instead of the two copies found in other classes of alleles.

5.5 Discussion

5.5.1 Interpopulation variation of normal *FRA16A* allele size

The initial aim of this study was to determine, then compare, the size variation and distribution of normal *FRA16A* allele types (not associated with fragile site expression) in the European, Japanese and Indian populations. The results of this analysis revealed significant differences in the distribution and frequency of *FRA16A* allele types between the European and the other two populations. A subsequent study of *FRA16A* allele

distributions (and alleles at other dynamic mutation loci) in a more diverse range of populations reconfirmed and extended these findings, as a high level of polymorphism at the *FRA16A* locus was still detectable only within the European population (Richards *et al.*, 1996).

5.5.2 Variation of *FMRI* CCG repeat tract size

In contrast to the *FRA16A* results, the initial population studies of *FRAXA* alleles did not identify any significant differences in the allele distribution patterns between normal populations of European, African, Hispanic and Asian origin (Fu *et al.*, 1991).

Kunst *et al.* (1996) more recently performed a genotyping and sequencing study, similar to that for *FRA16A*, with the aim of examining the relative stabilities of normal *FMRI* (or *FRAXA*) alleles in nine ethnic populations. Normal *FMRI* alleles were found to be very heterogeneous on the basis of both repeat tract length and the variation in haplotypes associated with specific alleles.

In addition, the level of *FMRI* allele size variation within a population appeared to be directly correlated with its age and/or genetic history. Population studies of *FRA16A* alleles on a much larger scale than have currently been performed would be required in order to ascertain any similar correlations.

The data from two separate studies on *FMRI* substructure were subjected to parsimony analysis in order to clarify the mutational mechanism leading to the variation. The independent, occasional loss of an AGG, leading to the formation of longer repeat tracts on many fragile X associated haplotypes, was predicted (Eichler *et al.*, 1996; Kunst *et al.*, 1996).

5.5.3 Molecular basis for instability at the *FRA16A* locus

The sequencing of normal *FRA16A* alleles revealed the *FRA16A* repeat tract to be both complex (i.e. composed of more than one sort of tandem repeating unit) and imperfect (i.e. containing interruptions, or variant repeats, within the repeat tract) in composition. Each different region within the complex *FRA16A* repeat appeared to contribute, by

varying degrees, to the level of *FRA16A* polymorphism. Based on the presence or absence of particular features in these different regions, each sequenced allele was assigned to one of the four different configurational groupings (A, B, C, D) (Fig. 5.6).

Comparisons of the differences in complex repeat structure between these four configurational allele groups made some relationships between them apparent. For example, through the occurrence of a single duplication mutation event configuration C' (represented by one allele) appeared to be derived from a configuration C allele. Additionally, the relationship between degree of instability and repeat tract composition (i.e. presence or absence of interruptions) appeared more obvious when the *FRA16A* alleles were grouped into the configurations.

A and B configuration alleles (Fig. 5.6) were distinguishable from configuration C alleles by the inclusion of a CTG interrupting sequence within the region III CCG repeat tract. Assuming the CTG interruption to have a stabilising effect, the absence of this sequence in configuration C alleles offered an explanation for the observation that approximately three quarters of the European allele types were of configuration C, yet this configuration accounted for only ~25% of the European chromosomes typed. This contrasted greatly with configuration B, which accounted for ~70% of European chromosomes but only one quarter of the allele types. The configuration C allele types therefore appeared have a relatively high mutation frequency, at least compared to configuration B alleles. In other words, the greater instability of configuration C may predispose to the generation of new allele types with a greater frequency than the more stable configuration B alleles.

In addition, the sequencing of *FRA16A* alleles in the Japanese and Indian population samples, previously established to contain a limited range of allele types, found that all Japanese alleles and 97% of Indian alleles were of configuration B. It is possible that the single configuration C Indian allele found is of European origin, as India was formerly colonised by the British. The identification of identical B configuration allele types in all three populations suggested that these alleles were present before racial divergence occurred.

That the larger, and apparently more unstable alleles, have a propensity toward length increases was evident from the observation that almost all *FRA16A* alleles of more than 168 bp in size are of configuration C. The sequencing analysis also revealed that the increased size of configuration C alleles (uninterrupted region III repeat tract) over configuration B alleles (interrupted CCG repeat tract) is solely due to increases of CCG repeat number in the configuration C region III. This finding indicated that the presence of an interruption in the repeat tract has had a highly stabilising influence, as in the populations with almost exclusively configuration B alleles (Indian and Japanese) there appears to have been a very low frequency of mutation events at the *FRA16A* locus.

Although the greatest size variation was evident within the configuration C region III CCG repeat tract, other smaller uninterrupted repeat tracts (in all configurations) within the complex repeat region were found to exhibit comparatively limited repeat number variations. For example, within the interrupted region III CCG repeat tract of configuration B, repeat number variation was evident in the short perfect repeat tracts on either side of the interruption (Figs. 5.4 and 5.6).

5.5.4 Stabilising effect of interruptions in simple tandem repeat tracts

A population survey of *FMRI* allele sequences found that highly interrupted CCG repeat alleles (>2 AGG interruptions) occur preferentially in alleles of modal repeat length, providing confirmatory evidence that the presence of AGG repeat interruptions confers stability (Eichler *et al.*, 1995a). This finding corresponds with that from *FRA16A*, as the configuration B alleles are the modal length alleles in the European population.

The stabilising effect of interruptions in microsatellite repeat tracts was confirmed by studies with a *S. cerevisiae* model. In *S. cerevisiae* a 51 bp poly(GT) tract was found to alter in length at a rate of 10^{-5} per cell division. Insertion of a single variant repeat into the centre of the poly(GT) tract resulted in a 100-fold stabilisation, providing that a functional DNA mismatch repair system was present (Heale *et al.*, 1995).

5.5.5 Models explaining the stabilisation of simple repetitive tracts by interruptions

In order to explain the observation that interruptions appear to stabilise simple repetitive tracts (di-, tri- or tetra-nucleotide), three alternative mechanisms have been proposed: (1) long repetitive tracts possess unusual structural characteristics that predispose them to alterations. Disruption of the continuity of the tracts by interruptions causes the loss of these structural characteristics (Bichara *et al.*, 1995; Gacy *et al.*, 1995; Kang *et al.*, 1995); (2) following dissociation between primer and template DNA strands (for replication), the presence of an interruption may contribute to perfect realignment, preventing the formation of displaced repeats (Heale *et al.*, 1995); (3) when the primer and template DNA strands dissociate during replication of a poly (GT) tract with an interruption, reassociation of these strands may result in 3 mismatches, 1 involving GT repeat displacement, and 2 involving the variant bases. The recognition and repair of a least 1 of the 3 mismatches may be more probable than recognition of a single displaced GT repeat in isolation, thus leading to a decreased rate of instability within interrupted repeat tracts (Heale *et al.*, 1995).

5.5.6 Polarity of mutation events

The polarity of short, uninterrupted repeat tract variation (i.e the differences in the CCG repeat copy number present in the short perfect repeat tracts between interruptions) evident in the *FRAXA* CCG repeat tract (Eichler *et al.*, 1994; Hirst *et al.*, 1994; Snow *et al.*, 1994) was not apparent within *FRA16A* alleles. This difference between the two loci, although not directly comparable because of differences in allele substructure (or configuration), suggested that the polarity effect is locus specific and therefore not an obligatory component or result of the mutational process. However, studies of interrupted repeat tracts in *S. cerevisiae* found polarity of repeat number alteration (Heale *et al.*, 1995), indicating that a universal mechanism may exist.

Comparisons of the sequences of *FRAXA* and *FRA16A* CCG repeat tract alleles revealed the *FRA16A* alleles to be less diverse than *FRAXA* alleles with regard to the number and position of the repeat interruptions. The apparent lack of *FRA16A* repeat tract length

polarity could be attributed to this relative lack of complexity, as it would be unclear (at least in configuration C alleles which are uninterrupted) which end of the repeat tract is affected in a positive or negative way. Another possibility, mentioned previously, is that there is variation in the polarity of complex repeat tract mutations (in a non-repeat sequence dependent manner) between loci. If this is the case, the mutational process must be influenced by *cis*-acting factors, such as the surrounding DNA sequences, DNA conformations, or the distance from an origin of replication.

5.5.7 Influence of adjacent sequences on repeat tract instability

In addition to the absence of the region III repeat tract interruption, *FRA16A* configuration C alleles were found to differ from configuration B alleles by the absence of one copy of the region II CCT repeat unit (configuration B has two copies). On the basis of this observation, it could be proposed that a *cis*-acting mechanism, whereby reduction in CCT repeat unit number either enables or precludes the binding of a specific protein, causes repeat tract destabilisation. Such a mechanism would probably be uninfluenced by the CCG repeat tract composition.

Such a *cis*-acting mutational mechanism has been proposed to explain the allele specific instabilities of a microsatellite repeat (Jeffreys *et al.*, 1994). As an explanation for the *FRA16A* findings, a hypothesis invoking a *cis*-acting mechanism could not easily be refuted. However, the absence of a similar, apparently predisposing, polymorphic sequence associated with the more unstable *FRAXA* haplotypes (i.e haplotypes including *FRAXA* alleles with long perfect repeat tracts), would tend to indicate that the (region II) *FRA16A* CCT one repeat allele is more likely to be in linkage disequilibrium with the perfect region III CCG repeat tract (configuration C), than to be the causative factor.

Interestingly, at the Machado-Joseph Disease (MJD) locus a polymorphism at the 3' end of the CAG repeat, where there is either a CGG or GGG trinucleotide present (although this is a sequence polymorphism, not a repeat unit number polymorphism, as with *FRA16A*), has been associated with the degree of instability in both the expanded unstable alleles and normal alleles. In *cis*, the CGG allele of normal chromosomes was

found to be associated with larger CAG repeat tracts than the GGG allele of normal chromosomes. By contrast, the GGG allele was found to influence the stability of MJD disease alleles in *trans*, suggesting the occurrence of inter-allelic reactions (Igarashi *et al.*, 1996). Another finding suggesting the existence of a *cis*-interaction has been that unstable CAG repeat Huntington's disease alleles are almost invariably associated with a specific number of CCG repeat units located immediately 3' to the CAG repeat tract (Andrew *et al.*, 1994; Barron *et al.*, 1994).

In the case of *FRA16A*, it seems most probable that the linkage disequilibrium (similar to the repeats at the HD locus) between the region II and III repeat tracts is the result of a single mutation event that caused loss of both the CCG repeat tract interruption and a CCT repeat unit, and thus formation of a configuration C allele. Such an event is most likely to be a deletion. The existence of a shorter, configuration C allele, containing eight repeats in region III, is consistent with this explanation. Alternatively, the small allele may have occurred as a result of a deletion in the CCG repeat tract of an existing larger and unstable configuration C allele.

5.5.8 Transitions between configurations

Because of the *in vivo* nature and apparently rare occurrence of transitional mutations, direct experimental evidence to support the assignments of *FRA16A* configurational origin depicted in Fig. 5.6 could not be obtained (i.e. the concept that one configuration (D) is the intermediary between B and C configurations is unproven). However, the pathways depicted between the configurations are those requiring the minimum number of mutational events, and are therefore considered to be the most likely to occur.

5.5.9 Frequency of *FRA16A* in the populations studied

The fragile site, *FRA16A*, is rare in the European population and has not been observed in other populations, although this low rate of ascertainment could be due, in part, to lack of clinical significance. There has been insufficient data acquired for even an estimation of the population frequency of *FRA16A*, or indeed any other autosomal folate sensitive

fragile site. However, if normal high copy number perfect CCG repeat tract alleles are the unstable precursors of fragile sites, a comparatively high incidence of the *FRA16A* fragile site would be expected in the European population, compared with Indian and Japanese populations, as this would be consistent with the interpopulation differences in normal *FRA16A* CCG repeat allele frequency and distribution.

5.5.10 Correlations between dynamic mutation loci allele size and disease frequency within human populations

A study of the structures of *FMRI* alleles derived from four broad ethnic groups reported no significant difference in the distribution of alleles with long (>20) pure repeat tracts. The authors therefore concluded that all the ethnic groups should be equally susceptible to development of the disease (Eichler *et al.*, 1995a).

However, a direct link between the occurrence of instability and long perfect repeat tracts at the *FMRI* locus was identified by studies of the Tunisian Jew ethnic subgroup population, which has both a high prevalence of the fragile X syndrome and an unusually high incidence of the larger normal *FMRI* CCG repeat alleles lacking AGG interruptions. That is, the proportion of alleles beyond the *FMRI* 35 repeat instability threshold in this population was significantly greater than the proportion in the control population (Falik-Zaccai *et al.* 1997).

Another example of the association between the frequency of large repeat tract alleles and disease was provided by studies of normal DRPLA alleles in several populations. It was found that longer perfect repeats are most prevalent amongst the Japanese, where DRPLA is most frequent, and are absent among Caucasians, where DRPLA is very rare (Deka *et al.*, 1995). Watkins *et al.* (1995) also found evidence of such a link at the DM, DRPLA and HD loci. Typing of alleles at these loci in African, Asian and Caucasian populations revealed that disease prevalence within a population corresponds directly with the number of allele types in the upper tail of the allele size distribution.

These population studies provided a clear indication that the variation of normal dynamic mutation locus allelic frequencies between populations may predict the relative incidence of the corresponding dynamic mutation disease within these populations.

5.5.11 Comparison of *FRA16A* with other dynamic mutation loci

The alleles at all known human loci at which dynamic mutations occur have been found to exhibit specific ranges of repeat copy number in normal individuals. The majority of these characterised loci are associated with genetic diseases. However, there may be a bias of ascertainment toward the identification of loci associated with a disease phenotype. In support of this suggestion, repeat expansion detection (RED) assays have detected a number of other trinucleotide expansions that are not apparently associated with a phenotype (Schalling *et al.*, 1993).

It has been observed that the incidence of dynamic mutation events (causing cytogenetic expression of a fragile site and/or phenotypic manifestation of a disease) at a given locus appears to be primarily influenced by repeat tract length and composition of alleles at that locus. The observation of founder effects for fragile X syndrome, myotonic dystrophy and Huntington's disease has led to the proposal that there exists pools of relatively high risk alleles (with longer or uninterrupted repeats tracts) at these loci, derived from one, or a small number of founder mutations (Richards *et al.*, 1992; Arinami *et al.*, 1993; Buyle *et al.*, 1993; Imbert *et al.*, 1993; Oudet *et al.*, 1993a; Barron *et al.*, 1994).

For the fragile X, it has been proposed that the evolution of full mutations could involve four definable stages: (1) ancestral events leading to the production of predisposing alleles with a large total repeat length and 1 or zero interruptions; (2) gradual slippage of these predisposing alleles to small premutations (S alleles); (3) conversion from S alleles to larger premutations (Z); (4) massive expansion from Z allele to full mutation (L). (Chakravarti, 1992; Morton and Macpherson, 1992; Macpherson *et al.*, 1994).

A phenotype, deleterious or otherwise, has not yet been correlated with *FRA16A* expression, therefore the incidence of the fragile site is not apparently constrained in this way. As *FRA16A* has never been found in a homozygous state, it is currently unknown if

a phenotype would result. However, given the apparent rarity of the *FRA16A* fragile site, genetic lethality in the homozygous state would have a negligible effect on its population frequency.

The striking similarities in the molecular basis of the *FRA16A*, SCA1 and *FRAXA* repeat tract instabilities indicate the existence of a common mutational mechanism at these loci (Chung *et al.*, 1993; Hirst *et al.*, 1994). The common features characterising the more stable or unstable alleles at all three loci appear to be only the repeat composition (i.e. presence or absence of interruptions), and the normal range in copy number which manifests as the degree of polymorphism within a given population. These features may therefore be a useful pointer to assess the possible instability of a given trinucleotide repeat tract, or even the degree of frequency of any associated genetic disease.

5.5.12 Degree of polymorphism appears directly related to perfect repeat copy number

The results of the analysis of *FRA16A* CCG repeat tract alleles presented in this chapter are supportive of the view that instability, which for one conformational group (C) of *FRA16A* alleles is manifest as a comparatively high degree of polymorphism, has a direct relationship to perfect repeat copy number. This relationship was first proposed as an explanation for the relative frequencies of mutation events within dinucleotide repeat tracts (Weber *et al.*, 1990). The subsequent studies discussed in this chapter have indicated that a general property for microsatellite repeat tracts of varying length and/or composition is that an increased probability of mutation (i.e. change in repeat unit number) bears a direct relationship to a greater existing perfect repeat copy number. This is in preference to a threshold model, in which below a given repeat number threshold level, transmission of a microsatellite allele is always stable, whereas above this level it is always unstable.

Chapter 6

*Identification and
Characterisation of a
FRA16A associated
Transcript*

Chapter 6: Contents

Identification and Characterisation of a FRA16A Associated Transcript

6.1	Summary	145
6.2	Introduction	146
6.3	Materials and Methods	147
6.3.1	Northern blot hybridisations.....	147
6.3.2	Identification of cosmid clones overlapping the <i>FRA16A</i> locus.....	147
6.3.3	Subcloning of <i>FRA16A</i> probe pf16A1 sonicated fragments for sequence analysis.....	148
6.3.4	Sequencing of pf16A1 sonicated fragment subclones	149
6.3.5	Searches for homologies and gene related features by computer analysis of the pf16A1 sequence.....	149
6.3.6	Hybridisation of <i>FRA16A</i> cosmid clones to zooblots	149
6.3.7	Subcloning cosmid c312B11 and c310A11 restriction fragments	149
6.3.8	Sub-localisation of conserved sequences within p312B11	150
6.3.9	Direct selection of cDNA library clones with <i>FRA16A</i> region DNA probes.....	150
6.3.10	Screening of fetal brain cDNA library pools.....	151
6.3.11	Completion of <i>FRA16A</i> CpG island sequence	152
6.3.12	Analysis of DNA sequences from regions adjacent to the <i>FRA16A</i> CCG repeat tract	153
6.3.13	Analyses of the predicted <i>FRA16A</i> associated polypeptide sequence to determine possible protein characteristics.....	153
6.3.14	MegaYAC information.....	154
6.4	Results.....	155
6.4.1	Strong cross-species homology of sequences near <i>FRA16A</i>	155

6.4.2	Initial detection of homologous transcripts on a Northern blot of multiple fetal tissues.....	155
6.4.3	Detection of cDNA clones homologous to <i>FRA16A</i> region sequences by a direct selection method	156
6.4.4	Isolation of a human fetal kidney cDNA clone	156
6.4.5	Isolation of additional clones from fetal brain cDNA library pools, and construction of a cDNA contig extending from <i>FRA16A</i>	157
6.4.6	Exon detection within cDNA sequences	158
6.4.7	Sequential hybridisation of cDNA sequence probes to a fetal tissue Northern blot	158
6.4.8	Identification of a putative coding sequence	159
6.4.9	Prediction of promoter regions near <i>FRA16A</i>	159
6.4.10	Prediction of coding exons within the 7407 basepair cDNA/genomic DNA sequence	160
6.5	Discussion	161
6.5.1	Cross-species homology of DNA sequences in the vicinity of <i>FRA16A</i>	161
6.5.2	Significance of the hybridisation of probes representing regions in the vicinity of <i>FRA16A</i> to differently sized transcripts	162
6.5.3	Occurrence of CCG repeat tracts in the 5' untranslated region of human genes.....	163
6.5.4	Significance of <i>FRA16A</i> CpG island.....	165
6.5.5	Function of the genes associated with characterised rare folate sensitive fragile sites	166
6.5.6	Features of the putative <i>FRA16A</i> associated gene.....	168
6.5.7	The putative <i>FRA16A</i> associated gene is a candidate disease gene for PXE.....	169

6.1 Summary

In an attempt to gain further insight into the nature of the functional role of fragile site related CCG repeat tracts, a search for transcribed sequences in close proximity to *FRA16A* was performed.

Northern blot hybridisation analysis with three genomic DNA probes in the region identified a number of transcripts, but none that were clearly in common. Screening of three different fetal tissue cDNA libraries identified fifteen cDNA clones. These were sequenced and arrayed in a contig. Comparison of the cDNA to genomic DNA revealed no evidence of RNA splicing events, therefore it was not conclusively shown that the cDNAs were representative of transcribed sequence.

The sequence of the cDNA contig was combined with genomic sequences in the vicinity of the *FRA16A* CCG repeat tract, producing a total of 7407 bp of DNA sequence. This DNA sequence was subjected to a number of computer analyses in order to determine the presence of gene associated features. The most interesting results of these analyses were the predictions of promoter regions and an exonic sequence located in close proximity to the *FRA16A* CCG repeat tract. Comparison of the predicted exon sequence against protein and DNA sequence databases found no significant homologies to other genes, but identified two short amino acid sequence motifs that are found on extracellular proteins. The putative *FRA16A* gene is a candidate for pseudoxanthoma elasticum disease gene, as the locus for this disorder has recently been localised to the same region.

6.2 Introduction

As discussed in previous chapters of this thesis, five rare folate sensitive fragile sites (*FRAXA*, *FRAXE*, *FRAXF*, *FRA11B* and *FRA16A*) have been cloned and characterised. Two of the fragile site loci (*FRAXA* and *FRA11B*) have been definitively localised within the 5' untranslated regions of the genes *FMR1* and *CBL2*, respectively (Verkerk *et al.*, 1991; Jones *et al.*, 1995). The situation for *FMR2*, the gene associated with *FRAXE*, is less clear. Only one cDNA clone containing the *FRAXE* CCG repeat was isolated, and it was not possible to determine whether this clone represented a transcribed sequence (Gu *et al.*, 1996). However, considering the close proximity (40 bp to ~80 bp) of the cDNA sequences (characterised independently by Gecz *et al.*, 1996 and Gu *et al.*, 1996) to the *FRAXE* CCG repeat, and since methylation of the expanded repeat silences the *FMR2* gene, it seems probable that the repeat is present on *FMR2* transcripts, or is within the promoter region of the gene.

Since at least two out of five fragile site loci are associated with genes, it is probable that the CCG repeat tract, which is potentially unstable and detrimental, has a biologically important function. Therefore, although *FRA16A* expression has not been associated with any genetic disorder, it was considered that valuable insights into the function of the CCG repeat tract and the nature of the associated genes might be gained from characterisation of transcribed sequences near this fragile site. Another point considered was that, although the *FRAXA* (*FMR1*), *FRAXE* (*FMR2*) and *FRA11B* (*CBL2*) genes appear to bear no relationship to one another in terms of function or expression, it is possible the characterisation of more fragile site associated genes may reveal a distinct pattern of similarities.

6.3 Materials and Methods

6.3.1 Northern blot hybridisations

Commercial Northern blots of multiple adult human tissue mRNAs (cat. # 7760-1) and fetal human tissue mRNAs (cat. # 7756-1) were obtained from Clontech Laboratories Inc. (Palo Alto, California, USA). The adult tissue Northern blot contained polyA⁺ mRNA derived from heart, placenta, lung, liver, skeletal muscle, kidney and pancreatic tissue sources. The fetal tissue Northern blot contained mRNAs derived from brain, lung, liver and kidney tissue sources.

The adult tissue Northern blot was probed with pre-reassociated pf16A1 insert DNA using hybridisation methods described in section 2.3.4. The fetal tissue Northern blot was probed with p312B11 (a 1.4 kb subclone of cosmid c312B11), using ExpresshybTM hybridisation buffer (Clontech cat. # 8015) according to the protocol provided with the buffer.

6.3.2 Identification of cosmid clones overlapping the *FRA16A* locus

Filters containing high-density gridded arrays of cosmid DNAs (representing ~10 times coverage of human chromosome 16) were obtained from Los Alamos National Laboratories (LANL). These filters were probed with pre-reassociated pf16A1 insert DNA (section 2.3.4). Three true positive cosmids (c312B11, c310A11, c337C11) corresponding to the hybridisation signal positions (i.e. a black dot after autoradiography - data not shown) on the filters were identified and obtained from LANL.

Confirmation that the cosmids spanned the *FRA16A* repeat tract region, and mapping of the cosmid inserts with respect to each other on the basis of the relative position of the *FRA16A* CpG island *Not* I site, was accomplished by probing Southern blotted *Not* I digests of the 3 cosmids with a PCR product probe (from primers pgdna16A-1 and pgdna16A1-4). The sequence and location of these primers, and all those used in the project, are given in Appendix 1. The localisation of the DNA sequence homologous to this probe to within specific *Not* I restriction fragments from each cosmid enabled the

orientation of the cloned human DNA inserts with respect to the *FRA16A* *Not* I site. The cosmid contig is shown in Fig. 6.1.

6.3.3 Subcloning *FRA16A* probe pf16A1 sonicated fragments for sequence analysis

A preparation of plasmid clone pf16A1 (10 µg in 1 ml of TE) was subjected to sonication at 40% power for 5 seconds (1 burst) on a Heat Systems sonicator in order to produce DNA fragments ranging in size from approximately 2.8 kb to 0.7 kb. The sonicated DNA fragments were incubated with T4 DNA polymerase in order to blunt or 'polish' the ends (Manatis *et al.*, 1989). The blunt ended fragments were Prep-a-Gene purified and concentrated (section 2. 3.2.3), then ligated to dephosphorylated *Hinc* II digested (blunt ended) pUC19 vector (section 2.3.3). The pf16A1/pUC19 ligation reaction was transformed into commercially prepared Epicurian Coli® SURE® 2 super competent cells (Stratagene cat. # 200152), in accordance with the protocol provided with the cells. The transformation mixture was plated onto ampicillin and IPTG/X-Gal plates for blue/white colour selection (section 2.3.3.4).

To eliminate clones containing pBluescript® vector derived DNA inserts (as pf16A1 insert DNA was not isolated from the pBluescript® vector before sonication) single white colonies were picked from the transformation plates and spotted in a grid pattern onto fresh ampicillin plates. The plates were incubated at 37°C overnight, and colony lifts performed. The colony lift filters were probed with pf16A1 insert and the colonies corresponding to very strong hybridisation signals on the filters were streaked and grown at 37°C overnight. For each positive clone, single colonies were used for the inoculation of 10 ml L. Broth + ampicillin cultures. The cultures were incubated at 37°C overnight and the plasmid DNA extracted using a QIAGEN tip 20 kit (cat. # 12125, QIAGEN Pty Ltd, Victoria, Australia), according to the protocol provided by the manufacturer.

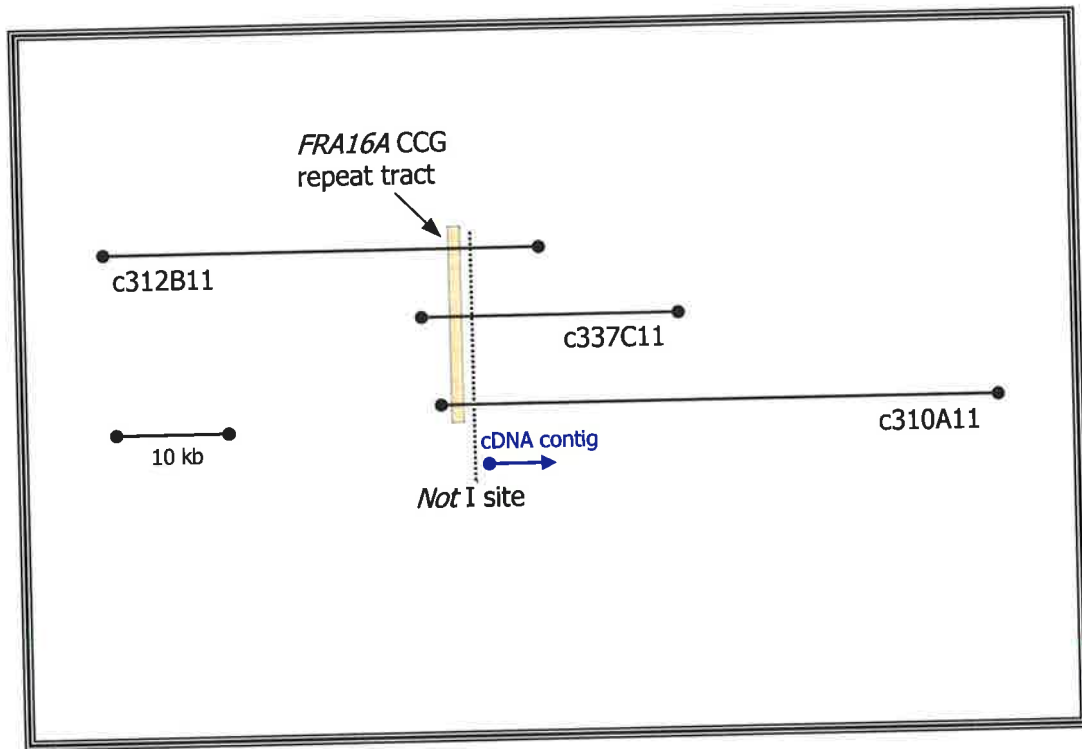


Figure 6.1 LANL cosmid contig overlapping the *FRA16A* CCG repeat tract.

6.3.4 Sequencing of pf16A1 sonicated fragment subclones

Automated fluorescent sequencing, using Perkin Elmer PRISM Dye Primer sequencing kits, was performed in both the forward (Dye Primer -21M13) and reverse (Dye Primer M13Rev) directions in order to acquire sequence data from both ends of the insert of each sonicated fragment subclone. The Lasergene SEQMAN program was used to assemble the pf16A1 subclone sequences into contigs.

6.3.5 Searches for sequence homologies and gene related features by computer analysis of the pf16A1 sequence

All pf16A1 insert derived sequences (both singly and in contigs) were compared to the Genbank nucleotide sequence database using the sequence comparison program BLAST. This analysis was designed to identify those subclones containing vector derived sequences from pBluescript®, many low and high copy number human repetitive sequences, as well as sequences with homology to characterised genes or expressed sequence tags (ESTs). For the prediction of possible exonic coding sequences, the sequence data were also analysed with Grail2, fexh, and fgenh (Table 6.1a).

6.3.6 Hybridisation of *FRA16A* cosmid clones to zooblots

The presence of evolutionarily conserved DNA regions can be indicative of the presence of coding sequences within genes. Two cosmids (c312B11, c310A11) spanning *FRA16A* were used as hybridisation probes to screen ~50 kb regions on either side of the locus for the presence of cross species DNA sequence homologies. Two filters were prepared, containing human, mouse, rat and dog *Pst* I digested DNAs (section 2.3.4). One filter was probed with c312B11, and the other with c310A11. After hybridisation, the filters were washed under conditions of normal stringency (section 2.3.4.9).

6.3.7 Subcloning cosmid c312B11 and c310A11 restriction fragments

A 1.436 kb *Eco* RI - *Sac* I fragment containing evolutionarily conserved sequences from cosmid c312B11 was subcloned into *Eco* RI and *Sac* I digested pUC19 (sections 2.3.3).

Table 6.1(a) World Wide Web DNA analysis programs

program	WWW address (http://...)	analytical function for human DNA sequence data
Promoter Prediction by Neural Network (NNPP)	www-hgc.lbl.gov/projects/ promoter.html	recognition and prediction of DNA polymerase II (Pol II) promoters in genomic DNA sequences
PROSCAN (PROMOTER SCAN II)	biosci.cbs.umn.edu/ cgi-bin/proscan	
TSSG	dot.imgen.bcm.tmc.edu: 9331/gene-finder/gf.html	recognition of human Pol II promoter regions and transcriptional start sites
TSSW		
BLAST-Basic Local Alignment Search Tool	www.ncbi.nlm.nih.gov/ cgi-bin/BLAST/nph-blast	Set of similarity search programs designed to explore all available sequence databases, regardless of whether the query sequence is protein or DNA
GRAIL2	avalon.epm.ornl.gov/ Grail-bin/EmptyGrailForm	Coding recognition program that uses a neural network which combines a series of coding prediction programs
fexh	dot.imgen.bcm.tmc.edu: 9331/genefinder/gf.html	searches for potential 5', internal, and 3' coding exons
fgeneh		prediction of exons, then construction of a gene model by exon assembling
Genie	www-hgc.lbl.gov/projects/ genie.html	gene finder based on hidden Markov models
Genview	litba.itba.mi.cnr.it/ ~webgene/wwwgene.html?	prediction of protein coding genes
GENSCAN	gnomic.stanford.edu/ ~chris/GENSCANW.html	prediction of complete gene structures, including exons, introns, promoter and poly-adenylation signals in genomic sequences

Table 6.1(b) World Wide Web gene feature and sequence similarity analysis programs

program	WWW address (http://...)	analytical function for human protein sequence data
BLASTP+BEAUTY	dot.imgen.bcm.tmc.edu: 9331/seq-search/protein- search.html	gapped BLASTP amino acid sequence similarity search with BEAUTY post processing that adds annotated domain information
MOTIF (MotifFinder)	www.genome.ad.jp.SIT/ MOTIF.html	searches for protein sequence motifs against the PROSITE protein motif database
BLOCKS	www.blocks.fhcrc.org/ blocks_search.html	detection and verification of protein sequence homology- compares a protein or DNA sequence to the current database of protein blocks (blocks are short multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins)
PSORT	psort.nibb.ac.jp/form.html	analysis and prediction of protein sorting signals coded in amino acid sequence, and thus protein localisation sites in cells

The resulting subclone, p312B11, mapped immediately distal to pf16A1, with a common *Sac* I site. To compare further genomic sequences with those of the isolated cDNAs, a 4 kb *Eco* RI fragment immediately distal, and sharing a common *Eco*RI site with p312B11, was subcloned into the *Eco* RI site of pUC19.

6.3.8 Sub-localisation of conserved sequences within p312B11

Subclone p312B11 was partially sequenced using fluorescent automated sequencing methods (section 2.3.7.1) and the resulting sequence data analysed with GRAIL2. In order to generate PCR product probes for finer localisation of the conserved sequence(s) within the subclone, two primers (pg312B11-5 and pcdna68a-1) were designed from the sequence.

PCR reactions with combinations of the two insert and two vector specific primers pUCF and pUCR were used to generate 3 overlapping PCR products from a p312B11 template. The PCR products generated were from the following primer pairs: pcdna68a-1 and pUCF; pg312B11-5 and pcdna68a-1; pg312B11-5 and pUCR. The relative position of these primers is shown in Appendix 1. Three further zooblot filters were made, containing *Pst* I digested dog, hamster, rat and human DNAs, and each was probed with one of the three overlapping PCR products.

For screening a human fetal brain cDNA library, the primer pair pg312B11-3 and pg312B11-4 was designed from a highly conserved region of DNA sequence detected by Southern blot analysis (primer sequence and location shown in Appendix 1). These primers enabled amplification of PCR products (of divergent sizes) from both mouse and human DNA templates (data not shown).

6.3.9 Direct selection of cDNA library clones with *FRA16A* region DNA probes

Probes consisting of: (1) p312B11 insert; (2) c310A11 (overlapping p312B11); (3) a unique PCR product from primers pgdna16A1-1 and pgdna16A-4 (Appendix 1), representing DNA sequences located on both sides of the *FRA16A* CCG repeat tract,

were used to screen Clontech adult brain (λ gt10), adult liver (λ gt11), adult kidney (λ gt11) and fetal brain (λ gt11) cDNA libraries by the direct selection method of Parimoo *et al.* (1991; 1993). In brief, cloned inserts from four cDNA libraries (1 μ l aliquot of each library in 100 μ l reactions) were PCR amplified with λ vector specific primers. For selection of the amplified cDNA insert PCR products, $\sim$ 0.2 ng of each DNA sample was immobilised on a separate 5 mm diameter nylon filter disc (Hybond N+). Discs were prehybridised and quenched at 65°C for 1 day. Quenching agents used were the following DNAs: P15 library (human chromosome 15 genomic library); XLRI (genomic repetitive sequence library derived from human X chromosome); pR7.3 and pR5.8 (human ribosomal RNA 45S precursor region clones); RK3535 (*E.coli* ribosomal RNA operon clone); sonicated human placental DNA. The composition of the solution used for both prehybridisation and hybridisation stages was 5 x SSPE, 5 x Denhart's solution, 0.5% SDS. After prehybridisation, the filter discs were washed briefly in 5 x SSPE, 0.1% SDS solution, then transferred to fresh hybridisation buffer with amplified cDNA inserts and quenching reagents. Filters were incubated in the hybridisation solution at 65°C for 48 hours, then subjected to 12 washing steps in 2 - 0.1 x SSC/0.1% SDS and 0.1 x SSC solutions at 65°C and room temperature. Amplified cDNA inserts bound to the filter were eluted in water at 4°C overnight, and a 5 μ l aliquot used as template for PCR amplification with λ gt10/ λ gt11 vector specific primers. Samples (10 μ l) of the resulting direct selection PCR products were electrophoresed (1.5% agarose gel), then Southern blotted and hybridised with a selecting probe.

6.3.10 Screening of fetal brain cDNA library pools (pools constructed by S. Whitmore)

The construction of pools of fetal brain cDNA clones for efficient PCR screening was performed by the following method: (1) dilutions of a Clontech fetal brain cDNA library were plated onto 100 plates (137 mm) at a density of $\sim$ 15000 pfu/plate (total of 1.5 million clones); (2) lambda phages were eluted from each plate into 10 ml SM, producing 100 stocks; (3) stocks were pooled into groups of five, which were then pooled again according to a grid system, thus producing 40 pools. Pooling of clones was done in such a

way that any true positive cDNA clone would be detectable by the amplification of PCR products from two pool templates within a grouping (A, B, C or D); (4) cDNA pools were PCR screened with primers designed from previously characterised cDNA or conserved sequences, using 2 µl of each cDNA library pool as template in a total 20 µl reaction volume. Ten µl of each PCR reaction was electrophoresed on 2% agarose gels (0.5 x TBE) in order to visualise the PCR products obtained from positive pools; (5) positive stocks were plated out at a density of 15000 pfu/plate, and plaque lifts performed. PCR product DNA probes were used to probe plaque lift filters for the several rounds of screening required to isolate each positive phage clone; (6) inserts of the λ cDNA clones were PCR amplified with λ gt11 forward and reverse primers in 50 µl reactions. PCR products were purified using a Prep-a-Gene kit (section 2.3.2.3), then sequenced with vector and locus specific primers.

In order to identify overlapping cDNAs within the fetal brain cDNA pools, PCR primer pairs were designed from the DNA sequence at both ends of each clone for further screening. The full sequences of cDNAs were obtained from vector and locus specific primers in a primer walking strategy (Voss *et al.*, 1993).

The sequence chromatograms of cDNA clones mapping within pf16A1 were aligned with those of the sonicated pf16A1 DNA subclones, using the Lasergene SEQMAN DNA sequence contig assembly program. Sequence chromatograms from those cDNA clones overlapping regions spanned by clone p312B11 and pf16A1 were also compared manually in order to identify any sequence divergences arising from splicing of the corresponding transcript.

6.3.11 Completion of *FRA16A* CpG island sequence

Sequencing of the remainder of the *FRA16A* CpG island (as the full sequence was not determined initially – see Chapter 4) was primarily completed by the use of primers PFRA16A-9 and PCLONE38(3)R as sequencing primers, and a *Pfu* polymerase cycle sequencing kit (section 2.3.7.2).

6.3.12 Analysis of DNA sequences from regions adjacent to the *FRA16A* CCG repeat tract

All cDNA sequences obtained, and the C and G rich genomic sequences adjacent to the *FRA16A* CCG repeat tract, were arrayed in a contig. The 7407 bp DNA sequence of the total contig was analysed by a number of gene feature recognition and prediction programs (through the Internet World Wide Web - WWW) in order to identify putative exons and promoter regions. The WWW programs used for the prediction of human Pol II promoter regions were: TSSG; TSSW; NNPP; PROSCAN. The WWW programs used for prediction of coding exons were: GRAIL2; fexh; fgeneh; Genie; Genview; GENSCAN 1.0. Table 6.1a lists the functions of programs utilised for the gene feature analyses with the corresponding WWW site.

6.3.13 Analyses of the predicted *FRA16A* associated polypeptide to determine protein characteristics

The amino acid sequence of the exon predicted by all utilised exon recognition algorithms was subjected to analysis with Lasergene PROTEAN, a computer program designed to predict protein characteristics from primary amino acid sequence data. Various analytical methods are employed to plot protein characteristics according to particular scientific concepts, such as hydropathy, secondary structure, antigenicity, amphiplicity, flexibility, charge density and surface probability (Lasergene User's Guide, 1994).

As given in Lasergene User's Guide (1994), the methods used for analysis of the *FRA16A* amino acid sequence were: (1) Garnier-Robson and Chou-Fasman algorithms for the prediction of α -helix, β -pleated sheet, β -turn and coil secondary structures. Both algorithms are based on statistical predictions derived from crystallographic analysis of a number of proteins with known sequence, although using different approaches (Chou and Fasman, 1978; Garnier *et al.*, 1978); (2) Kyte-Doolittle method for the prediction of the regional hydropathy of proteins from their amino acid sequences. The definition of hydropathy is a strong reaction toward water, where positive is hydrophilic and negative is hydrophobic (Kyte and Doolittle, 1982); (3) Eisenberg method for the prediction of

hydrophobic moments. Hydrophobic moments describe the distribution and asymmetry of hydrophilic, hydrophobic and amphipathic residue groups in a protein, and are semi-empirical quantities based on computational and experimental measurements (Eisenberg *et al.*, 1984); (4) Karplus-Shulz Flexibility method for the prediction of backbone chain flexibility. This algorithm classifies amino acid residues as either flexible or rigid, then calculates the nearest neighbour interactions between the residues. The results of the analysis may be of use for the identification of antigenic sites, as these regions are often the most flexible in a polypeptide sequence (Karplus and Shultz, 1985); (5) Jameson-Wolf method for the prediction of potential antigenic determinants. This method combines existing methods for protein structural predictions to determine antigenic index, or surface contours values (Jameson and Wolf, 1988); 6. Emini surface probability method for calculation of the probability that a given region lies on the surface of a protein. Comparison of the surface probability plot of the protein of interest with those from proteins similar in structure, function or origin, but have no similarities at the primary or secondary levels, may be of value (Emini *et al.*, 1985).

6.3.14 MegaYAC information

The megaYAC contig map and STS mapping data for the PXE region were obtained from the Whitehead Institute of Technology (<http://www-genome.wi.mit.edu/>) and Los Alamos National Laboratory (<http://www-ls.lanl.gov/>) WWW sites. Information regarding megaYAC size and integrity was also obtained from these sources.

6.4 Results

6.4.1 Strong cross-species homology of sequences in the vicinity of *FRA16A*

The cosmids c310A11 and c312B11, in combination, span *FRA16A* and the ~ 40 – 50 kb genomic DNA regions immediately proximal and distal to the fragile site. When the cosmids were hybridised to a zooblot, strong sequence homology to both hamster and mouse DNA was detected. This homology appeared to be located within the region of cosmid overlap (Fig. 6.1).

The region of strong cross-species homology was narrowed down to a 1.4 kb *Eco* RI - *Pst* I fragment located 3 kb from the *FRA16A* CCG repeat tract. This fragment was subcloned and titled p312B11. PCR products containing three overlapping sections of the *Eco* RI fragment were produced (section 6.2.8) and used as probes on zooblots containing human, mouse, hamster and dog DNAs. The results of these Southern blot hybridisations (Fig. 6.2) indicated that the homology extended right across the fragment, with an ~100 bp region of very high homology immediately adjacent to the *Sac* I site at one end of the fragment. However, exon prediction analysis of the complete sequence of the fragment did not reveal any features of interest, such as putative exons or homologies to other genes.

6.4.2 Initial detection of homologous transcripts on a Northern blot of multiple fetal tissues

The hybridisation of a Northern blot of human fetal tissue mRNAs with probe p312B11 detected a relatively abundant homologous 1.35 kb mRNA that is expressed in all tissues (Fig. 6.3). However, the most intense signal was present in the fetal kidney mRNA lane. The cDNA library of choice for an attempt to isolate a cDNA corresponding to this transcript was therefore one derived from a human fetal kidney tissue source.

6.4.3 Detection of cDNA clones homologous to *FRA16A* region sequences by a direct selection method

The results of the direct selection assay described in section 6.2.9 are shown in Fig. 6.4. The strongest signal was observed for the fetal brain cDNA library. However, the adult brain and lung cDNA libraries also gave signals, but of much lower intensity. The pg16A1 primer PCR product (representing sequences on the opposite side of the fragile site) did not appear to be homologous to any clones within the libraries (data not shown). This result suggested that the cDNA libraries were not extensively contaminated with genomic DNA sequences.

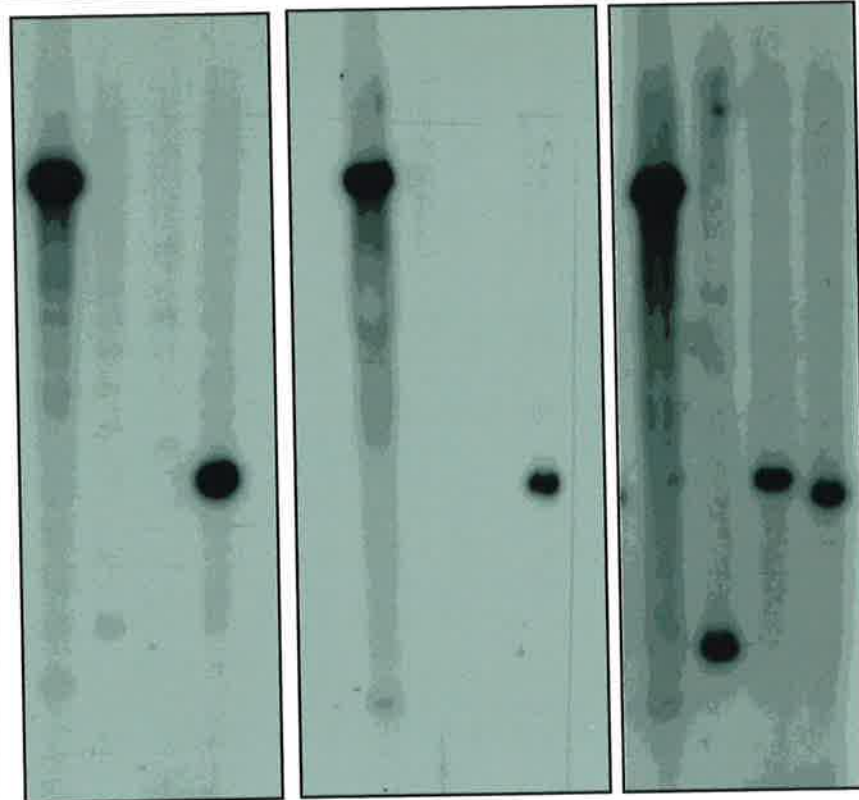
The small cDNA inserts that were enriched by the selection procedure were not further analysed as larger *FRA16A* region specific clones could be directly isolated from the pooled fetal brain cDNA library (the same library as was used for direct selection).

6.4.4 Isolation of a human fetal kidney cDNA clone

Screening of the human fetal kidney random and oligo-dT primed cDNA library isolated 8 positive clones of identical size and sequence, indicating this library to be over-amplified. The larger 1.8 kb size of the fetal kidney cDNA clone indicated that, although apparently containing very high sequence homology (if not an identical sequence) it did not directly correspond to the 1.35 kb mRNA identified on the fetal tissue Northern blot.

It was considered possible that the 1.35 kb mRNA detected on the fetal tissue Northern blot was the mature, spliced form of an inefficiently spliced *FRA16A* associated transcript. Therefore, as a mature mRNA is more likely to have a poly-A tail than the unspliced form, a total fetal oligo-dT primed library (theoretically derived only from poly-A⁺) was screened with p312B11 insert. One positive cDNA clone of 1.6 kb was isolated, but comparison with a cDNA clone sequence previously obtained demonstrated that (in this instance) an oligo-dT primer had primed off an A rich tract of either genomic DNA or transcribed RNA origin, not a poly-A tail. This approach was not therefore successful in isolating a mature transcript with a true poly-A tail from the cDNA library.

Figure 6.2 (A) Zooblot filter hybridisation results for PCR product probes derived from DNA sequences near *FRA16A*. Lanes contained *Pst* I digested DNAs from 4 different mammalian species, in the order: A. human; B. mouse; C. hamster; D. dog. PCR probes contained overlapping human DNA sequence from a region shown in (B) overlined in red and bordered by *Eco* RI and *Sac* I sites. The restriction fragment was subcloned from c312B11 into pUC19, producing plasmid clone p312B11. The primer pairs used to amplify the PCR product probes for each autoradiograph were: 1. pcdna68a-1 and pUCF; 2. pg312B11-5 and pcdna68a-1; 3. pg312B11-5 and pUCR. The position of the primers relative to the *FRA16A* CCG repeat tract is shown in (B). Blue arrows represent the pUC19 vector primers (pUC19F and pUC19R), the red arrow represents primer pcdna68a-1, and the yellow arrow represents primer pg312B11-5.



A

A B C D

1

probe: pcdna68a-1
to pUCF PCR product

A B C D

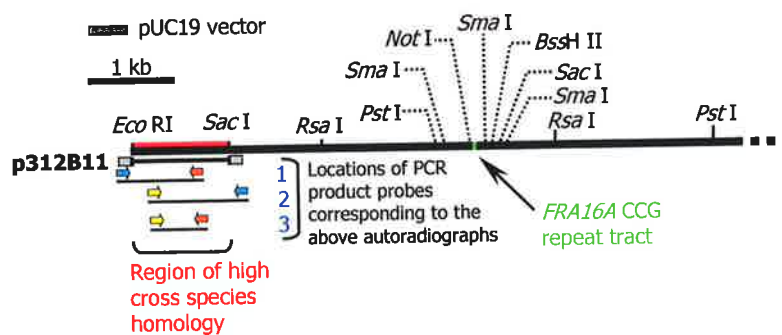
2

probe: pg312B11-5
to pcdna68a-1 PCR
product

A B C D

3

probe: pg312B11-5
to pUCR PCR product



B

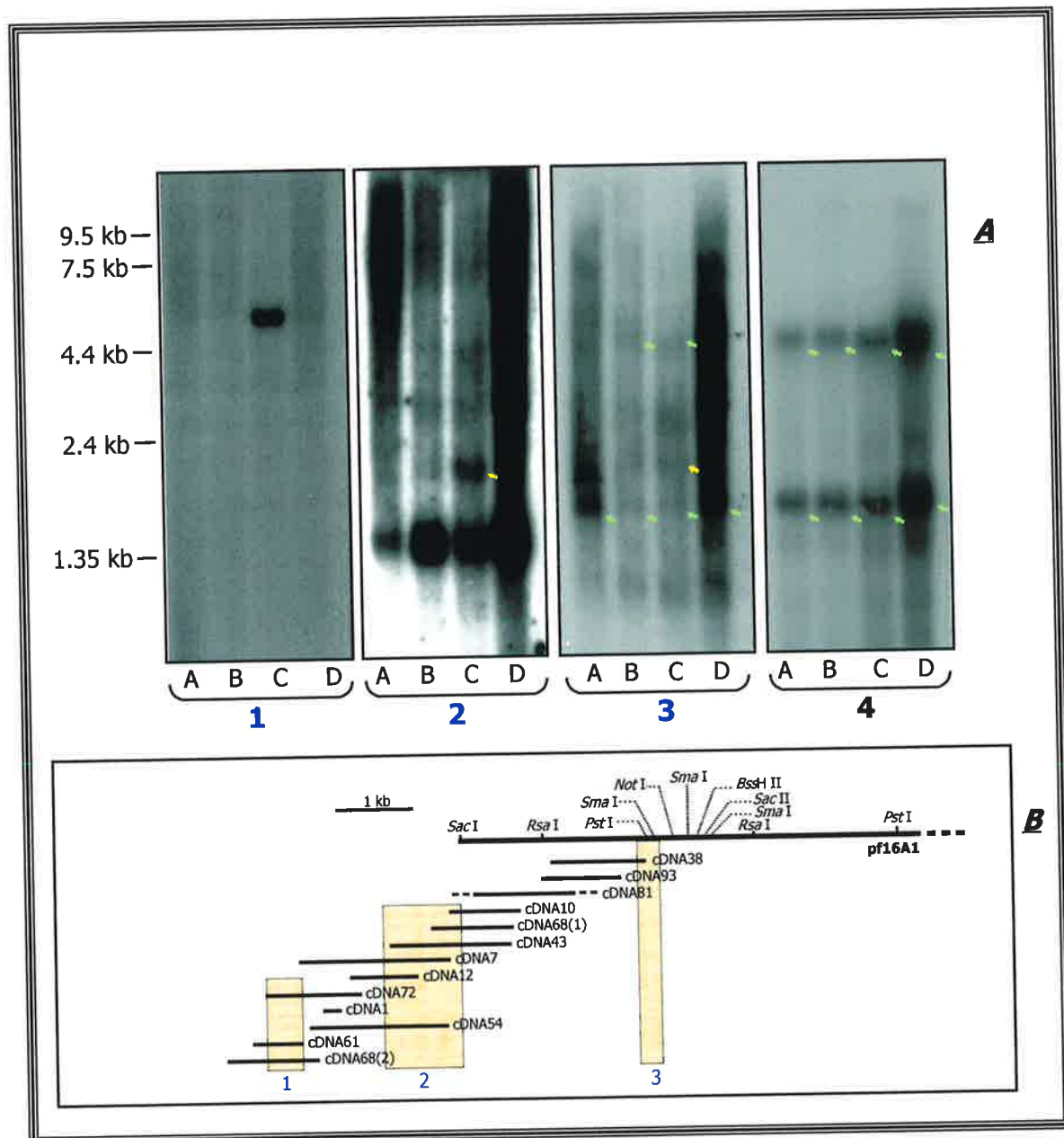


Figure 6.3 (**A**) The results of sequential hybridisations of a human fetal multiple tissue Northern blot. Probes used were: 1. pcdna72-1 - lambdagt11F PCR product; 2. p312B11 insert; 3. pgdna16A1-5 - pcdna38-1 PCR product; 4. Cloned 45S ribosomal RNA gene. The rRNA gene probe was included to ensure the positive hybridisation signals observed were not due to an rRNA homology (frequently a problem when using G and C rich probes). The green and yellow arrows mark possible common bands. Lanes A - D contained (in order) mRNAs from human fetal brain, lung, liver and kidney. (**B**) A map of the genomic and cDNA clones showing the positions of the Northern blot probes used for the sequential hybridisations that produced autoradiographs 1, 2 and 3. The yellow-brown shaded areas indicate the regions of DNA spanned by the probes. The numbers 1, 2 and 3 below the shaded areas correspond to the numbers below the autoradiographs.

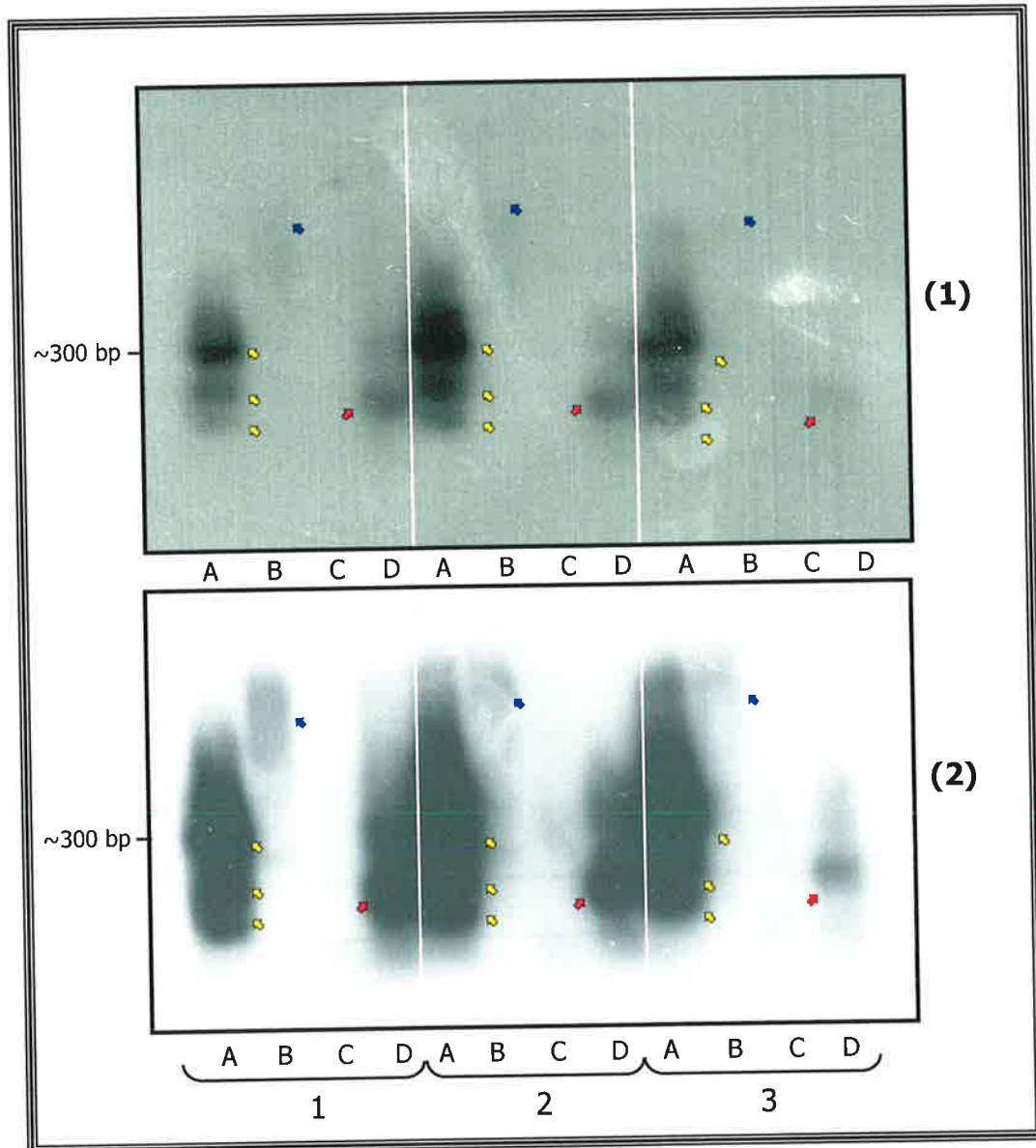


Figure 6.4 Results of cDNA selection. **(1)** is a short exposure autoradiograph and **(2)** represents a longer exposure of the same filter. After one round of hybridisation and PCR amplification, cDNA selection products from four different cDNA libraries were immobilised on a nylon filter and probed with p312B11 insert (1.4 kb *Eco* RI - *Sac* I fragment). Selecting probes were: **1.** p312B11; **2.** c310A11; **3.** c312B11. cDNA libraries used for selection were: **A.** fetal brain; **B.** adult brain; **C.** adult liver; **D.** adult lung. Yellow arrows mark the positions of fetal brain derived products, blue arrows mark the position of adult brain derived products, and red arrows mark the positions of adult lung derived products. No products were derived from the adult liver cDNA library.

6.4.5 Isolation of additional clones from fetal brain cDNA library pools, and construction of a cDNA contig extending from *FRA16A*

As the attempts to identify a cDNA corresponding to the 1.35 kb transcript detected in fetal tissues had been unsuccessful, it was decided to continue sequentially screening the human fetal brain cDNA library pools. This approach would add further clones to the contig and therefore had the potential to isolate cDNA clones showing evidence of splicing or with a poly-A tail (at the 3' end of the contig).

As a result of screening, thirteen unique cDNA clones were obtained from the pooled random and oligo-dT primed fetal brain cDNA library. One cDNA clone each was obtained from the random and oligo-dT primed fetal kidney and total fetal oligo-dT primed cDNA libraries. Therefore, in total, fifteen cDNA clones were obtained from the three cDNA libraries screened, all containing DNA sequences identical to those in the vicinity of *FRA16A*. A contig assembled from thirteen of the cDNA clones is shown in Fig. 6.5.

As comparison of sequences obtained from genomic and cDNA sources did not reveal any differences, the possibility that all clones isolated from the cDNA libraries were of genomic origin could not be excluded. However, Northern blot hybridisation results (Fig. 6.3) showed clear evidence of transcribed sequences in the region, although the size of the transcripts detected was smaller than would be predicted, based on the 5341 bp size of the characterised cDNA contig (assuming this sequence was representative of a true mRNA).

Sequence data from the fifteen cDNAs was assembled manually to form a sequence contig extending distally from *FRA16A*, with the proximal (5') end located 392 bp from the *FRA16A* CCG repeat tract. Repeated PCR based screenings of the fetal brain cDNA pools did not identify any cDNAs in closer proximity to the CCG repeat tract. In addition, no evidence of a poly-A tail was found in the most distal (3') clone of the contig, indicating the complete *FRA16A* transcript was not represented.

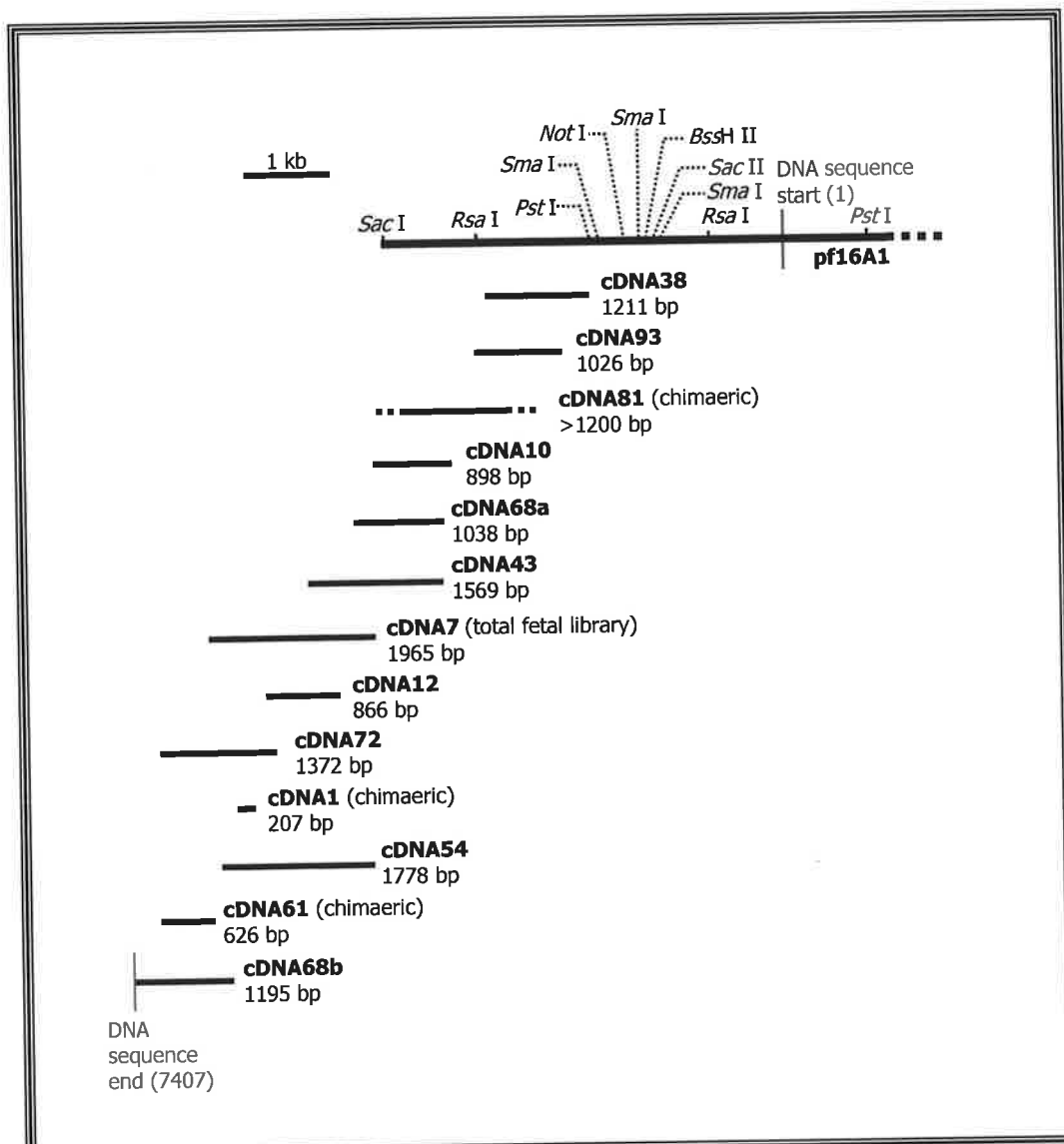


Figure 6.5 Assembled contig of 13 cDNA clones homologous to a region of DNA adjacent to *FRA16A*. The orientation and overlap of the cDNA clones with respect to each other and the *FRA16A* CCG repeat tract is shown. Red lines indicate the start and end of the sequenced region of DNA.

6.4.6 Exon detection within cDNA sequences

Analysis of the entire cDNA contig sequence by GRAIL 2, fexh, fgench, Genie, Genview, GENSCAN 1.0 failed to identify any putative exons. This result suggested that if the region of DNA spanned by the contig is transcribed, the resulting transcript is not subsequently spliced. The exon identification programs therefore provided no evidence to indicate that the isolated cDNAs were derived from an unspliced transcript.

6.4.7 Sequential hybridisation of cDNA sequence probes to a fetal tissue Northern blot

Autoradiographs resulting from the Northern blot hybridisation experiments are shown in Fig. 6.3. Observations from these results were: (1) the p312B11, 1.436 kb *Eco* RI-*Sac* I fragment detected a 1.35 kb transcript in all the fetal tissues on the blot (section 6.3.2); (2) a PCR product probe (primers pcdna72-1 – lambdagt11F) derived from cDNA72 sequences detected a ~6 kb transcript in fetal liver; (3) a PCR product probe from primers designed from both genomic and cDNA sequences (primers pgdnaA1-5 – pcdna38-1) detected ribosomal RNAs, probably due to the probe's CG richness. Other faint bands were detected, but were difficult to distinguish from the lane background. One such faint band, which is discussed below, was the same size as a band detected by the p312B11 probe.

Clear evidence of transcripts detected by more than one probe was not found, even after long exposure, although one band (marked by yellow arrows in Fig.6.3) appeared to be of similar size but not intensity. Only the size of the ~6 kb transcript detected by cDNA 72 was greater than the size of the cDNA contig.

BLASTN searches for sequence similarities in the Genbank, EST and TIGR databases to the cDNA contig and *FRA16A* sequences revealed no clear sequence homologies, indicating a low level of expression for the putative gene, and an absence of repetitive sequences.

6.4.8 Identification of a putative coding sequence

Three other folate sensitive fragile sites, *FRA11B*, *FRAXA* and *FRAXE*, are known to originate from unstable CCG (or GGC) repeat tracts located in the 5' untranslated regions of a genes. Consequently, it was considered possible (given the *FRAXA* and *FRA11B* findings) that the putative *FRA16A* transcript extends further toward (from one end of the cDNA contig), or encompasses the *FRA16A* CCG repeat tract. As the regions immediately distal and proximal to the CCG repeat tract were refractory to DNA sequencing with the fluorescent and manual methods previously employed, another sequencing method, more suitable for GC rich/high secondary structure regions, was required. Manual sequencing with a Stratagene *Pfu* cycle sequencing kit (enabling a higher DNA denaturing temperature to be used than can be used for *Taq* I, for improved sequencing of regions with high secondary structure) was used to complete the sequence characterisation of the region from the presumed 5' end of the cDNA contig to the *FRA16A* CCG repeat. The contiguous sequence of 1.4 kb proximal to the repeat was obtained by combining the new *Pfu* manual sequencing data with the automated fluorescent sequencing data obtained from the pf16A1 sonication subclones. The combined 7407 bp DNA sequence of this region, and the positions and sequences of the primers used for sequencing and walking, is shown in Appendix 1.

6.4.9 Prediction of promoter regions near *FRA16A*

Analysis of the 7407 basepair DNA sequence with the promoter prediction programs PROSCAN and NNPP (Table 6.1a) identified a number of regions as putative promoters. The locations of these regions are shown in Fig. 6.6.

6.4.10 Prediction of coding exons within the 7407 basepair cDNA/genomic DNA sequence

Analysis of the 7407 basepair DNA sequence, representing both the genomic and cDNA sequences combined, with the exon identification programs (GRAIL2, fexh, fgeneh, Genview, GENESCAN and Genie), predicted the existence of four separate exons (Fig. 6.6). However, the only region consistently predicted to be an exon by the programs started at a position 112 bp distal to the *FRA16A* CCG repeat tract, with a ATG (methionine) start codon. This ATG codon appeared to form part of a short stretch of sequence (motif) conforming with the Kozak consensus sequence, a feature indicating that the predicted exon may represent the most 5' coding region of the putative gene.

In-frame translation of the DNA sequence of the predicted exon region, if it was assumed that no splicing of the transcript occurs, indicated that the sequence encodes a polypeptide of 143 amino acids. A sequence similarity search of the amino acid sequence against Swissprot protein database sequences with the BLASTP program revealed no clearly significant homologies. However, analysis with the similarity search program BCM BLASTP+BEAUTY identified very limited homology to the collagen and fibrillin proteins of 3 different species (Fig. 6.7). Comparison of the amino acid sequence to protein motif databases, using the MOTIF and BLOCKS search programs revealed homology to a specific cell attachment sequence and the EGFBLOOD2 protein motifs (as well as several others that did not appear potentially significant). Further details of the similarity search programs are shown in Table 6.1(b).

The putative 143 amino acid polypeptide sequence was subjected to analysis by the program Protean (Lasergene) which was designed to identify structural features encoded by the sequence, such as alpha, beta sheet, turn, flexible and coiled regions (Fig. 6.8). In addition this program also identified regions of differing charge and hydrophobicity/hydrophilicity. Within the 143 amino acid sequence, the program identified distinct hydrophobic and hydrophilic domains within the amino acid sequence (Fig. 6.8). The results of the structural analysis appeared inconclusive, as no patterns

1 AACAGGCATG AGGCCCCCCC GCCCAGGCC TATCTCACTA TATTTTAATG
 51 ATTAACATCC ACATCAGGCT TCTCTAGACC TCAGGCTTGC TGAAGATGAG
 101 ATTAACCTC ATTCACCTTT GGATGCGTGC ACCTGGCCCA GTGTCTGACA
 151 CAAGCTTTGT ATTCACGGAT GCTTGTGGAG TGGATGAAAC AATCTCTTCC
 201 CACTCCCTGC CTCATTGTCT GATGAGCCTT CATTACACTC CACTTTCTTC
 251 TCTTTGTAAA TTAGGATGAG GTGAAAAAGC ACAGCAGACA CAATTCACAA
 301 ACTGATTTGA TCTGCTGTTG TATTCCCAGA GTGCTAGGTC CCGTATGGAC
 351 CACCCAAACG CCATCCCATA TGCTCACTGT TCTGCGCTGC TGTGTATAAA
 401 GCCCTGGGCA AGAGGACAAG GGAGGAAAA ACCAGAAAGA CTGAGTCAGA
 451 GAAGGCTCGG GATGAAATTT TGAGTCCATT GGGATCATTC TGCCTAAACG
 501 ACAGAGGAGA AAGCTTGGGC TCAGGGAGGA GTCCTCCTCC TCCTGGCTCA
 551 AGGTCACCGG GACAGACCAG TCAGGATGGG AAAGAACTCC ACTTTCAAGA
 601 GCAAGACCTG TTTCCCCAGG ACCTGGGCCC TGTGCAGACA GGAAGGGGGA
 651 CTTCAGGATG GATGGGGAAA AGGACACGCT GGGAGGCTGC GGGGCCAGCC
 701 TTTGGGGCTT GCATCCTGCG AAGGTTGAGT AACTCCGACG GTCTCCCTTT
 751 TGGCGCTATT CCACTTGCCA ACTGGGTTTG GGAGTCTCAT GGCTTGTGAG
 801 TCCTCAGGGA CCAGAGAAGT GACTCAGTGA ACACTTAGTA CATCGTTGGC
 851 ACCGGGGATT GTGTGTGATT CAGGAGTCAC TCACAGAGCT TGTGCTGATT
 901 AACTCCAGAA AATCAACCAA AGCCTCCTCC CCACCCCTTC AAACCCAGGA
 951 TGAAGGGATC AGACTGTGGC TCTGACACTC CTGAAATGTT TCTCACGTGT
 1001 CCACCTTCCA ACCGGTCTGG CACCGACAGA AGATCCCTGG ACCGCGAACT
 1051 CCGGGGCTAG GAGCTTCCTC CACGCACTGT CTCCTTCCAG AGGACCACAG
 1101 GGAAAACAGG ACCTCGGGAC CCCCTCTCTT CCTGACCTCC CCTGGATTCT
 1151 GAGCGGGCCC ACCCTCAACC AAAGGTCCCC TGTTTCGCGG GCCCTGCCTC
 1201 TCTTTTTCCTG CCTTGGGAAT TGACCCCTC GGCCCCCATC CTACTCCCAC
 1251 TACGTCCTGA GGGGTGTCTG TCTTCTCTGA TCGCCCCCAC CCCCTTCCTT

promoter
regions
predicted
by
PROSCAN
program

promoter
regions
predicted
by NNPP
program.
Bases
shown
in green
indicate
predicted
trans-
criptional
start sites.

Genie
predicted
exon

Genie and
Grail2
predicted
exon

fexh
predicted
exon

TATA box

tssw

Kozak consensus sequence

USF binding site

NF1 binding site

1301 TCCCTCCTCC TTTTCCTCCC TCGGCTCCTC CCCCAGGCCC GGCCTCTCC
 1351 GGCTCGGCCC GCGTCCCCCC GGGGCCTTCC CCATCACCTT CCCCTCCAGA
 1401 GGGGACAGGG GTGTGGGGAG GGGGCGCCGC GGGCCAGCGC CCGCCCTCCC
 1451 TCGCGCCCCG TCCCCGAGCG CGCCCGGCGC CCCCCTCCGC GGCCTCCCCG
 1501 CTCCCGCCTC CCGCCTTCTC CTCCTCCTGG TTGACGTCCC CGGGTCCCTG
 1551 CCCGTCTGAA AACACGCGGC GCGGCGGTCC AGGCGGCGGT GGAGGCGGCG
 1601 GCGGCGGCGG CCGCAGCGGC GCGGCGGCGG GCGGAGGAGG AGCAGCGGCG
 1651 ACGCGAGGCG GCGGCGGCGG CGGCCCGAGC GGGAGCCCGA GCGGCAGCCG
 1701 GCGGCGGCGG GAGCTGCGGG GAGCGCGGCG GCGGCCCGGA GCGTGCCGGG
 1751 GTCCCCGCTC CTCGCTCGCC TGCCGCGCTC CGAAGATGCT GCGGCGGCGG
 1801 TCGGCCCCGA GGCTGGCCCC GCGCTCGCAC TCGGCGCTGC TCGGCGGCGT
 1851 CGCGGTGCTG CTGCACACGC TGGTCGTGTG GAATTTTCAGC AGCCTCGACT
 1901 CCGGGGCGCG GGAGCGCCGC GGGGCGGCGG TCGGCGGCGG GGAGCAGCCG
 1951 CCGGCGGCGG CCGGCGGCGG CCGAGCCCGG GACCTGCCCC CCGAGCCGCT
 2001 P P A P tssg A P R R A R D L P A E P L
 2051 G C C C C A G G C G C G G C G G G G A G G C G G C C C G G A A A C C G C G G G A C A G C
 2101 A G C C G G C C A G C C G G G G G C A C T G C C C G C C G G G C T C T G G T T A G A T G C C C T
 2151 S R P A G G H C P tssw P G L W L D A L
 2201 T T C G G T C G C G C C C G C G G G A G G G A G G A A G G C A G C C C G A C C C C A C
 2251 S A R A P R R G R E E G S P T P
 2301 G C C C C C T A G C C C T T G A A C T C G T C A T C A A A C T G G G A C T C T G C A A G C C T A
 2351 R P P S P stop CCGACTCCAC ATCGCCTCTT GCCCGGGGAA GCCGGGTGGG GTCCCTGCCA
 2401 AGGCTCAGGG GTCGTGGGCT GCGGCTTTCA GAGTCTGAGC ACGGACCCCT
 2451 TGGGGTCTGG GGGAAGTTT CTCCAGCGGG GCGCTTTGG GGGTTACCGG

CpG island start (blue arrow pointing to TTTTCCTCCC)

BssH II (above TCCCCGAGCG)

TATA box (above AACACGCGGC)

tssg (red arrow pointing to A in AGGCGGCGGT)

tssw (red arrow pointing to P in GGAGGCGGCG)

NotI (above GCGGCGGCGG)

Kozak consensus sequence (above CGAAGATGCT)

open reading frame (blue arrow pointing to TCGGCCCCGA)

FRA16A CCG repeat tract (in italics) (referring to the repeat tract in the sequence)

direction of cDNA contig (blue arrow pointing right)

fgeneh, fexh, GenView, GEN-SCAN1.0, Grail2 predicted exon/open reading frame sequence (referring to the amino acid sequence)

Figure 6.6 Section of DNA sequence spanning the *FRA16A* CCG repeat tract and CpG island in which a high density of gene promoters and exons were predicted by computer analysis. A red arrow indicates the starting point of the cDNA contig. Bases shown in red indicate the transcriptional start sites predicted by TSSG and TSSW. Details of computers programs used for the analyses are given in Table 6.1 (a) and (b).

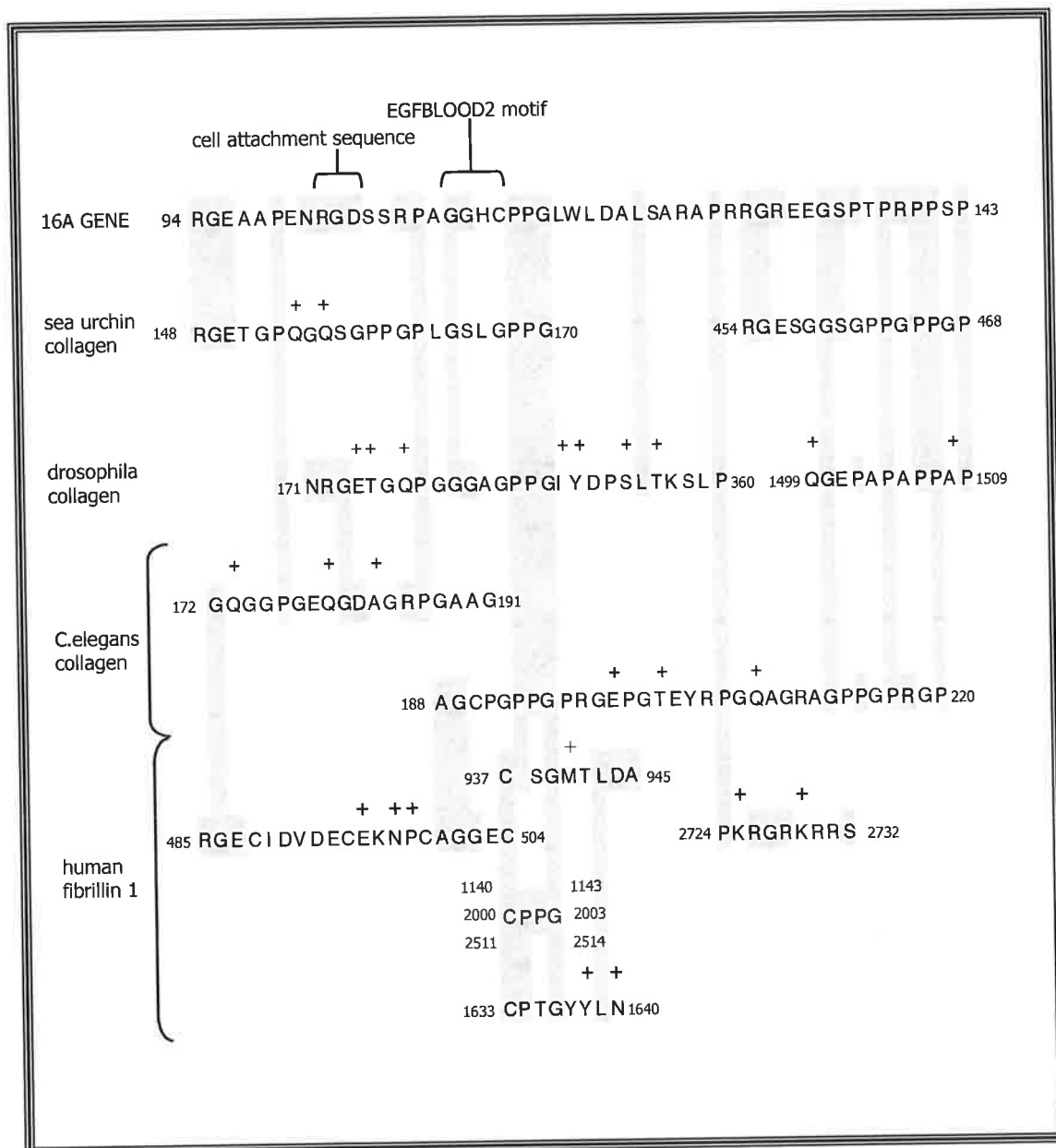


Figure 6.7 BLAST analysis of the predicted *FRA16A* gene amino acid sequence (residues 94 - 143) detected limited homologies with the extracellular matrix proteins collagen and fibrillin. Searches against protein motif databases identified a cell attachment sequence (present in some collagens) and EGFBLOOD2 (present in fibrillins) motifs within the amino acid sequence.

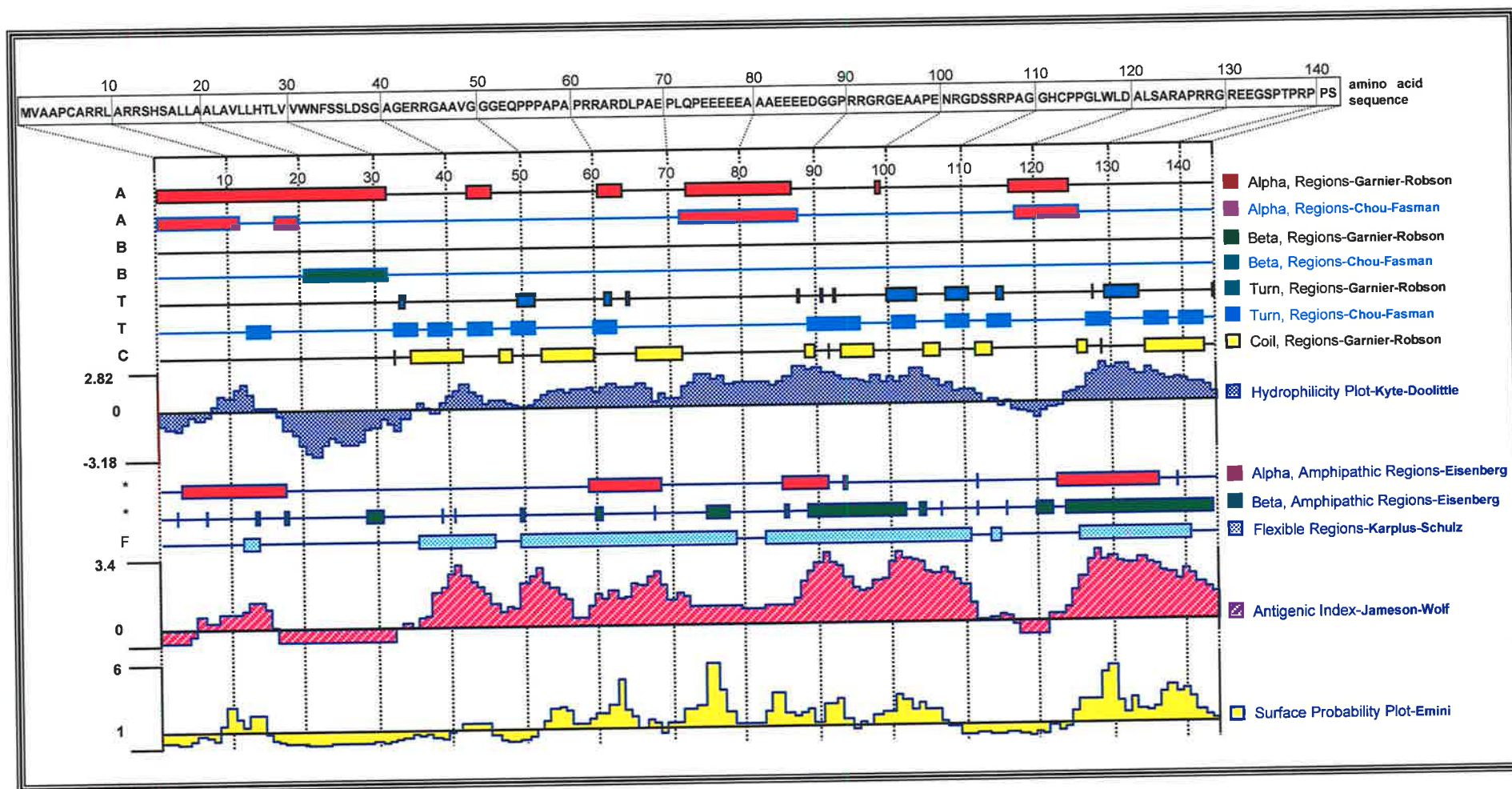


Figure 6.8 Analysis of the putative *FRA16A* gene predicted amino acid sequence with the computer program PROTEAN (Lasergene). The possible significance of these results is discussed in section 6.4.

were evident. However, this may be the case with proteins that do not have repeating structural features.

In addition, the predicted 143 amino acid polypeptide sequence was analysed with PSORT (Table 6.1b), a program designed to predict the protein localisation, based purely on the submitted amino acid sequences with no other structural data. PSORT predicted the location of the protein was most likely to be extracellular, with a signal sequence and possible cleavage site at the 28th amino acid in the sequence.

6.5 Discussion

6.5.1 Cross-species homology of DNA sequences in the vicinity of *FRA16A*

The zooblot assay results (section 6.3.1) indicated the presence of DNA sequences in the vicinity of *FRA16A* that are highly conserved between species, and thus pointed toward the existence of a gene within this region. The strong cross-species sequence homology was localised to within a ~1.4 kb region, spanned by the probe p312B11 insert (Fig. 6.2B), that is located ~3.3 kb from the *FRA16A* CCG repeat tract. However, computer analysis of DNA sequences in this region found no evidence of exonic sequences.

Such strong sequence homologies were not detected within DNA sequences immediately adjacent to the repeat tract (contained within the pgdna16A1-5 – pcdna38-1 PCR product, Fig. 6.3), where the putative exon coding for a 143 amino acid sequence is located (section 6.4 10). However, this could have been due to the obscuring effect of a high level of lane background on the zooblots (data not shown), probably created by the elevated GC content of the probe.

The zooblot assay method for the identification of possible gene coding sequences within genomic DNA probes stems from observations that protein-coding sequences appear subject to considerable selection pressure toward the conservation of biologically

important sequences. Non-coding sequences, however, accumulate mutations comparatively rapidly and are not well conserved between species (Strachan and Read, 1996). Therefore gene sequences are expected to have a higher level of cross species homology.

6.5.2 Significance of the hybridisation of probes representing regions in the vicinity of *FRA16A* to differently sized transcripts

The results obtained from the hybridisation of two cDNA sequence probes (representing two stretches of sequence within the contig) to a fetal tissue Northern blot were inconsistent with the premise that all members of the contig were derived from one gene with a single transcriptional start site. If all cDNAs in the contig were derived from one transcript, both probes should detect the same signal (one or more bands), unless alternate splicing has occurred. Alternative explanations for the results are: (1) the existence of an overlapping gene in the same or opposite orientation; or (2) multiple transcriptional start sites for a gene with the 5' end at *FRA16A*.

If it is assumed that the isolated clones are true cDNAs (and not contaminating genomic sequences), the relatively large number of cDNAs isolated indicates that the *FRA16A* transcript they represent is relatively highly expressed, at least in fetal brain, but perhaps inefficiently spliced. This suggestion corresponds with the findings from the Northern blot hybridisations, which were indicative of a high expression level. cDNAs representing the spliced transcript may therefore not be well represented within cDNA libraries, at least in comparison with those from the unspliced RNA transcript.

The absence of any apparent divergence of the cDNA sequences from the genomic sequence (i.e. there are no identifiable exons and introns), in addition to the inconclusive Northern blot results, did not serve to clarify which regions (if any) of the cDNA contig/genomic sequence are transcribed and/or translated. However, considering the large number of cDNA clones isolated, it seems unlikely that they are all of genomic DNA, not mRNA, origin although this possibility cannot be completely discounted.

The human fetal brain library from which the *FRA16A* associated clones were isolated has been PCR screened a number of times by different researchers for cDNAs representing a diverse range of genes. The results of these screenings have provided no evidence to indicate that the library is extensively contaminated with genomic DNA clones, as positive clones were not always identified with genomic DNA probes derived from exon trapping experiments (S. Whitmore, K. Lower, personal communication). One library screen, however, identified clones from a known *FMR2* intronic region, indicating either genomic contamination or the inclusion of unspliced *FMR2* transcripts within the library (J. Gecz, personal communication).

6.5.3 Occurrence of CCG repeat tracts in the 5' untranslated region of human genes

DNA sequences consisting of repeating CCG motifs have been identified within a number of human genes, frequently in 5' untranslated regions. However, an assessment of the overall frequency of occurrence of these repeat tracts in genes, and therefore the importance of their role or function has been problematic for two reasons. These are: (1) direct screening of cDNA libraries with a CCG repeat probe will detect contaminating 28S rRNA clones, of which there are often high numbers in cDNA libraries (such a high background can make the isolation of other CCG repeat tract containing clones a time consuming process, K. Harris, personal communication); (2) in general, it appears that G and C rich sequences are cloned at a lower efficiency and are more subject to deletions than most other sequences (with the exception of many tandem repetitive sequences), possibly due to secondary structure differences. Many GC rich sequences, in which CCG rich repeat tracts are often embedded may therefore not be well represented within cDNA libraries (unpublished observations).

Riggins *et al.* (1992) identified ten genes containing CCG repeat tracts by means of direct cDNA library screening methods. None of these genes co-localised with a known folate sensitive fragile site. However, as with the fragile site genes, the CCG repeat tracts were localised, or predicted to be in, the 5' untranslated region in 6 of these genes.

In addition to *FMRI*, the gene associated with the *FRAXA* fragile site and fragile X syndrome (discussed in chapter 1), two other genes associated with folate sensitive fragile sites have been identified to date. Contrasting approaches were used for the identification and/or isolation of these genes. Fragile site *FRA11B* associated gene *CBL2* was characterised before the fragile site itself, and was identified as a fragile site associated gene on the basis of a known CCG repeat tract within the gene, and its proximity to the fragile site (Jones *et al.*, 1995). The *FRAXE* fragile site was the feature initially characterised (Knight *et al.*, 1993), but the corresponding gene, *FMR2*, was not isolated until some time later (Chakrabarti *et al.*, 1996; Gecz *et al.*, 1996; Gu *et al.*, 1996).

Interestingly, *FMR2* cDNAs were initially identified, not from the presence of conserved DNA sequences within a few kilobases of the fragile site, but from conserved sequences located hundreds of kilobases away, within a submicroscopic DNA deletion present in a developmentally delayed boy (Gedeon *et al.*, 1995; Gecz *et al.*, 1996). After sequentially screening libraries with DNA sequence probes, the isolated *FMR2* cDNAs were arrayed to form a 6.2 kb contig, representing approximately 65% of the 9.5 kb transcript detected on a Northern blot (for placental tissue) (Gecz *et al.*, 1996). Within the cDNA contig sequence a 3903 bp ORF was identified, starting at a Kozak consensus sequence with an AUG codon located 394 bp from the *FRAXE* CCG repeat tract.

Reverse transcription-PCR analysis found that expansion of the *FRAXE* CCG repeat in *FRAXE* males and methylation of the CpG island of which it forms part, is correlated with loss of expression of the transcript represented by the cDNA contig. It was thus demonstrated that the gene (*FMR2*) encoding the transcript is associated with the CpG island adjacent to *FRAXE* and thus contributes to *FRAXE* associated mild mental retardation.

In a parallel project, Gu *et al.* (1996) used the exon prediction program XGRAIL to analyse the sequence of a 1.6 Mb region of DNA between the *FRAXA* and *FRAXE* fragile sites. This sequence had been previously determined by large-scale genomic sequencing at Baylor College of Medicine. PCR product probes based predicted exon sequences were then used for cDNA library screening, resulting in the identification of *FMR2* cDNA

clones. This approach was of use when the other methods (i.e. exon trapping, cDNA library screening) for identifying transcribed sequences provided no evidence for a gene near *FRAXE*. However, it is only applicable to genes within intensively sequenced DNA regions.

Another research group succeeded in identifying *FMR2* cDNAs by a more conventional approach of hybridising cDNA libraries with a DNA probe containing evolutionarily conserved sequences from a region of DNA immediately adjacent to the *FRAXE* CCG repeat tract (Chakrabarti *et al.*, 1996). This successful approach was identical to that used for the isolation of several cDNAs homologous to sequences near *FRA16A*. However, in the case of *FRA16A*, the conserved sequence cDNAs did not appear to contain exonic sequences (see section 6.5.1).

Coding sequences and cDNA clones of the *FMR1* gene were also initially identified by the use of conserved sequence probes to screen cDNA libraries (Verkerk *et al.*, 1991), further confirming the validity of this approach.

6.5.4 Significance of the *FRA16A* CpG island

A CpG island is defined as a short region of DNA, 1 - 2 kb in size, that is 60 - 70% CG-rich. These regions have been identified at the 5' ends of all housekeeping genes, as well as a large proportion of genes displaying tissue specific expression. The first exon of the associated gene is generally located within the CpG island (Cross and Bird, 1995). These general observations are consistent with the computer-generated predictions of a short open reading frame and exon and 5' to 3' orientation within the *FRA16A* CpG island.

It has been observed that relatively long, GC rich 5' untranslated regions, such as those of *FMR1* and *fmr1*, are associated with the occurrence of post-transcriptional control mechanisms (Kozak, 1987; 1988; 1991), indicating that the repeat tract may have a functional role in this process. Evidence for another, or associated function for CCG

repeat tract sequences came from protein binding assays that showed CCG repeat oligonucleotides to bind specific nuclear proteins *in vitro* (Richards *et al.*, 1993).

6.5.5 Functions of the genes associated with characterised rare folate sensitive fragile sites

Based on the assignment of the *FMRI* repeat tract as a non-coding sequence, Ashley *et al.* (1993a) postulated that this region has a regulatory function, either as a DNA binding site, or a site of interaction for mRNA binding proteins involved with post-transcriptional regulation. Any such function identified for the *FMRI* repeat tract may also apply to other folate sensitive fragile site CCG repeat tracts, or indeed any gene with a 5' untranslated region CCG repeat tract sequence.

There is no apparent link or similarity found in the known function and/or properties of the proteins encoded by the genes associated with the *FRAXA*, *FRA11B* and *FRAXE* fragile sites (*FMRI*, *CBL2*, *FMR2*). As a phenotype has not been associated with *FRAXF* expression in males, it is possible that if a gene is associated with this fragile site its function is redundant or non-essential. Such a conclusion could not be drawn for the putative *FRA16A* gene, as an individual homozygous for the fragile site has not been found.

Sequence analysis of *FMRI* RT-PCR products has revealed that alternate splicing of the transcript occurs, indicating that the *FMRI* gene codes for a heterogeneous set of proteins (FMRP) which are widely expressed in human and mouse tissues. Whether the alternate splicing affects the function of the protein is currently unknown. However, experimental evidence has indicated that alternate splicing of exon 14 influences sub-cellular localisation of FMRP to the nucleus or cytoplasm (Sittler *et al.*, 1996). Similarly, a number of mRNA isoforms of *FMR2* have been characterised which appear to be the products of alternate splicing events, but as with *FMRI*, the functional significance of this is not evident (J. Gecz, personal communication). Such alternate splicing events have been identified in a diverse range of transcripts, and in a number of instances found to be clearly functionally significant (O'Shaughnessy *et al.*, 1996; Roy *et al.*, 1996; Lorenzo *et al.*, 1997).

Comparative analyses of the FMRP sequence with protein sequence motifs of known function identified two conserved RNA binding domains (Ashley *et al.*, 1993b; Siomi *et al.*, 1993). *In vitro* protein binding studies found that FMRP was able to bind ~4% of human fetal brain mRNAs, including its own message (Ashley *et al.*, 1993b). The authors suggested that the absence of normal interaction between FMRP and a particular subset of RNA molecules might cause the pleiotropic phenotype associated with the fragile X syndrome. Subsequent binding studies found that FMRP is also associated with the 60S subunit of cytoplasmic ribosomes, indicating that the fragile X phenotype may be due to a defect in the translational machinery caused by the absence of FMRP (Khandjian *et al.*, 1996).

The precise cellular function of the *FMR2* protein product, as with FMRP, is not yet defined. However, an indication of the function was provided by sequence database searches with *FMR2* sequences. These searches revealed significant similarity to an amino acid sequence encoded by the proto-oncogene AF-4, the gene product of which is a putative transcription factor. This indicates that the *FMR2* protein may have a similar role (Chakrabarti *et al.*, 1995; Gecz *et al.*, 1996; Gu *et al.*, 1996).

The gene found to be associated with the *FRA11B* fragile site was the *CBL2* proto-oncogene, which had been previously characterised. This formed the basis of the molecular characterisation of *FRA11B*. Although the *CBL2* gene is the human homologue of a murine oncogene, the cellular function of the protein product is currently unknown (Jones *et al.*, 1995).

The three genes (*FMR1*, *FMR2*, *CBL2*) known to be associated with rare folate sensitive fragile sites have no apparent sequence homology at either the DNA or protein levels, with the exception of the unstable CCG repeat tracts in their 5' untranslated regions. The polypeptide sequence homology of the *FMR2* protein to a proto-oncogene (which is also the classification of *CBL2*) implies a functional similarity concerning DNA binding and/or repair, but whether this is related to the presence of the untranslated repeat tracts, has not been determined. However, the RNA/ribosome binding function of the FMRP does not appear to be related to the proposed transcription factor function of the *FMR2*

protein, which would involve interaction with genomic DNA (unless it is part of a protein complex). Although expression of both the *FRAXA* and *FRAXE* fragile sites has been associated with mental retardation of varying degrees of severity, a functional similarity cannot be assumed on this basis, as it is a feature of many pleiotropic phenotypes resulting from inheritance of a genetic mutation.

The identification of similarities in either sequence or protein product function between fragile site genes (or genes with CCG repeat tracts in their 5' untranslated regions) would be most interesting as it may elucidate the function of the common, potentially unstable CCG repeat tract. As a CCG repeat tract may have a biologically deleterious effect if it becomes expanded and causes CpG island hypermethylation, (causing a reduction in viability or reproductive capability), it would be expected that these relatively unstable sequences would be progressively lost from the population through mutation to a more stable non-repetitive sequence. In conclusion, when taken together with the findings that the *FMRI* CCG repeat tract is retained in many other mammalian species (although generally in a shorter form) (Deelen *et al.*, 1994; Eichler *et al.*, 1995b), such strong sequence conservation seems to indicate an important function for this sequence at either the DNA or RNA level.

6.5.6 Features of the putative *FRA16A* associated gene

A striking feature revealed by the sequence characterisation of the *FRA16A* associated cDNA contig was that, where direct comparisons were made, no divergence from the genomic sequence of the region was evident. Sequence divergences between genomic and cDNA sequences are consistent with the occurrence of heterogeneous nuclear RNA (hnRNA), splicing events and hnRNAs are generally short-lived and unstable molecules (Alberts *et al.*, 1983). As most mammalian genes contain introns, the vast majority of cDNAs that are representative of true mature mRNAs show evidence of splicing. In the absence of other data, the lack of sequence divergence found could therefore be taken as an indication that the *FRA16A* associated clones are the result of extensive genomic DNA

contamination of the fetal brain cDNA library (and, to a lesser extent, the two other libraries).

Two alternative explanations that may account for the *FRA16A* results are: (1) the transcript is very inefficiently spliced, so that the majority of *FRA16A* associated cDNA clones in the fetal brain library were reverse transcribed from the unprocessed hnRNA transcript. The detection of a transcript on a Northern blot of 1.35 kb, a size smaller than that of the contig (5.341 kb) is supportive of this view. Alternatively, the mature mRNA may be unstable, or after reverse transcription, the cDNA molecules (from spliced RNA) are unclonable; (2) the *FRA16A* gene is one of a minority of genes that are intronless. A distinct band indicating expression of a potentially intronless ~6 kb transcript was detected by a cDNA 72 probe in fetal liver and kidney tissue mRNA, although this result is difficult to interpret as the other probes did not detect this transcript. The transcript could also be highly unstable, such that only a smear is detectable in the fetal brain mRNA track (Fig. 6.3). Alternatively, the mRNA from which the cDNAs were derived was of too great a size for resolution on the agarose gel run prior to Northern blotting.

6.5.7 The putative *FRA16A* associated gene is a candidate disease gene for PXE

Pseudoxanthoma elasticum (PXE) is a rare (1 in 70000 - 100000) inherited disorder of connective tissue that is characterised by the progressive calcification of elastic fibres. The skin, eye and cardiovascular system are typically involved, resulting in skin lesions, deteriorated vision, and vascular disease.

Both autosomal recessive and, less commonly, autosomal dominant, patterns of inheritance have been described for PXE (Viljoen *et al.*, 1998). In order to map the disease locus, a genome wide screen with microsatellite markers was conducted on affected sib pairs. The results were then analysed with computer programs designed to detect allele sharing. Conventional linkage analysis confirmed the resulting localisation (Struk *et al.*, 1997). Van Soest *et al.*, (1997) performed a similar PXE mapping study that also mapped the gene to the PXE region.

Linkage/sib pair analysis localised the PXE locus to the 16p13.1 region between the genetic markers D16S405 (AFM070yal) and D16S499 (AFM259xb9), an area without any apparent candidate gene (Struk *et al.*, 1997). The position of these markers relative to somatic cell hybrid breakpoints had been previously determined by LANL, and indicated that *FRA16A* lies within the interval to which PXE was mapped.

The entire 16p13.1 region has been saturated with ordered megaYAC contigs (Struk *et al.*, 1997). Referral to the Whitehead Institute megaYAC contig containing My769H1 (from which *FRA16A* was cloned) identified a 570 kb YAC (My647G7) that appeared to span the distance between the closest flanking markers to the PXE locus (D16S405 and D16S499).

Initial uncloned genomic DNA pulsed field mapping data, that was obtained with the DNA probe 1.79 and is presented in Chapter 3, indicated that the PXE interval has a maximum size of 2 megabases, and few CpG islands (suggesting a low gene density). Subsequent assembly of the data obtained from other genomic DNA mapping methods produced a more detailed map spanning ~ 500 kb and depicting the probable relative locations of YACs, somatic cell hybrid breakpoints, and genetic markers within this interval. This map is shown in Fig. 6.9.

Any genes of unknown function (or with features of connective tissue proteins) mapped within the region are therefore strong candidates for the PXE disease gene.

Based on the information discussed above, the putative *FRA16A* gene appears to be a candidate for the PXE locus. In addition, searches with the amino acid sequence (predicted by all exon detection algorithms utilised) against protein motif databases found a cell attachment sequence (likelihood of random occurrence ~1 in 10). This cell attachment sequence is a tripeptide found in adhesive proteins such as collagens, fibrinogen, fibronectin, vitronectin, osteopontin. As adhesive proteins are a component of elastic fibres, there is a possible functional relationship between the *FRA16A* gene and PXE. An additional indication that the putative *FRA16A* gene is a reasonable candidate for the PXE disease gene is the prediction of extracellular localisation by PSORT. This

prediction corresponds with the localisation of other proteins (i.e. collagens, fibrillins) involved in connective tissue disorders.

Full characterisation of the gene, followed by mutation analysis in PXE families may either prove or disprove a relationship between the putative *FRA16A* gene and the PXE disease gene. Alternatively, the identification of PXE associated mutations in another gene within the region would also disprove any relationship.

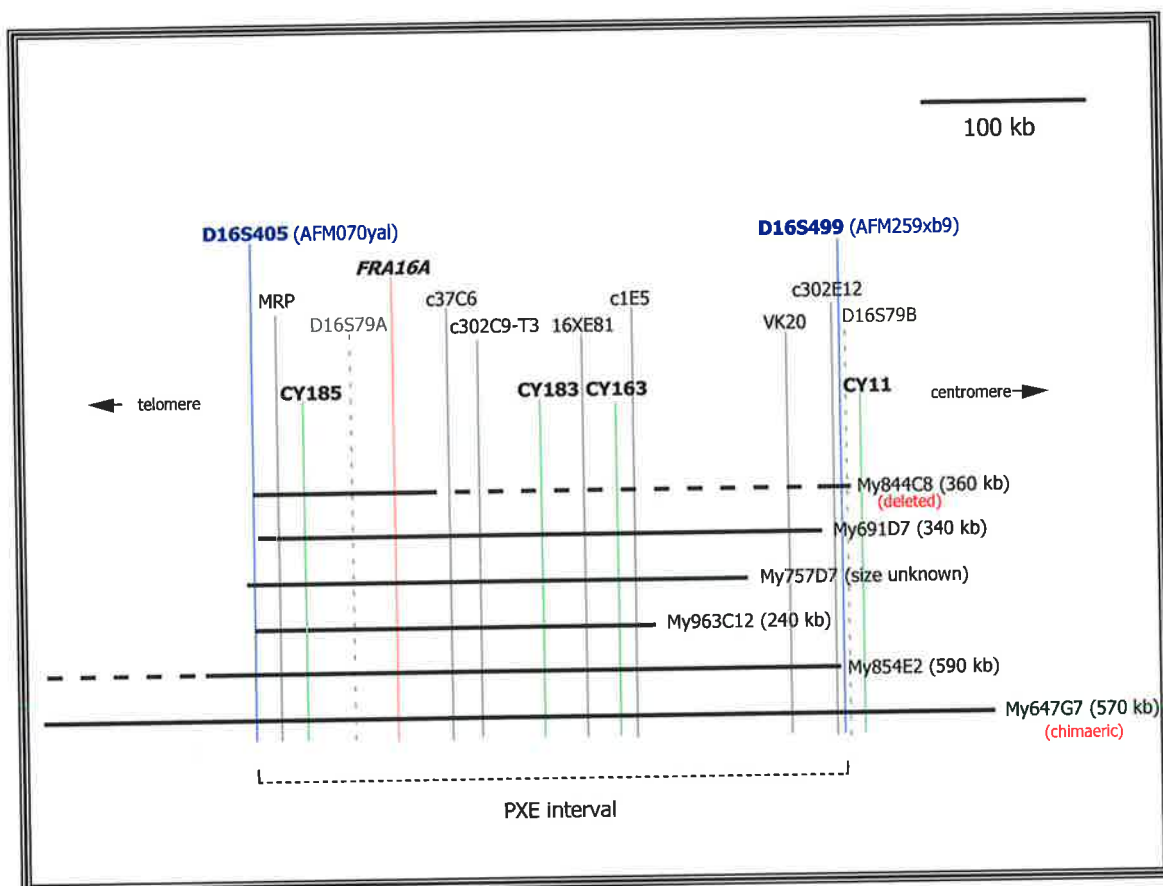


Figure 6.9 Map of MegaYACs and DNA markers in the Pseudoxanthoma Elasticum (PXE) locus region

Chapter 7

*Construction of a
MegaYAC Contig
across the FRA16D
Fragile Site*

Chapter 7: Contents

Construction of a MegaYAC Contig Across the FRA16D Fragile Site

7.1	Summary	174
7.2	Introduction	175
7.3	Materials and Methods	177
7.3.1	Physical somatic cell hybrid breakpoint map of the <i>FRA16D</i> region.....	177
7.3.2	FISH mapping of DNA markers and megaYACs with respect to <i>FRA16D</i>	177
7.3.3	Database searches for megaYAC contigs spanning the <i>FRA16D</i> region.....	177
7.3.4	Preparation, restriction digestion and Southern blot hybridisation of high molecular weight megaYAC DNAs	178
7.4	Results	178
7.4.1	Assignment of human chromosome 16 DNA cosmid and plasmid clones in the 16q23 region as proximal or distal to <i>FRA16D</i>	178
7.4.2	Identification of YAC clones in <i>FRA16D</i> interval	178
7.5	Discussion	180
7.5.1	Cytogenetic analysis of the common fragile site <i>FRA3B/3p14.2</i> region.....	180
7.5.2	Mapping of the <i>FRA3B</i> region	180
7.5.3	Characterisation of 3p14.2 deleted regions in both tumour and normal tissues.....	181
7.5.4	Further evidence of a role for <i>FHIT</i> in tumour suppression.....	182
7.5.5	Features of <i>FRA3B</i> region DNA sequence.....	182
7.5.6	Evidence that fragile sites are regions of instability liable to chromosomal breakage by exposure to DNA damaging substances	183
7.5.7	Characterisation of other fragile site classes	183
7.5.8	Comparisons between the <i>FRA16D</i> and <i>FRA3B</i> regions	184

7.1 Summary

FRA16D is a common, aphidicolin inducible fragile site located within band 16q23.1 on human chromosome 16. The DNA region containing *FRA3B*, a fragile site of the same class, has been characterised. However, unusual sequence features, such as the CG or AT rich tandem repeats at the *FRAXA* or *FRA16B* loci, were not identified. Therefore, in order to determine whether specific sequence patterns or features are present in all common fragile sites, the molecular characterisation of *FRA16D* has been attempted.

FRA16D had been previously localised to within a hybrid breakpoint interval. DNA markers mapping to the same interval were used for Whitehead Institute of Technology and Los Alamos National Laboratory database searches in order to identify YACs and YAC contigs within the region of interest. FISH analysis was then used to localise each clone with respect to *FRA16D*. Using this approach, a megaYAC contig was identified that appears to span the fragile site.

7.2 Introduction

Chromosomal fragile sites classified as common are present in all individuals and it has therefore been proposed that they are a macromolecular manifestation of a conserved feature (e.g. chromatin structure) or event (e.g. gene expression; DNA replication/repair) (Wilke *et al.*, 1996).

It has been proposed that all classes of fragile sites represent cytogenetic markers for genomic regions that are replicated late in the DNA synthesis phase of the cell cycle (Laird *et al.*, 1987). Such late replication has been observed to occur in association with the expression of *FRAXA* (Hansen *et al.*, 1993). Wilke *et al.* (1996) therefore considered that a greater understanding of the molecular basis of fragile sites on all levels may lead to advances in the understanding of DNA replication and repair processes, and the relationship of these mechanisms to chromatin structure.

The subject of the study presented in this chapter is the common, aphidicolin inducible fragile site *FRA16D*, which is located within dark band 16q23.1 on human chromosome 16. *FRA16D* has been identified as the second most highly expressed common fragile site in human chromosomes, after the common, aphidicolin inducible fragile site *FRA3B* (Tedeschi *et al.*, 1992). Aphidicolin, an inducer of common fragile site expression, is an inhibitor of both DNA polymerases α and δ (Glover *et al.*, 1984), and it has therefore been proposed that this chemical agent may induce fragile site expression by causing further delays in DNA replication within regions that are already late or slow replicating.

An interesting feature of common fragile sites is that they are expressed under conditions of folate stress (Glover *et al.*, 1984). This similarity between folate sensitive and common fragile sites implies that their DNA sequences and/or mechanisms of fragility may have common characteristics.

The most prominent reason for attempting this study was to gain further insight into the

mechanism(s) responsible for chromosomal fragility. In view of this aim, the characterisation of *FRA16D* may assist the understanding of results obtained from studies of *FRA3B*, which indicated that the fragile site spans a large DNA region prone to rearrangements but with no discernible distinctive sequence features (Boldog *et al.*, 1997). If related DNA sequences are a feature of common fragile sites, comparison of DNA sequences across the two most highly expressed fragile sites (*FRA16D* and *FRA3B*) would be the most likely to provide some insight into their mechanism of induction and/or molecular basis for fragility.

Wilke *et al.* (1996) suggested that a conservation of function was implied by the constancy of common fragile sites, and that such constancy could be due to an invariant nucleotide and/or chromatin structure. This situation would be in contrast with rare fragile site findings, in which the molecular instability observed at the nucleotide level appears to be translated to a chromatin structural change observable at the level of the chromosome (Wilke *et al.*, 1996).

Comparative analysis of rare fragile sites has been instrumental in gaining some understanding of the common physical features and mechanisms underlying their genesis. Such an analysis of common fragile sites is therefore intended to identify common physical features that might constitute essential elements for the expression and genesis of all classes of fragile sites.

7.3 Materials and Methods

7.3.1 Physical somatic cell hybrid breakpoint map of the *FRA16D* region

Details of the human chromosome 16 somatic cell hybrid map, illustrating the portion of this chromosome present in each hybrid cell line with a breakpoint in the vicinity of *FRA16D*, are shown in Fig. 7.1.

The position of *FRA16D* within a map of the 16q23.1 region somatic cell hybrid breakpoints is shown in Fig. 7.2.

S. Whitmore provided an updated version of the somatic cell hybrid breakpoint map reported by Callen *et al.* (1992).

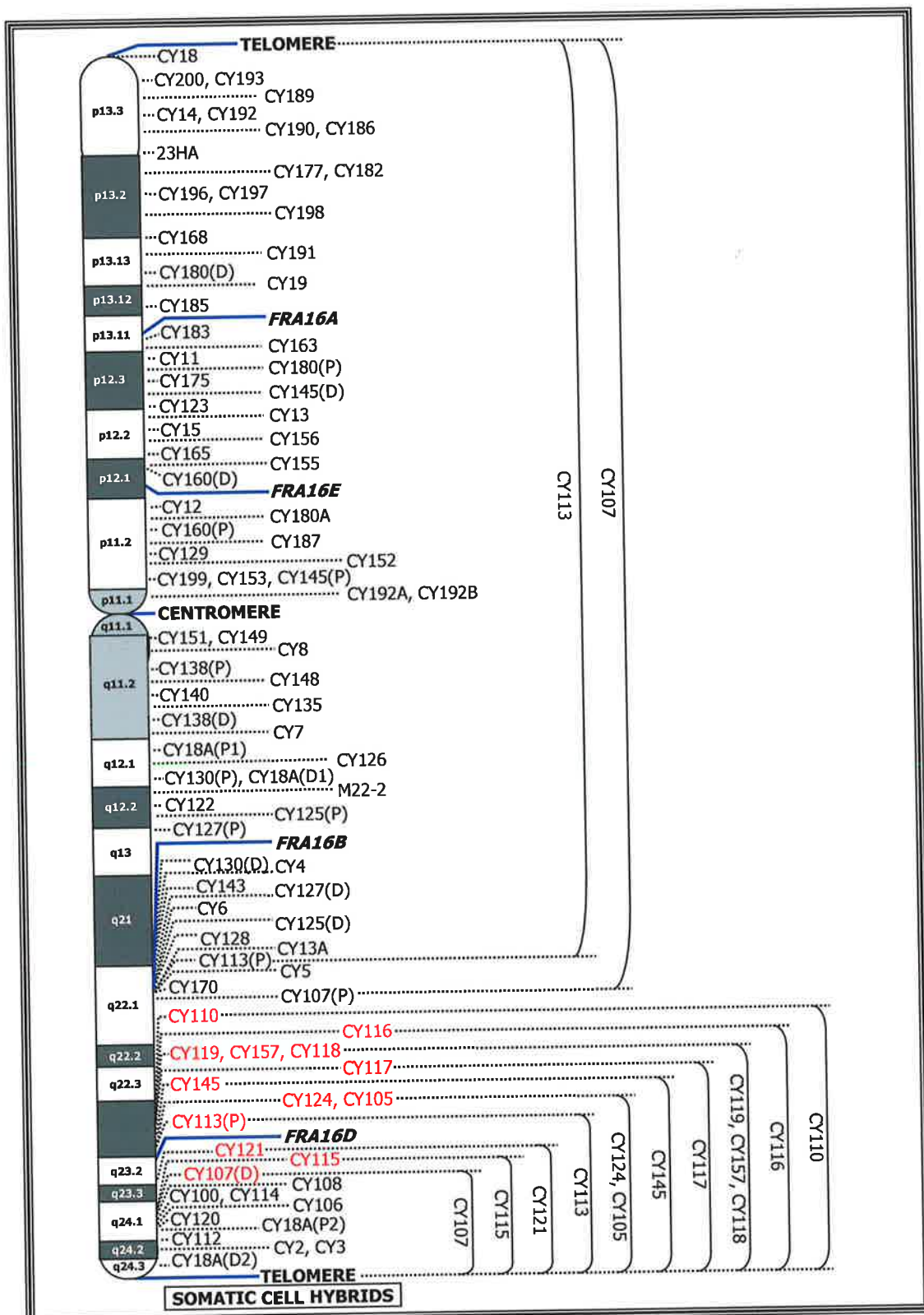
7.3.2 FISH mapping of DNA markers and megaYACs with respect to *FRA16D* (performed by J. Nicholl and E. Woollatt)

In order to refine the localisation of *FRA16D*, cosmid and plasmid clones that had been previously mapped within somatic cell hybrid intervals in the 16q23.1 region were used as FISH probes against chromosomes expressing *FRA16D*. Details of the probes (16/08, CRI-0119, c307A12 and c306D2) used for the localisation of *FRA16D* are given in Table 7.1. MegaYACS mapped to the *FRA16D* region were also subjected to FISH analysis

7.3.3 Database searches for megaYAC contigs spanning the *FRA16D* region

Two of the DNA markers (c307A12 and c306D2) within the *FRA16D* interval had been previously mapped, and their localisations were shown on an existing megaYAC contig map of the region constructed at Los Alamos National Laboratories (Doggett *et al.*, 1995). The identification of a number of megaYACs that had been localised to within the region of interest enabled further searches of the Whitehead Institute of Technology Human Genome Mapping database (<http://-genome.wi.mit.edu/>) to identify additional megaYACs and markers overlapping these clones. The megaYACs used for the localisation of *FRA16D* are listed in Table 7.2.

Figure 7.1 Map of somatic hybrid cell line breakpoints on human chromosome 16. The cytogenetic 'anchor points' of the map are indicated by blue lines. The broken lines indicate the position of hybrid breakpoints relative to each other. The cytogenetic localisations of these breakpoints shown in this figure are only approximations, as they cannot be defined accurately in these cell lines. The cell lines used for mapping *FRA16D* are shown in red text. Large brackets define the portions of human chromosome 16 within the somatic cell hybrids CY110, CY116, CY119, CY118, CY157, CY117, CY124, CY105, CY113(P), CY121, CY115 and CY107(D). Details of the construction and constitution of the somatic cell hybrids were published in Callen *et al.* (1986; 1989; 1990; 1992).



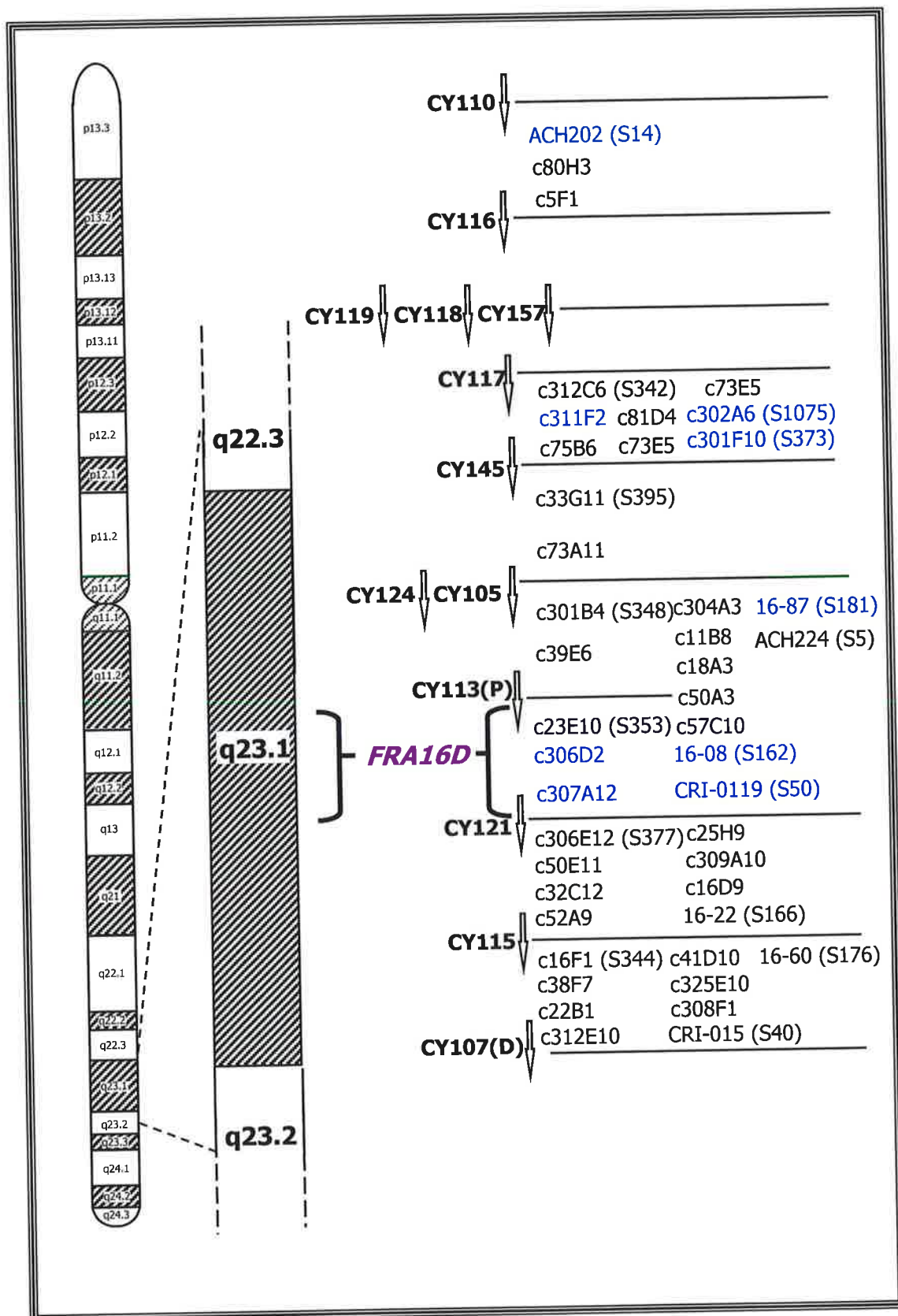


Figure 7.2 Schematic diagram depicting the position of *FRA16D* on human chromosome 16 and within the somatic cell hybrid breakpoint map of the chromosome. Markers used for *FRA16D* mapping are shown in blue.

Table 7.1 DNA markers used for the localisation of *FRA16D*

probe name/locus	human DNA insert size/insert site	vector/vector type	reference
16/08/D16S162	1 kb/ <i>Hind</i> III	Bluescribe/ phagemid	Harris <i>et al.</i> , 1989
CRI-0119/D16S50	40 kb/ <i>Mbo</i> I	c2RB/cosmid	Donis-Keller <i>et al.</i> , 1987
c307A12/D16S1218	36.954 kb/ <i>Eco</i> RI	sCos-1/cosmid	Stallings <i>et al.</i> , 1991; 1992
c306D2/D16S1203	33.691 kb/ <i>Eco</i> RI	sCos-1/cosmid	Stallings <i>et al.</i> , 1991; 1992

Table 7.2 CEPH YACs used for *FRA16D* study. YAC clones were derived from a total human megaYAC library constructed by Dr. D. Le Paslier and colleagues at the Centre d'Etude du Polymorphisme Humain (CEPH), Paris, France. The method for YAC library construction was reported by Burke *et al.*, (1987) and Abidi *et al.*, (1990). MegaYAC information was obtained from the Whitehead Institute of Biomedical Research website (<http://www-genome.wi.mit.edu/>).

YAC	human DNA insert size (kb)
My801B6	1440
My845D9	1570
My891F3	1190
My903D9	1690
My912D2	1210
My933H2	1530
My944D8	1330
My972D3	1690

7.3.4 Preparation, restriction digestion and Southern blot hybridisation of high molecular weight megaYAC DNAs

High molecular weight DNA was prepared as described in section 2.3.5.2. After purification, restriction digestion and Southern blot hybridisation of the agarose-immobilised DNA was performed as described in section 2.3.5.

7.4 Results

7.4.1 Assignment of human chromosome 16 DNA cosmid and plasmid clones in the 16q23 region as proximal or distal to *FRA16D* (performed by J. Nicholl and E. Woollatt)

FISH analysis with nine 16q23 region cosmid and plasmid clones localised the *FRA16D* locus to an interval of unknown size between the somatic cell hybrid breakpoints of CY113(P) and CY121 (Fig. 7.2). The results of the FISH analyses are summarised in Fig. 7.3, which depicts the location of *FRA16D* with respect to the FISH probes and the somatic cell hybrid breakpoints in the vicinity of the fragile site. Four DNA markers within the CY113(P)⇒CY121 interval, c306D2, c307A12, 16-08 and CRI-0119, were used as FISH probes against chromosomes expressing *FRA16D*. The results of this analysis mapped c306D2 proximal to *FRA16D*, and the other three markers (16-08, CRI-0119, c307A12) distal to *FRA16D*, thereby clearly localising the fragile site to the interval (CY113(P)⇒CY121) in which these probes had previously been mapped.

7.4.2 Identification of YAC clones in *FRA16D* interval

In view of the fact that relatively detailed YAC contig mapping information was available from the Whitehead Institute and LANL, these sources were utilised to identify megaYACs in the region in preference to PCR screening of the CEPH megaYAC library with markers from the CY113(P)⇒CY121 interval. The four markers (in the *FRA16D* region) previously mapped to YACs were not included in the STS panel used for the Whitehead Institute YAC contig assembly. Therefore this data source was not of

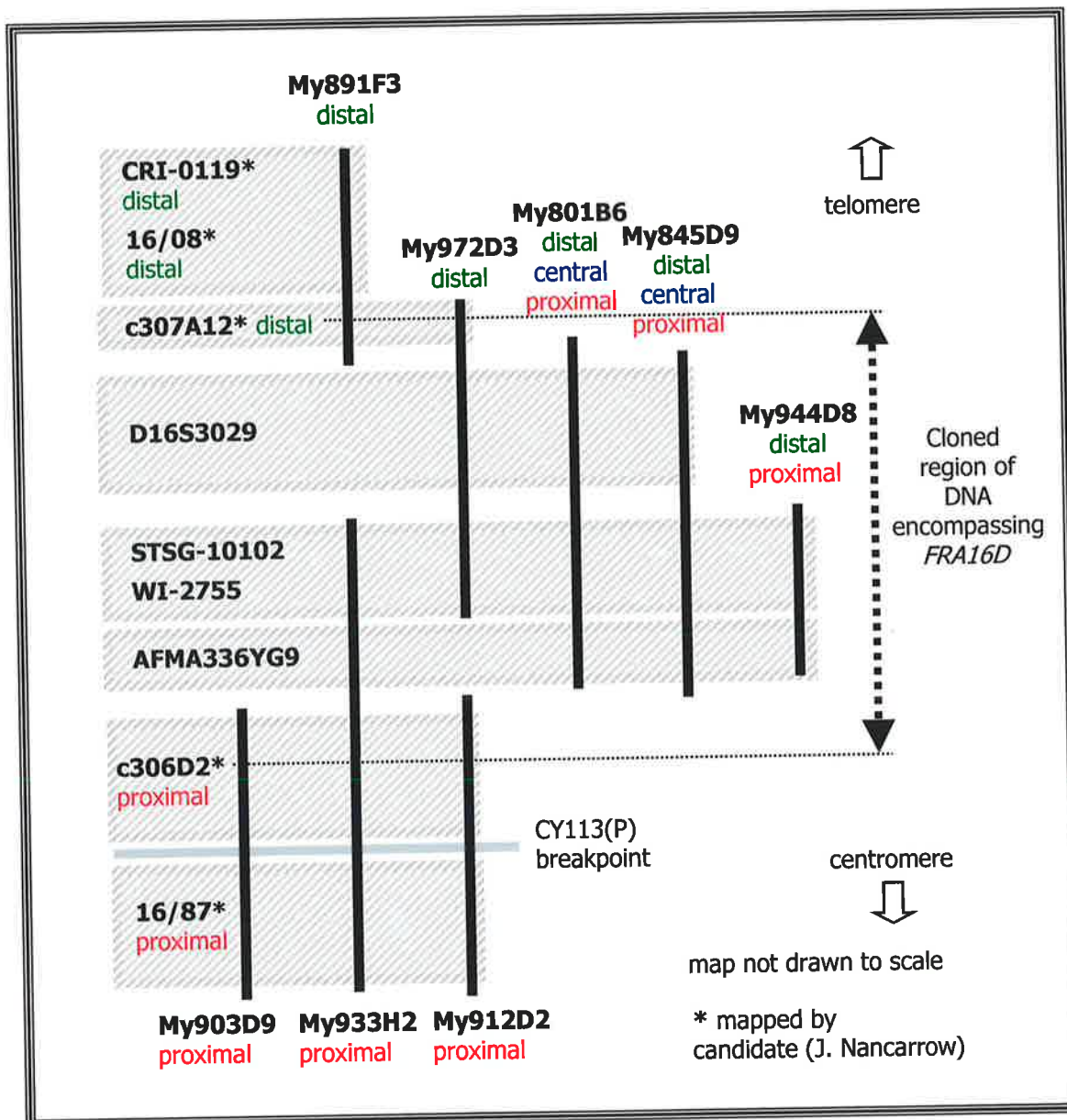


Figure 7.3 MegaYAC contig spanning the *FRA16D* region. FISH localisations were performed by J. Nicholl and E. Woollat.

immediate use. Information obtained from the LANL database, however, included megaYACs containing c306D2 (proximal) and c307A12 (distal) DNA sequences. An STS sequence from c306D2 was found to be contained within megaYACs: My903D9; My905G3; My912D2; My950B3. The c307A12 STS was found to be contained within a completely different set of YACs. These YACs were: My701H1; My798A11; My798A3; My891F3; and My972D3.

In order to ensure the fidelity of the LANL database YAC hits, the YACs My903D9, My912D2, My701H1, My798A11, My798A3, My891F3 and My972D3 were obtained from CEPH. High molecular weight DNA was prepared from each YAC and subjected to *Pst* I digestion. The digests were Southern blotted, and probed with 16-08, 16-87, CRI-0119, c306D2 and c307A12 in succession. Examples of the autoradiographs produced from these Southern blot hybridisations are shown in Fig. 7.4. The proximal probes c306D2 and 16-87 were found to hybridise to the YACs My903D9 and My912D2 DNAs, while the distal probes CRI-0119 and 16-08 hybridised to My891F3 DNAs. Cosmid c307A12 hybridised to both My891F3 and My972D3. Confirmation of these results was provided by FISH analysis, with the YAC DNAs as probes. The results of this assay are summarised in Fig. 7.3.

As LANL database search and the direct hybridisations had identified no YACs containing both proximal c306D2 and distal 307A12 STS hits (thus potentially spanning *FRA16D*) a further database search was conducted, and the results integrated with those previously obtained. Additional markers and YACs mapped by WIT were added, forming the basis of a megaYAC contig (shown in Fig.7.3) that appeared to span the *FRA16D* fragile site.

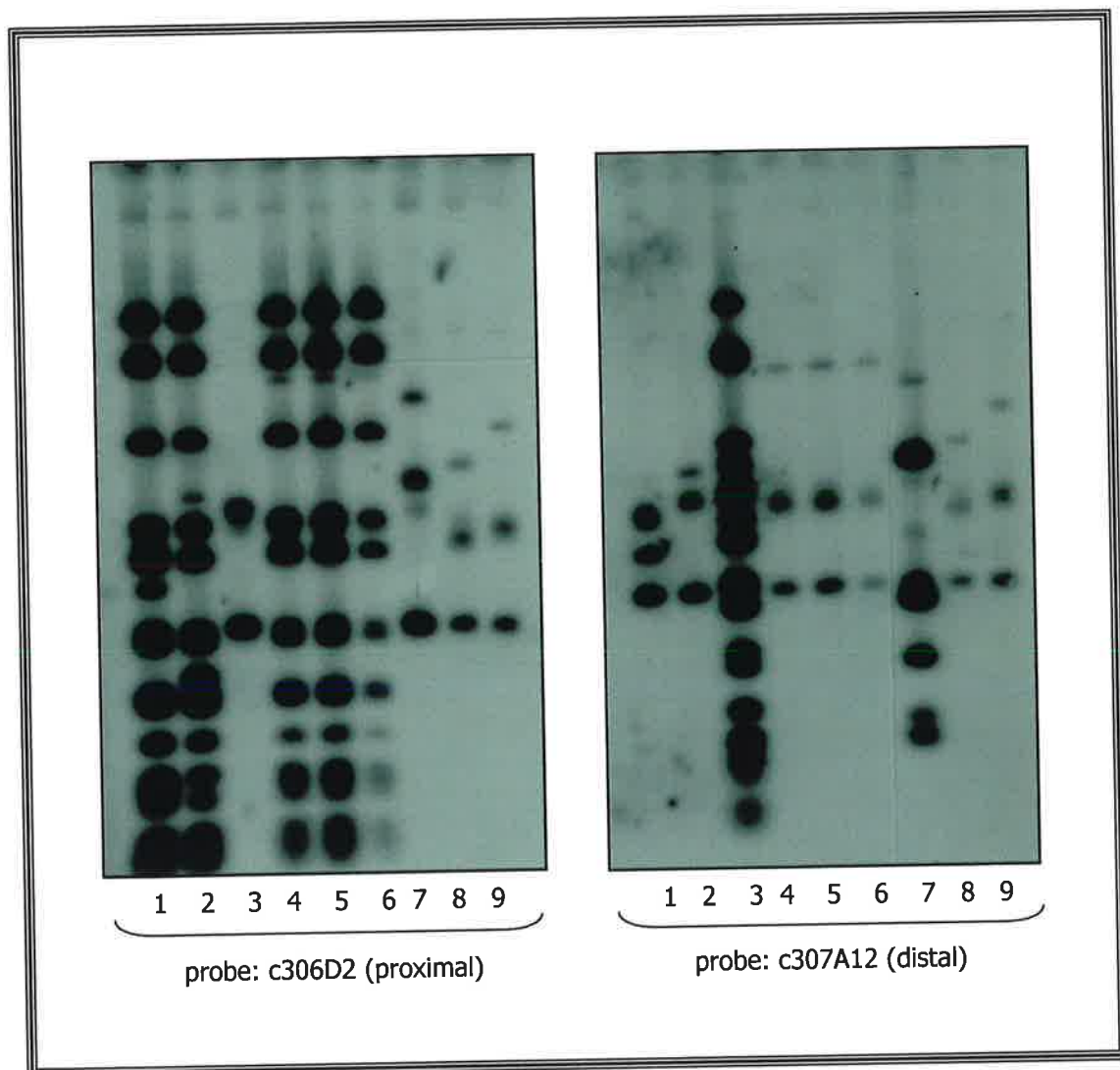


Figure 7.4 Confirmation of the presence of cosmids c306D2 and c307A12 sequences within a set of megaYACs. *Pst* I digested YAC samples were: lane 1. My903D9; lane 2. My933H2; lane 3. My891F3; lanes 4, 5 and 6. My912D2; lane 7. My701H1; lane 8. My798A11; 9. My922G4. My701H1 contains DNA sequences homologous to cosmid c307A12, but the more complex pattern of bands found in lane 3 is absent. This finding suggested three possibilities: 1. The insert of My 701H1 contains a deletion; 2. Cosmid c307A12 overlaps one end of the YAC insert; 3. My701H1 contains DNA sequences that have some homology to those of c307A12, but are derived from a different region of the genome.

7.5 Discussion

7.5.1 Cytogenetic analysis of the common fragile site *FRA3B*/3p14.2 region

Cytogenetic analysis has identified the common fragile site *FRA3B* as the most sensitive site on normal human chromosomes with respect to the formation of gaps or breaks that are produced under conditions of folate or aphidicolin stress (Smeets *et al.*, 1986; Tedeschi *et al.*, 1992). Further interest in the molecular characteristics of *FRA3B* was generated by findings of 3p14.2 region chromosomal deletions and rearrangements in many tumours and tumour cell lines (Smeets *et al.*, 1986; Kok *et al.*, 1987; Kovaks *et al.*, 1988; Daly *et al.*, 1991; Kovaks and Kung, 1991; Yokoyama *et al.*, 1992; Pandis *et al.*, 1993), which indicated a potential involvement in the development of malignancies. The first efforts at cloning and characterising a common, aphidicolin inducible fragile site were therefore directed toward *FRA3B*.

In one approach to cloning *FRA3B*, Wilke *et al.*, (1996) performed FISH analyses with λ subclones of a 1.3 Mb megaYAC spanning the fragile site. When used for FISH analyses against tumour cell line chromosomes, these clones detected gaps and breaks over a region of at least 50 kb. As there were no CCG repeat tracts detectable in the YACs, the authors proposed that a different mechanism of fragility might be responsible for common fragile site expression.

Another interesting finding was that the characterised *FRA3B* region (containing frequent gaps and breaks) spans a site for HPV16 (human papilloma virus 16) integration. HPV16 is associated with the formation of primary cervical carcinomas, and co-localises with the smallest commonly deleted region of 3p in these tumours. The authors suggested that these findings support the concept of a link between viral integration sites and fragile sites.

7.5.2 Mapping of the *FRA3B* region

For the purpose of studying *FRA3B*, several research groups constructed chromosomal breakpoint maps of the *FRA3B* region using rodent-human hybrid cell lines. For these

studies, hybrid cells retaining human chromosome 3 were treated with aphidicolin to induce fragile site expression before subcloning. Subclones containing portions of chromosome 3 with region 3p14.2 breaks were further characterised with respect to the loss or retention of specific 3p markers, and thus the relative positions of the breaks (LaForgia *et al.*, 1991, 1993; Paradee *et al.*, 1995). The 3p14.2, *FRA3B* associated breakpoints were subsequently mapped within the region of DNA that could be spanned by a single megaYAC.

Ohta *et al.* (1996) reported the finding of 200 - 300 kb homozygous deletions (not cytogenetically visible) within the 3p14.2 region in multiple tumour derived cell line DNAs. Utilising the YAC/breakpoint mapping information, a cosmid contig covering the homozygously deleted region was developed and the clones used for exon trapping experiments (Ohta *et al.*, 1996). This approach led to the detection and characterisation of a ubiquitously expressed gene. Sequence analysis revealed the gene to be a member of the histidine triad family, and it was thus titled *FHIT* as a shortened form of Fragile Histidine Triad. *FHIT* was found to have 10 exons distributed over a region of at least 500 kb. In addition, the finding of aberrant transcripts of the gene in ~50% of the oesophageal, stomach, and colon carcinomas analysed indicated a degree of clinical importance (Ohta *et al.* 1996).

A subsequent expression/deletion study of the *FHIT* gene found that two tumour cell lines with entirely intronic deletions expressed transcripts of normal size (Ong *et al.*, 1997). These results suggested that some homozygous deletions within the *FHIT* gene are without phenotypic effect, and are thus a result of the region's genomic instability.

7.5.3 Characterisation of 3p14.2 deleted regions in both tumour and normal tissues

The stability of the 3p14.2 *FRA3B* region, as measured by the detection and extent of deletions in both tumour cell lines and normal tissues, was determined by Boldog *et al.* (1997). Homozygous deletions were detected in cervix, breast, lung, and colorectal carcinoma cell lines. However, deletions (both continuous and discontinuous) were also detected in non-tumour DNAs, a finding suggesting that the *FHIT* gene is not a selective

target for mutation events. In addition, it was found that some reported aberrations of the *FHIT* transcript are attributable to variations of normal splicing.

On this basis, the authors proposed that the selective loss of the *FHIT* tumour suppressor gene does not occur within tumour cells. However putative targets, including the *FRA3B* region, may have deletion events as a consequence of a general genomic instability (but not mismatch repair deficiency).

7.5.4 Further evidence of a role for *FHIT* in tumour suppression

In order to study the functional role of the *FHIT* gene, Siprashvili *et al.*, (1997) transfected wild type and mutant *FHIT* genes into cancer cell lines lacking endogenous *FHIT*. The results indicated that *in vitro* cell growth is not consistently affected by expression of the gene. However, the result of a further experiment that was performed to ascertain the effects of *in vivo* *FHIT* transfection found that the FHIT protein suppressed tumourgenicity in nude mice.

7.5.5 Features of *FRA3B* region DNA sequence

Sequence analysis of a 110 kilobase portion of the *FRA3B* region revealed a high A-T content, as well as a high concentration of LINE and MER repeats (Boldog *et al.*, 1997). There appeared to be a paucity of expressed sequences (i.e. a gene poor region). Other features of the region characterised by the analysis were a long polypurine tract and two polymorphic extended dinucleotide repeats flanking a cluster of aphidicolin induced deletion breakpoints. A cervical carcinoma HPV integration site (associated with an interstitial deletion) and a sequence with similarity to small polydispersed circular (spc) DNA were also identified (Boldog *et al.*, 1997).

Boldog *et al.* (1997) suggested that the spc finding might be of biological significance, as spcDNA numbers are increased by exposure to DNA damaging agents and inhibitors (including aphidicolin). Small polydispersed circular DNAs are derived from chromosomal sequences, and are associated with clustered repeats. Their numbers are elevated in some conditions associated with genomic instability (e.g. Fanconi's anaemia).

The *FRA3B* findings discussed above suggest that a similar lack of distinguishing features will be evident for *FRA16D*. However, a comparison between the two fragile sites may reveal common elements essential for the expression of common fragile sites.

7.5.6 Evidence that fragile sites are regions of instability that are liable to chromosomal breakage by exposure to DNA damaging substances

The precise mechanism(s) of drug selected intrachromosomal amplification by breakage-bridge-fusion (BFB) cycles in mammalian cells has not been determined. However, it has been found that only clastogenic, or DNA breaking, drugs induce drug resistance in this way. Coquelle *et al.* (1997) demonstrated strict correlation of BFB cycles to the induction of fragile site expression, and therefore proposed that fragile sites play a significant role in the process. The authors also observed that the positions of fragile sites relative to the amplicon boundaries found in some human cancers seems to support proposals that fragile sites play a key role in the amplification of at least some oncogenes during tumour progression.

7.5.7 Characterisation of other fragile site classes

To date, members of only three types of fragile site have been characterised at the molecular level. These are: common, aphidicolin inducible; rare, folate sensitive; and rare, distamycin A inducible. The first two classes of fragile site are the central topics in this and preceding chapters, whereas distamycin fragile sites have not been previously discussed.

As with *FRA16A*, the distamycin A sensitive fragile site *FRA16B* was isolated by positional cloning methods. The unstable, expanded DNA sequence was found to consist of the 33 bp AT-rich minisatellite repeat p(ATATATTATATATTATATC TAATAATATAT^C/_ATA)_n (Yu *et al.*, 1997). This finding was thought to be consistent with the DNA sequence binding affinities of the chemical agents known to induce expression of *FRA16B*. The most interesting insight gained, however, was from the finding that repeat sequence copy number can proceed by similar mechanisms, and

appears to be independent of repeat motif length (Yu *et al.*, 1997). No such repeat motif has been found in the *FRA3B* region, suggesting a different mechanism of chromosomal fragility exists for common fragile sites.

7.5.8 Comparisons between the *FRA16D* and *FRA3B* regions

No evidence has been found to indicate that the *FRA16D* region within band 16q23.1 region is a common site of deletions in tumours. This observation cannot, however, exclude the possibility that many deletions in the region may be cytogenetically undetectable (as at 3p14.2). There have been few genes mapped to the region, suggesting that *FRA16D* is located within a gene poor region. However, only extensive sequence/expressed sequence characterisation could determine whether this is so.

As an unstable repetitive DNA sequence or some other unusual sequence feature was not found in the *FRA3B* region, it appears unlikely that such characteristics will be found in the *FRA16D* region. However, if a DNA sequence similarity between the *FRA16D* and *FRA3B* fragile site sequences was found, further insight into the mechanism of fragility at the DNA sequence level may be obtained.

If, as proposed, common fragile sites are involved in tumourgenicity and/or chemotherapy drug resistance, characterisation of these sequences may eventually lead to more effective strategies and/or treatments to prevent and treat these occurrences.

Chapter 8

Concluding Remarks

Chapter 8: Contents

Concluding Remarks

8.1	Approaches to cloning unstable repeat tracts	187
8.2	Dynamic mutations.....	188
8.3	Functional role of CCG repeat tracts in genes.....	189
8.4	Genesis of fragile sites	190
8.5	Hypermethylation at fragile sites	191
8.6	Molecular basis of repeat tract instability	191
8.7	The putative <i>FRA16A</i> gene.....	192
8.8	Conclusion.....	193

8.1 Approaches to cloning unstable repeat tracts

The first reports detailing the molecular characterisation of *FRAXA*, the rare folate sensitive fragile site associated with fragile X syndrome, were published in 1991. Pedigree analysis of fragile X families revealed the existence of a previously unknown mutational mechanism, later termed dynamic mutation (Richards and Sutherland, 1992).

Owing to improvements in positional cloning technology and resources available, such as YACs and expressed sequence tag databases, the years since the discovery of the molecular basis of the fragile X syndrome have been highly productive in terms of the characterisation of genes associated with human genetic diseases. A proportion of these studies identified dynamic mutations of microsatellite CAG repeat tracts as the causative factor in a number of dominantly inherited neurological disorders. These potentially unstable disease associated repeat tracts were found to be translated, and in all cases, to code for polyglutamine tracts within otherwise (apparently) unrelated proteins.

These findings led to more rapid approaches for identifying disease associated CAG repeat tracts, such as EST (expressed sequence tag) database searches, and direct hybridisation screening of cDNA clones and/or genomic DNA clones mapping to regions containing disease loci (identified by linkage analysis). Characterisation of fragile sites *FRA11B* and *FRAXF* and were achieved by similar methods, with the direct identification of potentially unstable CCG repeat tracts at these loci (Parrish *et al.*, 1994; Ritchie *et al.*, 1994; Jones *et al.*, 1995). These approaches rendered largely unnecessary the preliminary detailed physical mapping and construction of cosmid or lambda clone contigs across the regions of interest that are a usual feature of positional cloning. In contrast with the situation for *FRA11B* and *FRAXF*, isolation and characterisation of *FRA16A* was achieved after extensive linkage and physical mapping of the 16p13.1 fragile site region. The initial physical mapping was performed using the somatic cell hybrid panel of human chromosome 16, however the tool that proved most valuable was pulsed field gel electrophoresis (PFGE). This technique provided the means to obtain a relatively accurate map of the region, and to also detect the fragile site associated hypermethylation at a long range, thus facilitating the identification of clones containing *FRA16A* fragile site

sequences. However, with relatively well characterised genomic regions, such as the human X-chromosome, such mapping may not be necessary.

8.2 Dynamic mutations

The successful identification of dynamic mutations at twelve disease genes containing CAG repeat tracts, and at five CCG repeat tracts corresponding to rare folate sensitive fragile sites initially suggested that the ability to undergo this type of mutation may be a specific property of CAG/CCG repeat tracts. It was, however, considered possible that the relative ease of finding these repeat tracts had led to a bias in the discovery of dynamic mutation loci toward these sequences. That other trinucleotide repeats tract may also be unstable was indicated by the results obtained from repeat expansion detection assays, which detected massive expansions of all ten of the possible trinucleotide repeats in the human genome (Lindblad *et al.*, 1994).

That tandem repeats other than those composed of CAG or CCG (or indeed of a trinucleotide) were able to undergo dynamic mutation has been confirmed by several studies. In 1996, Campuzano *et al.* reported that a AAG trinucleotide repeat tract expansion was the major causative mutation for the autosomal recessive disorder Friedrich's ataxia. The AAG repeat tract was localised to an intron within the disease gene, and is therefore non-coding and without an apparent function. These findings were in contrast with those for other dynamic mutations, where the repeat tract was localised to the coding region (CAG repeat tracts) and 5' untranslated region (CCG repeat tracts).

Molecular characterisation of the distamycin inducible fragile site *FRA16B* revealed that the dynamic mutation mechanism can be a property of minisatellite repeat tracts, as a heritable, massive expansion of a 33 basepair AT rich repeat was found (Yu *et al.*, 1997). In addition, it has been demonstrated that the molecular basis of *EPM1*, a form of recessively inherited epilepsy, is the expansion of a CG rich dodecanucleotide (12 mer) minisatellite repeat tract. These results were of great interest, as they indicated that the same mutational mechanism does not depend on the size of the individual units forming repeat tracts (i.e. 3 bp versus 33 bp), as much as their uniformity. On this basis, almost

any tandem repetitive sequence could be a potential target for dynamic mutation. Richards and Sutherland (1992) further suggested that, as repeat sequences of variant length and composition can undergo dynamic mutations and thereby give rise to both dominant and recessive genetic disorders, this form of mutation may be a significant cause of human disease.

In summary, dynamic mutations have been associated with a variety of human genetic diseases to date, with the most recent findings at the *FRDA*, *EPN1* and *FRA16B* loci indicating a more universal role for the mutation mechanism than previously thought.

Characterisation of an expanded AT rich minisatellite in association with *FRA16B* expression (Yu *et al.*, 1997) has indicated that dynamic mutation of either micro- or minisatellite repeat tracts may have a universal role in the expression of rare, heritable fragile sites. Given the lack of such a feature in association with the common fragile site *FRA3B*, it seems less likely that fragile sites of the same class will be associated with expanded repeat tracts of similar (if not identical) composition.

8.3 Functional role of CCG repeat tracts in genes

Studies of *FRA16B* and the rare folate sensitive fragile sites *FRAXA*, *FRAXE*, *FRAXF*, *FRA11B* and *FRA16A* have gathered much information regarding the nature of fragile sites at the nucleotide level. However, other aspects regarding the nature of fragile sites are yet to be addressed. One of these is an elucidation of the functional role of CCG tandem repeat sequences in the genome, as this may explain why such potentially deleterious features have been retained throughout evolution. As discussed in other chapters, some if not all, rare folate sensitive fragile sites are located in the 5' untranslated regions of genes, within CpG islands. The evolutionary conservation of the CCG repeat tracts within this region of the *FMRI* gene (Ashley *et al.*, 1993) suggests an important function, possibly relating to gene expression or post transcriptional control. Protein binding assays have indicated that certain proteins specifically bind the repeat *in vitro* (Richards *et al.*, 1993), but whether this is functionally significant is unknown. Another approach to defining the functional role of fragile site associated CCG repeat tract within

genes may be to compare the expression patterns and function of a number of such genes. It was for this reason that the isolation and characterisation of a gene associated with *FRA16A* was attempted.

8.4 Genesis of fragile sites

Characterisation of three classes of fragile site (common, aphidicolin inducible; rare, folate sensitive; rare, distamycin inducible) has failed to fully clarify the nature of the molecular events leading to the formation of a microscopically visible chromosome gap or break, under particular conditions of cell culture. Additionally, it has not been determined whether fragile site genesis and expression is affected only by *cis* acting factors (i.e. the length and presence/absence of interruptions in the tandem repeat tract) or if *trans* acting factors also play a role. The destabilisation of microsatellite repeat tracts in hereditary non-polyposis colon cancer (HNPCC) disorders, which are caused by mutations in DNA mismatch repair enzymes (Ionov *et al.*, 1993; Wooster *et al.*, 1994), indicate that this type of *trans* acting mutation may account for at least some of the mutations predisposing to the development of a rare fragile site, or an expansion at any other dynamic mutation locus. Zhong *et al.* (1995) proposed the existence of such an underlying mutation mechanism that may simultaneously affect multiple repeat loci, the resulting mutations of which are then inherited as a haplotype (and may or may not ultimately lead to the formation of an intergenerationally unstable repeat tract). Evidence in favour of this mechanism was provided by the identification of positive allele size associations between the *FRAXA* CCG repeat tract and nearby microsatellite loci (Brown *et al.*, 1996). In other words, the larger alleles at each microsatellite locus tend to be on the same haplotype more often than would be expected by chance alone, and similarly for the smaller alleles.

8.5 Hypermethylation at fragile sites

One of the foremost reasons for embarking upon the project of cloning and characterisation of *FRA16A* was to resolve the question of whether the hypermethylation associated with expression of *FRAXA* was the result of a failure to cancel the X-inactivation imprint in the region, as proposed by Laird (1987). The successful characterisation and methylation analysis of the autosomal fragile *FRA16A* conclusively demonstrated that the X-inactivation process is not a necessary component of rare folate sensitive fragile site hypermethylation, which may therefore be a property of the expanded CCG repeat tract itself. However, although there has been a number of proposals put forward regarding the mechanism of expanded CCG repeat tract hypermethylation (discussed in chapter 4), there have been no findings implicating the involvement of a specific enzyme or pathway. As the *FRA16A* CCG repeat tract is not normally methylated, unlike *FRAXA* which is subject to X-inactivation, this fragile site may provide a resource for experiments designed to elucidate the molecular events causing hypermethylation.

8.6 Molecular basis of repeat tract instability

Part of the project documented in this thesis was an elucidation of the molecular basis the interpopulation variation of normal *FRA16A* CCG repeat tract allele sizes. This study found that alleles of a particular configuration (mainly found in the European population) were of greater and more variable size than the alleles of different configuration. The molecular basis of this size difference was the absence of a single CCG repeat tract interruption within the larger alleles. This finding corresponded with those from studies of alleles at the *FRAXA*, spinocerebellar ataxia 1 (SCA1) and AC repeat loci, which also found that loss of interruption causes a predisposition to the formation of longer repeat tracts, ultimately resulting in disease states for *FRAXA* and SCA1. The value of the *FRA16A* study has therefore been to contribute to an increased understanding of the potentially deleterious tendency toward instability in microsatellite repeat tracts.

8.7 The putative *FRA16A* gene

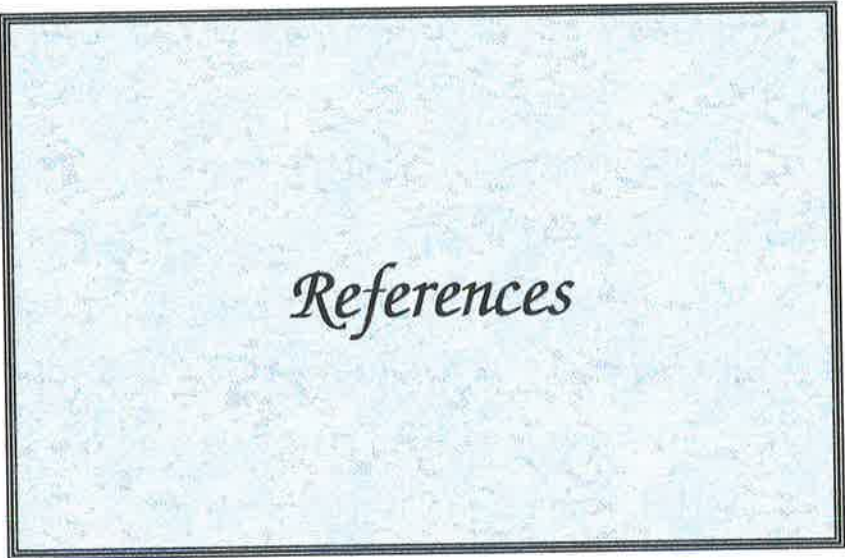
The last stage of the *FRA16A*-related project described in this thesis was an attempt to identify a *FRA16A* associated gene. The approaches used were: the identification of regions near *FRA16A* with cross-species homology; Northern blot hybridisations with probes from the *FRA16A* region; screening cDNA libraries with DNA probes containing transcribed sequences or sequences with cross-species homology; and DNA sequencing of *FRA16A* genomic region and the isolated cDNA clones, followed by computer analysis to identify gene related features.

While the results obtained from the approaches listed above strongly indicated the presence of transcribed sequences within the *FRA16A* region, they did not provide complete proof of this, as no evidence of splicing could be found in the cDNA sequences. This finding indicated that the transcript from which the cDNAs were derived may be poorly spliced, the transcribed gene may be intronless, or the isolated cDNAs may be due to genomic DNA contamination.

In view of the results obtained from screening cDNA, different experimental approaches may be required to determine the existence of a *FRA16A* associated gene. One such approach, for the identification of splicing events, may be to perform extensive RT-PCR analysis of the region in a number of different tissues, as it is possible that differential splicing may occur between tissues. Another approach, which is independent of manipulating transcribed sequences (prone to contamination with genomic DNA), may be to directly identify and/or isolate a *FRA16A* gene associated protein. In order to accomplish this, it would be necessary to raise monoclonal or polyclonal antibodies against the computer predicted amino acid sequence of the gene. These antibodies could then be used for assays to determine the presence of a corresponding protein product. However, if a protein were identified by this strategy, it would still be necessary to determine the genomic structure of the gene.

8.8 Conclusion

The molecular characterisation of *FRA16A* has contributed to the understanding of the properties of rare folate sensitive fragile sites, and thus dynamic mutations. It is to be hoped that, in the not too far distant future, the success of the project described in this thesis, and others like it, will also contribute to a more global understanding of the precise cellular processes involved, and hence the development of therapeutic, or preventative strategies, for at least some of the human genetic disorders caused by dynamic mutations.



References

- Abidi, F. E., Wada, M., Little, R. D., and Schlessinger, D. (1990). Yeast artificial chromosomes containing human Xq24-Xq28 DNA: library construction and representation of probe sequences. *Genomics* 7, 363-376.
- Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. (1983). *Molecular biology of the cell* (New York: Garland Publishing Inc.).
- Anderson, C. (1993). Genome shortcut leads to problems [news] [see comments]. *Science* 259, 1684-7.
- Andrew, S. E., Goldberg, Y. P., Theilmann, J., Zeisler, J., and Hayden, M. R. (1994). A CCG repeat polymorphism adjacent to the CAG repeat in the Huntington disease gene: implications for diagnostic accuracy and predictive testing. *Hum Mol Genet* 3, 65-7.
- Arinami, T., Kondo, I., and Nakajima, S. (1986). Frequency of the fragile X syndrome in Japanese mentally retarded males. *Hum Genet* 73, 309-12.
- Arinami, T., Asano, M., Kobayashi, K., Yanagi, H., and Hamaguchi, H. (1993). Data on the CGG repeat at the fragile X site in the non-retarded Japanese population and family suggest the presence of a subgroup of normal alleles predisposing to mutate. *Hum Genet* 92, 431-6.
- Ashley, C. T., Sutcliffe, J. S., Kunst, C. B., Leiner, H. A., Eichler, E. E., Nelson, D. L., and Warren, S. T. (1993a). Human and murine FMR-1: alternative splicing and translational initiation downstream of the CGG-repeat. *Nat Genet* 4, 244-51.
- Ashley, C. T., Jr., Wilkinson, K. D., Reines, D., and Warren, S. T. (1993b). FMR1 protein: conserved RNP family domains and selective RNA binding. *Science* 262, 563-6.
- Barron, L. H., Rae, A., Holloway, S., Brock, D. J., and Warner, J. P. (1994). A single allele from the polymorphic CCG rich sequence immediately 3' to the unstable CAG trinucleotide in the IT15 cDNA shows almost complete disequilibrium with Huntington's disease chromosomes in the Scottish population. *Hum Mol Genet* 3, 173-5.
- Bell, M. V., Hirst, M. C., Nakahori, Y., MacKinnon, R. N., Roche, A., Flint, T. J., Jacobs, P. A., Tommerup, N., Tranebjaerg, L., Froster-Iskenius, U., Kerr, B., Turner, G., Lindenbaum, R. H., Winter, R., Pembrey, M., Thibodeau, S., and Davies, K. E. (1991). Physical mapping across the fragile X: hypermethylation and clinical expression of the fragile X syndrome. *Cell* 64, 861-6.

Benton, W. D., and Davis, R. W. (1977). Screening lambda gt recombinant clones by hybridisation to single plaques in situ. *Science* 196, 180-182.

Bichara, M., Schumacher, S., and Fuchs, R. P. (1995). Genetic instability within monotonous runs of CpG sequences in *Escherichia coli*. *Genetics* 140, 897-907.

Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature* 321, 209-13.

Boggs, B. A., and Nussbaum, R. L. (1984). Two anonymous X-specific human sequences detecting restriction fragment length polymorphisms in region Xq26----qter. *Somat Cell Mol Genet* 10, 607-13.

Boldog, F., Gemmill, R. M., West, J., Robinson, M., Robinson, L., Li, E., Roche, J., Todd, S., Waggoner, B., Lundstrom, R., Jacobson, J., Mullokandov, M. R., Klinger, H., and Drabkin, H. A. (1997). Chromosome 3p14 homozygous deletions and sequence analysis of FRA3B. *Hum Mol Genet* 6, 193-203.

Breuning, M. H., Snyderwint, F., Brunner, H., Verwest, A., Kievits, T., Saris, J., Dauwerse, H., Reeders, S. T., Keith, T., Callen, D. F., Hyland, V. J., Xiao, H. G., Scherer, G., Higgs, D. R., Nakamura, Y., White, R., Bachner, L., Blonden, L. A. J., van Ommen, G. J. B., and Pearson, P. L. (1989). Map of 13 polymorphic loci on the short arm of chromosome 16 close to the gene for polycystic kidney disease (PKD1). *Cytogenet Cell Genet* 51, 969.

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., Sohn, R., Zemelman, B., Snell, R.G., Rundle, S. A., Crow, S., Davies, J., Shelbourne, P., Buxton, J., Jones, C., Juvonen, V., Johnson, K., Harper, P. S., Shaw, D. J., and Houseman, D. E. (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 [published erratum appears in *Cell* 1992 Apr 17;69(2):385]. *Cell* 68, 799-808.

Brown, R. M., Fraser, N. J., and Brown, G. K. (1990). Differential methylation of the hypervariable locus DXS255 on active and inactive X chromosomes correlates with the expression of a human X-linked gene. *Genomics* 7, 215-21.

Brown, T. (1993). Analysis of DNA sequences by blotting and hybridisation. In *Current protocols in molecular biology*, F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith and K. Struhl, eds. (New York: John Wiley & Sons, Inc.), pp. 2.9.1-2.9.20.

Brown, T., Zhong, N., and Dobkin, C. (1996). Positive fragile X microsatellite associations point to a common mechanism of dynamic mutation evolution. *Am J Hum Genet* 58, 641-3.

Buchanan, J. A., Buckton, K. E., Gosden, C. M., Newton, M. S., Clayton, J. F., Christie, S., and Hastie, N. (1987). Ten families with fragile X syndrome: linkage relationships with four DNA probes from distal Xq. *Hum Genet* 76, 165-72.

Burke, D. T., Carle, G. F., and Olson, M. V. (1987). The cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* 236, 806-812.

Buyle, S., Reyniers, E., Vits, L., De Boulle, K., Handig, I., Wuyts, F. L., Deelen, W., Halley, D. J., Oostra, B. A., and Willems, P. J. (1993). Founder effect in a Belgian-Dutch fragile X population. *Hum Genet* 92, 269-72.

Callen, D. F. (1986). A mouse-human hybrid cell panel for mapping human chromosome 16. *Ann Genet* 29, 235-9.

Callen, D. F., Hyland, V. J., Baker, E. G., Fratini, A., Gedeon, A. K., Mulley, J. C., Fernandez, K. E., Breuning, M. H., and Sutherland, G. R. (1989). Mapping the short arm of human chromosome 16. *Genomics* 4, 348-54.

Callen, D. F., Baker, E., Eyre, H. J., and Lane, S. A. (1990). An expanded mouse-human hybrid cell panel for mapping human chromosome 16. *Ann. Genet.* 33, 190-95.

Callen, D. F., Doggett, N. A., Stallings, R. L., Chen, L. Z., Whitmore, S. A., Lane, S. A., Nancarrow, J. K., Apostolou, S., Thompson, A. D., Lapsys, N. M., Eyre, H. J., Baker, E. G., Shen, Y., Holman, K., Phillips, H., Richards, R. I. and Sutherland, G. R. (1992). High-resolution cytogenetic-based physical map of human chromosome 16. *Genomics* 13, 1178-85.

Camerino, G., Mattei, M. G., Mattei, J. F., Jaye, M., and Mandel, J.-L. (1983). Close linkage of fragile X-mental retardation syndrome to haemophilia B and transmission through a normal male. *Nature* 306, 701-4.

Campuzano, V., Montermini, L., Molto, M. D., Pianese, L., Cossee, M., Cavalcanti, F., Monros, E., Rodius, F., Duclos, F., Monticelli, A., Zara, F., Canizares, J., Koutnikova, H., Bidichandani, S. I., Gellera, C., Brice, A., Trouillas, P., De Michelle, G., Filla, A., De Frutos, R., Palau, F., Patel, P. I., Di Donato, S., Mandel, J.-L., Coccozza, S., Koenig, M., and Pandolfo, M. (1996). Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion [see comments]. *Science* 271, 1423-7.

Caskey, C. T., Pizzuti, A., Fu, Y.-H., Fenwick Jr., R. G., and Nelson, D. L. (1992). Triplet repeat mutations in human disease. *Science* 256, 784-789.

Chakrabarti, L., Knight, S. J., Flannery, A. V., and Davies, K. E. (1996). A candidate gene for mild mental handicap at the FRAXE fragile site. *Hum Mol Genet* 5, 275-82.

Chakravarti, A. (1992). Fragile X founder effect? [news]. *Nat Genet* 1, 237-8.

Chaplin, D. D., and Brownstein, B. H. (1995). Yeast Artificial Chromosome Libraries. In *Current protocols in molecular biology*, F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith and K. Struhl, eds. (New York: John Wiley & Sons, Inc.), pp. 6.10.4-6.10.5.

Chou, P.Y. and Fasman, G.D. (1978) Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol* 47, 45-148.

Chu, G., Vollrath, D., and Davis, R. W. (1986). Separation of large DNA molecules by contour-clamped homogeneous electric fields. *Science* 241, 1203-1205.

Chung, C. T., Niemela, S. L., and Miller, R. H. (1989). One step preparation of competent *Escherichia coli* : Transformation and storage of bacterial cells in the same solution. *Proc. Natl. Acad. Sci. USA* 86, 2172-2175.

Chung, M. Y., Ranum, L. P., Duvick, L. A., Servadio, A., Zoghbi, H. Y., and Orr, H. T. (1993). Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I. *Nat Genet* 5, 254-8.

Consortium, T. D.-B. F. X. (1994). *Fmr1* knockout mice: a model to study fragile X mental retardation. *Cell* 78, 23-33.

Coquelle, A., Pipiras, E., Toledo, F., Buttin, G., and Debatisse, M. (1997). Expression of fragile sites triggers intrachromosomal mammalian gene amplification and sets boundaries to early amplicons. *Cell* 89, 215-25.

- Craig-Holmes, A. P., Strong, L. C., Goodacre, A., and Pathak, S. (1987). Variation in the expression of aphidicolin-induced fragile sites in human lymphocyte cultures. *Hum Genet* 76, 134-7.
- Cross, S. H., and Bird, A. P. (1995). CpG islands and genes. *Curr Opin Genet Dev* 5, 309-14.
- Daniel, A. (1986). Clinical implications and classification of the constitutive fragile sites. *Am J Med Genet* 23, 419-27.
- Daly, M. C., Douglas, J. B., Bleehe, N. M., Hastleton, P., Twentyman, P.R., Sundaresan, V., Carritt, B., Bergh, J. and Rabitts, P.H. (1991). An unusually proximal deletion on the short arm of chromosome 3 in a patient with small cell lung cancer. *Genomics* 9, 113-119.
- Day, E. J., Marshall, R., MacDonald, P. A., and Davidson, W. M. (1967). Deleted chromosome 18 with paternal mosaicism. *Lancet* 2, 1307.
- De Boulle, K., Verkerk, A. J. M. H., Reyniers, E., Vits, L., Hendrickx, J., Van Roy, B., Van den Bos, F., de Graaff, E., Oostra, B. A. and Willems, P. J. (1993). A point mutation in the FMR-1 gene associated with fragile X mental retardation. *Nature Genet* 3, 31-35.
- De Braekeleer, M., Smith, B., and Lin, C. C. (1985). Fragile sites and structural rearrangements in cancer. *Hum Genet* 69, 112-6.
- Deelen, W., Bakker, C., Halley, D. J., and Oostra, B. A. (1994). Conservation of CGG region in FMR1 gene in mammals. *Am J Med Genet* 51, 513-6.
- Deka, R., Miki, T., Yin, S. J., McGarvey, S. T., Shriver, M. D., Bunker, C. H., Raskin, S., Hundrieser, J., Ferrell, R. E., and Chakraborty, R. (1995). Normal CAG repeat variation at the DRPLA locus in world populations [letter]. *Am J Hum Genet* 57, 508-11.
- Devys, D., Biancalana, V., Rousseau, F., Boue, J., Mandel, J.-L., and Oberlé, I. (1992). Analysis of full fragile X mutations in fetal tissues and monozygotic twins indicate that abnormal methylation and somatic heterogeneity are established early in development. *Am J Med Genet* 43, 208-16.
- Doerfler, W. (1992) DNA methylation: eucaryotic defence against the transcription of foreign genes? *Microb Patholog* 12, 1-8.

Doggett, N. A., Goodwin, L. A., Tesmer, J. G., Meincke, L. J., Bruce, D. C., Clark, L. M., Altherr, M. R., Ford, A. A., Chi, H. C., and Marrone, B. L., (1995). An integrated physical map of human chromosome 16. *Nature* 377, 335-65.

Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T. P., Bowden, D. W., Smith, D.R., Lander, E. S., Botstein, D., Akots, G., Rediker, K. S., Gravius, T., Brown, V. A., Rising, M. B., Parker, C., Powers, J. A., Watt, D. E., Kauffman, E. R., Bricker, A., Phipps, P., Muller- Kahle, H., Fulton, T. R., Ng, S., Schumm, J. W., Braman, J. C., Knowlton, R. G., Barker, D. F., Crooks, S. M., Lincoln, S. E., Daly, M. J., and Abrahamson, J. (1987). A genetic linkage map of the human genome. *Cell* 51, 319-37.

Drayna, D., Davies, K., Hartley, D., Mandel, J. L., Camerino, G., Williamson, R., and White, R. (1984). Genetic mapping of the human X chromosome by using restriction fragment length polymorphisms. *Proc Natl Acad Sci U S A* 81, 2836-9.

Drayna, D., and White, R. (1985). The genetic linkage map of the human X chromosome. *Science* 230, 753-8.

Eichler, E. E., Holden, J. J., Popovich, B. W., Reiss, A. L., Snow, K., Thibodeau, S. N., Richards, C. S., Ward, P. A., and Nelson, D. L. (1994). Length of uninterrupted CGG repeats determines instability in the FMR1 gene. *Nat Genet* 8, 88-94.

Eichler, E. E., Hammond, H. A., Macpherson, J. N., Ward, P. A., and Nelson, D. L. (1995a). Population survey of the human FMR1 CGG repeat substructure suggests biased polarity for the loss of AGG interruptions. *Hum Mol Genet* 4, 2199-208.

Eichler, E. E., Kunst, C. B., Lugenbeel, K. A., Ryder, O. A., Davison, D., Warren, S. T., and Nelson, D. L. (1995b). Evolution of the cryptic FMR1 CGG repeat. *Nat Genet* 11, 301-8.

Eichler, E. E., Macpherson, J. N., Murray, A., Jacobs, P. A., Chakravarti, A., and Nelson, D. L. (1996). Haplotype and interspersed analysis of the FMR1 CGG repeat identifies two different mutational pathways for the origin of the fragile X syndrome. *Hum Mol Genet* 5, 319-30.

Eisenberg, D., Weiss, R.M. and Terwilliger, T.C. (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc Nat Acad Sci* 81(1), 140-144.

Emini, E.A., Hughes, J., Perlow, D. and Bolger, J. (1985) Induction of hepatitis A virus-neutralising antibody by a virus specific synthetic peptide. *J Virology* 55, 836-839.

Evans, G. A. (1993). "MegaYAC" library [letter; comment]. *Science* 260, 877.

Falik-Zaccai, T. C., Shachak, E., Yalon, M., Lis, Z., Borochowitz, Z., Macpherson, J. N., Nelson, D. L., and Eichler, E. E. (1997). Predisposition to the fragile X syndrome in Jews of Tunisian descent is due to the absence of AGG interruptions on a rare Mediterranean haplotype. *Am J Hum Genet* 60, 103-12.

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

Feng, Y., Zhang, F., Lokey, L. K., Chastain, J. L., Lakkis, L., Eberhart, D., and Warren, S. T. (1995). Translational suppression by trinucleotide repeat expansion at FMR1. *Science* 268, 731-4.

Filippi, G., Rinaldi, A., Archidiacono, N., Rocchi, M., Balazs, I., and Siniscalco, M. (1983). Brief report: linkage between G6PD and fragile-X syndrome. *Am J Med Genet* 15, 113-9.

Finney, M. (1994). Pulsed-field gel electrophoresis. In *Current protocols in molecular biology*, F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith and K. Struhl, eds. (New York: John Wiley & Sons, Inc.), pp. 2.5.9-2.5.17.

Follette, P. J., and Laird, C. D. (1992). Estimating the stability of the proposed imprinted state of the fragile- X mutation when transmitted by females. *Hum Genet* 88, 335-43.

Froster-Iskenius, U., Schulze, A., and Schwinger, E. (1984). Transmission of the marker X syndrome trait by unaffected males: conclusions from studies of large families. *Hum Genet* 67, 419-27.

Fry, M., and Loeb, L. A. (1994). The fragile X syndrome p(CGG)_n nucleotide repeats form a stable tetrahelical structure. *Proc Natl Acad Sci U S A* 91, 4950-4.

Fryns, J. P., Kleczkowska, A., Kubien, E., Petit, P., and Van den Berghe, H. (1984). Inactivation pattern of the fragile X in heterozygous carriers. *Hum Genet* 65, 400-1.

Fu, Y. H., Kuhl, D. P., Pizzuti, A., Pieretti, M., Sutcliffe, J. S., Richards, S., Verkerk, A. J., Holden, J. J., Fenwick, R. G., Jr., Warren, S. T., Oostra, B. A., Nelson, D. L., and Caskey, C. T. (1991). Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell* 67, 1047-58.

- 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.
- Garnier, J., Osguthorpe, D.J., and Robson, B. (1978). Analysis of the accuracy and implications of a simple method for predicting the secondary structure of globular proteins. *J Mol Biol* 120, 97-120.
- 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.
- Gedeon, A. K., Mulley, J. C., and Breuning, M. H. (1989). XmnI, HincII and BclI RFLPs at D16S79. *Nucleic Acids Res* 17, 4905.
- Gedeon, A. K., Baker, E., Robinson, H., Partington, M. W., Gross, B., Manca, A., Korn, B., Poustka, A., Yu, S., Sutherland, G. R., and Mulley, J. C. (1992). Fragile X syndrome without CCG amplification has an FMR1 deletion. *Nat Genet* 1, 341-4.
- Gedeon, A. K., Meinanen, M., Ades, L. C., Kaariainen, H., Gecz, J., Baker, E., Sutherland, G. R., and Mulley, J. C. (1995). Overlapping submicroscopic deletions in Xq28 in two unrelated boys with developmental disorders: identification of a gene near FRAXE. *Am-J-Hum-Genet* 56, 907-14.
- Giraud, F., Ayme, S., Mattei, J. F., and Mattei, M. G. (1976). Constitutional chromosomal breakage. *Hum Genet* 34, 125-36.
- Glover, T. W., Berger, C., Coyle, J., and Echo, B. (1984). DNA polymerase alpha inhibition by aphidicolin induces gaps and breaks at common fragile sites in human chromosomes. *Hum Genet* 67, 136-42.
- Glover, T. W., Coyle-Morris, J., and Morgan, R. (1986). Fragile sites: overview, occurrence in acute nonlymphocytic leukemia and effects of caffeine on expression. *Cancer Genet Cytogenet* 19, 141-50.
- Glover, T. W., and Stein, C. K. (1988). Chromosomal breakage and recombination at fragile sites. *Am J Hum Genet* 43, 265-73.
- Goonewardena, P., Gustavson, K. H., Holmgren, G., Tolun, A., Chotai, J., Johnsen, E., and Pettersson, U. (1986). Analysis of fragile X-mental retardation families using flanking polymorphic DNA probes. *Clin Genet* 30, 249-54.

Grunstein, M., and Hogness, D. S. (1975). Colony hybridisation: A method for the isolation of cloned DNAs that contain a specific gene. *Proc. Natl. Acad. Sci. USA* 72, 3961-3965.

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.

Haataja, R., Vaisanen, M.L., Li, M. Ryyanen, M., and Leisti, T. (1994). Demonstration of a founder effect by analysis of microsatellite haplotypes. *Hum Genet* 94, 479-483.

Hagerman, R., and Silverman, A. (1991). *Fragile X Syndrome: diagnosis, treatment, and research*: The John Hopkins University Press, Baltimore).

Haldane, J. (1948). The Formal Genetics of Man. *Proc Roy Soc Lond* 35, 147-170.

Hansen, R. S., Gartler, S. M., Scott, C. R., Chen, S. H., and Laird, C. D. (1992). Methylation analysis of CGG sites in the CpG island of the human FMR1 gene. *Hum Mol Genet* 1, 571-8.

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

Harley, H. G., Brook, J. D., Floyd, J., Rundle, S. A., Crow, S., Walsh, K. V., Thibault, M. C., Harper, P. S., and Shaw, D. J. (1991). Detection of linkage disequilibrium between the myotonic dystrophy locus and a new polymorphic DNA marker. *Am J Hum Genet* 49, 68-75.

Harvey, J., Judge, C., and Wiener, S. (1977). Familial X-linked mental retardation with an X chromosome abnormality. *J Med Genet* 14, 46-50.

Heale, S. M., and Petes, T. D. (1995). The stabilization of repetitive tracts of DNA by variant repeats requires a functional DNA mismatch repair system [published erratum appears in *Cell* 1996 May 31;85(5):following 779]. *Cell* 83, 539-45.

Hecht, F., and Glover, T. W. (1984). Cancer chromosome breakpoints and common fragile sites induced by aphidicolin. *Cancer Genet Cytogenet* 13, 185-8.

Hecht, F., and Hecht, B. K. (1986). Chromosome 9 in acute lymphoblastic leukemia: breaks in band 9p21-22 and a fragile site [editorial]. *Cancer Genet Cytogenet* 21, 1-3.

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., Grewal, P. K., and Davies, K. E. (1994). Precursor arrays for triplet repeat expansion at the fragile X locus. *Hum Mol Genet* 3, 1553-60.

Hofker, M. H., Bergen, A. A., Skraastad, M. I., Carpenter, N. J., Veenema, H., Connor, J. M., Bakker, E., van Ommen, G. J., and Pearson, P. L. (1987). Efficient isolation of X chromosome-specific single-copy probes from a cosmid library of a human X/hamster hybrid-cell line: mapping of new probes close to the locus for X-linked mental retardation. *Am J Hum Genet* 40, 312-28.

Hornstra, I. K., Nelson, D. L., Warren, S. T., and Yang, T. P. (1993). High resolution methylation analysis of the FMR1 gene trinucleotide repeat region in fragile X syndrome. *Hum Mol Genet* 2, 1659-65.

Howard-Peebles, P. N., and Stoddard, G. R. (1979a). X-linked mental retardation with macro-orchidism and marker X chromosomes. *Hum Genet* 50, 247-51.

Howard-Peebles, P. N., Stoddard, G. R., and Mims, M. G. (1979b). Familial X-linked mental retardation, verbal disability, and marker X chromosomes. *Am J Hum Genet* 31, 214-22.

Hudson, T. J., Clark, C. D., and Gschwend, M. (1994). PCR methods of genotyping. In *Current protocols in human genetics*, N. C. Drocopoli, J. L. Haines, B. R. Korf, D. T. Moir, C. C. Morton, C. S. Seidman, J. G. Siedman, D. R. Smith eds. (New York: John Wiley & Sons, Inc.), pp. 2.5.1-2.5.20.

Hyland, V. J., Callen, D. F., Fernandez, K. E. W., MacKinnon, R. N., Friend, K. L., Mulley, J. C., Fratini, A. F., Baker, E., Breuning, M. H., Keith, T., and Sutherland, G. R. (1989a). Anonymous DNA probes to specific intervals of human chromosome 16. *Cytogenet Cell Genet* 51, 1017.

Hyland, V. J., Fernandez, K. E., Callen, D. F., MacKinnon, R. N., Baker, E., Friend, K., and Sutherland, G. R. (1989b). Assignment of anonymous DNA probes to specific intervals of human chromosomes 16 and X. *Hum Genet* 83, 61-6.

- Igarashi, S., Takiyama, Y., Cancel, G., Rogaeva, E. A., Sasaki, H., Wakisaka, A., Zhou, Y. X., Takano, H., Endo, K., Sanpei, K., Oyake, M., Tanaka, H., Stevanin, G., Abbas, N., Durr, A., Rogaev, E. I., Sherrington, R., Tsuda, T., Ikeda, M., Cassa, E., Nishizawa, M., Benomar, A., Julien, J., Weissenbach, J., Wang, G.-X., Agid, Y., St. George-Hyslop, P. H., Brice, A., and Tsuji, S. (1996). Intergenerational instability of the CAG repeat of the gene for Machado- Joseph disease (MJD1) is affected by the genotype of the normal chromosome: implications for the molecular mechanisms of the instability of the CAG repeat. *Hum Mol Genet* 5, 923-32.
- Imbert, G., Kretz, C., Johnson, K., and Mandel, J.-L. (1993). Origin of the expansion mutation in myotonic dystrophy. *Nat Genet* 4, 72-6.
- Imbert, G., and Mandel, J.-L. (1995) The fragile X mutation. *Mental Retardation and Developmental Disabilities Research Reviews* 1, 251-262.
- Innis, M. A., Gelfand, D. H., Sninsky, J. J., and White, T. J. (1989). PCR protocols. A guide to methods and applications. (San Diego, California: Academic Press, Inc.).
- Ionov, Y., Peinado, M. A., Malkhosyan, S., Shibata, D., and Perucho, M. (1993). Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature* 363, 558-61.
- Jacobs, P. A., Glover, T. W., Mayer, M., Fox, P., Gerrard, J. W., Dunn, H. G., and Herbst, D. S. (1980). X-linked mental retardation: a study of 7 families. *Am J Med Genet* 7, 471-89.
- Jacobs, P. A., Mayer, M., Matsuura, J., Rhoads, F., and Yee, S. C. (1983). A cytogenetic study of a population of mentally retarded males with special reference to the marker (X) syndrome. *Hum Genet* 63, 139-48.
- Jacobs, P. A., Sherman, S., Turner, G., and Webb, T. (1986). The fragile (X) syndrome: the mutation problem. *Am J Med Genet* 23, 611-7.
- Jameson, B. A. and Wolf, H. (1988). The antigenic index: a novel algorithm for predicting antigenic determinants. *CABIOS* 4, 181-186.
- Jeffreys A.J, T. K., MacLeod A., Monckton D.G., Neil D.L., and Armour J.A. (1994). Complex gene conversion events in germline mutation at human minisatellites. *Nature Genetics* 6, 136-145.

- 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.
- Kang, S., Ohshima, K., Shimizu, M., Amirhaeri, S., and Wells, R. D. (1995). Pausing of DNA synthesis in vitro at specific loci in CTG and CGG triplet repeats from human hereditary disease genes. *J Biol Chem* 270, 27014-21.
- Karplus, P. A. and Schultz, G. A. (1985). Prediction of chain flexibility in proteins. *Naturwissenschaften* 72, 212-213.
- Kenwrick, S., Patterson, M., Speer, A., Fischbeck, K., and Davies, K. (1987). Molecular analysis of the Duchenne muscular dystrophy region using pulsed field gel electrophoresis. *Cell* 48, 351-7.
- Khalifa, M. M., Reiss, A. L., and Migeon, B. R. (1990). Methylation status of genes flanking the fragile site in males with the fragile-X syndrome: a test of the imprinting hypothesis. *Am J Hum Genet* 46, 744-53.
- Khandjian, E. W., Corbin, F., Woerly, S., and Rousseau, F. (1996). The fragile X mental retardation protein is associated with ribosomes. *Nat Genet* 12, 91-3.
- Knight, S. J. L., Flannery, A. V., Hirst, M. C., Campbell, L., Christodoulou, Z., Phelps, S. R., Pointon, J., Middleton-Price, H. R., Barnicoat, A., Pembrey, M. E., Holland, J., Oostra, B.A., Bobrow, M., and Davies, K. E. (1993). Trinucleotide repeat amplification and hypermethylation of a CpG island in FRAXE mental retardation. *Cell* 74, 127-134.
- 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.
- Knoll, J. H., Chudley, A. E., and Gerrard, J. W. (1984). Fragile (X) X-linked mental retardation. II. Frequency and replication pattern of fragile (X)(q28) in heterozygotes. *Am J Hum Genet* 36, 640-5.
- Kogan, S. C., Doherty, B. S., and Gitsher, J. (1987). An improved method for prenatal diagnosis of genetic diseases by analysis of amplified DNA sequences. *N Eng J Med* 317, 985-90.

- Kok, K., Osinga, J., Carritt, B., Davis, M. B., van der Hout, A. H., van der Veen, A. Y., Landsvater, R. M., der Leij, L. F. M. H., Berendsen, H. H., Postmus, P. E., Poppema, S. and Buys, C. H. C. M. (1987). Deletion of a DNA sequence at the chromosomal region 3p21 in all major types of lung cancer. *Nature* 330, 578-581.
- Kovacs, G. and Kung, H. (1991). Nonhomologous chromatid exchange in hereditary and sporadic renal cell carcinomas. *Proc Natl Acad Sci USA* 88, 194-198.
- Kozak, M. (1987). An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res* 15, 8125-48.
- Kozak, M. (1988). Leader length and secondary structure modulate mRNA function under conditions of stress. *Mol Cell Biol* 8, 2737-44.
- Kozak, M. (1991). Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J Biol Chem* 266, 19867-70.
- Kozman, H., Phillips, H., Sutherland, G., and Mulley, J. (1992). A multipoint genetic linkage map around the fragile site FRA16A on human chromosome 16. *Genet (Life Sci Adv.)* 11, 229-233.
- Kremer, E. J., Yu, S., Pritchard, M., Nagaraja, R., Heitz, D., Lynch, M., Baker, E., Hyland, V. J., Little, R. D., Wada, M., Toniolo, D., Vincent, A., Rousseau, D., Schlessinger, D., Sutherland, G. R., and Richards, R. I. (1991a). Isolation of a human DNA sequence which spans the fragile X. *Am J Hum Genet* 49, 656-61.
- 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. (1991b). Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)_n. *Science* 252, 1711-4.
- Kuhl, D. P., and Caskey, C. T. (1993). Trinucleotide repeats and genome variation. *Curr Opin Genet Dev* 3, 404-7.
- Kunst, C.B. and Warren, S.T. (1994). Cryptic and polar variation of the fragile X repeat could result in predisposing normal alleles. *Cell* 77, 853-861.
- Kyte, J. and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 157, 105-132.

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.

LaForgia, S., Morse, B., Levy, J., Barnea, G., Cannizzaro, L. A., Li, F., Nowell, P. C., Boghosian-Sell, L., Glick, J., Weston, A., and et al. (1991). Receptor protein-tyrosine phosphatase gamma is a candidate tumor suppressor gene at human chromosome region 3p21. *Proc Natl Acad Sci U S A* 88, 5036-40.

LaForgia, S., Lasota, J., Latif, F., Boghosian-Sell, L., Kastury, K., Ohta, M., Druck, T., Atchison, L., Cannizzaro, L. A., Barnea, G., Schlessinger, J., Modi, W., Kuzmin, I., Tory, K., Zbar, B., Croce, C. M., Lerman, M., and Huebner, K. (1993). Detailed genetic and physical map of the 3p chromosome region surrounding the familial renal cell carcinoma chromosome translocation, t(3;8)(p14.2;q24.1). *Cancer Res* 53, 3118-24.

Laird, C. D. (1987). Proposed mechanism of inheritance and expression of the human fragile-X syndrome of mental retardation. *Genetics* 117, 587-99.

Laird, C., Jaffe, E., Karpen, G., Lamb, M., and Nelson, R. (1987). Fragile sites in human chromosomes as regions of late replicating DNA. *Trends Genet* 3, 274-81.

Laird, C. D., Lamb, M. M., and Thorne, J. L. (1990). Two progenitor cells for human oögonia inferred from pedigree data and the X-inactivation imprinting model of the fragile-X syndrome. *Am J Hum Genet* 46, 696-719.

Laird, C. D. (1991). Possible erasure of the imprint on a fragile X chromosome when transmitted by a male. *Am J Med Genet* 38, 391-5.

Le Beau, M. M. (1986). Chromosomal fragile sites and cancer-specific rearrangements. *Blood* 67, 849-58.

Lindblad, K., Zander, C., Schalling, M., and Hudson, T. (1994). Growing triplet repeats. *Nat Genet* 7, 124.

Lorenzo, M. J., Gish, G. D., Houghton, C., Stonehouse, T. J., Pawson, T., Ponder, B. A. and Smith, D. P. (1997). RET alternate splicing influences the interaction of activated RET with the SH2 and PTB domains of Shc, and the SH2 domain of Grb2. *Oncogene* 20, 763-71.

Lubs, H. A. (1969). A marker X chromosome. *Am J Hum Genet* 21, 231-44.

Lubs, H., Lujan, J., Donahue, R., and Lubs, M. (1984a). Diminished frequency of marker X and mental retardation after transmission through males. *American Journal of Human Genetics* 36, 102s.

Lubs, H., Travers, H., Lujan, E., and Carroll, A. (1984b). A large kindred with X-linked mental retardation, marker X and macroorchidism. *Am J Med Genet* 17, 145-57.

Lugenbeel, K.A., Carson, N., Chudley, A., and Nelson, D. L. (1995) Intragenic loss of function mutations demonstrate the primary role of FMR1 in fragile X syndrome. *Nature Genet* 10, 483-495.

MacDonald, M. E., Novelletto, A., Lin, C., Tagle, D., Barnes, G., Bates, G., Taylor, S., Allitto, B., Altherr, M., Myers, R., and et al. (1992). The Huntington's disease candidate region exhibits many different haplotypes. *Nat Genet* 1, 99-103.

Macpherson, J. N., Bullman, H., Youings, S. A., and Jacobs, P. A. (1994). Insert size and flanking haplotype in fragile X and normal populations: possible multiple origins for the fragile X mutation. *Hum Mol Genet* 3, 399-405.

Mandel, J.-L., and Heitz, D. (1992). Molecular genetics of the fragile-X syndrome: a novel type of unstable mutation. *Curr Opin Genet Dev* 2, 422-430.

Martin, J., and Bell, J. (1943). A pedigree of mental defect showing sex linkage. *J Neurol Psychiatry* 6, 154-157.

Michels, V. V. (1985). Fragile sites on human chromosomes: description and clinical significance. *Mayo Clin Proc* 60, 690-6.

Monaco, A. P., Lam, V. M., Zehetner, G., Lennon, G. G., Douglas, C., Nizetic, D., Goodfellow, P. N., and Lehrach, H. (1991). Mapping irradiation hybrids to cosmid and yeast artificial chromosome libraries by direct hybridization of Alu-PCR products. *Nucleic Acids Res* 19, 3315-8.

Moore, D. (1993). Preparation and analysis of DNA. In *Current protocols in molecular biology*, F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith and K. Struhl, eds. (New York: John Wiley & Sons, Inc.), pp. 2.0.5-2.1.9.

Morton, N. E., and Macpherson, J. N. (1992). Population genetics of the fragile-X syndrome: multiallelic model for the FMR1 locus. *Proc Natl Acad Sci U S A* 89, 4215-7.

Morris, A., Morton, N.E., Collins, A., et al. (1995) Evolutionary dynamics of the FMR1 locus. *Ann Hum Genet* 59, 283-289.

Mulley, J. C., and Sutherland, G. R. (1987). Fragile X transmission and the determination of carrier probabilities for genetic counseling. *Am J Med Genet* 26, 987-90.

Mulley, J. C., Gedeon, A. K., Thorn, K. A., Bates, L. J., and Sutherland, G. R. (1987). Linkage and genetic counselling for the fragile X using DNA probes 52A, F9, DX13, and ST14. *Am J Med Genet* 27, 435-48.

Mulley, J. C., Yu, S., Loesch, D. Z., Hay, D. A., Donnelly, A., Gedeon, A. K., Carbonell, P., Lopez, I., Glover, G., Gabarron, I., and et al. (1995). FRAXE and mental retardation. *J Med Genet* 32, 162-9.

Nielsen, K. B., and Tommerup, N. (1981). Macroorchidism, mental retardation, and the fragile X [letter]. *N Engl J Med* 305, 1348.

Nielsen, K. B., Tommerup, N., Poulsen, H., Jacobsen, P., Beck, B., and Mikkelsen, M. (1983). Carrier detection and X-inactivation studies in the fragile X syndrome. Cytogenetic studies in 63 obligate and potential carriers of the fragile X. *Hum Genet* 64, 240-5.

Nussbaum, R. L., Airhart, S. D., and Ledbetter, D. H. (1986). Recombination and amplification of pyrimidine-rich sequences may be responsible for initiation and progression of the Xq27 fragile site: an hypothesis. *Am J Med Genet* 23, 715-21.

Oberlé, I., Heilig, R., Moisan, J. P., Kloepper, C., Mattei, G. M., Mattei, J. F., Boue, J., Froster-Iskenius, U., Jacobs, P. A., Lathrop, G. M., Lalouel, J. M., and Mandel, J. L. (1986). Genetic analysis of the fragile-X mental retardation syndrome with two flanking polymorphic DNA markers. *Proc Natl Acad Sci U S A* 83, 1016-20.

Oberlé, 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.

Ohta, M., Inoue, H., Cotticelli, M. G., Kastury, K., Baffa, R., Palazzo, J., Siprashvili, Z., Mori, M., McCue, P., Druck, T., Croce, C. M., and Huebner, K. (1996). The FHIT gene, spanning the chromosome 3p14.2 fragile site and renal carcinoma-associated t(3;8) breakpoint, is abnormal in digestive tract cancers. *Cell* 84, 587-97.

- Ong, S. T., Fong, K. M., Bader, S. A., Minna, J. D., Le Beau, M. M., McKeithan, T. W., and Rassool, F. V. (1997). Precise localization of the FHIT gene to the common fragile site at 3p14.2 (FRA3B) and characterization of homozygous deletions within FRA3B that affect FHIT transcription in tumor cell lines. *Genes Chromosomes Cancer* 20, 16-23.
- Oostra, B. A., Hupkes, P. E., Perdon, L. F., van Bennekom, C. A., Bakker, E., Halley, D. J., Schmidt, M., Du Sart, D., Smits, A., Wieringa, B., and van Oost, B. A. (1990). New polymorphic DNA marker close to the fragile site FRAXA. *Genomics* 6, 129-32.
- Orita, M., Suzuki, Y., Sekiya, T., and Hayashi, K. (1989). Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction. *Genomics* 5, 874-9.
- 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.
- O'Shaughnessy, P.J., Dudley, K., and Rajapaksha, W. R. (1996). Expression of follicle stimulating hormone-receptor mRNA during gonadal development. *Mol Cell Endocrinol* 125, 169-75.
- Oudet, C., Mornet, E., Serre, J. L., Thomas, F., Lentes-Zengerling, S., Kretz, C., Deluchat, C., Tejada, I., Boue, J., Boue, A., and Mandel, J-L. (1993a). Linkage disequilibrium between the fragile X mutation and two closely linked CA repeats suggests that fragile X chromosomes are derived from a small number of founder chromosomes. *Am J Hum Genet* 52, 297-304.
- Oudet, C., von Koskull, H., Nordstrom, A. M., Peippo, M., and Mandel, J. L. (1993b). Striking founder effect for the fragile X syndrome in Finland. *Eur J Hum Genet* 1, 181-9.
- Overhauser, J., and Radic, M. Z. (1989). Encapsulation of cells in agarose beads for use with pulsed-field gel electrophoresis. *Focus* 9, 8-9.
- Pandis, N., Jin, Y. S., Limon, J., Bardi, G., Idvall, I., Mendahl, N., Mittleman, F. and Heim, S. (1993). Interstitial deletion of the short arm of chromosome 3 as the primary chromosomal abnormality in carcinomas of the breast. *Genes Chrom Cancer* 6, 151-55.
- Paradee, W., Mullins, C., He, Z., Glover, T., Wilke, C., Opalka, B., Schutte, J., and Smith, D. I. (1995). Precise localization of aphidicolin-induced breakpoints on the short arm of human chromosome 3. *Genomics* 27, 358-61.

Parimoo, S., Patanjali, S. R., Shukla, H., Chaplin, D. D., and Weissman, S. M. (1991). cDNA selection: efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. *Proc Natl Acad Sci U S A* 88, 9623-7.

Parimoo, S., Patanjali, S. R., and Weissman, S. M. (1993). Normalization and selection with short fragment cDNAs. In *Methods in molecular genetics*, K. W. Adolph, ed. (San Diego, California: Academic Press. Inc.), pp. 23-50.

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 [see comments]. *Nat Genet* 8, 229-35.

Pembrey, M. E., Winter, R. M., and Davies, K. E. (1985). A premutation that generates a defect at crossing over explains the inheritance of fragile X mental retardation. *Am J Med Genet* 21, 709-17.

Peterson, M. G. (1988). DNA sequencing using Taq polymerase. *Nucleic Acids Res* 16, 10915.

Phillips, H. A., Hyland, V. J., Holman, K., Callen, D. F., Richards, R. I., and Mulley, J. C. (1991). Dinucleotide repeat polymorphism at D16S287. *Nucleic Acids Res* 19, 6664.

Pieretti, M., Zhang, F. P., Fu, Y. H., Warren, S. T., Oostra, B. A., Caskey, C. T., and Nelson, D. L. (1991). Absence of expression of the FMR-1 gene in fragile X syndrome. *Cell* 66, 817-22.

Poustka, A., Dietrich, A., Langenstein, G., Toniolo, D., Warren, S. T., and Lehrach, H. (1991). Physical map of human Xq27-qter: localizing the region of the fragile X mutation. *Proc Natl Acad Sci U S A* 88, 8302-6.

Reed, K. C., and Mann, D. A. (1985). Rapid transfer of DNA from agarose gels to nylon membranes. *Nucl. Acids Res.* 13, 7207-7221.

Reyniers, E., Vits, L., De Boulle, K., Van Roy, B., Van Velzen, D., de Graaff, E., Verkerk, A. J., Jorens, H. Z., Darby, J. K., Oostra, B., and Willems, P. J. (1993). The full mutation in the FMR-1 gene of male fragile X patients is absent in their sperm [see comments]. *Nat Genet* 4, 143-6.

Richards, R. I., Holman, K., Friend, K., Kremer, E., Hillen, D., Staples, A., Brown, W. T., Goonewardena, P., Tarleton, J., Schwartz, C., and Sutherland, G. R. (1992). Evidence of founder chromosomes in fragile X syndrome. *Nat Genet* 1, 257-60.

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., Holman, K., Yu, S., and Sutherland, G. R. (1993). Fragile X syndrome unstable element, p(CCG)_n, and other simple tandem repeat sequences are binding sites for specific nuclear proteins. *Hum Mol Genet* 2, 1429-35.

Richards, R.I., Kondo, I., Holman, K., Yamauchi, M., Seki, N., Kishi, K., Staples, A., Sutherland, G. R. and Hori, T. (1994). Haplotype analysis at the FRAXA locus in the Japanese population. *Am J Med Genet* 51, 412-416.

Richards, R. I., Crawford, J., Narahara, K., Mangelsdorf, M., Friend, K., Staples, A., Denton, M., Easteal, S., Hori, T. A., Kondo, I., Jenkins, T., Goldman, A., Panich, V., Ferakova, E., and Sutherland, G. R. (1996). Dynamic mutation loci: allele distributions in different populations. *Ann Hum Genet* 60, 391-400.

Richards, R. I. and Sutherland, G. R. (1997). Dynamic mutation: possible mechanisms and significance in human disease. *Trends Biochem Sci* 22, 432-6

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 [published erratum appears in *Nat Genet* 1993 Mar;3(3):273]. *Nat-Genet* 2, 186-91 issn: 1061-4036.

Ritchie, R. J., Knight, S. J., Hirst, M. C., Grewal, P. K., Bobrow, M., Cross, G. S., and Davies, K. E. (1994). The cloning of FRAXF: trinucleotide repeat expansion and methylation at a third fragile site in distal Xqter. *Hum Mol Genet* 3, 2115-21.

Rosenberg, C., Vianna-Morgante, A. M., Otto, P. A., and Navajas, L. (1991). Effect of X inactivation on fragile X frequency and mental retardation. *Am J Med Genet* 38, 421-4.

Rousseau, F., Vincent, A., Oberlé, I., and Mandel, J-L. (1990). New informative polymorphism at the DXS304 locus, a close distal marker for the fragile X locus. *Hum Genet* 84, 263-6.

- Rousseau, F., Heitz, D., Biancalana, V., Blumenfeld, S., Kretz, C., Boue, J., Tommerup, N., Van Der Hagen, C., DeLozier-Blanchet, C., Croquette, M.F., Gilgenkrantz, S., Jalbert, P., Voelckel, M-A., Oberlé, I., and Mandel, J.-L. (1991) Direct diagnosis by DNA analysis of the fragile X syndrome of mental retardation *N Engl J Med* 325, 1673-81.
- Roy, K., Mitsugi, K. and Sirotnak, F. M. (1996). Organization and alternate splicing of the murine folylpolyglutamate synthetase gene. Different splice variants in L1210 cells encode mitochondrial or cytosolic forms of the enzyme. *J Biol Chem* 271, 23820-7.
- Saiki, R.K. (1989). The design and optimization of the PCR. In *PCR technology: principles and applications for DNA amplification*, Erlich, H.A., ed. (New York: Stockton Press), pp. 7-16.
- Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). *Molecular cloning: A laboratory manual*, Volume 1 (New York: Cold Spring Harbor Laboratory Press).
- 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.
- 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.
- Schlessinger, D., Little, R. D., Freije, D., Abidi, F., Zucchi, I., Porta, G., Pilia, G., Nagaraja, R., Johnson, S. K., Yoon, J. Y., Srivastava, A., Kere, J., Palmeri, G., Ciccodicola, A., Montanaro, V., Romano, G., Casamassimi, A., and D'Urso, M. (1991). Yeast artificial chromosome-based genome mapping: some lessons from Xq24-q28. *Genomics* 11, 783-93.
- Schmidt, M., Du Sart, D., Kalitsis, P., Fraser, N., Leversha, M., Voullaire, L., Foster, D., Davies, J., Hills, L., Petrovic, V., and Hutchinson, R. (1991). X chromosome inactivation in fibroblasts of mentally retarded female carriers of the fragile site Xq27.3: application of the probe M27 beta to evaluate X inactivation status. *Am J Med Genet* 38, 411-5.
- Schwartz, D. C., Saffron, W., Welsh, J., Haas, R., Goldenberg, M., and Cantor, C. R. (1982). New techniques for purifying large DNAs and studying their properties and packaging. *Cold Spring Harbor Symp. Quant Biol* 47, 189-195.
- Sealy, P. G., Whittaker, P. A., and Southern, E. M. (1985). Removal of repeated sequences from hybridisation probes. *Nuc. Acids Res.* 13, 1905-1922.

Shabtai, F., Klar, D., Hart, J., and Halbrecht, I. (1985). On the meaning of fragile sites in cancer risk and development. *Cancer Genet Cytogenet* 18, 81-5.

Sherman, S. L., Morton, N. E., Jacobs, P. A., and Turner, G. (1984). The marker (X) syndrome: a cytogenetic and genetic analysis. *Ann Hum Genet* 48, 21-37.

Sherman, S. L., Jacobs, P. A., Morton, N. E., Froster-Iskenius, U., Howard-Peebles, P. N., Nielsen, K. B., Partington, M. W., Sutherland, G. R., Turner, G., and Watson, M. (1985). Further segregation analysis of the fragile X syndrome with special reference to transmitting males. *Hum Genet* 69, 289-99.

Sherman, S. L., Rogatko, A., and Turner, G. (1988). Recurrence risks for relatives in families with an isolated case of the fragile X syndrome. *Am J Med Genet* 31, 753-65.

Short, J. M., Fernandez, J. M., Sorge, J. A., and Huse, W. D. (1988). Lambda ZAP: a bacteriophage lambda expression vector with in vivo excision properties. *Nucleic Acids Res* 16, 7583-600.

Simmers, R. N., Sutherland, G. R., West, A., and Richards, R. I. (1987). Fragile sites at 16q22 are not at the breakpoint of the chromosomal rearrangement in AMMoL. *Science* 236, 92-4.

Siomi, H., Siomi, M. C., Nussbaum, R. L., and Dreyfuss, G. (1993). The protein product of the fragile X gene, FMR1, has characteristics of an RNA-binding protein. *Cell* 74, 291-8.

Siprashvili, Z., Sozzi, G., Barnes, L. D., McCue, P., Robinson, A. K., Eryomin, V., Sard, L., Tagliabue, E., Greco, A., Fusetti, L., Schwartz, G., Pierotti, M. A., Croce, C. M., and Huebner, K. (1997). Replacement of Fhit in cancer cells suppresses tumorigenicity. *Proc Natl Acad Sci U S A* 94, 13771-6.

Sittler, A., Devys, D., Weber, C., and Mandel, J. L. (1996). Alternative splicing of exon 14 determines nuclear or cytoplasmic localisation of fmr1 protein isoforms. *Hum Mol Genet* 5, 95-102.

Slatko, B. E., and Albright, L. M. (1991). Denaturing gel electrophoresis for sequencing. In *Current protocols in molecular biology*, F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith and K. Struhl, eds. (New York: John Wiley & Sons, Inc.), pp. 7.6.1-7.6.13.

Smeets, D. F. C. M., Scheres, J. M. J. C., Scheres, J. M. J. C. and Hustinx, T. W. J. (1986). The most common fragile site in man is 3p14. *Hum Genet* 72, 215-20.

Smeets, H. J., Smits, A. P., Verheij, C. E., Theelen, J. P., Willemsen, R., van de Burgt, I., Hoogeveen, A. T., Oosterwijk, J. C., and Oostra, B. A. (1995). Normal phenotype in two brothers with a full FMR1 mutation. *Hum Mol Genet* 4, 2103-8.

Smith, S. S., Kan, J. L., Baker, D. J., Kaplan, B. E. and Dembeck, P. (1991). Recognition of unusual DNA structures by human DNA (cytosine-5) methyltransferase. *J Mol Biol* 217, 39-51

Smith, S. S., Laayoun, A., Lingeman, R. G., Baker, D. J., and Riley, J. (1994). Hypermethylation of telomere-like foldbacks at codon 12 of the human c- Ha-ras gene and the trinucleotide repeat of the FMR-1 gene of fragile X. *J Mol Biol* 243, 143-51.

Smits, A., Smeets, D., Hamel, B., Dreesen, J., and van Oost, B., (1992) High prevalence of the fra(X) syndrome cannot be explained by a high mutation rate. *Am J Med Genet* 1, 345-52

Snow, K., Tester, D. J., Kruckeberg, K. E., Schaid, D. J., and Thibodeau, S. N. (1994). Sequence analysis of the fragile X trinucleotide repeat: implications for the origin of the fragile X mutation. *Hum Mol Genet* 3, 1543-51.

Stallings, R. L., Torney, D. C., Hildebrand, C. E., Longmire, J. L., Deaven, L. L., Jett, J. H., Doggett, N. A., and Moyzis, R. K. (1990). Physical mapping of human chromosomes by repetitive sequence fingerprinting. *Proc Natl Acad Sci U S A* 87, 6218-22.

Stallings, R. L., Doggett, N. A., Callen, D., Apostolou, S., Chen, L. Z., Nancarrow, J. K., Whitmore, S. A., Harris, P., Michison, H., Breuning, M., Saris, J. J., Fickett, J., Cinosky, M., Torney, D. C., Hildebrand, C. E., and Moyzis, R.K. (1992). Evaluation of a cosmid contig physical map of human chromosome 16. *Genomics* 13, 1031-9.

Stallings, R. L., Whitmore, S. A., Doggett, N. A., and Callen, D. F. (1993). Refined physical mapping of chromosome 16-specific low-abundance repetitive DNA sequences. *Cytogenet Cell Genet* 63, 97-101.

Steen, A. M., Marcus, S., Sahlen, S., Nielsen, K. B., and Lambert, B. (1991). The fragile X mutation does not have any major effect on the expression of the hypoxanthine phosphoribosyltransferase (HPRT) locus in human fibroblasts. *Hum Genet* 87, 503-5.

Strachan, T., and Read, A. P. (1996). *Human Molecular Genetics* (Oxford: BIOS Scientific Publishers Ltd).

Struhl, D. (1987). Construction of hybrid DNA molecules. In *Current protocols in molecular biology*, F. M. Ausubel, R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith and K. Struhl, eds. (New York: John Wiley & Sons, Inc.), pp. 3.16.1-3.16.11.

Sutcliffe, J. S., Nelson, D. L., Zhang, F., Pieretti, M., Caskey, C. T., Saxe, D., and Warren, S. T. (1992). DNA methylation represses FMR-1 transcription in fragile X syndrome. *Hum Mol Genet* 1, 397-400.

Sutherland, G. R. (1977). Fragile sites on human chromosomes: demonstration of their dependence on the type of tissue culture medium. *Science* 197, 265-6.

Sutherland, G. (1979). Heritable fragile sites on human chromosomes I. Factors affecting expression in lymphocyte culture. *Am J Hum Genet* 31, 125-135.

Sutherland, G., and Hecht, F. (1985). *Fragile Sites on Human Chromosomes* (Oxford University press, New York).

Sutherland, G. (1985). The enigma of the fragile X chromosome. *Trends in Genetics* 1, 108-111.

Sutherland G.R., Gedeon, A., Kornman, L., Donnelly, A., Byard, R.W., Mulley, J.C., Kremer, E., Lynch, M., Pritchard, M., Yu, S. and Richards, R. I. (1991a). Prenatal diagnosis of fragile X syndrome by direct detection of the unstable sequence (comments). *N Eng J Med* 325, 1720-1722.

Sutherland, G. R., Haan, E. A., Kremer, E., Lynch, M., Pritchard, M., Yu, S., and Richards, R. I. (1991b). Hereditary unstable DNA: a new explanation for some old genetic questions? *Lancet* 338, 289-92.

Sutherland, G. R., and Richards, R. I. (1992). Anticipation legitimized: unstable DNA to the rescue [editorial]. *Am J Hum Genet* 51, 7-9.

Sutherland, G. R., and Richards, R. I. (1995a). Simple tandem DNA repeats and human genetic disease. *Proc Natl Acad Sci U S A* 92, 3636-41.

Sutherland, G. R., and Richards, R. I. (1995b). The molecular basis of fragile sites in human chromosomes. *Curr Opin Genet Dev* 5, 323-7.

Suthers, G. K., Callen, D. F., Hyland, V. J., Kozman, H. M., Baker, E., Eyre, H., Harper, P. S., Roberts, S. H., Hors-Cayla, M. C., Davies, K. E., Bell, M. V. and Sutherland, G. R. (1989). A new DNA marker tightly linked to the fragile X locus (FRAXA). *Science* 246, 1298-300.

Suthers, G. K., Hyland, V. J., Callen, D. F., Oberlé, I., Rocchi, M., Thomas, N. S., Morris, C. P., Schwartz, C. E., Schmidt, M., Ropers, H. H., Baker, E., Oostra, B. A. Dahl, N., Wilson, P. J., Hopwood, J. J., and Sutherland, G. R. (1990). Physical mapping of new DNA probes near the fragile X mutation (FRAXA) by using a panel of cell lines. *Am J Hum Genet* 47, 187-95.

Suthers, G. K., Mulley, J. C., Voelckel, M. A., Dahl, N., Vaisanen, M. L., Steinbach, P., Glass, I. A., Schwartz, C. E., van Oost, B. A., Thibodeau, S. N., Haites, N. E., Oostra, B. A., Schinzel, A., Carballo, M., Morris, C. P., Hopwood, J. J., and Sutherland, G. R. (1991a). Linkage homogeneity near the fragile X locus in normal and fragile X families. *Genomics* 10, 576-82.

Suthers, G. K., Mulley, J. C., Voelckel, M. A., Dahl, N., Vaisanen, M. L., Steinbach, P., Glass, I. A., Schwartz, C. E., van Oost, B. A., Thibodeau, S. N., Haites, N. E., Oostra, B. A., Gine, R., Carballo, M., Morris, C. P., Hopwood, J. J. and Sutherland, G. R. (1991b). Genetic mapping of new DNA probes at Xq27 defines a strategy for DNA studies in the fragile X syndrome. *Am J Hum Genet* 48, 460-7.

Szabo, P., Purrello, M., Rocchi, M., Archidiacono, N., Alhadeff, B., Filippi, G., Toniolo, D., Martini, G., Luzzatto, L., and Siniscalco, M. (1984). Cytological mapping of the human glucose-6-phosphate dehydrogenase gene distal to the fragile-X site suggests a high rate of meiotic recombination across this site. *Proc Natl Acad Sci U S A* 81, 7855-9.

Tedeschi, B., Vernole, P., Sanna, M. L., and Nicoletti, B. (1992). Population cytogenetics of aphidicolin-induced fragile sites. *Hum Genet* 89, 543-7.

Thibodeau, S. N., Dorkins, H. R., Faulk, K. R., Berry, R., Smith, A. C., Hagerman, R., King, A., and Davies, K. E. (1988). Linkage analysis using multiple DNA polymorphic markers in normal families and in families with fragile X syndrome. *Hum Genet* 79, 219-27.

Turner, G., Till, R., and Daniel, A. (1978). Marker X chromosomes, mental retardation and macro-orchidism [letter]. *N Engl J Med* 299, 1472.

Turner, G., Daniel, A., and Frost, M. (1980). X-linked mental retardation, macro-orchidism, and the Xq27 fragile site. *J Pediatr* 96, 837-41.

- Turner, G., Brookwell, R., Daniel, A., Selikowitz, M., and Zilibowitz, M. (1980). Heterozygous expression of X-linked mental retardation and X-chromosome marker fra(X)(q27). *N Engl J Med* 303, 662-4.
- Turner, G., and Jacobs, P. (1984). Marker (X)-linked mental retardation. *Adv Hum Genet* 13, 83-112.
- Turner, G., Webb, T., Wake, S., and Robinson, H. (1996). Prevalence of fragile X syndrome. *Am J Med Genet* 64, 196-7.
- Veenema, H., Carpenter, N. J., Bakker, E., Hofker, M. H., Ward, A. M., and Pearson, P. L. (1987). The fragile X syndrome in a large family. III. Investigations on linkage of flanking DNA markers with the fragile site Xq27. *J Med Genet* 24, 413-21.
- 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., Eussen, B. E., van Ommen, G. J. B., Blonden, L. A. J., Riggins, G. J., Chastain, J. L., Kunst, C. B., Galjaard, H., Caskey, C. T., Nelson, D. L., Oostra, B. A., and Warren, S. T. (1991). 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.
- Viljoen, D. (1988). Pseudoxanthoma elasticum (Gronblad-Strandberg syndrome). *J Med Genet* 25, 488-90.
- Vincent, A., Dahl, N., Oberlé, I., Hanauer, A., Mandel, J. L., Malmgren, H., and Pettersson, U. (1989). The polymorphic marker DXS304 is within 5 centimorgans of the fragile X locus. *Genomics* 5, 797-801.
- Vincent, A., Heitz, D., Petit, C., Kretz, C., Oberlé, I., and Mandel, J. L. (1991). Abnormal pattern detected in fragile-X patients by pulsed-field gel electrophoresis. *Nature* 349, 624-6.
- Voelckel, M. A., Mattei, M. G., N'Guyen, C., Philip, N., Birg, F., and Mattei, J. F. (1988). Dissociation between mental retardation and fragile site expression in a family with fragile X-linked mental retardation. *Hum Genet* 80, 375-8.
- Vogel, F. (1984). Mutation and selection in the marker (X) syndrome. A hypothesis. *Ann Hum Genet* 48, 327-32.

Voss, H., Wiemann, S., Grothues, D., Sensen, C., Zimmermann, J., Schwager, C., Stegemann, J., Erfle, H., Rupp, T., and Ansorge, W. (1993). Automated low-redundancy large-scale DNA sequencing by primer walking. *Biotechniques* 15, 714-21.

Wada, M., Little, R. D., Abidi, F., Porta, G., Labella, T., Cooper, T., Della Valle, G., D'Urso, M., and Schlessinger, D. (1990). Human Xq24-Xq28: approaches to mapping with yeast artificial chromosomes. *Am J Hum Genet* 46, 95-106.

Warren, S. T., Zhang, F., Licameli, G. R., and Peters, J. F. (1987). The fragile X site in somatic cell hybrids: an approach for molecular cloning of fragile sites. *Science* 237, 420-3.

Warren, S. T., Zhang, F. P., Sutcliffe, J. S., and Peters, J. F. (1988). Strategy for molecular cloning of the fragile X site DNA. *Am J Med Genet* 30, 613-23.

Warren, S. T., Knight, S. J., Peters, J. F., Stayton, C. L., Consalez, G. G., and Zhang, F. P. (1990). Isolation of the human chromosomal band Xq28 within somatic cell hybrids by fragile X site breakage. *Proc Natl Acad Sci U S A* 87, 3856-60.

Warren, S. T. (1996). The expanding world of trinucleotide repeats [comment]. *Science* 271, 1374-5.

Watkins, W. S., Bamshad, M., and Jorde, L. B. (1995). Population genetics of trinucleotide repeat polymorphisms. *Hum Mol Genet* 4, 1485-91.

Webb, T., Butler, D., Insley, J., Weaver, J. B., Green, S., and Rodeck, C. (1981). Prenatal diagnosis of Martin-Bell syndrome associated with fragile site at Xq27-28 [letter]. *Lancet* 2, 1423.

Webb, G. C., Rogers, J. G., Pitt, D. B., Halliday, J., and Theobald, T. (1981). Transmission of fragile (X) (q27) site from a male [letter]. *Lancet* 2, 1231-2.

Webb, G. C., Halliday, J. L., Pitt, D. B., Judge, C. G., and Leversha, M. (1982). Fragile (X)(q27) sites in a pedigree with female carriers showing mild to severe mental retardation. *J Med Genet* 19, 44-8.

Webb, T., Thake, A., and Todd, J. (1986). Twelve families with fragile X(q27). *J Med Genet* 23, 400-6.

Webb, T. (1991). Expression of fragile-X in a female fetus diagnosed after chorionic villus sampling. *Prenat Diagn* 11, 333-8.

Weber, J. L. (1990). Informativeness of human (dC-dA)_n.(dG-dT)_n polymorphisms. *Genomics* 7, 524-30.

Wilke, C. M., Hall, B. K., Hoge, A., Paradee, W., Smith, D. I., and Glover, T. W. (1996). FRA3B extends over a broad region and contains a spontaneous HPV16 integration site: direct evidence for the coincidence of viral integration sites and fragile sites. *Hum Mol Genet* 5, 187-95.

Winter, R. M., and Pembrey, M. E. (1986). Analysis of linkage relationships between genetic markers around the fragile X locus with special reference to the daughters of normal transmitting males. *Hum Genet* 74, 93-7.

Winter, R., and Pembrey, M. (1987). Interpretation of the heterogeneity in the linkage relationships of DNA markers around the fragile X locus [letter]. *Hum Genet* 77, 297-8.

Wöhrle, D., Kotzot, D., Hirst, M. C., Manca, A., Korn, B., Schmidt, A., Barbi, G., Rott, H. D., Poustka, A., Davies, K. E., and Steinbach, P. (1992). A microdeletion of less than 250 kb, including the proximal part of the FMR-I gene and the fragile-X site, in a male with the clinical phenotype of fragile-X syndrome. *Am J Hum Genet* 51, 299-306.

Wooster, R., Cleton-Jansen, A. M., Collins, N., Mangion, J., Cornelis, R. S., Cooper, C. S., Gusterson, B. A., Ponder, B. A., von Deimling, A., Wiestler, O. D., Cornelisse, C. J., Devilee, P., and Stratton, M. R. (1994). Instability of short tandem repeats (microsatellites) in human cancers. *Nat Genet* 6, 152-6.

Wyman, A. R., and White, R. (1980). A highly polymorphic locus in human DNA. *Proc Natl Acad Sci USA* 77, 6754-6758.

Yanisch-Peron, C., Vieira, J., and Messing, J. (1985). Improved M13 phage cloning vectors and host strains: Nucleotide sequence of the M13mp18 and pUC19 vectors. *Genet* 33, 103-19.

Yokoyama, S., Yamakawa, K., Tsuchiya, E., Murata, M., Sakiyama, S. and Nakamura, Y. (1992). Deletion mapping on the short arm of chromosome 3 in squamous cell carcinoma and adenocarcinoma of the lung. *Cancer Res* 52, 873-77

Yu, W. D., Wenger, S. L., and Steele, M. W. (1990). X chromosome imprinting in fragile X syndrome. *Hum Genet* 85, 590-4.

Yu, S., Pritchard, M., Kremer, E., Lynch, M., Nancarrow, J., Baker, E., Holman, K., Mulley, J. C., Warren, S. T., Schlessinger, D., Sutherland, G. R., and Richards, R. I. (1991). Fragile X genotype characterized by an unstable region of DNA. *Science* 252, 1179-81.

Yu, S., Mulley, J., Loesch, D., Turner, G., Donnelly, A., Gedeon, A., Hillen, D., Kremer, E., Lynch, M., Pritchard, M., Sutherland, G. R., and Richards, R. I. (1992). Fragile X syndrome: unique genetics of the heritable unstable element. *Am J Hum Genet* 50, 968-80.

Yu, S., Mangelsdorf, M., Hewett, D., Hobson, L., Baker, E., Eyre, H. J., Lapsys, N., Le Paslier, D., Doggett, N. A., Sutherland, G. R., and Richards, R. I. (1997). Human chromosomal fragile site FRA16B is an amplified AT-rich minisatellite repeat. *Cell* 88, 367-74.

Yunis, J. J., and Soreng, A. L. (1984). Constitutive fragile sites and cancer. *Science* 226, 1199-204.

Zhong, N., Yang, W., Dobkin, C., and Brown, W. T. (1995) Fragile X gene instability: anchoring AGGs and linked microsatellites. *Am J Hum Genet* 57, 351-61.

Appendix 1

*FRA16A region
primers and DNA sequence*

Table of oligonucleotide primers used for *FRA16A* studies

no.	primer name	primer sequence (5' to 3' orientation) reverse orientation	position within 7407 bp sequence
1	pgdna16A1-1	CAG TCA GGA TGG GAA AGA ACT CCA C	568-592
2	pgdna16A1-2	GGA AAC AGG TCT TGC TCT TGA AAG T*	590-615
3	pgdna16A1-3	CTC TGA CAC TCC TGA AAT GTT TCT C	970-994
4	pgdna16A1-4	ACA GTG CGT GGA GGA AGC TCC TAG C*	1056-1080
5	<i>d</i>	ATC CTA CTC CCA CTA CGT CCT GAG G	1238-1262
6	<i>e</i>	GCC TTC CCC ATC ACC CTC CCC TCC A	1374-1398
7	<i>c</i>	TTT TCA GAC GGG CAG GGA CCC G*	1541-1562
8	<i>b</i>	CCG GGT CCC TGC CCG TCT GAA AA	1541-1562
9	<i>a</i>	CCG GCT GCC GCT CGG GCT CCC GCT*	1678-1701
10	<i>f</i>	CAC GCT CCG GGC CGC CGC CGC GCT C*	1721-1745
11	pgdna16A1-5	GAG CGC GGC GGC GGC CCG GAG CGT G	1721-1745
12	pcdna38-1	CCG GCT GCT GTC CCC GCG GTT TTC*	2083-2106
13	pcdna38-2	GTA AAA ATG CCC TTT CGG CTC GCG	2139-2161
14	pcdna38-3	GCG ATG TGG AGT CGG TAG GCT TGC*	2242-2265
15	pcdna81-1	CCT AAC TTC TCC AGG GAT ACC TTT	3033-3056
16	pcdna81-2	CTG AGA CCT TGG CTC TGA AAT GGG*	3131-3154
17	pcdna81-3	GGA TGG TTT CAG GAT CTG AAG ACT	3278-3302
18	pcdna81-4	AAG CAT CAC GCT GTG TAC TAA GGC*	3352-3375
19	pcdna10-1	CTC TTG GAG TTG ATA GTG ATG	3631-3651
20	pcdna10-2	GGC TTA CTG TAT CAT TTG CAT CTG	3784-3807
21	pcdna10-3	CAG ATG CAA ATG ATA CAG TAA GCC*	3784-3807
22	pcdna10-4	CAG GAG GTA CCT TTA GGA TG	3824-3843
23	pg312B11-1	GAG CTC CAG GAG ACC AGA AGG	4420-4440
24	pg312B11-2	CCT TGC CTG CAA CCC AAG CT*	4474-4493

no.	primer name	primer sequence (5' to 3' orientation) *reverse orientation	position within 7407 bp sequence
25	pg312B11-3	GAG AGG ACC AGA CCT TTT CGC A	4493-4514
26	pg312B11-4	TCC TGA ATC TCC CTC CAA AG*	4579-4598
27	pcdna68a-1	TCC AGA TGG CTC TTG TAG AGA G	4633-4652
28	pcdna68a-2	GCC TGC CTG ACA CTC TCT GCA*	4785-4805
29	pcdna68a-3	ATT TGG CTT AGG AAA TGC CC*	4816-4835
30	pg312B11-5	TTC TGA GAT GAC AGT AAA ATC CAC AGC*	5118-5144
31	pcdna43-1	GGA GAA CTG TCA CAA GGA TCA	5190-5210
32	pcdna43-2	ACT TCA AGA AGA GTC ACT AAG*	5291-5311
33	pg312B11-6	ACT TAC TGA GTA ACA GTT TAT CTA A*	5335-5359
34	cdna12-1	TCT GAG ACC ACA GAA AAC AGG	5628-5648
35	cdna12-2	CTG TAT ACA AAG TGT AAT GCG T*	5713-5734
36	cdna54-1	ACC TAG GAA GCC AGC CAT GGC GTA AG	5989-6014
37	cdna54-2	GGA AAT TTG GTG GCT TGA GTT AAA GC	6066-6091
38	cdna54-3	CTT ATG TCA GAG ATA TGC CTA AGG*	6154-6177
39	cdna54-4	CCT TAG GCA TAT CTC TGA CAT AAG	6154-6174
40	pg310A11-1	ATG GAT TTA TGA CCT TAT GGA TC*	6231-6253
41	pcdna72-1	TGC GGA CCC AGA TCC CCA TCT ACC AG	6406-6431
42	pcdna72-2	TTA GAA TGC AGG TTT TGT TAA GAT C*	6485-6509
43	pcdna61-1	GGT TCA CCT CCT GGC TCG GCC	6614-6634
44	pcdna61-2	TAC AGC CCT ATG AGG TAG GTC AG*	6690-6712
45	pcdna61-3	TTC GTG CAG TTA TCC ATT TCC CCG	6826-6849
46	pcdna68b-1	CCC CAG CTC TGT GAA CTT CAT GAG*	7268-7391

Explanation of primer names: 1. The prefixes 'pcdna' and 'pg' indicate the primer was designed from cDNA or genomic sequences, respectively; 2. Subsequent numbers and/or letters indicate the origin of the cDNA or genomic clone from which the primer sequence was derived (i.e. isolate from cDNA pool, or cosmid).

Combined sequence (7407 bp) of *FRA16A* region genomic and cDNA clones

1	ACAGGCATG	AGGCCCCCCC	GCCCCAGGCC	TATCTCACTA	TATTTTAATG	50
	ATTAACATCC	ACATCAGGCT	TCTCTAGACC	TCAGGCTTGC	TGAAGATGAG	100
	ATTAAACCTC	ATTACACCTT	GGATGCGTGC	ACCTGGCCCA	GTGTCTGACA	150
	CAAGCTTTGT	ATTACACGAT	GCTTGTGGAG	TGGATGAAAC	AATCTCTTCC	200
	CACTCCCTGC	CTCATTGTCT	GATGAGCCTT	CATTACACTC	CACTTTCTTC	250
	TCTTTGTAAA	TTAGGATGAG	GTGAAAAAGC	ACAGCAGACA	CAATTCACAA	300
	ACTGATTTGA	TCTGCTGTTG	TATTCCCAGA	GTGCTAGGTC	CCGTATGGAC	350
	CACCCAAACG	CCATCCCATA	TGCTCACTGT	TCTGCGCTGC	TGTGTATAAA	400
	GCCCTGGGCA	AGAGGACAAG	GGAGGGAAAA	ACCAGAAAGA	CTGAGTCAGA	450
	GAAGGCTCGG	GATGAAATTT	TGAGTCCATT	GGGATCATT	TGCCTAAACG	500
	ACAGAGGAGA	AAGCTTGGGC	TCAGGGAGGA	GTCTCCTCC	TCCTGGCTCA	550
	AGGTCACCGG	GACAGACCAG	TCAGGATGGG	AAAGAACTCC	<u>ACTTTCAAGA</u>	600
	<u>GCAAGACCTG</u>	<u>TTTCCCCAGG</u>	ACCTGGGCCC	TGTGCAGACA	<u>GGAAGGGGGA</u>	650
	CTTCAGGATG	GATGGGGAAA	AGGACACGCT	GGGAGGCTGC	GGGGCCAGCC	700
	TTTGGGGCTT	GCATCCTGCG	AAGGTTGAGT	AACTCCGACG	GTCTCCCTTT	750
	TGGCGCTATT	CCACTTGCCA	ACTGGGTTTG	GGAGTCTCAT	GGCTTGTGAG	800
	TCCTCAGGGA	CCAGAGAAGT	GACTCAGTGA	ACACTTAGTA	CATCGTTGGC	850
	ACCGGGGATT	GTGTGTGATT	CAGGAGTCAC	TCACAGAGCT	TGTGCTGATT	900
	AACTCCAGAA	AATCAACCAA	AGCCTCCTCC	CCACCCCTTC	AAACCCAGGA	950
	TGAAGGGATC	AGACTGTGGC	TCTGACACTC	CTGAAATGTT	<u>TCTCACGTGT</u>	1000
	CCACCTTCCA	ACCGGTCTGG	CACCGACAGA	AGATCCCTGG	ACCGCGAACT	1050
	<u>CCGGGGCTAG</u>	<u>GAGCTTCCTC</u>	<u>CACGCACTGT</u>	CTCCTTCCAG	AGGACCACAG	1100
	GGAAACAGG	ACCTCGGGAC	CCCCTCTCTT	CCTGACCTCC	CCTGGATTCT	1150
	GAGCGGGCCC	ACCCTCAACC	AAAGGTCCCC	TGTTTCGCGG	GCCCTGCCTC	1200
	TCTTTTTCCG	CCTTGGAAT	TGACCCCTC	GGCCCCC	<u>CTACTCCAC</u>	1250
	<u>TACGTCCTGA</u>	GGGGTGTCTG	TCTTCTCTGA	TCGCCCCAC	CCCCTTCCTT	1300
	TCCCTCCTCC	TTTTCTCTCC	TCGGCTCCTC	CCCCAGGCC	CGCCCTCTCC	1350
	GGCTCGGCCC	GCGTCCCCC	GGGGCCTTCC	<u>CCATCACCT</u>	<u>CCCCTCCAGA</u>	1400
	GGGGACAGGG	GTGTGGGGAG	GGGGCGCCGC	GGGCCAGCGC	CCGCCCTCCC	1450
	TCGCGCCCCG	TCCCCGAGCG	CGCCCGGCGC	CCCCCTCCGC	GGCCTCCCCG	1500
	CTCCCGCCTC	CCGCCTTCTC	CTCCTCCTGG	TTGACGTCCC	<u>CGGGTCCCTG</u>	1550
	<u>CCCGTCTGAA</u>	<u>AACACGCGGC</u>	GCGGCGGTCC	AGGCGGCGGT	GGAGGCGGCG	1600
	GCGGCGGCGG	CGGCAGCGGC	GGCGGCGGCG	GCGGAGGAGG	AGCAGCGGCG	1650
	ACGCGAGGCG	GCGGCGGCGG	CGGCCCGAGC	GGGAGCCCGA	<u>GCGGCAGCCG</u>	1700
	<u>GGCGGCCGCG</u>	GAGCTGCGGG	<u>GAGCGCGGCG</u>	<u>GCGGCCCGGA</u>	<u>GCGTGCCGGG</u>	1750
	GTCCCCGCTC	CTCGCTCGCC	TGCCGCGCTC	CGAAGATGGT	GGCGGCGCCG	1800
	TGCGCCCGGA	GGCTGGCCCG	GCGCTCGCAC	TCGGCGCTGC	TCGCGGCGCT	1850
	CGCGGTGCTG	CTGCACACGC	TGGTCGTGTG	GAATTTACAGC	AGCCTCGACT	1900
	CCGGGGCCCG	GGAGCGCCGC	GGGGCAGCGG	TCGGCGGCGG	GGAGCAGCCG	1950
	CCCCCGGCC	CGGCCCGCG	CCGAGCCCGG	GACCTGCCCG	CCGAGCCGCT	2000
	GCAGCCCGAG	GAGGAGGAGG	AGGCGGCGGC	GGAGGAGGAG	GAGGACGGGG	2050

GCCCCAGGCG	CGGGCGCGGG	GAGGCGGCCC	<u>CGGAAAACCG</u>	CGGGGACAGC	2100
<u>AGCCGGCCAG</u>	CCGGGGGGCA	CTGCCC GCCC	GGGCTCTGGT	<u>TAGATGCCCT</u>	2150
<u>TTCGGCTCGC</u>	<u>GCCCCGCGGC</u>	GAGGGAGGGA	GGAAGGCAGC	CCGACCCAC	2200
GCCCCCCTAG	CCCTTGA ACT	CGGTCATCAA	ACTGGGACTC	<u>TGCAAGCCTA</u>	2250
<u>CCGACTCCAC</u>	<u>ATCGCCTCTT</u>	GCCCCGGGAA	GCCGGGTGGG	GTCCCTGCCA	2300
AGGCTCAGGG	GTCGTGGGCT	GCCCCTTTCA	GAGTCTGAGC	ACGGACCCCT	2350
TGGGGTCTGG	GGGAAGTTTT	CTCCAGCGGG	GGCGCTTTGG	GGGTACCGG	2400
AGCCTGGCTG	GGTTCCTGTT	CCCCATCACT	CTTGCTGGGG	ACCCAGGTG	2450
ACTTGCTCCG	TGCGTGTAGG	TCTAGGTGAC	TCGCTCCCTG	CCTGAGTGAG	2500
TGGATCTCCC	CGGATCTTCA	GCCGCTGTTC	CTGGACGGTT	GTGCGAGCAG	2550
CATCTCTCTG	GTTTTTGCCG	AGCTGGGTGC	AGACAGTTGA	GCCCTCTGTT	2600
CGTCCTGGCC	TGCGATGAGG	GCTGAGCTGG	GCGTCCCAAG	TCCCGACCA	2650
TGAGCCTGAA	TCAATACTCA	GTCCGGGGTG	TCTGGTCTAA	CTTGGGGTCT	2700
CCTGTTTCTC	TCCTCTGGAT	GTTTGAGCCT	TCCGTCTTCA	CCTTCTCTGT	2750
CTCTAACCTT	GCCGTCTCCC	GTGCCCTCTG	TGACTTTATC	TCCTCCCCAT	2800
CTCCAGTTTC	AGGGCATGGC	TCCTTCTCCT	AGCTGGATGT	CTCTGTCTTT	2850
CTGTCTTGTT	CGGCCGGTTC	AGGTCACAGC	TGTGTCCGAC	CACACCGTCA	2900
GGGCTTCCAC	TCCTATGGGA	TTACTTTGCC	CTTCGTGGGT	GTGAACAGTC	2950
AGGGCTCTGA	AACTGTTTTT	TCCGCAGATC	TGGGGCCATC	TGGGTGGCTG	3000
GGTGGGGAAC	TGACTTGTCT	TCCTACCCTG	<u>TCCCTAACTT</u>	<u>CTCCAGGGAT</u>	3050
<u>ACCTTTCCCA</u>	CCTTCTCTGG	GCTGCCTTGC	CCAGCCTACT	GAGCTAGGAG	3100
CAGGAGCACA	TGGGGCAGGA	GGCCGGAGTC	<u>CCCATTT CAG</u>	<u>AGCCAAGGTC</u>	3150
<u>TCAGTTCTCC</u>	AGGGGCGTGA	GTTCTGCCTT	AGGGCATAGC	CAATGTGTTG	3200
TGTGTCTGTC	TGTTTCTTCA	TCACACTACT	TGTCTGCAAG	GATTTGCCTA	3250
GCCGTTAAAG	GATATTAGAC	<u>ACCAGGGGGA</u>	<u>TGGTTTCAGG</u>	<u>ATCCTGAAGA</u>	3300
<u>CTGAATGCAT</u>	ATTTCTATCA	TTTGTCACTT	TGGAAGTCCT	CACTGAAGAC	3350
<u>TGCCTTAGTA</u>	CACAGCGTGA	<u>TGCTTTCTGA</u>	GTTGCTAAAC	TTCTTCGGAG	3400
TCTGTGCGAC	CCTTTCTGCC	GGAGTGCAGC	TCTTCATGCC	TGAGAACTGA	3450
GCCAGTCCCA	AGACGTGATG	TGATGGCATT	TCACGTTGCT	GTGTGGGTAG	3500
CAGGGAGAAG	TCTTTCCCTC	ATTTGCGGGG	AGAATAGTGG	CCTGAGCTGC	3550
CTGACGGAGA	ACCAGAACAT	TCCTAGTGTC	CTGCGATCTC	TTCTTTACAG	3600
AGAGGGGTAA	TTTGAAACGC	TGCCTTTGAA	<u>CTCTTGGAGT</u>	<u>TGATAGTGAT</u>	3650
<u>GGGTTCCCGC</u>	AGAAATGCAA	TTAGAAGTTC	TTTGTGCATC	CAGAGCCCAG	3700
GAGGTGGGGT	GGGAGGGTTG	TGGGTCGTGT	CCTGCTGGTC	CAGTGTAAGA	3750
ACGTGTAGGG	GAGACATGGT	GGGGGAAGCC	<u>TGGGGCTTAC</u>	<u>TGTATCATTT</u>	3800
<u>GCATCTGGGG</u>	GTCAGAGTGG	<u>GTT CAGGAGG</u>	<u>TACCTTTAGG</u>	<u>ATGCATGTAC</u>	3850
<u>CACCTGTGTA</u>	AACAAATGAA	GAGAAGGAAG	GAGGCAGGGC	TAGCTGATTT	3900
CTTCATAGGT	CCCCGGCATT	GTGAGCATCC	ACCCCCAGCC	CCTCTCTGTG	3950
TGCTCGGCTC	AGTGTTTTAC	TCTCTCACAC	CCAGCTCAGC	CTCTGAGAGA	4000
TCTGCGGTTG	GGGAGGTAGG	TGGGAACACA	CTATTTTCCT	TTTTACTTCC	4050
TTTTGGAATT	GCAGTTCCAG	CAGCGGAGGT	GCTTTGCTTT	TCTGAAACTC	4100

TGCTTCCAAC	TGCAAGTTCA	GTTACTCTCC	CCAAATCTGC	ATGTCCTAAT	4150
GTCTCAATTT	CTTTACCTTC	CAAGATGGAC	ATTTAAGTAT	TTTTTTTTTC	4200
TTTTGGTTCT	CTTCTCTTCT	TGGTGGCTGT	CTACCATGCT	GGTGACTGTT	4250
GTCTCTTGGG	GCGTTGTATT	TTATTTTCATT	TGCTTGCTCTC	CTTTCTTTCC	4300
TCTCCCCCA	CACCTACTCC	ATGACCTGCT	AAATCCTGAA	ATCACAGTAG	4350
AGACAGGCTG	GCTGGAGCCC	GGGGTCTGAC	CTACTCTGAC	TGTGCCCACC	4400
CAGTTCCTGC	CCTTCCATGG	AGCTCCAGGA	GACCAGAAGG	GGGCAGGCTG	4450
TTCCCAAAGT	CTGGTATAGC	TGGAGCTTGG	GTTGCAGGCA	AGGAGAGGAC	4500
CAGACCTTTT	CGCAAGCAAT	AGAGGCCGTT	GAAATGTGCA	GTTTGTTC	4550
AGACCCTGCG	GAGGCTGCCA	GGATGATGCT	TTGGAGGGAG	ATTCAGGAGT	4600
GGGACATTTA	AAGAAACAAA	TCCCCGTGAT	TTTCCAGATG	GCTCTTGTAG	4650
AGAGGTGGTT	AGACACATGC	TTAGAAGAAG	ATGCTTTAGC	TCAGTGCTCT	4700
GTTGCAGGTG	GGGATGGGGT	GGGTGAACAC	CCTTGGCTCT	TGCACGTAGA	4750
GAGAACCTGT	GTCCTGGGTT	GCTGGTGGGT	GTGTTGCAGA	GAGTGTGAGG	4800
CAGGCATTGT	GTATAGGGCA	TTTCCTAAGC	CAAATCTCCC	TCCTTCCTGA	4850
TGCCCCACAGC	CTCAAAGCAT	TTGTCCTTTT	TATGTGGGGG	GACCTGCGCT	4900
TCGGTGAAG	GCCCTGATTG	TATTTTTGGT	GCACTTAAGA	AGCCAAGAGG	4950
AAACACCCCG	TTTCTCAAGT	AGACAGAGTC	AGAGGGCATG	TCCATAAAAG	5000
ATTTGCAAGC	TCTCCTTTTC	CTCTCAGCCT	CTTTTAAAGC	TCTTAAATAT	5050
TAGAATTATT	GGGTGGATTG	CAGCTGTTTT	GTGGCTGACA	CTGATTCCTT	5100
CACTGGGTCT	CCAGTAAGCT	GTGGATTTTA	CTGTCATCTC	AGAATTTAAG	5150
AAACAGATGG	ACTCCCTTAT	AAAGGGAGGA	AAAATATTGG	GAGAACTGTC	5200
ACAAGGATCA	GCCTACAGGT	TCTGAATAGA	TTTTGGAGAA	AATGAAAGAG	5250
AGCTGCGTGT	GTTAATGTAA	TGAATGTTTA	TTTTCTTTCC	CTTAGTGACT	5300
CTTCTTGAAG	TCCCCACATT	TATTGCTCGT	GTTCTTAGAT	AAACTGTTAC	5350
TCAGTAAGTT	TCATTTAAGA	AATCTAATTT	TTTATTACCC	TTCCCACCCC	5400
CCGCTGCCG	CCCTACATAA	GCACACAAAG	TAGAATGGGG	TTTCTGTATT	5450
ATCCGACGCC	TTTGAAATGT	TGATTTTTTCG	GCAGAAATTC	CTCATGCTTC	5500
ACATTACTCA	GATAACTCAG	GACTCTTTTA	GACATCAGCA	GTGACTTTGA	5550
AAGATCGGTT	CACTGGCTTT	AGATCACCTC	TGCTTCTAAC	AGCCTTGGAG	5600
GTATAGCTTT	TCTTCCTATC	CAGATGTTCT	GAGACCACAG	AAAACAGGGT	5650
ATGTTATTGA	TGGTTCGTTT	TTTTTTTTTCA	CACTTTCAAT	GGAGTTTTGT	5700
TTTGTTTGAG	CGACGCATTA	CACTTTGTAT	ACAGAAATAT	GCATCATTGA	5750
ATAATATTTA	TCTTACTGCA	AATGTTGCCA	TCTGTTTATC	ATGGAAAGTG	5800
TTAAATAGAG	TAAATAAAAG	TAGAAATAAT	ATACGTAGGT	ATAGATATGT	5850
TTGTAATATG	AATTCTAATA	TATCTATGTA	TTCACTTTCC	TTCCGGTCTT	5900
TTCATACTTT	AGACGTGCTC	TTCAGTTTAT	CAATTCATTT	TTTAAAAATC	5950
TTCCAGAAGG	GTGGGTGGTC	AAAGGCTTGC	TATTTTTTAC	CTAGGAAGCC	6000
AGCCATGGCG	TAAGTTCTTA	TTTACTGGTG	ACTGTTCAAG	TTTGATTTCAT	6050
GGTTTTAATA	GAAGAGGAAA	TTTGGTGGCT	TGAGTTAAAG	CAGAGATTTG	6100
ACATCATCGC	GTCATTGTGT	ATGTGTGTGT	GTATCCTCAG	AAAGCAAAGA	6150

TATCCTTAGG	CATATCTCTG	ACATAAGAGT	CTGCCATGCT	GCCGAAGCTG	6200
38		39			
CCCCCCACAC	CCCGCTCCAG	AGCTGTTTAT	GATCCATAAG	GTCATAAATC	6250
CATCTATATG	AGTGTTTCAG	TGCAGAAGAA	CTTAACACAC	ACACCTGTTG	6300
40					
GACTGCACAG	GTGAGGTTGA	GAGTCTCTAT	CTTTTGACCA	GTGCTACCTT	6350
AAGAGCTGCC	ACTTATGATC	TCTCTCTTTC	CTTTTCCTTT	GCATTTCCCC	6400
ATTTCTGCGG	ACCCAGATCC	CCATCTACCA	GAAACAAAGA	GAAAATGCCT	6450
41					
TGGGGAAAAT	AAAAAATAAC	TAAAAAACAA	ACTCGATCTT	AACAAAACCT	6500
GCATTCTAAG	GAAGGGGTGT	GGGAACAGGA	AAAAACGTCC	CAACAAGTAT	6550
42					
GTGCGAGGCC	CCGGCACCAA	TGAAAGAGGC	CTCTGTTGAG	TACCCTCAGG	6600
CCTGACATAC	CTGGGTTTAC	CTCCTGGCTC	GGCCCTTTCT	CATGCCCCTA	6650
43					
ATCTTGGGCA	AGTCACTGTA	ATTTCGGTCC	TCTGTGAATC	TGACCTACCT	6700
CATAGGGCTG	TATGAGGATG	AAATGTAGCA	ATATATGTGA	ACTCATGTGG	6750
44					
GCCTGTCCCT	TAATCAGTGT	TCAAAAATGA	GCTGTTCTTA	AAAGGAAATT	6800
CTTGGTATGT	TTGAAAATAA	GGAATTTTCGT	GCAGTTATCC	ATTTCCCCGT	6850
45					
ACAGTGTAGG	GTGCGAGTAA	ATCCAAGGCA	GTAGGCAATG	AATATAGCGT	6900
TCGTGTGAGC	AGGACGGAAT	GTGGTAGAGG	AAGGAGTACT	GACATTTATT	6950
GAGTGCCTGC	TGTATGTCAG	GTGTCGTTTA	GGTATTTTGC	ATACGTTAGG	7000
TTATTAAATG	CTCACCCCCC	CCCCCCCCAA	AAAAAAAACA	CAAAAATTTT	7050
CAGGAAGGAA	GACGGTATTA	CACAACTGAT	AAAAAGTTAA	AACTAAGGAC	7100
ACCCAAC TAG	AAATCGCTGG	TCTTATGATG	GCTTGTTATG	TGGCAGGAAT	7150
AATGGAGGGA	TTTGTAATG	TTACCTTACT	GTCCTACTGA	TTTCTCCCAA	7200
CTTTGTGGGA	TAAGACTTTT	GCTGTCCCCT	GGTGGAAGCT	GAAGCTCAAA	7250
GAAGTAGAAG	CTGCTTGCTC	ATGAAGTTCA	CAGAGCTGGG	GAGTAGCAGA	7300
46					
ACTGAGATCC	AAACCCAGGC	CTGCCCCCTT	CCAAAGATGC	CACCTTTCCG	7350
AGTTGCGTTC	CTTTTGCATG	GTTTGTTGAA	AAACCCTAAA	TGCAACCATG	7400
CATTCCC	7407				

Blue arrows represent the position, direction and size of primers complementary to the sequence. Red arrows represent the same for primers reverse complementary to the sequence. Arrow-heads are shown at the 3' ends and dots at the 5' ends of primers.

Appendix 2
Publications

Publications arising from the results presented in this thesis are given below. Reference to the appropriate chapter(s) in the thesis is indicated. Following this page, copies of the papers are bound in the thesis.

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. (Chapters 3 and 4)

Nancarrow, J. K., Holman, K., Mangelsdorf, M., Hori, T., Denton, M., Sutherland, G. R., and Richards, R. I. (1995). Molecular basis of p(CCG)_n repeat instability at the *FRA16A* fragile site locus. *Hum Mol Genet* 4, 367-72. (Chapter 5)

Nancarrow, J.K., Kremer, E., Holman, K., Eyre, H., Doggett, N.A., (et al.), (1994)
Implications of FRA16A structure for the mechanism of chromosomal fragile site
genesis.
Science, v. 264 (5167), pp. 1938-1341.

NOTE:

This publication is included in the print copy
of the thesis held in the University of Adelaide Library.

It is also available online to authorised users at:

<http://dx.doi.org/10.1126/science.8009225>

Nancarrow, J.K., Holman, K., Mangelsdorf, M., Hori, T., (et al.), (1995) Molecular basis of p(CCG)_n repeat instability at the *FRA16A* fragile site locus. *Human Molecular Genetics*, v. 4 (3), pp. 367-362.

NOTE:

This publication is included in the print copy
of the thesis held in the University of Adelaide Library.

It is also available online to authorised users at:

<http://dx.doi.org/10.1093/hmg/4.3.367>