MOLECULAR ANALYSIS

OF A RARE FOLATE SENSITIVE FRAGILE SITE

LOCATED AT 19p13.1, *FRA19B*

by

Rhonda L. Mogk

A Thesis
Submitted to the Faculty of Graduate Studies
in Partial Fulfillment of the Requirements
for the Degree of

MASTER OF SCIENCE

Department of Human Genetics
University of Manitoba
Winnipeg, Manitoba

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

*****

COPYRIGHT PERMISSION PAGE

Molecular Analysis of a Rare Folate Sensitive Fragile Site Located at 19p13.1, *FRA19B*

BY

Rhonda L. Mogk

A Practicum submitted to the Faculty of Graduate Studies of The University of Manitoba in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

TABLE OF CONTENTS

ABSTRACT

Fragile sites are chromosomal loci that stain poorly and exhibit uncondensed gaps and DNA breaks on metaphase chromosomes *in vitro.* To date, over 100 fragile sites have been identified. They are divided into two classes, rare and common, based on their frequency within the population and their method of induction (folate stress, excess thymidine, etc.). Thirty rare fragile sites have been identified, three of which have been associated with a clinical phenotype. Currently, expansions of $p(CGG)_n$ trinucleotide repeats have been identified as the underlying molecular etiology of rare folate sensitive fragile sites. Trinucleotide expansions of other repeat structures (CAG, GCG and GAA), while not associated with fragile sites, have also been implicated in various neurodegenerative diseases. Anticipation, defined as an increased severity and decreased age of onset in subsequent generations, is a hallmark of trinucleotide expansion diseases and has also been indicated in certain psychiatric disorders. Thus, it has been speculated that the molecular etiology of psychiatric disorders also involves trinucleotide expansions.

Two families with the rare folate sensitive fragile site *FRA19B*, located at 19p13.1, have been identified previously. Clinical phenotypes, including schizophrenia, autism and mental retardation, exist within one of these families; however, a clinical phenotype is not apparent in the second family. It is hypothesized that the molecular basis for *FRA19B* is an expanded CG-rich trinucleotide repeat, which may affect a gene(s) important for normal mental

function. Thus, molecular analysis of *FRA19B* was undertaken in an attempt to identify: 1.) the basis of chromosomal fragility at the *FRA19B* locus and 2.) a potential gene(s) within *FRA19B*.

Molecular analysis of *FRA19B* was made possible by the availability of cosmids 25308 and 16864, shown previously to span this fragile site. However, attempts to clone *FRA19B* by classic methodologies were unsuccessful, due to a high proportion of *Alu* repetitive elements within these cosmids. The presence of *Alu* repeats was used advantageously by a technique termed *Alu*-splice PCR, originally designed to isolate potential exons. Twenty-seven *Alu*-splice PCR products were sequenced. Eleven containing unique DNA were utilized as molecular probes and localized to various restriction fragments of cosmid 25308. Six probes provided a unique hybridization signal against human genomic DNA. None, however, identified a DNA expansion. Nonetheless, one probe hybridized to two mRNA transcripts, approximately 4.5 and 7.0 kb, which are expressed in at least eight human tissues including heart, brain, placenta, lung, liver, skeletal muscle, kidney, and pancreas. The structure and function of this gene, and the effect of *FRA19B* on its expression, have yet to be determined.

ACKNOWLEDGEMENTS

To begin, I would like to thank Dr. Angie Dawson, my supervisor, for encouraging me to become a M.Sc. student. This thesis would not have been possible without you. In addition, thanks to you I've had the opportunity to work in the field of human genetics since 1994. Thanks Angie for all you've done for me over the last 5½ years. I've learned a lot and I appreciate it all.

I would like to extend my gratitude to Dr. Angie Dawson, Dr. Barb Triggs-Raine and Dr. Terry Zelinski for providing me with an encouraging and supportive committee. Thanks for all your input, efforts, comments, and time.

Kim, I know you tried. I just don't listen do I? But here it is and its all over with. Today I can say I'm glad I didn't scare easily. I want to thank you for your support, encouragement, thoughtfulness, perspective, and friendship...not only for the duration of my M.Sc. project, but for the last 8 years. By the way, a great big thanks to the Department of Human Genetics for leading me to such a wonderful friend!

I would also like to take this opportunity to tell my family how much they mean to me. Thanks to Mom who has shown me, by example, how to be independent, supportive, understanding, giving, strong, and above all else, loving. I am the person I am today because of you Mom. (That is credit, not blame!) Also, thanks for making home where the heart is. Thanks to Verna for being my favorite "seester". Thank you for accepting the role as my grammar guru and, subsequently, all of the time you spent reading and editing this thesis. In addition, thanks for always being at the other end of the line when I need you the most! Thanks to Wayne for your brotherly pride in everything I do. I would also like to extend a big thanks to my favorite Uncle Don and Aunt Valma. Thanks for taking me into your home for the second time. In addition, thanks for your support and kind words on all those yucky days I had. Regardless of how much education we receive, the most important lesson any of us will learn is that nothing is more important than family!

Last but not least, thanks to Dan. To begin, thank you for being a member of my committee for the first year. In addition to everything you do here on the 3rd floor, thank you for all of those little extra things from your lab that made my life easier. I owe you more than gratitude and thanks for your encouragement, support and friendship. Without you, I'm sure I wouldn't have had the strength, will power or desire to complete this project or to even stay in science. Thanks for being there especially on the bad days. Above all else, thank you for coming into my life. With you, I feel like I've come home.

LIST OF ABBREVIATIONS

A	adenine
α-^{32}P	alpha labeled phosphorous
APS	ammonium persulfate
BLAST	Basic Local Alignment Search Tool
bp	base pair(s)
BrdC	bromodeoxycytidine
BrdU	bromodeoxyuridine
BSA	bovine serum albumin
C	cytosine
cDNA	complementary deoxyribonucleic acid
Ci	Curie
cm	centimetre (10^{-3} metre)
cpm	counts per minute
DAPI	4,6-diamidino-2-phenylindole
DNA	deoxyribonucleic acid
dNTPs	mixture of all four deoxynucleotides
dATP	deoxyadenosine triphosphate
dCTP	deoxycytidine triphosphate
C	cytosine
°C	degrees Celsius
dGTP	deoxyguanosine triphosphate
DRPLA	dentatorubral-pallidoluysian atrophy
DTT	dithiothreitol
dTMP	deoxythymidine monophosphate
dTTP	deoxythymidine triphosphate
dUMP	deoxyuridine monophosphate
E. coli	*Escherichia coli*
EDTA	ethylene diaminetetracetic acid
FISH	fluorescent *in situ* hybridization
FrdU	5-fluorodeoxyuridine
γ-^{32}P	gamma labeled phosphorous
g	gram
G	guanine
HD	Huntington disease
HRS	Haw River syndrome
kb	kilobase (10^3 bp)
KOAc	potassium acetate
kV	kilovolt (10^3 V)
kW	kilowatt
l	litre
LB	Luria Bertani
M	molar
Mb	megabase (10^6 bp)

MD	myotonic dystrophy
µF	microfarad
µg	microgram (10^{-6} g)
µl	microlitre (10^{-6} l)
min.	minute(s)
MJD	Machado-Joseph disease
mM	millimolar (10^{-3} M)
MR	mental retardation
ms	millisecond
NaCl	sodium chloride
NaOAc	sodium acetate
ng	nanogram (10^{-9} g)
NH_4OAc	ammonium acetate
nm	nanometre (10^{-9} m)
NTMs	nonpenetrant transmitting males
Ω	ohms
PCR	polymerase chain reaction
PEG	polyethylene glycol
PFGE	pulse field gel electrophoresis
pmol	picamole (10^{-12} mole)
rpm	revolutions per minute
s	seconds
SBMA	spinal and bulbar muscular atrophy
SCA	spinocerebellar ataxia
SDS	sodium dodecyl sulfate
SSC (20X)	3 M NaCl, 0.3 M $Na_3C_6H_2O_7 \cdot 2H_2O$, pH 7.0
T	thymine
TAE (1X)	40 mM Tris·acetate, 1 mM EDTA
TBE (1X)	90 mM Tris·borate, 1 mM EDTA
TE	10 mM Tris·HCl (pH 8.0), 1 mM EDTA
TEMED	*N'*, *N'*, *N'*, *N'*-tetramethylethylene diamine
Tris	tris(hydroxymethyl)-aminomethane
UTS	untranslated sequence
UV	ultraviolet
V	volts
w/v	weight per volume
X-Gal	5-bromo-4-chloro-3-indoyl-β-D-galactopyranoside

LIST OF FIGURES

LIST OF TABLES

1.0 INTRODUCTION

Fragile sites are defined as chromosomal loci that stain poorly and exhibit uncondensed gaps and breaks on metaphase chromosomes *in vitro* (Wang and Griffith, 1996; Griffith and Wang, 1998). The expression of these fragile sites, which does not generally occur spontaneously, requires the growth of cells under conditions that inhibit DNA replication or repair (Glover, 1998a; 1998b). To date, over 100 fragile sites have been identified, but only 10 have been cloned and characterized. Each fragile site is classified according to both its frequency within the general population (rare and common) and the tissue culture conditions required for its expression (presence of folate, thymidine, etc.) (Sutherland *et al.*, 1996). Five rare folate sensitive fragile sites have been cloned and found to be associated with CGG trinucleotide expansions and, in some cases, genetic disease (Glover, 1998a; 1998b). Cloning of three common fragile sites has not identified any associated expansions and, thus, their molecular basis remains unknown. Evidence suggests that the common fragile sites may play a role in chromosome breakage and rearrangements involved in cancer and the integration of viruses (Engleman *et al.*, 1998; Mishmar *et al.*, 1998; Wilke *et al.*, 1996).

To date, two unrelated families have been identified that express a rare folate sensitive fragile site, *FRA19B*, at 19p13.1 (Figure 1). One family is apparently clinically unaffected while some members of the other family display schizophrenia, autism and mental retardation. To further our understanding of

the mechanism and consequences of chromosome fragility, an attempt has been made to molecularly characterize the rare folate sensitive fragile site *FRA19B*.

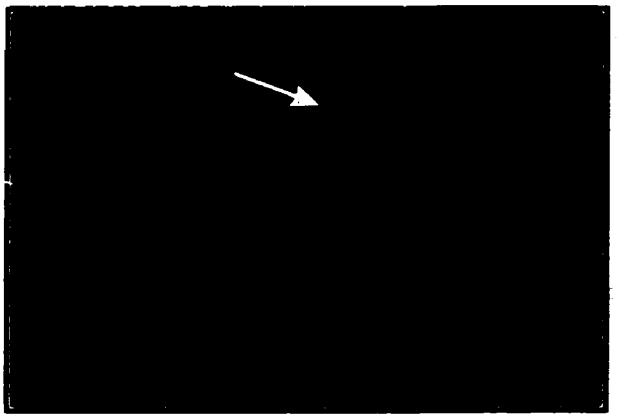

Figure 1: Partial R-banded metaphase spread illustrating *FRA19B* (arrow) located at 19p13.1. R-banding performed with fluorescence using acridine orange. (Courtesy of the Cytogenetic Laboratory, Health Sciences Centre, Winnipeg, Manitoba.)

1.1 History of Fragile Sites

In 1965, Dekaban *et al.* reported a gap or constriction on the long arm of an unidentified C-group chromosome (speculated to be chromosome 9) of a woman who had previously undergone irradiation treatment. This was described as a "weak" region susceptible to breakage. In 1968, Lejeune *et al.* observed a similar aberration in chromosome region 2q1 of individuals from three unrelated families (Magenis *et al.*, 1970). In 1970, Magenis *et al.* reported a large family in which an isochromatid break on the long arm of chromosome 16 was observed in 31 relatives. This region, due to its susceptibility to breakage, was designated as a "fragile site". Pedigree analysis determined that this fragile site was segregating in a simple Mendelian co-dominant fashion with full penetrance. It was noted that the break at this fragile site was not observed in 100% of the cells

and that the proportion of cells containing the site varied, depending on the type of tissue that was cultured. Magenis *et al.* (1970) speculated that heritable fragile sites were minor variations likely due to the intrinsic chromosome structure specific to that region.

By 1977, fragile sites had been identified on human chromosomes 2, 12, 16, 17, X, and unidentified C-group chromosomes (Sutherland, 1977). However, when the fragile site at 2q1 was reinvestigated by Sutherland (1977), the site appeared to have vanished. It was then realized that the expression and frequency of the fragile site was dependent on the type of culture medium utilized. This was a significant finding since one of the fragile sites located on the terminal long arm of the X chromosome (the fragile X) had been associated with X-linked mental retardation. Sutherland (1979) proved that fragile sites at 2q11, 10q23, 11q13, 16p12, 20p11, and Xq27 were inhibited by the addition of folic acid, thymidine or bromodeoxyuridine (a thymidine analogue) to the culture medium. On the other hand, the addition of methotrexate (an inhibitor of folate metabolism) induced these sites (Sutherland, 1979). However, the fragile site at 16q22 was unaffected by any of these additives, but was later found to be induced by distamycin A (an oligopeptide antibiotic). To account for the *in vitro* behavior of the fragile sites, Sutherland (1979) proposed that the isochromatid breaks, which appeared to be heritable, were likely dependent on the actual base composition of the DNA within these regions.

1.2 Classification of Fragile Sites

To date, there have been over 100 fragile sites identified in the human genome. Fragile sites are divided into two classes: 1.) rare fragile sites which are inherited in a co-dominant Mendelian fashion and, typically, are found in less that 1% of the population and 2.) common fragile sites which are thought to be intrinsic to chromosome structure and are found in 100% of the population (Sutherland and Richards, 1999). Further divisions of rare and common fragile sites are based on the conditions or agents required to induce their expression (Table 1).

Table 1: Classification of fragile sites (Sutherland and Richards, 1999).

Class	Subclass	Number
Rare	folate sensitive	23
	distamycin A inducible	5
	BrdU inducible	2
Common	aphidicolin inducible	76
	5-azacytidine inducible	4
	BrdU inducible	6
	adenovirus 12 inducible	4

Note: FrdU, 5-fluorodeoxyuridine; BrdU, bromodeoxyuridine; BrdC, bromodeoxycytidine.

Currently 30 rare fragile sites have been identified (Table 1). Rare folate sensitive fragile sites can be induced by a deficiency of folate and thymidine; the addition of inhibitors of thymidylate synthase such as 5-fluorodeoxyuridine (FrdU); the addition of anti-folates such as methotrexate; or the addition of high levels of thymidine (Sutherland and Richards, 1999). The rare fragile sites that can be induced by either oligopeptide antibiotics that bind to the minor groove of DNA, such as distamycin A and berenil, or the thymidine analogue

bromodeoxyuridine (BrdU) are referred to as rare distamycin A inducible fragile sites (Sutherland, 1988; Sutherland and Richards, 1999). Rare fragile sites that can be induced by either the addition of BrdU or bromodeoxycytidine (BrdC) are referred to as BrdU inducible rare fragile sites.

To date, approximately 90 common fragile sites have been identified (Table 1). These sites are inducible by aphidicolin (an inhibitor of DNA polymerases α and δ), 5-azacytidine (a cytidine analog used as a demethylation agent), BrdU, or adenovirus 12 (Sutherland, 1988).

1.3 Molecular Basis and Biological Significance of Fragile Sites

1.3.1 Rare Fragile Sites

While 30 rare fragile sites have been identified to date, only seven have been cloned: five folate sensitive fragile sites (*FRAXA*, *FRAXE*, *FRA16A*, *FRAXF*, *FRA11B*), one distamycin induced fragile site (*FRA16B*) and one BrdU induced fragile site (*FRA10B*) (Table 2). Based on these seven fragile sites, the method of induction may indicate the general sequence composition of rare fragile sites.

1.3.1.1 *FRAXA*

Many fragile sites had been located within the human genome by 1991. However, the molecular basis for their fragility was still unknown. At that time, the only fragile site associated with a clinical phenotype, the FRAXA syndrome, was the rare folate sensitive fragile site *FRAXA*, located at Xq27.3. Due to the clinical relevance of *FRAXA*, a priority was placed on determining the molecular

Table 2: A summary of rare fragile sites (Sutherland et al., 1996).

Induction Method	Location	Fragile Site	Molecular Basis	Associated Phenotype
Folic Acid Deficiency	1p21.3	*FRA1M*	unknown	
	2q11.2	*FRA2L*	unknown	
	2q11.2	*FRA2A*	unknown	
	2q13	*FRA2B*	unknown	
	2q22.3	*FRA2K*	unknown	
	5q35	*FRA5G*	unknown	
	6p23	*FRA6A*	unknown	
	7p11.2	*FRA7A*	unknown	
	8p22.3	*FRA8A*	unknown	
	9p21.1	*FRA9A*	unknown	
	9q32	*FRA9B*	unknown	
	10q23.3	*FRA10A*	unknown	
	11q13.3	*FRA11A*	unknown	
	11q23.3	*FRA11B*	$p(CCG)_n$	Jacobsen syndrome
	12q13.1	*FRA12A*	unknown	
	12q24.13	*FRA12D*	unknown	
	16p13.11	*FRA16A*	$p(CGG)_n$	
	19p13.1	*FRA19B*	unknown	
	20p11.23	*FRA20A*	unknown	
	22q13.1	*FRA22A*	unknown	
	Xq27.3	*FRAXA*	$p(CGG)_n$	FRAXA syndrome
	Xq28	*FRAXE*	$p(CGG)_n$	FRAXE syndrome
	Xq28	*FRAXF*	$(GCCGTC)_n(CGG)_n$	
Distamycin A Treatment	8q24.1	*FRA8E*	unknown	
	11q15.1	*FRA11I*	unknown	
	16p12.1	*FRA16E*	unknown	
	16q22.1	*FRA16B*	33bp $(AT)_n$	*cis* effects
	17p12	*FRA17A*	unknown	
BrdU Treatment	10q25.2	*FRA10B*	AT rich	
	12q24.2	*FRA12C*	unknown	

nature of this fragile site. In 1991, Kremer *et al.* identified a restriction fragment which was found to be unstable in FRAXA syndrome pedigrees. This fragment contained two regions of interest: a methylation-sensitive, CpG-rich region and a $p(CGG)_n$ repeat region. *In situ* hybridization further demonstrated that the region containing the $p(CGG)_n$ repeat corresponded with the region containing the fragile site (Kremer *et al.*, 1991). Thus, the molecular basis of the

FRAXA syndrome was the first instance in which an expansion of a trinucleotide was identified (Kremer *et al.*, 1991). This was an important finding which led to further understanding fragile sites and the complex molecular basis of the FRAXA syndrome.

FRAXA syndrome results from the inactivation of the fragile X mental retardation (*FMR1*) gene, located on chromosome Xq27.3 (Verkerk *et al.*, 1991; Imbert *et al.*, 1998). Typically, this inactivation is caused by the amplification of a $p(CGG)_n$ repeat, found in the nontranscribed first exon of *FMR1*, and methylation of an adjacent 5' CpG island (Bell *et al.*, 1991; Verkerk *et al.*, 1991; Vincent *et al.*, 1991). However, patients have also been identified in which the *FMR1* gene is inactivated as a result of point mutations or deletions (Imbert *et al.*, 1998). The protein product of *FMR1*, FMRP, is hypothesized to be involved in the nuclear export of some mRNAs expressed in the brain and their subsequent presentation to the translation machinery (Imbert *et al.*, 1998).

There are four types of FRAXA alleles, which are grouped according to the copy number of the $p(CGG)_n$ repeat, and include normal, grey zone, premutation, and full mutation alleles (Imbert *et al.*, 1998). Molecular analysis of normal individuals has shown that the length of the normal $p(CGG)_n$ repeat varies from 6 to 45; within this range, the repeat is a stable polymorphism which segregates in a co-dominant, Mendelian fashion (Imbert *et al.*, 1998). A fragile site is not expressed in the normal individuals.

Alleles ranging in size between 45 to 60 triplets have varying degrees of stability and are aptly termed "grey zone" alleles (Eichler *et al.*, 1994; Zhong

et al., 1996; Imbert *et al.*, 1998). These alleles are not associated with chromosome fragility. Some grey zone alleles are stable, thereby remaining the same size through subsequent generations while others are unstable and can expand slowly by one to two repeats in subsequent generations. The probability of these grey zone alleles expanding to a full mutation in one generation is very low (Imbert *et al.*, 1998). Sequence composition, in addition to sequence length, was found to be a critical parameter in the stability of these alleles. Within the p(CGG)$_n$ repeat tract of normal individuals, two AGG interspersions (one 5' and one 3') are found that appear to confer a degree of stability to the allele. However, only the 5' AGG interspersion is found in most unstable alleles (Eichler *et al.*, 1994; Kunst and Warren, 1994; Nolin *et al.*, 1996; Zhong *et al.*, 1996; Imbert *et al.*, 1998). The loss of the 3' AGG interspersion lengthens the uninterrupted p(CGG)$_n$ repeat tract (Eichler *et al.*, 1994; 1995). This loss appears to be an important mutational event resulting in the decreased stability of an allele, which is subsequently at an increased risk of expansion to a full mutation (Eichler *et al.*, 1995; Zhong *et al.*, 1996; Mogk *et al.*, 1998).

Alleles that are expanded from 60 to 200 repeats also do not express *FRAXA* and are known as premutations (Imbert *et al.*, 1998). Such repeats can expand even further when transmitted through a female and result in full mutations of more than 200 repeats. Males and females with repeats in the premutation size range are referred to as nonpenetrant transmitting males and carrier females, respectively. Males affected with FRAXA syndrome have the full mutation which contains at least 200 repeats (Fu *et al.*, 1991; Warren and

Nelson, 1994; Nolin *et al.*, 1996). When the number of repeats reaches a threshold of approximately 200, the fragile site *FRAXA* is expressed, the adjacent CpG island generally becomes fully methylated and *FMR1* is inactivated (Imbert *et al.*, 1998; Nelson, 1998).

1.3.1.2 *FRAXE*

FRAXE, located distal to *FRAXA* at Xq27.3, was originally identified in families with mild mental impairment in which no expansion was observed at the *FRAXA* locus although a cytogenetic fragile site was observed (Flynn *et al.*, 1993). FRAXE syndrome was also found to be associated with a $p(CGG)_n$ repeat, found within the fragile X mental retardation 2 (*FMR2*) gene on chromosome Xq27.3, distal to *FMR1* (Flynn *et al.*, 1993). The size of the repeat is polymorphic, ranging in size from 7 to 35 repeats, in normal individuals. The number of repeats ranges from 130 to 150 in carriers. Affected individuals have size ranges between 230 to 750 repeats and express *FRAXE* (Wilmot and Warren, 1998). A CpG island, which lies 5' to the $p(CGG)_n$ repeat, is also methylated when an expansion reaches a threshold between 130 to 200 repeats. Expansion of the repeat and methylation of the CpG island is usually associated with down-regulation of the *FMR2* gene. *FMR2* encodes a nuclear protein with the capacity to bind DNA and potentially activate transcription (Gecz *et al.*, 1996; 1997; Gu *et al.*, 1996). The down-regulation of *FMR2* results in a mild non-specific mental handicap in males and mild to borderline mental retardation in females carrying a full mutation (Gecz *et al.*, 1996; Gu *et al.*, 1996; Nelson, 1998).

1.3.1.3 *FRA16A*

The molecular basis for the autosomal fragile site located at 16p13.11, *FRA16A*, has also been determined to be a $p(CGG)_n$ repeat with an associated 5' CpG island. The $p(CGG)_n$ repeat within *FRA16A* is not a pure repeat as it has been found to contain CTG interspersions which, when lost, decrease allelic stability (Nancarrow *et al.*, 1994; Nelson, 1998). This locus is normally polymorphic, ranging in size from 16 to 49 repeats, in individuals who do not express *FRA16A*. Individuals expressing this fragile site have methylation at the CpG island, in addition to a repeat ranging in size between 1000 to 2000 triplets (Nancarrow *et al.*, 1994; Nelson, 1996). These *FRA16A* positive individuals appear phenotypically normal. No phenotype or pathology has been associated with the heterozygous expression of *FRA16A*. However, if a phenotype does occur, it may only be evident in individuals homozygous for *FRA16A* which, given its frequency, is unlikely (Nelson, 1998).

1.3.1.4 *FRAXF*

The fourth rare fragile site to be cloned was *FRAXF*, which is located in Xq28 (Ritchie *et al.*, 1994). *FRAXF* was originally identified in individuals not positive for an expansion within *FRAXA* or *FRAXE* but displaying mental retardation or developmental delay and a cytogenetically visible fragile site (Nelson, 1998). Molecular analysis of this fragile site revealed an amplification of a compound repeat with the structure $(GCCGTC)_n(GCC)_n$. In most normal individuals there are three copies of the (GCCGTC) repeat at the 5' end of the *FRAXF* repeat while only a few individuals have four copies. Individuals not

expressing *FRAXF* have 12 to 26 copies of the GCC repeat. *FRAXF* positive individuals have greater than 300 of these triplets, which are also associated with methylation (Ritchie *et al.*, 1994). The association between *FRAXF* and mental function is weak. Only four families have been found to be segregating for *FRAXF*. A clinical phenotype is observed in some, but not all, individuals who express *FRAXF*. Individuals with the clinical phenotype have also been identified who do not express *FRAXF* (Nelson, 1998). Like its X-linked counterparts, *FRAXA* and *FRAXE*, *FRAXF* is also associated with methylation of a CpG island. However, *FRAXF* does not appear to be contained within a promoter and no down-regulated genes have been identified within the chromosomal region (Nelson, 1998).

1.3.1.5 *FRA11B*

FRA11B, located at 11q23.3 within the Casitas B-lineage lymphoma 2 (*CBL2*) proto-oncogene, is also associated with a $p(CGG)_n$ repeat and methylation of a 5' CpG island (Blake *et al.*, 1991; Jones *et al.*, 1994; 1995). This repeat is not highly polymorphic in the general population (Jones *et al.*, 1995). *FRA11B* is the only autosomal rare folate sensitive fragile site to be associated with a specific clinical phenotype, Jacobsen syndrome. This syndrome, while quite variable in phenotype, is affiliated with mental retardation, dysmorphism and growth retardation (Jones *et al.*, 1995; Nelson, 1998). Jacobsen syndrome is typically a direct result of a terminal deletion of 11q with breakpoints located in 11q23.3-q24.2, a breakpoint cluster region of approximately 13.5 Mb (Tunnacliffe *et al.*, 1999). It has been proposed that such

a wide variation in breakpoint sites is responsible for the high degree of phenotypic variability observed among Jacobsen syndrome patients. In most cases, the breakpoint lies distal to the *FRA11B* locus. However, in approximately 10% of the cases the deleted chromosome is derived from a parent with an expanded $p(CGG)_n$ repeat at *FRA11B* (Tunnacliffe *et al.*, 1999). This is the first indication of the involvement of a $p(CGG)_n$ expansion and a rare folate sensitive fragile site in the *in vivo* fragility of chromosomes. However, a direct effect of the expansion on the *CBL2* proto-oncogene is not known (Jones *et al.*, 1995).

1.3.1.6 *FRA16B*

FRA16B was the first rare non-folate sensitive fragile site to be cloned. This site, located at 16q22.1, is induced by distamycin A and has a frequency of approximately 1 in 40 chromosomes in European populations (Yu *et al.*, 1997). The cloning of this fragile site revealed an expansion of a 33 bp minisatellite repeat with the structure $p(ATATATTATATATTATATCTAATAATATAT^{C}/_{A}TA)_n$, where *n* normally varies from 7 to 12 with some expansions differing by less than a whole repeat unit (Yu *et al.*, 1997). Expansion of this minisatellite, up to 2000 repeats, gives rise to *FRA16B*. This AT repeat sequence is compatible with the A-T binding action of distamycin A. However, this finding emphasizes two aspects of expansion: 1.) the method required for induction of fragile sites is likely related to the sequence composition of the DNA within these sites, thereby explaining the various classes of fragile sites and 2.) the mechanism of trinucleotide repeat expansion may also be responsible for the variation

observed in minisatellite repeats and variable number tandem repeats (VNTRs) (Yu *et al.*, 1997).

FRA16B, although not implicated in a clinical phenotype, has been reported to exert *cis*-configuration effects on a neighboring gene, the haptoglobin gene. The two alleles of the haptoglobin locus, *HP1* and *HP2*, are associated with different levels of haptoglobin in the blood (Côté, 1990). When a *FRA16B* expansion is coupled with (or located on the same homologue as) *HP2*, the level of haptoglobin is not affected. However, when a *FRA16B* expansion is coupled with *HP1*, the expression level from this haptoglobin allele is reduced to approximately 20% (Côté, 1990). This suggests that fragile sites, in general, may have the capacity to disrupt neighboring genes in an allele specific manner. This neighboring effect may account for the wide variability observed in disorders associated with fragile sites (Côté, 1990).

1.3.1.7 *FRA10B*

FRA10B, located at 10q25.2, is inducible by BrdU or BrdC (Hewett *et al.*, 1998). The composition of the DNA repeat associated with *FRA10B* is much more complex than those associated with other fragile sites cloned to date. Southern blot analysis located the amplified region to a 1.9 kb *Pst*I fragment. Sequence analysis of this fragment from normal alleles revealed an AT-rich (91%) region less than 1 kb, which is highly polymorphic in length (Hewett *et al.*, 1998). A variety of repeat motifs, ranging in size from 16 to 52 bp, were identified within this polymorphic region. Based on the varying degrees of relatedness, the sequences of these repeat motifs were grouped into five

categories: TspI (T), proximal (P), expanded (E), SnaBI (S), and distal (D). Alleles containing a variety of different repeat motifs were grouped into four categories based on length: small normal (N), intermediate (INT), large normal (LN), and *FRA10B*-expressing. Sequence analysis of unstable alleles revealed at least four uninterrupted E repeats. Interestingly, a consensus sequence of E repeat motifs in LN and *FRA10B* alleles revealed that 26 bp were shared with the 33 bp repeat associated with *FRA16B* (Hewett *et al.*, 1998).

1.3.2 Common Fragile Sites

Common fragile sites are aptly named since they are found to be expressed *in vitro* at specific loci in the cells of all individuals. It has been postulated that these common fragile sites are part of the normal chromosome structure. Most common fragile sites, unlike the rare fragile sites, are induced by aphidicolin which inhibits the elongation activity of DNA polymerases α and δ. The common fragile sites are found predominantly in G-light bands suggesting an association with active gene regions (Glover, 1998a; 1998b; Mishmar *et al.*, 1998). The most frequently induced common fragile sites are *FRA3B* (3p14.2), *FRA16D* (16q23), *FRA6E* (6q26), *FRA7H* (7q32), and *FRAXB* (Xp22.3) (Glover, 1998a). The level of expression and the number of the common fragile sites expressed per cell is increased dramatically when stress is placed on either DNA replication or G_2 repair (Glover, 1998a; 1998b). It has been suggested that the expression of common fragile sites is controlled not only by genetic factors but also by environmental ones. For example, a higher level of expression is consistently observed in lymphocytes of individuals who smoke cigarettes

(Glover, 1998b). These observations suggest that, while rare and common fragile sites may have some related aspects, the mechanisms for their instability are different (Glover, 1998b).

At least 90 common fragile sites have been identified throughout the human genome, but very little is known about their molecular basis or their biological significance. However, based on their locations and behavior *in vitro*, it has been suggested that they may play a role in chromosome breakage and rearrangements involved in cancer, birth defects, evolution, recombination and sister chromatid exchange, and the integration of viruses into the human genome (Glover, 1998a; 1998b). Only three common fragile sites, all aphidicolin inducible, have been cloned: *FRA3B*, *FRA7G* and *FRA7H*.

1.3.2.1 *FRA3B*

FRA3B, which maps to 3p14.2, is the most frequently observed fragile site and is expressed in a large proportion of the cells treated with folate stress or aphidicolin (Wilke *et al.*, 1996; Glover, 1998b). This fragile site is approximately 160 kb telomeric to the t(3;8) translocation breakpoint that is associated with familial renal cell carcinoma (Wilke *et al.*, 1996). The fragile histadine triad (*FHIT*) gene, a frequently mutated tumor suppressor gene, spans both the translocation breakpoint and *FRA3B* (Glover, 1998a; 1998b).

Sequence analysis of *FRA3B* did not reveal any obvious repeat regions, thereby suggesting a different expression mechanism for "common" as compared to "rare" fragile sites. However, the DNA surrounding the breakpoints located within *FRA3B* can form secondary structures, similar to those formed by

p(CGG)$_n$ repeats (Wang *et al.*, 1997). Interestingly, an integration site for a human papillomavirus (HPV16), a well-known transforming factor in cervical cancer, has been identified within *FRA3B* (Wilke *et al.*, 1996; Le Beau *et al.*, 1998). In fact, a small deletion within 3p14.2 is involved in loss of heterozygosity (LOH) associated with cervical cancers (Boldog *et al.*, 1997). The location of this integration site coincides with an area that frequently exhibits gaps or breaks within *FRA3B* (Wilke *et al.*, 1996). This was the first *in vivo* viral integration event to be observed within a fragile site and, thus, may be the first clue as to the biological significance of common fragile sites (Wilke *et al.*, 1996). However, the breakage and fragility associated with the *FRA3B* locus is not limited to this integration site, but has been localized to a genomically unstable region of at least 500 kb (Le Beau *et al.*, 1998). In fact, the location of the aphidicolin induced breakage of *FRA3B* involves two tight clusters that flank either side of the HPV16 integration site. The two clusters are 100 and 300 kb distal to the t(3;8) translocation breakpoint. Molecular analysis of the two clusters has revealed the presence of 11 regions of high flexibility, defined as a high degree of fluctuation in the twist angle of the DNA double helix (Mishmar *et al.*, 1998). One of these regions of high flexibility colocalizes to an aphidicolin induced breakpoint cluster (Mishmar *et al.*, 1998). Sequence analysis of the *FRA3B* has revealed several features characteristic of matrix attachment regions (MARs), which act as initiation points for replication (Wang *et al.*, 1997).

1.3.2.2 *FRA7G*

FRA7G is located within 7q31.2. This common fragile site also appears to encompass a large region of at least 300 kb. A $(CA)_n$ minisatellite repeat marker which spans *FRA7G*, *D7S522*, is frequently deleted in breast, prostate, renal, stomach, colorectal, pancreatic, uterine, and ovarian cancers (Engelman *et al.*, 1998). Genes encoding caveolin-1 and caveolin-2 have been localized to this deletion region (Engelman *et al.*, 1998). The protein products of these two genes normally form hetero-oligomeric complexes which function in vesicular trafficking and signal transduction processes (Engelman *et al.*, 1998). However, down-regulation of caveolin-1 has been reported in various transformed cell lines and mammalian cancer cells and is, therefore, suspected to be a tumor suppressor gene (Engelman *et al.*, 1998). Thus, the second common fragile site to be cloned also appears to be involved in the manifestation of cancer. Molecular analysis was undertaken to account for the fragility of *FRA7G*. Eight regions, three of which cluster within 18 kb of *D7S522*, were identified as highly flexible (Mishmar *et al.*, 1998). These regions of high flexibility colocalize to regions of low stability, determined by the energy required to disrupt the DNA duplex (Breslauer *et al.*, 1986; Mishmar *et al.*, 1998).

1.3.2.3 *FRA7H*

Cloning of a simian virus 40 integration site on chromosome 7 led to the discovery that integration events can occur within the common aphidicolin induced fragile site *FRA7H*, located at 7q23.3 (Mishmar *et al.*, 1998). Molecular analysis has indicated that several regions within *FRA7H* have an unusual DNA

structure including high flexibility, low stability, and non-B-DNA forming sequences (Mishmar *et al.*, 1998). Currently, no genes have been localized to this region.

1.4 Induction Mechanism of Rare Folate Sensitive Fragile Sites

Spontaneous expression of fragile sites does not generally occur. However, manipulation of tissue culture conditions can be made to induce fragile sites *in vitro* (Sutherland and Richards, 1995). The process by which fragile sites are formed is unclear to this day. However, Sutherland proposed a preliminary theory in 1979 which, based on the nature of the inhibitors and inducers of the folate sensitive fragile sites, involves the biosynthesis of dideoxythymidine triphosphate (dTTP).

The *in vitro* expression of the folate sensitive fragile sites is speculated to arise during DNA synthesis (Sutherland, 1979). The metabolic pathway implicated involves the conversion of deoxyuridine monophosphate (dUMP) to deoxythymidine monophosphate (dTMP) (Figure 2). During DNA synthesis, dUMP is converted to dTMP which, in turn, is converted to deoxythymidine triphosphate (dTTP) which is subsequently incorporated into the growing DNA strand. If the conversion of dUMP to dTMP is inhibited by 5-fluorodeoxyuridine (FrdU, a thymidylate synthetase inhibitor) or methotrexate (an inhibitor of dihydrofolate reductase), the pool of dTMP is eventually depleted. Consequently, this inhibition results in an increase of cellular dUMP. Normally, dUMP is at a very low concentration relative to dTMP. However, treatment with FrdU results in the ratio of dUMP to dTMP increasing over 1000-fold (Reidy,

1988). Thymidine overload or starvation also results in the alteration of the dUMP:dTMP ratio (Meuth, 1984). A disturbance in the pool of nucleotides arrests DNA synthesis and leads to error-prone repair and mitotic recombination (Meuth, 1984; Sutherland, 1988). If the DNA within the fragile site is not replicated when the cell moves out of S phase or repaired during G_2 phase, packaging of this region into a condensed chromatin structure will not be complete prior to the entry into mitosis (Reidy, 1988; Sutherland and Richards, 1999). Thus, a gap or break appears in metaphase at the locations where the synthesis of DNA was not completed.

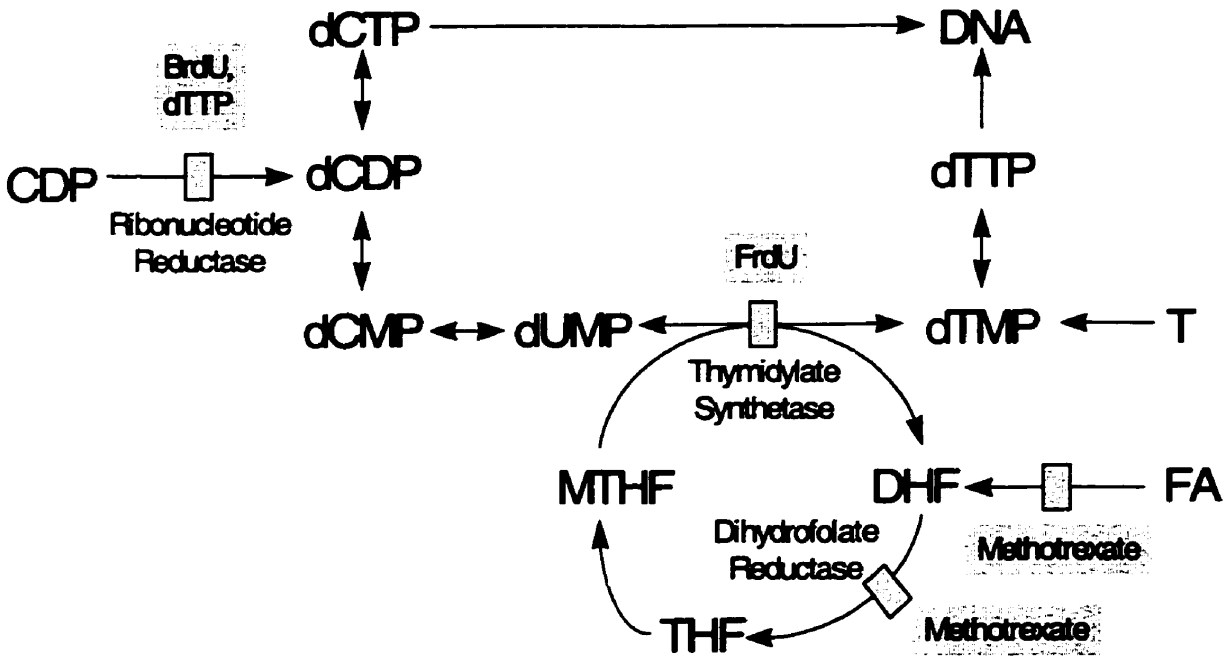

Figure 2: Metabolic pathway involving folic acid in DNA synthesis. BrdU, bromodeoxyuridine; CDP, cytidine diphosphate; dCDP, deoxycytidine diphosphate; dCMP, deoxycytidine monophosphate; dCTP, deoxycytidine triphosphate; DHF, dihydrofolate; dTMP, deoxythymidine monophosphate; dTTP, deoxythymidine triphosphate; dUMP, deoxyuridine monophosphate; FA, folic acid; FrdU, 5-fluorodeoxyuridine; MTHF, methylenetetrahydrofolate; T, thymidine; THF, tetrahydrofolate. (Glover, 1981; Friedberg, 1985; Sutherland, 1988)

1.5 Trinucleotides Not Associated With Fragile Sites

Rare folate sensitive fragile sites appear to arise only from expansions of CGG trinucleotide repeat tracts. The *FRAXA*, *FRAXE* and *FRA11B* are associated with syndromes characterized by inborn mental and developmental delay. However, expansions of trinucleotide repeats with other base compositions have been identified as the molecular basis of many neurodegenerative diseases. All trinucleotide repeat expansion diseases exhibit several common features including variability between different members of the same family, genetic anticipation (increased severity and decreased age of onset in subsequent generations), and a threshold number of trinucleotide repeats which, when surpassed, results in a clinical phenotype (Shastry, 1994; Warren, 1996; Korneluk and Narang, 1997). Unstable trinucleotide repeats, in general, have several fundamental differences including the location of the repeated sequence within the gene, the size range of the normal and expanded trinucleotide repeat, the preferential parent-of-origin of the expanded repeat, and the clinical phenotype associated with the expansion. However, the most important difference between the trinucleotide repeats is the actual base composition of the repeat, being either CGG, GAA, GCG, or CAG.

1.5.1 GAA Trinucleotide Repeats

As listed in Table 3, the expansion of a $p(GAA)_n$ repeat has been implicated in only one disorder to date, Friedreich ataxia. This multisystem neurodegenerative disorder primarily affects the spinal cord, peripheral nerves and heart (Wilmot and Warren, 1998). Friedreich ataxia is inherited in an

autosomal recessive fashion and does not display genetic anticipation, a hallmark of all the other trinucleotide repeat diseases (Warren, 1996). However, the age of onset and/or the clinical severity of Friedreich ataxia is related to the size of the expansion. The $p(GAA)_n$ expansion is located in 9q13-21.1 within the first intron of the *X25* gene, which codes for frataxin (Wilmot and Warren, 1998). It has been suggested that the partial loss of function of frataxin is due to the creation of numerous AG splice acceptor sites formed by the $p(GAA)_n$ expansion (Warren, 1996).

1.5.2 GCG Trinucleotide Repeats

To date, oculopharyngeal muscular dystrophy is the only disease in which the molecular etiology involves an expansion of a $p(GCG)_n$ repeat. A very small expansion occurs in the coding region of the poly(A) binding protein 2 (*PABP2*) gene located on chromosome 14q11 (Table 3). The disease presents after the age of 60 with progressive difficulties in swallowing, drooping of the eyelids and weakness of the proximal limbs (Brais *et al.*, 1998). Homozygosity of a $p(GCG)_7$ allele results in an autosomal recessive form of oculopharyngeal muscular dystrophy while compound heterozygotes for the $p(GCG)_7$ and $p(GCG)_9$ alleles present with symptoms of greater severity (Brais *et al.*, 1998). A gain of function of the poly(A) binding protein results from the expansion of this repeat. It has been suggested that the increased length of the polyalanine stretch within the protein may cause the accumulation of nuclear filaments observed in skeletal muscle fibers of patients (Brais *et al.*, 1998).

Table 3: Summary of trinucleotide repeat expansion diseases (Brais *et al.*, 1998; Imbert *et al.*, 1998; Wilmot and Warren, 1998; Koob *et al.*, 1999).

Disorder	Inheritance Pattern	Gene	Location	Protein	Normal	Premutation	Mutant	Repeat Location	Effect	PLE
Friedreich ataxia	AR	*X25*	9q13-21.1	Frataxin	$(GAA)_{6-34}$	$(GAA)_{80}$	$(GAA)_{112-1700}$	intron 1	PLOF	M
Oculopharyngeal muscular dystrophy	AD	*PABP2*	14q11	PABP2	$(GCG)_{6}$	—	$(GCG)_{8-13}$	coding	GOF	ND
Dentatorubral-pallidoluysian atrophy	AD	*DRPLA*	12p13.1.31	Atrophin-1	$(CAG)_{6-35}$	—	$(CAG)_{51-88}$	coding	GOF	P
Huntington disease	AD	*IT15*	4p16.3	Huntingtin	$(CAG)_{6-39}$	—	$(CAG)_{36-121}$	coding	GOF	P
Spinobulbar muscular atrophy	XLD	*AR*	Xq13-21	Androgen receptor	$(CAG)_{11-33}$	—	$(CAG)_{38-66}$	coding	GOF, PLOF	ND
Spinocerebellar ataxia type 1	AD	*SCA1*	6p23	Ataxin-1	$(CAG)_{6-39}$	—	$(CAG)_{41-81}$	coding	GOF	P
Spinocerebellar ataxia type 2	AD	*SCA2*	12q24.1	Ataxin-2	$(CAG)_{14-31}$	—	$(CAG)_{35-64}$	coding	GOF	P
Spinocerebellar ataxia type 3	AD	*SCA3*	14q32.1	Ataxin-3	$(CAG)_{12-41}$	—	$(CAG)_{40-84}$	coding	GOF	P
Spinocerebellar ataxia type 6	AD	*CACNA1A*	19p13	α_{1A} VDCCS	$(CAG)_{7-18}$	$(CAG)_{20-237}$	$(CAG)_{21-27}$	coding	GOF	ND
Spinocerebellar ataxia type 7	AD	*SCA7*	3p12-13	Ataxin-7	$(CAG)_{7-17}$	—	$(CAG)_{38-130}$	coding	GOF	P
Spinocerebellar ataxia type 8	AD	*SCA8*	13q21	ND	$(CTG)_{16-92}$	—	$(CTG)_{107-127}$	noncoding	ND	M
Myotonic dystrophy	AD	*DMPK*	19q13	DM protein kinase	$(CTG)_{5-37}$	—	$(CTG)_{50-3000}$	3' UTS	DN	M
FRAXA syndrome	XL	*FMR1*	Xq27.3	FMRP	$(CGG)_{6-45}$*	$(CGG)_{60-200}$	$(CGG)_{200-1000}$	5' UTS	LOF	M
FRAXE syndrome	XL	*FMR2*	Xq28	FMR2 protein	$(CGG)_{7-35}$	$(CGG)_{130-150}$	$(CGG)_{230-750}$	5' UTS	LOF	ND

Note: AD, autosomal dominant; AR, autosomal recessive; DN, dominant negative; GOF, gain of function; LOF, loss of function; M, maternal; ND, not determined; P, paternal; PABP2, poly(A) binding protein 2; PLE, parent limited expansion; PLOF, partial loss of function; UTS, untranslated sequence; VDCCS, voltage-dependent calcium channel subunit; XL, X-linked; XLD, X-linked dominant. **FRAXA* alleles having 45 to 60 repeats are known as grey zone alleles.

1.5.3 GAC Trinucleotide Repeats

The expansion of $p(CAG)_n$ repeats underlies the molecular etiology of many neurodegenerative diseases including the classic example of Huntington disease. This adult onset, autosomal dominant, progressive, neurodegenerative disease is associated with selective neuronal cell death causing abnormal involuntary movements, motor impairment, psychiatric disturbances, and dementia (Scherzinger *et al.*, 1997; Wilmot and Warren, 1998). Normally, Huntington disease is an adult onset disorder that is not fully penetrant. However, paternal inheritance often leads to a more severe juvenile onset form which includes dystonia, rigidity and seizures (Wilmot and Warren, 1998). The $p(CAG)_n$ expansion, located within exon 1 of the *IT15* gene on chromosome 4p16.3, results in an expanded polyglutamine tract within the protein product, huntingtin (Wilmot and Warren, 1998). Huntingtin, whose function is unknown, is expressed cytoplasmically in a wide range of cell types (Gusella and MacDonald, 1996). However, expanded alleles acquire a toxic gain of function causing the huntingtin protein to aggregate and accumulate in the nucleus of neurons, resulting in cell death (Difiglia *et al.*, 1997; Scherzinger *et al.*, 1997).

Dentatorubral-pallidoluysian atrophy (DRPLA) is a rare autosomal dominant disorder, most commonly found in Japan. This disorder results from a $p(CAG)_n$ expansion located on chromosome 12p13.31 within the *DRPLA* gene, which codes for atrophin-1 (Wilmot and Warren, 1998). Dentatorubral-pallidoluysian atrophy shows prominent anticipation involving neuronal loss resulting in an age of onset-dependent variation in degrees of ataxia, twitching,

myoclonic seizures, involuntary movement, and dementia with personality changes (Gusella and MacDonald, 1996; Wilmot and Warren, 1998). A second clinical phenotype, Haw River Syndrome, has been shown to result from the same expansion in *DRPLA*. This dominant neurodegenerative disease, identified in five generations of a single African-American family in rural North Carolina, is characterized by ataxia, spastic movements, seizures, dementia, mental retardation, and psychiatric disease (Burke *et al.*, 1994). Atrophin-1, the protein affected in both of these clinical phenotypes, contains extended polyglutamine tracts that result in abnormal formation of a complex containing disulfide bonds and ubiquitin (a protein which tags other proteins for proteosome degradation) (Yazawa *et al.*, 1999). It is speculated that the neurodegeneracy of dentatorubral-pallidoluysian atrophy and Haw River Syndrome is the result of the accumulation of these complexes in the nucleus of neurons (Igarashi *et al.*, 1998; Yazawa *et al.*, 1999).

Spinobulbar muscular atrophy, also called Kennedy disease, is an X-linked recessive disease caused by a $p(CAG)_n$ expansion in an exon of the androgen receptor gene, located on chromosome Xq11.2-q12 (Gisella and MacDonald, 1996; Wilmot and Warren, 1998). The DNA expansion encodes an expanded polyglutamine tract within the androgen receptor. This alteration in the androgen receptor is believed to cause its partial loss of function resulting in a mild insensitivity to androgen (Wilmot and Warren, 1998). In addition, this expansion also creates a toxic gain of function of the androgen receptor resulting

in neuronal involvement including degeneration of motor neurons and a mild sensory neuropathy (Wilmot and Warren, 1998).

The autosomal dominant spinocerebellar ataxias (SCA) type 1, 2, 3 (Machado-Joseph disease), 6, and 7 have been attributed to $p(CAG)_n$ expansions in the coding regions of *SCA1*, *SCA2*, *SCA3*, *CACNA1A*, and *SCA7*, respectively (Wilmot and Warren, 1998). The clinical phenotypes associated with these disorders overlap considerably and include features such as cerebellar ataxia, paralysis of the eye muscles, involuntary eye movement, parkinsonian signs, visual loss, exaggerated reflexes, and spasticity (Martin, 1999). The age of onset of spinocerebellar ataxia type 1 is correlated with, in addition to the length, the purity of the repeat within *SCA1*, which is normally interrupted by stabilizing CAT trinucleotides (Pearson *et al.*, 1998; Matsuyama *et al.*, 1999). Similarly, *SCA2* of unaffected individuals contains CAA interruptions (Matsuyama *et al.*, 1999). Four of the SCA genes (*SCA1*, *SCA2*, *SCA3*, and *SCA7*) code for soluble proteins called ataxins, the functions of which are unknown. Spinocerebellar ataxia type 6 involves the *CACNA1A* gene that codes for a membrane bound protein, the α_{1A}-voltage dependent calcium channel subunit (Wilmot and Warren, 1998). When a $p(CAG)_n$ expansion is present in any of these five SCA genes, the resulting proteins contain an enlarged polyglutamine tract which leads to a toxic gain of function. It is speculated that these glutamine stretches inhibit the protein degradation process performed by the proteosomes (Martin, 1999). Fragments of the partially degraded protein form a complex with portions of the proteosome and ubiquitin.

These complexes are translocated into the nucleus where they aggregate, forming intranuclear inclusions which may alter the transcription of other genes and eventually lead to neuronal death (Martin, 1999).

1.5.4 CTG Trinucleotide Repeats

Spinocerebellar ataxia type 8 is different than its counterparts since the molecular etiology involves an expansion of a untranslated $p(CTG)_n$ repeat (Koob *et al.*, 1999). This ataxia is also inherited in an autosomal dominant fashion and has clinical features and pathology much like the other spinocerebellar ataxias. The *SCA8* gene is not fully characterized and the protein product has not yet been identified.

Myotonic dystrophy is an autosomal dominant disease associated with progressive muscle weakness and spasms, cataracts, endocrine dysfunction, cardiac conduction system abnormalities, frontal balding, and mild mental impairment (Gusella and MacDonald, 1996; Wilmont and Warren, 1998; Timchenko, 1999). Myotonic dystrophy displays genetic anticipation with a wide range of severity including a congenital form with hypotonia, facial diplegia, swallowing difficulties, and severe mental retardation (Wilmont and Warren, 1998). The molecular etiology of this disease is a $p(CTG)_n$ expansion in the 3' UTS of a serine-threonine protein kinase, myotonic dystrophy protein kinase (DMPK) (Wilmont and Warren, 1998; Timchenko, 1999). Three possible roles of the expansion in the pathogenesis of myotonic dystrophy have been hypothesized: 1.) the expansion affects the expression of DMPK at the transcriptional or post-translational level; 2.) the expansion alters the expression

of a neighboring gene, the myotonic dystrophy-associated homeodomain gene; and/or 3.) the CUG repeats within the DMPK mRNA interact with certain proteins and, thereby, affect the processing of other mRNAs (Timchenko, 1999).

1.6 Why Are Only the CGG Trinucleotides Associated With Fragile Sites

If DNA synthesis is being interrupted by a disturbance in the nucleotide pool, why are only the $p(CGG)_n$ repeats associated with fragile sites? Since Sutherland's theoretical proposal, additional evidence has accumulated in support of the interrupted DNA synthesis theory. The progression of DNA polymerase through $p(CTG)_n$ and $p(CGG)_n$ repeats has, in fact, been proven to be interrupted during DNA synthesis (Kang *et al.*, 1995). The strength of this pausing is directly proportional to the number of triplets: a continuous length of at least 80 CTG and 61 CGG repeats is required for pausing to occur (Kang *et al.*, 1995). However, this pausing is eliminated when interruptions exist within the repeat tract, such as the AGG interspersions in the $p(CGG)_n$ tract of the *FMR1* gene (Samadashwily *et al.*, 1997).

The pausing during replication of expanded CGG repeat tracts is also temperature dependent. An increase in temperature decreases the strength of pausing, suggesting that DNA secondary structures, stabilized by hydrogen bonds, are formed within these expanded repeat tracts (Kang *et al.*, 1995). Such secondary structures are also formed within CTG repeat tracts, but are less stable. These secondary structures block replication and are speculated to be involved in the slipped-strand model proposed for the expansion mechanism of trinucleotide repeats (Moore *et al.*, 1999). The secondary structure that is

thought to be involved in trinucleotide repeats is the hairpin (Moore, 1999). A hairpin structure forms when a section of DNA is displaced from its complementary strand enabling a loop to form. The loop is stabilized by hydrogen bonds between complementary base pairs (Mitas, 1997). Further evidence for pausing within $p(CGG)_n$ expansions during DNA synthesis has been shown in expanded *FMR1* alleles of affected males in which the timing of replication is delayed from late S phase to G_2M phase (Hansen *et al.*, 1993).

Considerable delays in replication, followed by the recruitment of DNA repair proteins to remove secondary structures, may inhibit nucleosome formation and, thus, chromosome condensation prior to entry into G_2M. In addition to trinucleotide expansion, replication blocks resulting in delayed DNA synthesis may be important in the genesis of fragile sites (Usdin and Woodford, 1995).

Fragile site induction methods are not 100% efficient since gaps or breaks are visualized in only a fraction of metaphase spreads from an individual. For example, the rare folate sensitive fragile sites are induced in an average of approximately 30% of metaphase spreads (Sutherland and Richards, 1995). The mechanism currently proposed for the induction of rare folate sensitive fragile sites can also account for this inefficiency. In most cells, the DNA within the fragile site is likely replicated or repaired prior to the entry into mitosis. However, this process may be incomplete in a small proportion of cells. It is in these cells that the DNA packaging is not complete and fragile sites are

identifiable (Sutherland and Richards, 1999). Why then are only the $p(CGG)_n$ expansions associated with fragile sites?

Condensation of chromosomes requires packaging of the DNA into chromatin. The first level of packaging involves the creation of nucleosomes whereby DNA is wrapped around histones proteins. Differences at this level of packaging exist between the $p(CGG)_n$ and $p(CAG)_n$ repeats. Repeats containing as few as six CTG triplets (on the strand complementary to the CAG triplets) have been shown to preferentially bind to histones and form very stable nucleosome structures (Godde and Wolffe, 1996). However, blocks of more than 50 pure CGG repeats show a decreased efficiency in nucleosome assembly as compared to DNA of mixed sequence (Wang and Griffith, 1996). This decreased efficiency may be due, in part, to the stiffness (decreased flexibility of the twist angle of the double helix) of regions containing CGG triplets in contrast to the CTG triplets (Sarai *et al.*, 1989). When methylation exists within these repeats, this inefficiency is more dramatic (Godde *et al.*, 1996). If the assembly of nucleosomes is inhibited, condensation of DNA cannot occur, thereby increasing the accessibility of these regions, including upstream CpG islands, to DNA methyltransferase (Wang and Griffith, 1996). The resulting methylation, known to inactivate genes, decreases nucleosome assembly even more, resulting in a visible area of non-chromatization allowing the cytogenetic visualization of a gap in a metaphase chromosome - a fragile site (Wang *et al.*, 1996; Wang and Griffith, 1996)

1.7 Anticipation, Repeat Expansions and Psychiatric Disorders

Genetic anticipation is defined as an increase in severity or decreasing age of onset of a disease in successive generations within a pedigree (McInnis and Margolis, 1998). This concept, first documented in 1857, arose from studies of families displaying physical and mental degeneration involved in psychiatric diseases. Interest in genetic anticipation did not resurface until a century later when it was further demonstrated in myotonic dystrophy (Howeler *et al.*, 1989). Since that time, several disorders have been shown to display anticipation. In 1991, the cloning of the *FRAXA* locus implicated unstable trinucleotide repeats as the underlying etiology of the FRAXA syndrome. This etiology has since been proven in many other diseases displaying anticipation such as myotonic dystrophy, Huntington disease, spinobulbar muscular atrophy, spinocerebellar ataxia (types 1, 2, 3, 6, and 7), and dentatorubral-pallidoluysian atrophy. However, while certain psychiatric disorders (such as schizophrenia, bipolar disorder and autism) also display anticipation, no experimental evidence has yet proven the involvement of repeat expansions in their underlying etiologies (McInnis and Margolis, 1998).

Interestingly, although the genes involved in the trinucleotide repeat disorders are widely expressed, the clinical characteristics of the disorders appear to be limited to certain regions of the brain (McInnis and Margolis, 1998). In fact, all the trinucleotide repeat expansion disorders, except Friedreich ataxia, display some level of mental handicap, dementia, psychosis, or other abnormal cognitive difficulties (Orr, 1994; Wilmot and Warren, 1998). The clinical

phenotype associated with Friedreich ataxia, which is a multisystem degenerative disease, involves the heart, spinal cord and peripheral nerves. Thus, the central nervous system appears to be the most sensitive tissue to trinucleotide expansions. Why this would be true is still unknown. However, this tissue specificity also lends support to the belief that psychiatric disorders may result from trinucleotide repeat expansions.

Despite the high prevalence of psychiatric disorders, various factors have made it difficult to identify predisposing genes. The mode of inheritance of these disorders, while still unclear, is complicated by the phenotypic variability seen within and between families. A genetic component has been established through adoption, twin and family studies (McInnis and Margolis, 1998). For example, the risk of schizophrenia is higher among family members of patients than in the general population. The risk increases with each affected family member, up to approximately 50% when both parents are affected (Schultz, 1999). However, the concordance rate between monozygotic twins is only about 50% (Jennings, 1999). Thus, it is clear that the inheritance pattern is not a simple dominant or recessive pattern associated with a single gene. In fact, a polygenic mode of inheritance, involving genetic heterogeneity, has been postulated that involves the additive effects of several interacting genes and perhaps an environmental component (Smith *et al.*, 1997; Schultz, 1999).

1.8 Hypothesis and Project Aims

Fragile site *FRA19B*, located within human chromosome band 19p13.1, has been identified in two families: one Manitoban family affected with mental

retardation, autism and schizophrenia and one apparently clinically unaffected Danish family (Tommerup *et al.*, 1985; Chodirker *et al.*, 1987). The rarity of this site suggests that the DNA sequence involved in the fragility is normally quite stable. This stability may be due to the presence of intervening sequences which, when lost or altered, result in an expansion. The rarity of *FRA19B*, in combination with its location in an early-replicating and gene-rich chromosome band, suggests there is a high probability that the *FRA19B* locus contains one or more genes. However, the fragile site may not have a direct clinical effect on a gene, but may indirectly affect neighboring genes. Considering the clinical phenotypes that exist in the Manitoban family but not in the Danish family, an affected gene(s) may play a role in normal mental function. However, when affected by a mutation, in combination with additional mutations within the genome, psychiatric disorders may ensue.

A region of 19p13, which originally spanned 5 million base pairs (5 Mb), was under investigation to determine the human genomic DNA sequence of the *FRA19B* locus. This region was previously narrowed down to 30 kb within cosmid 25308 by utilizing fluorescent *in situ* hybridization (Appendix A). The first objective of this project was to identify the molecular basis of *FRA19B.* Based on previously cloned rare folate sensitive fragile sites, the molecular basis for fragility at the *FRA19B* locus is hypothesized to be due to an expansion of a CG-rich trinucleotide repeat that is normally stabilized by intervening sequences. The location of such a repeat is speculated to be within or near a gene involved

in mental function. Thus, the second objective was to identify a potential gene(s) within *FRA19B*.

2.0 MATERIALS AND METHODS

2.1 *FRA19B* Positive Human Genomic DNA Samples

To date, two families have been identified in which *FRA19B* is segregating. The first family, illustrated as A in Figure 3, was identified in Denmark (Tommerup *et al.*, 1985). DNA samples have been obtained from three *FRA19B* positive members of this family (I-1, II-2, and II-4). The second family, illustrated as B in Figure 3, was identified in Manitoba, Canada (Chodirker *et al.*, 1987). Blood samples had been obtained previously from two *FRA19B* positive members of this family (II-3 and II-4) and were transformed with an Epstein Barr virus into lymphoblastoid cell lines. Genomic DNA had been isolated from both the blood samples and the cell lines. These two cell lines were also utilized previously to prepare cytogenetic cultures in which the *FRA19B* was induced by folic acid deficiency.

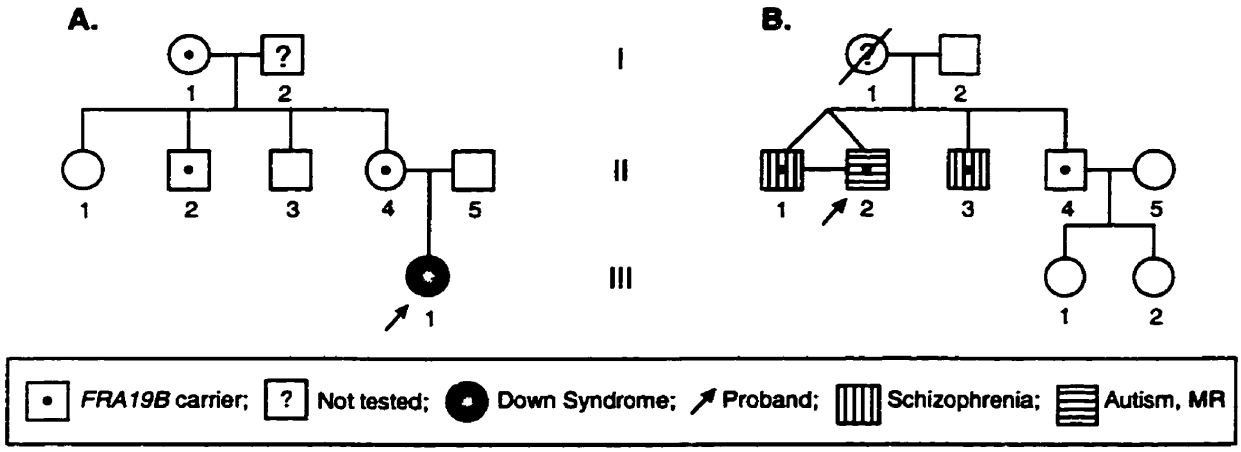

Figure 3: Segregation of *FRA19B* in A.) a Danish family and B.) a Manitoban family. MR indicates mental retardation. (Chodirker *et al.*, 1987; Tommerup *et al.*, 1985)

2.2 Cosmids Spanning *FRA19B*

Previous fluorescence *in situ* hybridization (FISH) on the cultures of individuals II-3 and II-4 from the Manitoban family, induced to express *FRA19B*, identified two cosmids, 25308 and 16864 (gifts from Dr. H. Mohrenweiser at Lawrence Livermore National Laboratory), which spanned the fragile site *FRA19B* located at 19p13.1 (Appendix A). The *Eco*RI restriction fragments of these two cosmids have been aligned based on various overlapping cosmids in this area (http://www-bio.llnl.gov/rmap-bin/attrib_select) (Figure 4).

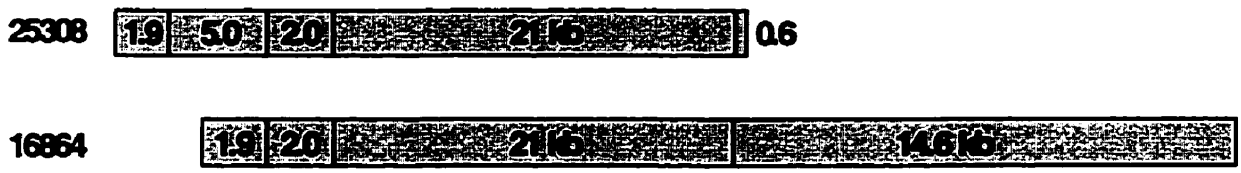

Figure 4: *Eco*RI restriction fragments of cosmids 25308 and 16864.

2.3 Isolating Cosmid DNA

2.3.1 Preparation of Superbroth

Superbroth was utilized as the growth medium in order to obtain adequate growth of cosmid cultures. The following ingredients were dissolved in 800 ml of NANOpure™ H_2O: 8 g tryptone, 5 g yeast extract and 1.25 g NaCl (Ausubel *et al.*, 1996). The medium was titrated to a pH of 7.0 with approximately 125 µl of 10 N NaOH and finalized to a volume of 1 l with NANOpure™ H_2O. Aliquots of 500 ml were transferred to 1 l Erlenmeyer flasks containing a magnetic stir bar. The superbroth was sterilized by autoclaving on a liquid cycle for 20 min. at

121°C and 15 psi (pounds per square inch). After the medium was allowed to cool, it was supplemented with 0.25 g sucrose and the antibiotic kanamycin to a final concentration of 20 μg/ml. The medium was then stored at 4°C.

2.3.2 Large-Scale Alkaline Lysis Method

Approximately 50 μl of cosmid 25308 (stored frozen in 50% glycerol) was utilized to inoculate 250 ml of superbroth plus 500 μl of 10 mg/ml kanamycin in a 1 l Erlenmeyer flask. Two individual cultures were set up and incubated in a 37°C shaker overnight. The cultures were transferred to 250 ml tubes and centrifuged for 5 min. at 10,000 rpm in a Beckman JA-17 rotor which had previously been cooled to 4°C. The medium was removed, the pellets were vortexed, combined, and then resuspended in 3 ml of GTE [50 mM glucose, 25 mM Tris·HCl (pH 8.0), 10 mM EDTA] for a total volume of approximately 5 ml. Eight hundred μl of 25 mg/ml hen egg white lysozyme (freshly made with NANOpure™ H_2O) was mixed in with the cells and then incubated at room temperature for 10 min. to assist in the degradation of the bacterial cell walls and subsequently increase the efficiency of the release of cosmid DNA. Ten ml of freshly prepared lysis buffer (1% SDS, 0.2 N NaOH) was added and gently mixed. This mixture was incubated on ice for 10 min. To neutralize the solution and precipitate the cellular proteins and bacterial chromosomes, 7.5 ml of 5 M potassium acetate (KOAc) was added. The solution was gently mixed. The mixture was incubated on ice for 10 min. and then centrifuged once for 10 min. at 10,000 rpm (JA-17) to pellet out the bacterial chromosomal DNA, SDS:protein

complexes and other cellular debris. The supernatant, containing the cosmid DNA, was decanted into a 50 ml tube to which 11.5 ml each of phenol and chloroform were added. After vigorous mixing, the two phases were separated by centrifugation for 10 min. at 5,000 rpm (JA-17). The upper aqueous layer was removed to a fresh 50 ml tube and 2.3 ml of NaOAc plus 24 ml of isopropanol were added. The tube was gently mixed by inversion and placed at -20°C overnight. The nucleic acids were then pelletted by centrifugation for 10 min. at 5,000 rpm (JA-17). The supernatant was decanted and the pellet was washed three times with 70% ethanol and allowed to air dry. The pellet was resuspended in 600 µl of TE [10 mM Tris·HCl (pH 8.0), 1 mM EDTA]. To digest the RNA, 1 µl of 25 mg/ml RNase A was added prior to an incubation at 37°C for at least 30 min. To quantify the cosmid DNA, a dilution of 5 in 300 was prepared with NANOpure™ H_2O and the absorbance was read at a wavelength of 260 nm. The absorbance was multiplied by the dilution factor (60) and then by 50 (50 µg of DNA has an absorbance of 1) to calculate the quantity of DNA (ng/µl).

2.4 Isolating Plasmid DNA

2.4.1 Preparation of Luria Bertani (LB) Medium and Agar Plates

The following ingredients were dissolved in 700 ml of NANOpure™ H_2O: 5 g yeast extract, 10 g tryptone and 10 g NaCl. The medium was titrated to a pH of 7.0 with 10 N NaOH and finalized to a volume of 1 l. The LB medium was sterilized by autoclaving as previously described. The medium was cooled to

55°C, supplemented with the antibiotic ampicillin to a final concentration of 50 μg/ml and stored at 4°C.

To prepare LB agar plates, 10 g of agar were added to the medium prior to autoclaving. As the medium was cooling after sterilization, it was mixed to distribute the agar. When cooled to 55°C, ampicillin was added to a final concentration of 50 μg/ml. The warm medium was quickly poured into plastic petri dishes in aliquots of approximately 30 ml, allowed to solidify, and dried at room temperature for approximately three days while protected from light.

For blue/white color selection of transformed *E. coli*, 70 μl of 25 μg/μl 5-bromo-4-chloro-3-indoyl-β-D-galactopyranoside (X-Gal) in dimethyl formamide (at a final concentration of 50 μg/ml) was aseptically spread onto the LB plates supplemented with ampicillin (50 μg/ml) and allowed to soak into the agar.

2.4.2 Alkaline Lysis Miniprep Method

Plasmid DNA was isolated from 2 ml *E. coli* overnight cultures by the alkaline lysis miniprep method (Birnboim and Doly, 1979). Approximately 1.5 ml of culture was transferred into a 1.7 ml microfuge tube. A 30 s centrifugation to pellet the bacterial cells was performed in a table top microcentrifuge. The LB medium was removed, the pellet vortexed and resuspended in 100 μl of cold GTE. The bacteria were lysed by the addition of 200 μl of lysis buffer (0.2 N NaOH, 1% SDS) followed by three inversions to gently mix the solution. The cellular proteins and the bacterial chromosomes were precipitated by the

addition of 150 µl of ice-cold 5 M KOAc (pH 4.8), followed by vigorous mixing. This debris was pelletted by centrifugation for 5 min. at 4°C. The supernatant was removed to a clean 1.7 ml microfuge tube and extracted with one volume of phenol:chloroform (1:1). The two phases were then separated by centrifugation for 1 min.. The aqueous (top) layer was removed to a clean 1.7 ml microfuge tube and the plasmid nucleic acids were precipitated with the addition of 50 µl of 3 M NaOAc and 1 ml of 100% ethanol. After a 30 min. incubation at -20°C, the nucleic acids were pelletted by centrifugation for 5 min., washed with 70% ethanol, air dried, and resuspended in 50 µl of TE. The RNA was digested by the addition of 2 µl of 10 mg/ml RNase A followed by an incubation at 37°C for at least 30 min.

2.5 Restriction Endonuclease Digestion of Human Genomic DNA

Approximately 10 µg of human genomic DNA was digested with 20 units of enzyme (Gibco BRL). The reaction was carried out in a final volume of 45 µl which included the appropriate 1X Buffer (Gibco BRL) and NANOpure™ H_2O. The reaction was incubated at the appropriate temperature for 1½ hours. To ensure complete digestion of the DNA, 5 µl of each digest was separated by electrophoresis in a 1% agarose gel. If the digestion was incomplete, an additional 20 units of enzyme were added to the remaining sample which was then incubated for an additional 1½ hours at the appropriate temperature. When the DNA samples were completely digested, they were separated by electrophoresis at 50 V for at least 18 hours (section 2.6).

2.6 DNA Agarose Gel Electrophoresis

2.6.1 Preparation of an Agarose Gel

A 1% agarose gel was prepared by adding 2.5 g of powdered electrophoresis grade agarose (Gibco BRL) to 250 ml of 1X TBE (2.5 mM NaOH, 89 mM Tris base, 89 mM boric acid, 2 mM EDTA). The agarose was dissolved by boiling (approximately 3 min. on high power) in a 1.25 kW microwave (Cimco Inc., Model MGC843) and then cooled to approximately 50°C. Ethidium bromide, which intercalates within the major groove of a DNA helix and fluoresces upon exposure to UV light (λ of 254 nm), was added to a final concentration of 0.6 μg/ml (Sambrook *et al.*, 1989). The agarose was then poured into a horizontal gel mold and allowed to solidify, creating a 24 x 13.5 cm gel with 16 wells each with a capacity of approximately 50 μl.

2.6.2 Electrophoresis of DNA Samples

A 1% agarose gel was placed in a horizontal electrophoretic unit containing 2 l of 1X TBE. DNA samples, previously digested with a restriction endonuclease, were mixed with 0.1 volumes of a 10X loading dye [0.25% (w/v) xylene cyanol, 50% (w/v) sucrose, 10 mM Tris·HCl (pH 8.0)] and loaded into individual wells. The DNA fragments were separated by electrophoresis at 30 to 200 V for 2 to 18 hours. Approximately 600 ng of λ/*Hind*III, 1 kb and 500 bp DNA ladders (Gibco BRL) were loaded alongside the digested DNA samples as size references. The migration of the DNA fragments was monitored by visualization with UV light (λ of 254 nm) from a Spectroline® transilluminator

(Spectronics Corporation, Model TR-254). For determination of fragment size, fluorescent rulers were placed along each side of the gel, which was then placed on a transilluminator. A photograph was taken using a Polaroid type 667 film and a Polaroid MP-4 land camera fitted with a Kodak #9 Wratten filter (f8, 10 s exposure).

2.7 Southern Blotting

To ensure the Southern transfer of fragments larger than 4 kb, the agarose gel was partially depurinated in 400 ml of 0.25 N HCl and gently mixed for 30 min. (after which time the xylene cyanol had changed in color from blue to green). The gel was then rinsed with NANOpure™ H_2O and placed in 400 ml of fresh 0.4 N NaOH and gently mixed for 20 min. This reduced the pH of the gel and subsequently denatured and imparted a negative charge on the DNA fragments.

The Southern transfer was performed using a TurboBlotter™ Rapid Downward Transfer System (Sclessinger and Schull). A stack of paper towels, cut to the size of the gel, was placed on the base of the TurboBlotter™. Two pieces of 3MM Whatman® chromatography paper (Whatman International Ltd.), cut to the size of the gel and pre-wet with fresh 0.4 N NaOH, were placed on the paper towels. A long plastic tube (2 cm in diameter) was used to roll out any air bubbles trapped between the paper. A piece of nylon Zeta-Probe® GT blotting membrane (Bio-Rad), cut to the size of the gel and pre-wet with NANOpure™ H_2O, was placed on the 3MM Whatman® chromatography paper. Air bubbles were again removed. The gel was carefully placed on the nylon membrane

ensuring no air bubbles were trapped between the gel and nylon membrane. Two pieces of 3MM Whatman® chromatography paper, cut to the size of the gel and pre-wet with fresh 0.4 N NaOH, were placed on the gel. To act as a wick, two pieces of 24 x 27 cm 3MM Whatman® chromatography paper were pre-wet with fresh 0.4 N NaOH and placed on the top of the blotting assembly. The two sides of these top pieces were placed in the middle part of the TurboBlotter™ which contained fresh 0.4 N NaOH. To avoid drying of the assembly, the top part of the TurboBlotter™ was placed on the wick. The transfer was allowed to proceed, with only the aid of gravity, for 3 hours. The blotting assembly was dismantled, the Southern blot was rinsed with 2X SSC and then air dried between two pieces of 3MM Whatman® chromatography paper.

2.8 Labeling DNA to Produce Radioactive Probes

2.8.1 Random Prime Method

Approximately 100 ng of probe DNA was labeled with [α-^{32}P]dCTP by the random priming method in a reaction volume of 25 µl (Feinberg and Vogelstein, 1983; 1984). The probe DNA was mixed with 1 to 5 µg of pd(N)$_6$ primer (Pharmacia) and NANOpure™ H_2O. To anneal the primer to the template, this mixture was heated in a Perkin Elmer Thermocycler at 100°C for 10 min. and immediately placed on ice. While still on ice, 2.5 µl of 0.5 mM deoxynucleotide mix (dATP/dGTP/dTTP), 2.5 µl of 10X Klenow Buffer [0.5 M Tris·HCl (pH 7.5), 0.1 M $MgCl_2$, 50 mM dithiothreitol (DTT), 0.5 M KCl, 0.5 mg/ml BSA], 50 µCi of [α-^{32}P]dCTP (NEN Life Science Products), and 5 units of Klenow fragment were

added. The labeling reaction was allowed to incubate at room temperature for 2 to 4 hours and then stopped with the addition of 1 µl of 0.5 M EDTA. The unincorporated nucleotides were removed from the labeled DNA as follows. The labeling reaction was added to a NICK® column containing Sephadex® G-50 (Pharmacia) which had been rinsed and equilibrated with one column volume of Elution Buffer (1X SSC, 0.1% SDS). After the mixture entered the gel bed, 400 µl of Elution Buffer was added. The eluate was discarded. An additional 400 µl of Elution Buffer was added and the eluate was collected in a 1.7 ml microfuge tube. This eluate, containing the labeled probe, was quantified by adding a 1 µl aliquot to 1 ml of ScintiSafe™ Econo 1 scintillation fluid (Fisher Scientific). A scintillation count was performed in duplicate in a scintillation counter (Beckman, Model LS1801). The specific activity of the probe (cpm/µg) was calculated by multiplying the average scintillation count (cpm) by 400 (the total volume in µl) and dividing by the amount of probe utilized (µg). The probe, having a specific activity of at least 1×10^8 cpm/µg, was mixed with 500 mg of salmon sperm DNA, denatured for 10 min. at 100°C and chilled briefly on ice.

2.8.2 Prehybridization of Repeat Containing Probe

Approximately 100 ng of DNA was labeled with [α-^{32}P]dCTP by random primed synthesis and cleaned with a NICK® column as described above. The specific activity, which had to be at least 1×10^9 cpm/µg for this method, was calculated. The probe was prehybridized with human DNA to suppress repetitive DNA elements (Ardeshir *et al.*, 1983). The 400 µl of labeled probe was mixed

with 10 µl of 10 mg/ml native sonicated salmon sperm DNA (average size approximately 1 kb) and 10 µl of 100 mM spermine tetrahydrochloride. This mixture was precipitated on ice for 1 hour and then centrifuged at 4°C for 30 min. After removing the supernatant, the pellet was resuspended in 34 µl of sodium phosphate buffer (0.5 M NaH_2PO_4, 0.5 M Na_2HPO_4). To aid resuspension, the tube was incubated at 65°C for 5 min. Two hundred and fifty µl of 2.5 mg/ml human cot-1 DNA, sonicated to an average size of 400 to 700 bp (Sigma), was mixed in. After a brief centrifugation, the mixture was heated at 100°C for 10 min. and chilled on ice. A quick centrifugation was performed prior to incubating the mixture at 65°C for 1½ to 2 hours. This prehybridized, radiolabeled probe was not denatured.

2.8.3 5' End Labeling

Approximately 0.5 pmol of a $(CGG)_5$ oligonucleotide (University of Manitoba DNA Laboratory) was mixed with 10 µCi of [γ-^{32}P]dATP (NEN Life Science Products), 10 units of T4 Polynucleotide Kinase (Gibco BRL), and 1X Exchange Buffer [250 mM imidazole·HCl (pH 6.4), 60 mM $MgCl_2$, 5 mM 2-mercaptoethanol, 350 µM ADP] (Gibco BRL) in a final volume of 5 µl. The labeling reaction was incubated at 37°C for 45 min. The labeled oligonucleotide was purified through a NAP™-5 column containing Sephadex® G-25 (Pharmacia) to remove the excess [γ-^{32}P]dATP. The NAP™-5 column was first equilibrated with three volumes of 1X SSC/1% SDS. The reaction was loaded onto the column and 1.0 ml of 1X SSC/1% SDS was then passed through

the column and the eluate was discarded. Subsequently, an additional 1.5 ml of 1X SSC/1% SDS was passed through the column and the resultant eluate was collected and used as a probe.

2.9 Southern Blot Hybridization

2.9.1 Random Prime Labeled Plasmid Probes

The Southern blot was pre-wet with 1X SSC/0.1% SDS and rolled into a Hybridization Tube (Robbins Scientific) with the aid of a 10 ml pipette. The blot was prehybridized with 5 ml of Westneat solution [7% SDS, 1 mM EDTA (pH 8.0), 0.263 M Na_2HPO_4, 1% BSA] (Westneat *et al.*, 1988) and 500 mg of denatured salmon sperm DNA for 1 hour at 65°C in a Hybridization Incubator (Robbins Scientific).

A denatured probe, radiolabeled by the random primed synthesis method, was added to the prehybridized membrane. Hybridization was performed overnight at 65°C. The hybridization mixture was poured out of the hybridization tube and, with the blot still in the tube, three washes were performed with approximately 30 ml of 4X SSC/0.05% SDS. The blot was then removed from the tube, placed in a plastic container, washed at room temperature for 30 min. each with 500 ml 2X SSC/0.05% SDS and 500 ml of 1X SSC/0.05% SDS. If the blot was still over 2,000 cpm, as monitored with a Geiger Counter, the stringency of the wash solution was increased. The following 20 to 30 min. individual washes were performed until the blot was below 2,000 cpm as measured by a Geiger counter: 1X SSC/0.05% SDS at 65°C, 0.1X SSC/0.1% SDS at room

temperature, 0.1X SSC/0.1% SDS from room temperature to 65°C, and 0.1X SSC/0.1% SDS at 65°C. An autoradiograph was obtained by placing the blot in a sealed bag and exposing it to a Biomax™ MS Scientific Imaging Film (Kodak) in a developing cassette containing intensifying screens at -80°C for a variable, pre-determined length of time for each probe.

2.9.2 Random Prime Labeled PCR Probes

The Southern blot was pre-wet with 1X SSC/0.1% SDS and rolled into a Hybridization Tube. The blot was prehybridized with 5 ml of ExpressHyb (Clontech) in the Hybridization Incubator (Robbins Scientific) for 30 min. at 65°C. The specific activity (cpm) of the probe was utilized to determine the volume required to have exactly 1 x 10^7 cpm (2 x 10^6 cpm per ml of ExpressHyb). This volume was transferred to a 0.65 ml microcentrifuge tube, heated at 100°C for 5 min., quickly chilled on ice, and added to the prehybridized blot. Hybridization was allowed to proceed for 1 hour at 65°C. The blot was washed as previously described. An autoradiograph was obtained at room temperature by placing the blot in a sealed bag and exposing it for 2 to 4 hours to a Biomax™ MS Scientific Imaging Film (Kodak) in a developing cassette containing intensifying screens.

2.9.3 5' End-Labeled Oligonucleotide Probe

The Southern blot was pre-wet with 1X SSC/0.1% SDS and rolled into a hybridization tube. The blot was prehybridized with 5 ml of Westneat solution [7% SDS, 1 mM EDTA (pH 8.0), 0.263 M Na_2HPO_4, 1% BSA] (Westneat *et al.*, 1988) and 500 mg of denatured salmon sperm DNA in the Hybridization

Incubator (Robbins Scientific) for 1 hour at 33°C. The 5' end-labeled $(CGG)_5$ oligonucleotide was then added and hybridization was performed overnight at 33°C. The hybridization mixture was poured out of the hybridization tube and three washes were performed with approximately 30 ml of 4X SSC/0.05% SDS. The blot was then removed from the tube and washed with 500 ml of 4X SSC/0.05% SDS at room temperature for 30 min. The blot was placed in a sealed bag and exposed to a Reflection™ NEF Autoradiography Film (DuPont) for 1 to 14 days.

2.10 Developing Autoradiographs

The film was removed from the developing cassette, in a darkroom, with the use of only a red safety light. The film was placed in GBX Developer (Kodak) and gently shaken for up to 7 min. The developing process was arrested by shaking the film in a stop solution (0.35 M acetic acid) for 30 s, Rapid Fixer solution (Kodak) for 2 min. and then rinsing with H_2O. Scanning of autoradiographs was performed with DeskScanII software on a Hewlett Packard ScanJetIIcx scanner connected to a Power Macintosh G3 computer. Adobe Photoshop™ was utilized to decrease the intensity of the background of some autoradiographs.

2.11 Stripping and Storing Southern Blots

The radioactive probe had to be removed or "stripped" to facilitate reuse of the Southern blot. A solution of 0.05% SDS was boiled and then poured over the radioactive blot. The blot was gently shaken in a 90°C water bath for at least

1 hour. This process was repeated. To ensure that the probe had been effectively removed, an autoradiograph was obtained by exposing the blot, sealed in a bag, to a Biomax™ MS Scientific Imaging Film (Kodak) overnight. If the autoradiograph was blank, the Southern blot was then stored at -80°C until required. If a signal was observed on the autoradiograph, the stripping procedure was repeated.

2.12 Single-Stranded DNA Sequencing

Template DNA for sequencing was prepared as follows. Four µg of plasmid DNA was aliquoted into a 1.5 ml microcentrifuge tube in a total volume of 18 µl. The DNA was denatured at room temperature for 5 min. with 2 µl of 2 N NaOH. The single-stranded DNA was precipitated with the addition of 8 µl of 5 M NH_4OAc and 100 µl of 100% ethanol, followed by a 1 hour incubation at -20°C. Following centrifugation for 10 min. at 4°C, the supernatant was removed and the pellet was washed once with 70% ethanol. The pellet was dried for 5 min. using a Speed Vac Concentrator (Savant) in which the vacuum was created by a Freeze Dry System (Labconco). The pellet could now be stored at -20°C. Prior to sequencing, the pellet was resuspended in 7 µl of NANOpure™ H_2O. (The DNA was thoroughly washed off the walls of the tube.) The DNA template was then sequenced with the aid of a Sequenase Version 2.0 DNA Sequencing Kit (Amersham).

The 7.0 µl of single-stranded DNA was mixed with 2.0 µl of 5X Sequenase reaction buffer and 1.0 µl of 0.5 pmol/µl primer. This mixture was transferred to

a 0.65 ml microcentrifuge tube. The primer was annealed to the template by heating for 5 min. at 65°C and slowly cooling to 30°C over 30 min. The mix was spun briefly and put on an ice-H_2O bath. While still on ice, the following were added and mixed in thoroughly by pipetting: 1.0 µl of 0.1 M DTT, 2.0 µl of 1:4 dilution of 5X dGTP labeling mix, 0.5 µl of 12.5 mCi/ml [α-^{35}S]dATP (NEN Life Science Products), and 2.0 µl of diluted Sequenase (3.25 µl of Sequenase dilution buffer, 0.25 µl of 5 units/ml inorganic pyrophosphatase, and 0.5 µl of 13 units/µl Sequenase version 2.0 T7 polymerase). A brief centrifugation was then performed. For maximum extension, the mix was incubated at room temperature for 5 min. (A shorter incubation time of approximately 2.5 min. was performed if sequence closer to primer was required.) During this incubation, 2.5 µl of the four dideoxynucleotides were added to four separate tubes and warmed to 40°C. A 3.5 µl aliquot of sequence mix was transferred to each of the four dideoxynucleotides, mixed well and incubated at 40°C for 5 min. The reaction was stopped with the addition of 4 µl of Stop Buffer.

2.13 Polyacrylamide Gel Electrophoresis of Sequencing Reactions

2.13.1 Preparation of a Sequencing Gel

A 6% denaturing polyacrylamide gel was prepared by dissolving 25.2 g electrophoresis grade urea (Gibco BRL) in a solution containing 6 ml 10X TBE, 12 ml 29% acrylamide:1% bisacrylamide and 12 ml of NANOpure™ H_2O. This mixture was warmed on a hotplate while stirring to aid in the dissolving. The volume was finalized to 60 ml with NANOpure™ H_2O after

the urea had dissolved. Sixty-three µl of fresh 25% ammonium persulfate (APS) (Sigma) and 50 µl of *N', N', N', N'*-tetramethylethylene diamine (TEMED) (Sigma) were added to approximately 10 ml of this gel mix. This activated gel mix was quickly used to "plug" the bottom of the gel chamber of a 21 x 50 cm Sequi-Gen® sequencing apparatus (Bio-Rad). The remainder of the gel solution was mixed with 67 µl of 25% APS and 50 µl TEMED. The activated gel mix was poured into the gel chamber with the aid of a 60 ml syringe and an inverted 24 well sharkstooth comb was inserted to form one large well across the top of the gel. The gel was then allowed to polymerize overnight at room temperature.

2.13.2 Polyacrylamide Gel Electrophoresis

The polyacrylamide gel was immersed in a 1X TBE buffer, pre-warmed to 50°C, and current was applied for approximately 30 min. at 2,000 V or until the gel reached 55°C. With the aid of a syringe, the large well of the gel was rinsed with running buffer. The four sequencing reactions were heated for 2 min. at 77°C. Three µl of each reaction were loaded immediately onto the gel with the use of a 24 well sharkstooth comb. The sequencing gel was run at 55°C for 1.5 to 2 hours or until the bromophenol blue dye reached the bottom of the gel. The sequencing apparatus was disassembled and the gel was carefully transferred to a piece of 3MM Whatman® chromatography paper and covered with plastic wrap. The gel was dried in a Gel Dryer (Bio-Rad) connected to a vacuum source (Labconco) for at least 2 hours at 80°C. The plastic wrap was

removed and the dried sequencing gel was exposed to an audioradiography film overnight. The film was developed manually as described previously.

2.14 *Alu*-Splice PCR

Cosmid 25308 was found to contain numerous *Alu* repeats and, thus, to produce DNA fragments which could be utilized as probes, these repetitive sequences had to be eliminated. This was accomplished by a technique developed by Fuentes *et al.* (1997) and termed "*Alu*-splice PCR". This technique takes advantage of the presence of *Alu* elements and conserved splice sites. Primers homologous to either ends of an *Alu* element and consensus sequences for 5' and 3' splice sites were produced (ACTG Inc.) (Table 4). A restriction enzyme recognition site was incorporated into each primer. Four PCR reactions were performed using the four possible primer combinations: 1.) *Sac*II-3'spl and *Sal*I-A33, 2.) *Sac*II-3'spl and *Sal*I-A44, 3.) *Not*I-5'spl and *Sal*I-A33, and 4.) *Not*I-5'spl and *Sal*I-A44. The PCR reactions contained 1X PCR Buffer [20 mM Tris·HCl (pH 7.5), 100 mM KCl, 1 mM DTT, 0.1 mM EDTA, 0.5% Tween 20, 0.5% Nonidet P40, 50% glycerol], 3 mM $MgCl_2$, 0.15 mM dNTPs, 20 pmol splice primer, 10 pmol *Alu* primer, 1 unit Expand™ High Fidelity enzyme mix (Boehringer Mannheim), and 300 ng cosmid 25308. The final reaction volume was 40 µl and each of the four reactions were performed in triplicate. The reactions were carried out on a Cetus DNA Thermal Cycler (Perkin Elmer) under the following conditions: an initial denaturation of 5 min. at 94°C; 30 cycles of 30 s at 94°C, 40 s at 55°C, 2 min. at 74°C; and a final elongation of 4 min. at 74°C. The triplicate PCR reactions were combined and 40 µl of each

were mixed with 5 µl of 10X loading buffer and run on a 1% agarose gel/1X TBE to check for quality and quantity. If suitable, the remainder of the PCR reactions were then purified.

Table 4: *Alu*-splice primers. (W represents A or T, Y represents T or C, and N represents any nucleotide.)

Name	5'→3' Sequence
*Sac*II-3'spl	CGCCCGCGGTCNCAGGT
*Not*I-5'spl	CGCGCGGCCGCACWYACCW
Sal-A33	CGCGTCGACCACTGCACTCCAGCCTGGGCG
Sal-A44	CGCGTCGACGGGGATTACAGGCGTGAGCCAC

2.15 PCR Purification

The PCR products were purified with a QIAquick™ PCR Purification Kit to remove primers, nucleotides, polymerases, and salts (Qiagen). The remaining 80 µl of the PCR reaction was mixed with five volumes of Buffer PB (proprietary composition). This mixture was added to a QIAquick™ column, previously placed in a 2 ml collection tube, and centrifuged for 30 to 60 s to bind the DNA. The eluate was removed from the collection tube. A wash of 750 µl of Buffer PE (proprietary composition) was added to the column followed by centrifugation for 30 to 60 s. The eluate was again removed. To remove residual ethanol, the column was centrifuged for an additional 1 min. The QIAquick™ column was placed in a clean 1.7 ml microfuge tube. The DNA was eluted by the addition of 50 µl of Buffer EB (10 mM Tris·HCl, pH 8.5) to the center of the QIAquick™ column, allowing absorption for 1 min., followed by centrifuged for 60 s.

2.16 Restriction Endonuclease Digestion of Plasmid DNA

All four PCR reactions and the vector BlueScript SK+ (a gift from Dr. R.D. Gietz, Department of Human Genetics, University of Manitoba) were digested with *Sal*I. The digests were carried out in a final volume of 60 µl containing 1X *Sal*I Buffer (New England BioLabs), 20 units *Sal*I, 1X BSA, and 50 µl of purified PCR product or vector. The reactions were incubated at 37°C for 8.5 hours and the reactions were purified by extraction with 60 µl phenol:chloroform (1:1) to remove the enzymes. The mixture was vortexed for 1 min. and spun for 3 min. The aqueous layer was removed, placed in a clean 1.7 ml microfuge tube and precipitated overnight by the addition of 6 µl of 3 M NaOAc and 132 µl 100% ethanol. Following centrifugation for 10 min. at 4°C, the supernatant was removed and the pellet washed with 70% ethanol and allowed to air dry. The four PCR reactions were each resuspended in 50 µl of NANOpure^{TM} H_2O. The vector was resuspended in 100 µl and divided into two aliquots of 50 µl each. One of these aliquots and the two PCR reactions produced using the *Sac*I-3'spl primer were digested with *Sac*I. The second aliquot of vector and the two PCR reactions produced using the *Not*I-5' spl primer were digested with *Not*I. These digestions were performed in the same manner as described above and were resuspended in 50 µl. The small DNA fragments restricted from the ends of the PCR products were removed by PEG exclusion (Sambrook *et al.*, 1989). Thirty-three µl of 2.5 M NaCl/20% PEG were added to each restriction digest. The digests were incubated in an ice-H_2O bath for 3 hours and centrifuged for 10 min. at 4°C. The supernatant, containing the

small DNA fragments, was removed. The pellet, containing digested PCR product or vector, was washed with 70% ethanol and allowed to air dry. The samples and vectors were resuspended in 10 and 50 µl of TE, respectively.

2.17 Cloning

The PCR products of each of the four reactions were then shotgun cloned into the appropriately digested BlueScript SK+ vector. The 20 µl ligation reactions consisted of 2 µl 10X ligation buffer (a gift from Dr. R.D. Gietz) [0.66 M Tris·HCl (pH 7.5), 50 mM $MgCl_2$, 50 mM DTT, 10 mM ATP], 1 unit T4 DNA Ligase (Boehringer Mannheim), 5 µl digested PCR product, and 0.5 µl digested BlueScript SK+. The four ligation reactions were incubated at 11°C overnight, ethanol precipitated and resuspended in 20 µl of TE.

2.18 Transformation of DH5α *E. coli* with Plasmid DNA

2.18.1 Preparation of Electrocompetent DH5α *E. coli* Cells

An LB plate was streaked with DH5α *E. coli* (Gibco BRL) and grown overnight at 37°C. The following day, a single colony was utilized as an inoculum for a 50 ml starter culture of LB medium which was grown overnight in a 37°C shaking incubator. A 1/100 dilution of the starter culture was prepared and used as an inoculum for a 2 l LB culture, pre-warmed to 37°C. This culture was grown in a 37°C shaking incubator until the absorbance at 600 nm measured between 0.5 and 1.0 (approximately 2¼ hours). The culture was transferred to a bottle and centrifuged at 8670 g with a JA-14 rotor in a Beckman

J2-21 centrifuge pre-cooled to 4°C. The LB medium was decanted and the cells were resuspended in 2 l of ice-cold sterile NANOpure™ H_2O. Centrifugation was repeated to pellet the cells, the NANOpure™ H_2O was decanted and the cells were resuspended in 1 l of ice-cold sterile NANOpure™ H_2O. After pelleting, the cells were again resuspended in a total volume of 50 ml of ice-cold sterile NANOpure™ H_2O and transferred to an ice-cold 50 ml centrifugation tube. The cells were pelletted and resuspended in 20 ml of ice-cold 10% glycerol (filter sterilized). They were again pelletted and resuspended in 2 ml of ice-cold 10% glycerol. Aliquots of 50 µl were transferred to 0.65 ml microfuge tubes, flash frozen in liquid nitrogen and stored at -80°C (Dower *et al.*, 1988).

2.18.2 Electroporation of Electrocompetent DH5α *E. coli* Cells

DH5α *E. coli* cells were transformed with plasmid DNA by electroporation (Dower *et al.*, 1988). One µl of a ligation product was mixed with 25 µl of electrocompetent DH5α *E. coli* cells that had been allowed to thaw on ice. This mixture was transferred to an ice-cold electroporation cuvette (Bio-Rad) with a 0.1 cm aperture. The cells were subjected to an electric pulse which was generated by a Bio-Rad Gene Pulser™ set to 1.25 kV with a 25 µF capacitor connected to a pulse controller set to 400 Ω (Miller, 1988; Dower *et al.*, 1988). The pulse time constant ranged from 7 to 8.5 ms. Immediately following the pulse, 1 ml of LB medium was added to the cells, mixed, transferred to a 15 ml culture tube, and incubated in a 37°C shaker for 30 min. Fifty µl and 100 µl aliquots were spread onto solid LB plates supplemented with ampicillin

(50 μg/ml) and topped with 70 μl of X-Gal (25 mg/ml) in dimethyl formamide. These plates were then incubated for approximately 30 min. at 37°C, inverted and incubated overnight.

From the plates of each of the four reactions, approximately 25 colonies, ranging in color from white to light blue, were selected and spread with a sterile wooden toothpick into 1 cm x 1 cm patches on LB plates supplemented with ampicillin (50 μg/ml). These were inverted and incubated at 37°C overnight. A sterile toothpick was utilized to scrape off a small amount of the colony patch which was then used to inoculate 2 ml of LB medium supplemented with ampicillin (50 μg/ml). Two cultures were set up per colony patch. These cultures were incubated in a 37°C shaker overnight. One culture was used to isolate the plasmid DNA via the alkaline lysis miniprep method while the bacterial cells were pelletted from the second culture, resuspended in 50% glycerol and stored at -20°C.

2.19 Clone Analysis

Two *Alu*-splice PCR reactions were set up for each of the shotgun cloned plasmids that had been extracted from the colony patches. The first reaction contained the original combination of *Alu* and splice primers while the second reaction contained only the original splice primer. Plasmids which amplified with only the splice primer were eliminated. Plasmids were further analyzed if they amplified only in the reaction containing both of the original primers.

2.20 Isolating the *Alu*-Splice Inserts

Using 10 µl of the plasmid miniprep DNA as a template, each insert was amplified with 140 pmol of each M13 -40 primer (5'-GTTTTCCCAGTCACTAC-3') and M13R primer (5'-TTCACACAGGAAACAG-3') (ACTG Inc.), 1X PCR Buffer [20 mM Tris·HCl (pH 7.5), 100 mM KCl, 1 mM DTT, 0.1 mM EDTA, 0.5% Tween 20, 0.5% Nonidet P40, 50% glycerol], 3 mM $MgCl_2$, 0.15 mM dNTPs, and 14 units Expand™ High Fidelity enzyme mix (Boehringer Mannheim). The total volume of 560 µl was divided equally between five tubes, a layer of mineral oil added to the top of each, and then placed in a Cetus DNA Thermal Cycler (Perkin Elmer). The reaction conditions were the same as those described in the *Alu*-splice PCR protocol (section 2.14). The samples were then purified by means of a QIAquick™ PCR Purification Kit (Qiagen) as previously described. The samples were eluted with 30 µl of Buffer EB. Each sample was combined with 5 µl of loading buffer and loaded into a wide well of a 0.8% agarose gel/1X TAE. The gel was run at 100 V until band separation was complete. The bands were visualized on a UV light box, sliced out of the gel with a clean scalpel, weighed, and placed in a 1.7 ml microcentrifuge tube. The PCR fragments were purified from the gel using a QIAquick™ Gel Extraction Kit (Qiagen). To dissolve the agarose, three volumes of Buffer QC (proprietary composition) were added to the volume of gel slice and incubated for 10 min. at 50°C with intermittent mixing. To ensure that the solution was at the appropriate pH (≤ 7.5), its color was monitored. If the yellow color of the solution changed to an orange or violet, 10 µl of 3 M NaOAc (pH 5.0) was added. In either case, one

gel volume of isopropanol was added. The sample was then loaded into a QIAquick™ spin column which had been placed in a 2 ml collection tube. The sample was centrifuged for 1 min. and the eluate was discarded. To remove all traces of agarose, 0.5 ml of Buffer QC was added to the QIAquick™ column followed by centrifugation for 1 min.. The eluate was again discarded. To wash the column, 0.75 ml of Buffer PE was added followed by centrifugation for 1 min. The eluate was discarded and an additional centrifugation of 1 min. was performed to remove any residual ethanol from Buffer PE. The QIAquick™ column was placed in a 1.7 ml microcentrifuge tube. The DNA was eluted by adding 30 µl of Buffer EB which was allowed to absorb for 1 min. This was followed by centrifugation for at least 1 min. To determine the concentration of each sample, a 4 in 40 µl dilution was prepared and the absorbance at 260 nm was measured.

2.21 Sequence Analysis of the *Alu*-Splice Inserts

Approximately 500 ng (in a volume of 5 µl) of purified PCR product was sent to University Core DNA Services (Calgary, AB, Canada) where 0.02 µg of template per 100 bp was utilized per 48 cm Cycle Sequencing reaction on a Perkin Elmer ABI 377 DNA Sequencer. Sequencing was performed with the M13 T7 primer, located on the end of the vector closest to the *Alu* sequence in the PCR insert, or with the M13 T3 primer, located on end of the vector closest to the splice sequence in the PCR insert. The sequencing data was transferred using the Internet and a File Transfer Protocol (FTP).

The cloning sites were identified in each of the sequences. The ends of the inserts that had homologies to the *Alu* and splice sequences were edited out. To search for similarities in the public databases, the edited sequence of each insert was then put through the BLAST (Basic Local Alignment Search Tool) program (http://www2.ncbi.nlm.nih.gov/BLAST/) made available by the National Center for Biotechnology Information. Nucleic acid sequences were searched in all non-redundant GenBank, EMBL (European Molecular Biology Laboratory), DDBJ (DNA Data Bank of Japan), and PDB (3-dimensional structure Brookhaven Protein Data Bank) databases. If an insert was found to be primarily composed of repetitive elements or the cloning vector, it was eliminated. Eleven inserts were found to contain non-repetitive, presumably unique sequences.

2.22 Producing Probes From the *Alu*-Splice Inserts

Internal primers (Table 5) were designed for each of the eleven inserts in such a way that all repetitive elements, including the original primer regions, would not be included in a resultant PCR product or "edited insert". Using the gel purified PCR fragments as templates, each edited insert was amplified using the same PCR conditions as those used in the *Alu*-splice PCR except that the *Alu* and splice primers were substituted with 10 pmol of the appropriate internal forward and reverse primers. Each 40 µl reaction was done in triplicate. After the amplification was completed, the three reactions were combined, purified by the QIAquick™ PCR Purification Protocol (Qiagen) (section 2.15), and eluted in a final volume of 30 µl. Absorbance at 260 nm of a 4 in 40 dilution of each of the

11 edited inserts was measured. Solutions containing 25 ng/µl of PCR product were then prepared to be utilized as a probes.

Table 5: Internal primers of the *Alu*-splice clones.

Primer Set	Colony #	Forward Primer 5'→3'	Reverse Primer 5'→3'
1	9	TAGCGAGATACTTTCGGTGC	AAACCACATTGTGACCTCCC
	13	CCATCCGCTTAAAAAGG	CTGTGAAGGTAATGATGGC
2	2	TCAAGACTTGGCTTCCCCAGCTT	AGCCCACTTTTCACCTCTCT
	15	ATCCCTTCATCCACCCAAGA	TAAAGGGGTGAGGGGTAACAAA
	16	TCAAGACTTGGCTTCCCCAA	TTCACCTCTCTATAGCTCCCT
3	2	TGACATCGGCTTTGGGACTGT	ACAAAGGGGCCCAAACCCTT
	8	AACTATACADDTGCCCCCTT	ACACATGCCAGGCCCTGTTAA
4	2	TGCTTTCCCCAAGATCAGCT	TTTGCCTGGCTTAGAGCTGT
	4	CACATATCATGAAATAGG	GTCCAATTAAAGGTTGTGG
	7	AGTGGCCTCAGTGTATGA	TCTTAGGTACAGGGTGAGCA
	12	TAGCTCTGGATACTCCGTGT	CGGCTTGAGAGCCTTTTTCT

2.23 Northern Blot Analysis

A Multiple Tissue Northern (MTN™) blot (Clontech) was prehybridized in a hybridization tube with 10 ml of pre-warmed ExpressHyb (Clontech) at 68°C for at least 30 min. Approximately 200 ng of an edited *Alu*-splice PCR probe, produced as previously described in section 2.22, was labeled with [α-^{32}P]dCTP by the random primed method. A scintillation count was performed to determine the number of cpm per µl. Exactly 2×10^7 cpms (2×10^6 cpms per ml of hybridization solution) were denatured and added to the prehybridized MTN™ blot. The blot was hybridized overnight at 68°C.

The MTN™ blot was rinsed in the hybridization tube three times with 2X SSC/0.05% SDS at room temperature and removed. The blot was washed in three 10 min. washes with 2X SSC/0.05% SDS at room temperature with

continuous shaking. A final wash with 0.1X SSC/0.1% SDS at 50°C was performed for 40 min. The blot was placed in a plastic bag, sealed and exposed to two films, one for 7 and the other for 14 days. To remove the probe, the MTN™ blot was stripped with 0.05% SDS at 90°C to 100°C for 1 hour.

3.0 RESULTS

3.1 Southern Blot Analysis Using Cosmid 25308

Prior to the commencement of this project, two overlapping cosmids, 16864 and 25308, were found to span or hybridize to both the proximal and distal sides of *FRA19B* (Appendix A). Thus, the fragile site was hypothesized to lie within these cosmids. To date, cloning of rare folate sensitive fragile sites has identified CG-rich trinucleotide expansions as the basis of fragility. Thus, *FRA19B* is hypothesized to also result from such a DNA expansion. Restriction fragments of cosmid 25308 where utilized as probes against Southern blots containing *FRA19B* positive and negative individuals to further narrow the region of fragility. It was postulated that one of these fragments should contain the normal sequence which, through dynamic mutation, gave rise to *FRA19B* in the two families presented. Therefore, upon Southern blot hybridization an increase in restriction fragment size should be observed in *FRA19B* positive individuals relative to *FRA19B* negative samples. However, *Bam*HI fragments of 25308, shotgun cloned into a pUC vector, did not detect distinct fragments on Southern blots but resulted in autoradiographs that had a ladder-like hybridization pattern consistent with that of a probe containing repetitive DNA (Figure 5). Southern blot analysis was, therefore, ineffective in narrowing down the region of fragility.

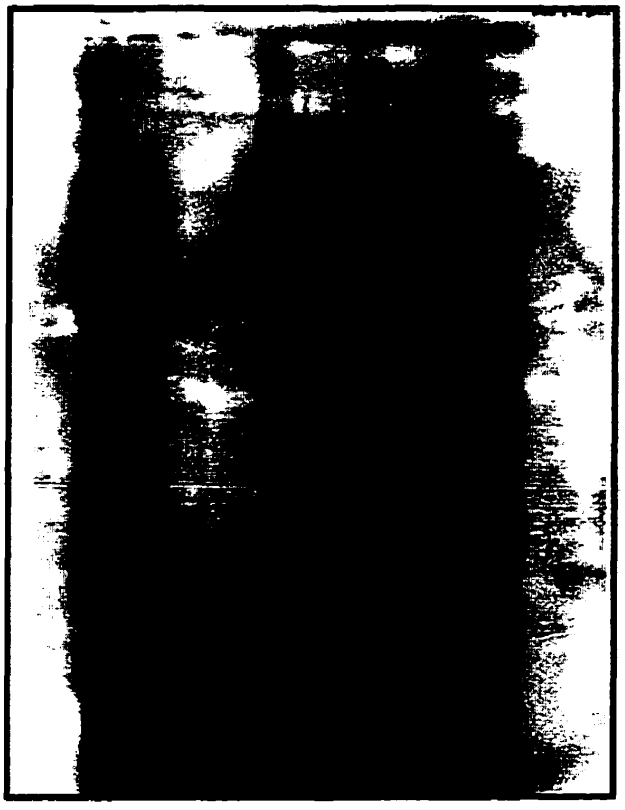

Figure 5: Southern blot containing human genomic DNA hybridized with a 2.0 kb *Bam*HI fragment of cosmid 25308.

3.2 Identification of *Alu* Repetitive Elements

A different approach was performed in a second attempt to narrow down the region of fragility of *FRA19B*. As previously stated, the fragility of rare folate sensitive fragile sites, cloned to date, is due to the expansion of CG-rich repeats. It was hypothesized that a CG-rich oligonucleotide may identify a CG-rich restriction fragment of cosmid 25308, which may contain the normal DNA sequence prone to expansion.

A $(CGG)_5$ oligonucleotide, used previously in the laboratory for *FRAXA* analysis (Mogk *et al.*, 1998), consistently hybridized under relaxed conditions, on three different Southern blots, to a single 2.0 kb *Bam*HI fragment of 25308

(Figure 6). Single-stranded DNA sequencing was performed bidirectionally on this fragment (sequence listed in Appendix B). The sequence of this fragment is not complete as there are two non-overlapping areas or gaps within the sequencing data. No CG-rich region was found within the 2.0 kb *Bam*HI fragment. However, a BLAST homology search revealed four sites within the 2.0 kb *Bam*HI fragment that were homologous to *Alu* repetitive elements. Sequence analysis also revealed an *Alu* repeat in the 1.8 kb *Bam*HI fragment (640 bp sequenced), but none in the sequenced portion of the 1.6 kb *Bam*HI fragment (555 bp sequenced) (Appendix B). A third *Bam*HI clone, approximately 700 bp, was sequenced and found to contain sequences homologous to vector DNA (Appendix B). Thus, at least five *Alu* repeats lie within these four *Bam*HI fragments of 25308 (Table 6).

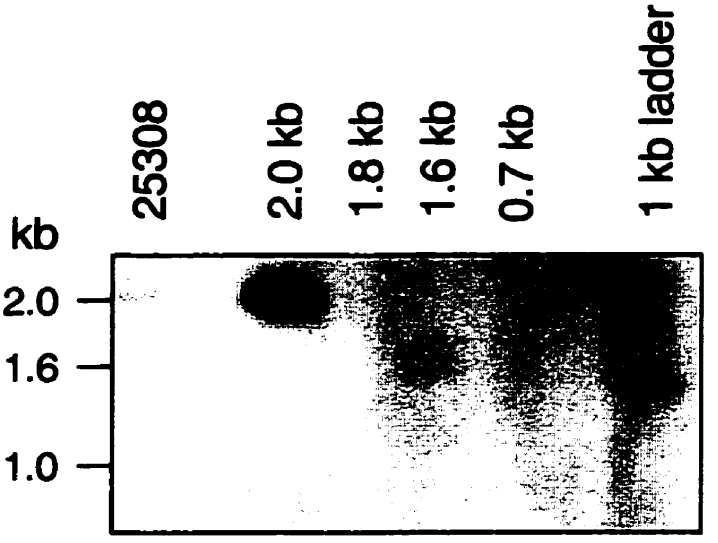

Figure 6: Southern blot containing *Bam*HI fragments of 25308 hybridized with a $(CGG)_5$ oligonucleotide.

Table 6: Summary of sequence analysis of four *Bam*HI fragments.

*Bam*HI Fragment (kb)	Amount Sequenced (bp)	DNA Content	*Alu* Repeats Identified
~2.0	2086	genomic	4
~1.8	640	genomic	1
~1.6	555	genomic	0
~0.7	566	vector	0

3.3 *Alu*-Splice PCR

Restriction fragments of cosmid 25308 were ineffective for Southern blot analysis due to the presence of *Alu* repeats, commonly located in close proximity to exons (Fuentes *et al.*, 1997). Thus, a third approach, involving *Alu*-splice PCR, was devised to narrow down the region of fragility. *Alu*-splice PCR products were amplified using cosmid 25308 as template DNA and primers homologous to the two ends of *Alu* repeats in combination with primers homologous to consensus sequences of 5' and 3' splice sites (Fuentes *et al.*, 1997). Four PCR reactions were performed based on the four possible primer combinations (Figure 7). The resulting *Alu*-splice PCR products potentially contained exons devoid of repetitive sequences and could, thus, be utilized as probes in Southern blot analysis.

Numerous *Alu*-splice PCR products, which ranged in size from approximately 100 bp to greater than 1200 bp, were produced in each of the four reactions in which cosmid 25308 was utilized as a template for *Alu*-splice PCR (Figure 7).

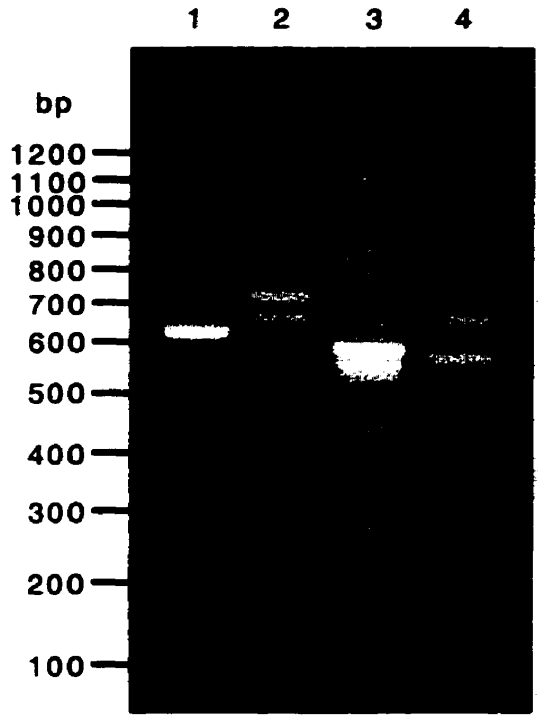

Figure 7: The products of *Alu*-splice PCR, using the four primer combinations with cosmid 25308 as the template, separated by electrophoresis through 1% agarose/1X TBE and stained with ethidium bromide.
Reaction 1 contained splice primer *Sac*II-3'spl and *Alu* primer *Sal*I-A33.
Reaction 2 contained splice primer *Sac*II-3'spl and *Alu* primer *Sal*I-A44.
Reaction 3 contained splice primer *Not*I-5'spl and *Alu* primer *Sal*I-A33.
Reaction 4 contained splice primer *Not*I-5'spl and *Alu* primer *Sal*I-A44.

The *Alu*-splice PCR products were subcloned directionally into BlueScript SK+. At least 25 white *E. coli* colonies were selected from the transformants of each PCR reaction to recover as many clones as possible (Table 7). A PCR reaction containing the original primer combination was performed on the selected colonies. This reaction ensured the presence of an insert and, in addition, enabled the size of the insert to be determined. A second PCR reaction was performed using only the original splice primer. This reaction

was performed to ensure the amplified product was not a result of non-specific priming events involving only the slice primer.

After screening the DNA from each colony by two PCR reactions, in which no amplification was ever seen in the reaction containing only the splice primer, a number of PCR products were recovered for reactions one and two. In contrast, only a few PCR products were recovered for reactions three and four. This inefficiency was speculated to result from incomplete digestion and, thus, the digestion of these PCR products was repeated along with the remaining screening process.

Based on size, twenty-seven presumably unique clones were chosen for further analysis including eight from reaction one, nine from reaction two, three from reaction three, and seven from reaction four (Table 7). Several clones which appeared to be duplicates, based on size, were eliminated from each reaction. Amplification products of each of these clones, produced using the -40 and reverse M13 primers, contained a portion of vector at both ends thereby providing universal primer sites for sequence analysis. These PCR products, ranging in concentration from 40 to 290 ng/µl, were purified by gel extraction.

Table 7: The number of *Alu*-splice clones analyzed.

	Primer Combination	Colonies Screened	Colonies Sequenced
1	*Sac*II-3'spl + *Sal*-A33	25	8
2	*Sac*II-3'spl + *Sal*-A44	25	9
3	*Not*I-5'spl + *Sal*-A33	68	3
4	*Not*I-5'spl + *Sal*-A44	56	7
Totals		**174**	**27**

Since the *Alu*-splice PCR products were amplified with -40 and reverse M13 primers, cycle sequencing could be performed on each of the 27 purified *Alu*-splice PCR products from either the splice end, using the T3 M13 primer, or the *Alu* end, using the T7 M13 primer. However, the data was reliable only if sequencing was performed from the splice end towards the *Alu* end using the T3 M13 primer. Several of the inserts were found to contain AT-rich regions towards the *Alu* end of the *Alu*-splice PCR products. Sequencing reactions using the T7 M13 primer had to process this AT-rich region prior to the majority of the *Alu*-splice PCR product. Sequencing through this region likely altered the nucleotide balance early in the sequencing reaction thereby resulting in poor sequencing data. Thus, the sequences (none of which contained a CG-rich region) reported in Appendix C were all produced using the T3 M13 primer and not the T7 M13 primer. The actual sizes of the *Alu*-splice inserts ranged from 68 to approximately 900 bp as stated in Table 8. Those inserts that have only an approximate size reported were not completely sequenced as the sequencing reaction was performed in only one direction, starting from the splice end of the insert.

The *Alu*-splice inserts that contained only repetitive DNA elements (1-1, 1-7, 1-11, 1-12, 1-22, 1-25, 2-8, 2-23, 2-24, 3-1, 4-1, 4-8, and 4-15) or vector DNA (2-12), as identified by BLAST homology searches, were eliminated from further analysis (Table 8). The inserts that contained only unique DNA (1-9, 2-16, 4-2, and 4-7) or large tracts of unique DNA interspersed with repetitive sequences (1-13, 2-2, 2-15, 3-2, 3-8, 4-4, and 4-12) were analyzed further. The

Table 8: Summary of the sequence analyses of the *Alu*-splice PCR clones.

Primer Set	Colony #	Insert Size (bp)	DNA Content	Probe (bp)
1	1	203	repetitive	
	7	253	repetitive	
	9	583		493
	11	269	repetitive	
	12	225	repetitive	
	13	~620[a]		158
	22	449	repetitive	
	25	275	repetitive	
2	2	433		109
	7	68		
	8	~700[a]	repetitive	
	12	73	vector	
	15	~800[a]		439
	16	159		108
	19	159		---[b]
	23	314	repetitive	
	24	449	repetitive	
3	1	~650[a]	repetitive	
	2	~850[a]		337
	8	~600[a]		213
4	1	~900[a]	repetitive	
	2	202		165
	4	~880[a]		222
	7	~750[a]		620
	8	156	repetitive	
	12	686		434
	15	~900[a]	repetitive	

[a] Inserts with approximate sizes do not have complete sequence data.
[b] PCR product 2-19 was not analyzed further as it is identical to 2-16.

consensus sequences of the *Alu* and splice primer sites had to be removed from the inserts in order to eliminate cross hybridization when used as probes in Southern blot analysis. Thus, internal primers were designed for each insert.

Insert 2-7 contained 68 bp of unique DNA, but after eliminating the *Alu* and splice primer sites, only 33 bp remained. Thus, due to size, insert 2-7 was

not used for further analysis. In addition, insert 2-19 was not utilized for further analysis since the sequencing data revealed that it was identical to insert 2-16. BLAST searches performed on the remaining 11 unique inserts did not reveal any significant matches.

3.4 Localization of *Alu*-Splice Probes on Cosmid 25308

Each of the 11 *Alu*-splice probes successfully hybridized to a unique *Eco*RI restriction fragment of cosmid 25308. Figure 8 illustrates the distribution of the probes along the cosmid in relation to the *Eco*RI restriction map made available by Lawrence Livermore National Laboratory. Two probes (1-9 and 3-2) hybridized to the 1.9 kb *Eco*RI fragment, six probes (1-13, 2-2, 2-16, 3-8, 4-4, and 4-7) hybridized to the 5.0 kb *Eco*RI fragment, and three probes (2-15, 4-2 and 4-12) hybridized to the 21 kb *Eco*RI fragment. No probes were localized to either the 0.6 or 2.0 kb *Eco*RI fragments.

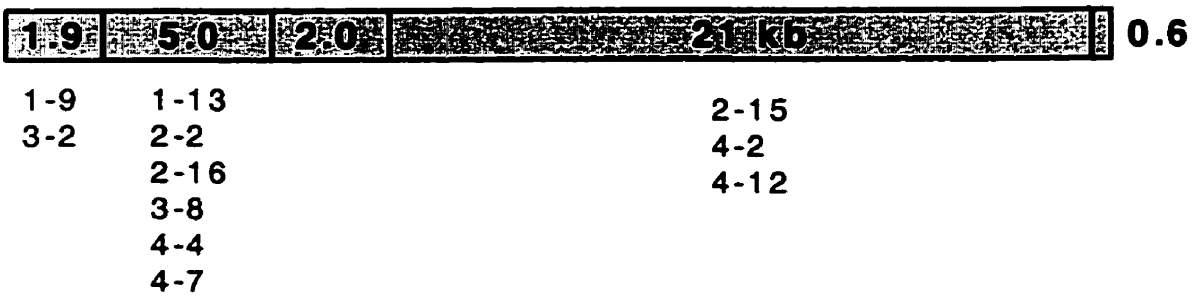

Figure 8: Distribution and localization of the *Alu*-splice probes on the *Eco*RI restriction fragments of 25308.

In addition, after sequence analysis was complete, two of the *Alu*-splice probes were localized to three of the *Bam*HI fragments of cosmid 25308 that were originally sequenced. The entire sequence of probe 4-12 was localized to

the 2.0 kb *Bam*HI fragment (illustrated in Appendix B). Probe 2-15, found to contain a *Bam*HI site (as indicated by italic type in Appendix C), was localized to both the 1.8 kb and 1.6 kb *Bam*HI fragments (as illustrated in Appendix B).

3.5 Mapping Analysis of *Alu*-Splice Probes to Cosmid 25308

Each *Alu*-splice probe detected a unique set of restriction fragments on Southern blots comprised of seven different restriction endonuclease digestions (*Bam*HI, *Eco*RI, *Hind*III, *Hin*fI, *Pst*I, *Rsa*I, and *Sma*I) of cosmid 25308. The data for each probe is summarized in Table 9. A typical autoradiograph produced from these analyses is illustrated in Figure 9.

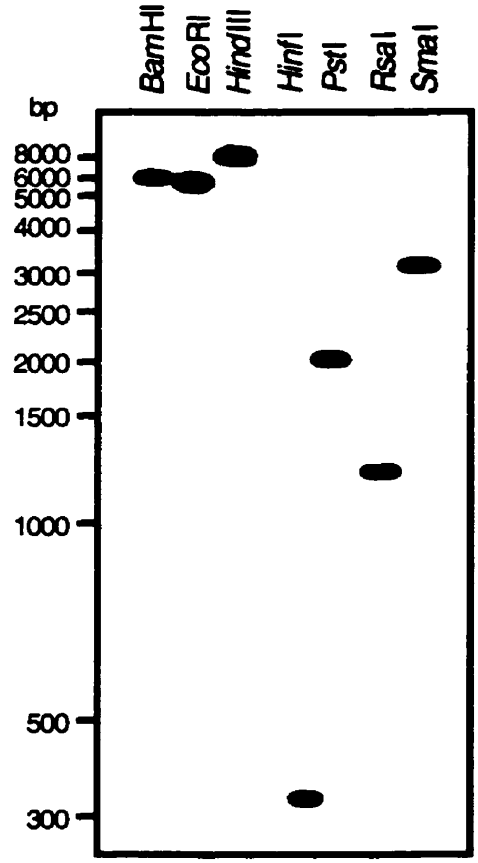

Figure 9: Southern blot localizing *Alu*-splice probe 3-8 to various restriction fragments of cosmid 25308.

Appropriate restriction digests (determined from the data in Table 9) were performed on human genomic DNA samples (*FRA19B* positive and negative) based on the optimum size restriction fragment size determined for each probe. The restriction fragments highlighted by bold type in Table 9 indicate the restriction endonucleases utilized in the digestion of the human genomic DNA samples for each probe.

Table 9: Localization of 11 *Alu*-splice probes to various restriction fragments of cosmid 25308. Restriction fragments of human genomic DNA, sizes shown in bold, were detected on Southern blots with the appropriate probes.

Primer Set	Colony #	Probe (bp)	*Bam*HI (kb)	*Eco*RI (kb)	*Hin*dIII (kb)	*Hin*fI (kb)	*Pst*I (kb)	*Rsa*I (kb)	*Sma*I (kb)
1	9	493	3.7	1.9	~20	0.60	**3.8**	4.2	3.5
	13	158	5.5	5.0	~8	0.60	**1.9**	2.0	2.8
2	2	109	**3.7**	5.0	~20	0.25	3.8	4.2	3.5
	15	439	**1.6, 1.8**	21	~20	1.2	1.18, 3.8	0.41, 4.5	1.55
	16	108	3.7	5.0	~20	0.28	**3.8**	4.5	3.5
3	2	337	3.7	**1.9**	~20	0.60	3.8	4.2	3.5
	8	213	6.0	5.0	~8	0.32	**2.0**	1.3	3.4
4	2	165	3.7	21	~20	1.0	0.37	**2.0**	5.0
	4	222	6.0	5.0	~8	0.60	2.0	**1.3**	3.5
	7	620	3.7, 6.0	5.0	~8	1.2	1.6	**1.7**	.65
	12	434	**2.0**	21	~20	0.55	0.37	1.8	5.5

3.6 Searching for a DNA Expansion

When hybridized against human genomic DNA, five of the probes (1-9, 1-13, 2-16, 4-4, and 4-12) resulted in repetitive hybridization patterns similar to that illustrated in Figure 5 and were, therefore, eliminated from further analysis. Six of the probes (2-2, 2-15, 3-2, 3-8, 4-2, and 4-7) hybridized to unique fragments. However, none of these probes demonstrated the presence of a larger restriction fragment in the DNA from *FRA19B* positive individuals

compared to the negative controls. A typical autoradiograph produced from this set of Southern blot analyses is illustrated in Figure 10.

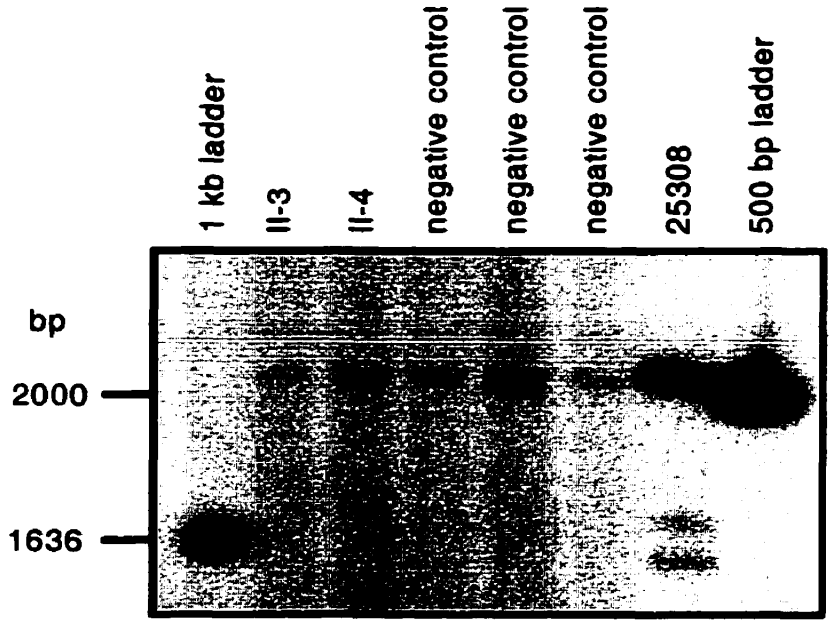

Figure 10: Southern blot containing *Rsa*I digestions hybridized with *Alu*-splice probe 4-2. Lanes labeled II-3 and II-4 represent the *FRA19B* positive individuals from the Manitoban family. Lanes labeled as negative controls represent individuals that are *FRA19B* negative. The lane labeled 25308 contains the positive control, cosmid 25308 digested with *Rsa*I.

3.7 Identification of Expressed Sequences

Five of the six *Alu*-splice probes that contained unique sequence (2-2, 2-15, 3-2, 3-8, and 4-7) did not detect any RNA fragments on a human Multiple Tissue Northern (MTN^{TM}) blot. However, probe 4-2 detected two transcripts with approximate sizes of 4.5 and 7.0 kb in each of the eight tissues including heart, brain, placenta, lung, liver, skeletal muscle, kidney, and pancreas (Figure 11).

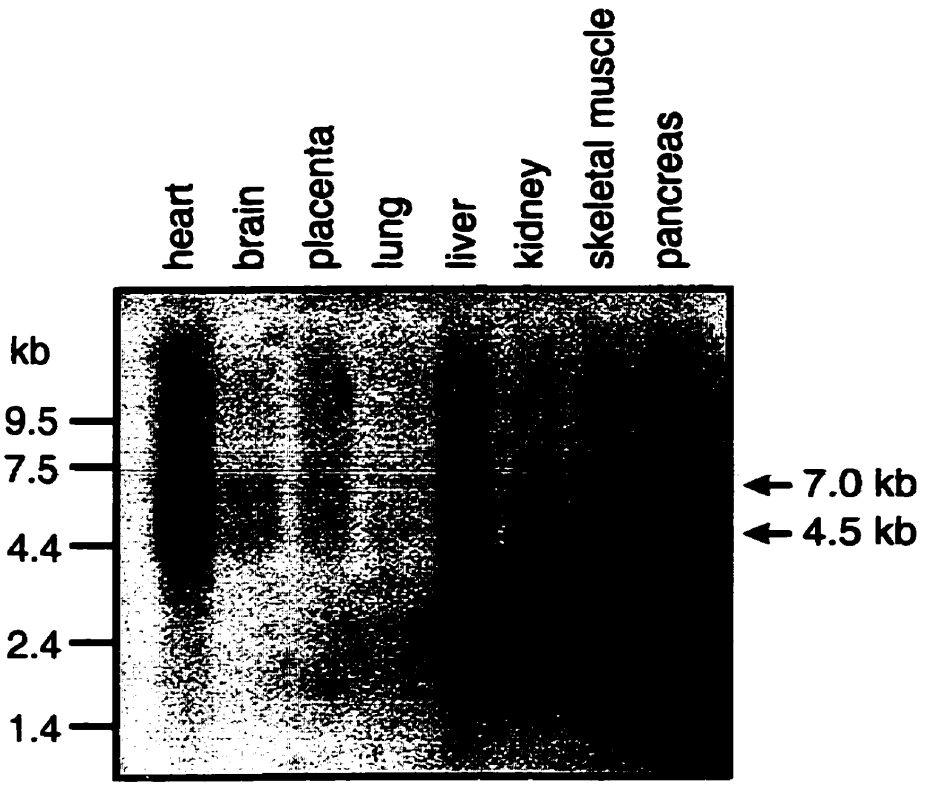

Figure 11: Human Multiple Tissue Northern (MTN™) blot hybridized with *Alu*-splice probe 4-2.

4.0 DISCUSSION

To date, similar methodologies have been utilized to molecularly characterize the rare folate sensitive fragile sites (Kremer *et al.*, 1991; Knight *et al.*, 1993; Ritchie *et al.*, 1994). Briefly, large fragments containing genomic DNA, such as yeast artificial chromosomes or cosmids, were found to span the specific fragile site using fluorescent *in situ* hybridization. This spanning fragment was then digested into smaller fragments that were utilized as probes against Southern blots resulting in distinct hybridization signals. One of the probes identified a size difference between the hybridization bands of the fragile site positive individuals and individuals who did not express the fragile site. The sequence of this altered fragment was determined and a DNA expansion was identified. However, the attempt to clone *FRA19B* has not been as straightforward or successful.

4.1 Southern Blot Analysis Using Cosmid 25308

Fluorescent *in situ* hybridization had previously identified two cosmids, 16864 and 25308, that spanned *FRA19B* (Appendix A). *Bam*HI fragments of cosmid 25308 were to be utilized as probes against Southern blots containing *FRA19B* positive and negative individuals in hopes of finding an increased fragment size in the positive individuals. However, Southern blot hybridization revealed that these clones did not provide a distinct hybridization signal, but resulted in a ladder-like hybridization pattern in each lane containing human genomic DNA. A protocol was then attempted to suppress the repetitive

elements within the clone. However, this protocol resulted in no hybridization signals. Thus, these genomic clones were found to be unusable and alternative methods had to be designed to begin the molecular analysis of *FRA19B*.

4.2 Identification of *Alu* Repetitive Elements

It was hypothesized that the molecular basis for *FRA19B*, induced by a deficiency in folic acid, is a CG-rich repeat as found in the other folate sensitive fragile sites cloned to date (Nelson, 1998). Thus, to narrow down the region of fragility, an attempt to identify a CG-rich fragment of cosmid 25308 was made. To accomplish this, a $(CGG)_5$ oligonucleotide, used previously in *FRAXA* analysis (Mogk *et al.*, 1998), was utilized as a probe against a Southern blot containing *Bam*HI fragments of cosmid 25308. Under conditions of very low stringency, the CGG oligonucleotide consistently hybridized to a 2 kb fragment which was therefore sequenced. While the sequence was found to be CG-rich, no repeat was identified. However, a BLAST search revealed four *Alu* repeats within the partially sequenced 2 kb fragment. Three other *Bam*HI clones, approximate sizes 0.7, 1.6 and 1.8 kb, were partially sequenced from both ends to determine if *Alu* repeats may have been responsible for the ladder-like hybridization pattern observed in the Southern blot analyses. An *Alu* repeat was also identified in the 640 bp of sequence obtained for the 1.8 kb fragment. Thus, five *Alu* repeats were identified in approximately 3.2 kb of cosmid 25308 (Table 6). *Alu* repeats constitute approximately 5% of the human genome with, on average, one *Alu* repeat every 3 to 6 kb (Strachan and Read, 1996). Therefore, cosmid 25308 appears to have a higher proportion of *Alu* repeats

than expected. It has been shown that *Alu* repeats are associated with genes, being located either in the 3' untranslated sequences, where they may act in the regulation of gene expression, or within introns (Wu *et al.*, 1990; Brini *et al.*, 1993; Hambor *et al.*, 1993; Hewitt *et al.*, 1995; Kochanek *et al.*, 1995; Norris *et al.*, 1995; Vansant and Reynolds, 1995). Thus, the presence of such a high proportion of *Alu* repeats within cosmid 25308 suggests that it may contain expressed sequences. This is consistent with the hypothesis that chromosome 19, which stains very lightly when G-banded (indicating a high proportion of euchromatin), is very gene dense (Therman and Susman, 1993).

4.3 *Alu*-Splice PCR

The presence of *Alu* repeats was used advantageously by a technique termed *Alu*-splice PCR, originally designed to isolate exon-containing fragments, to eliminate the repetitive DNA (Fuentes *et al.*, 1997). Primers were designed to be homologous to the consensus sequences of 5' donor and 3' acceptor splice sites and to the two ends of *Alu* repeats. Correct orientation of these primers would produce PCR products which could be potential exons. Utilizing cosmid 25308 as a template, *Alu*-splice PCR was speculated to create products that contained potentially expressed DNA sequences but no repetitive DNA, which is normally found within introns and nonexpressed sequences. Consequently, such PCR products could be utilized as probes in the molecular characterization of *FRA19B*.

Numerous *Alu*-splice PCR products were amplified in each of the four reactions (Figure 7). Cloning the products of reactions one and two was

successful since most products appeared to be recovered from the 25 *E. coli* colonies that were screen per reaction. However, *Alu*-splice PCR reactions three and four were problematic as each resulted in the recovery of only a few *Alu*-splice PCR products. To retrieve most of the products of reactions three and four, 68 and 56 *E. coli* colonies had to be screen, respectively. This inefficiency may have been a result of 1.) inefficient digestion of the PCR products and/or 2.) a high concentration of certain *Alu*-splice PCR products within the reaction which, when shotgun cloned, would be over-represented in the screened colonies.

Two screening PCR reactions, as described by Fuentes *et al.* (1997), were performed on each selected *E. coli* colony in order to eliminate potential nonspecific PCR products. For example, there are two orientations in which each DNA element (intron, exon or *Alu* repeat) may lie within the template. Thus, for example, two putative consensus sequences for 3' acceptor splice sites, inverted in relation to one another, may lie in close proximity. If this occurs within cosmid 25308, a PCR product from reaction one or two may have been amplified with only the 3' splice primer and would likely not contain a potential exon, but could possibly contain only repetitive DNA. Thus, by performing a PCR reaction using only the original splice primer such non-specific products could be quickly eliminated. Two such amplifications were carried out on cosmid 25308, each using one of the two *Alu*-splice primers (results not shown). Only very large PCR products were produced which, based on size, were not identified in any of the *Alu*-splice clones. Thus, no *Alu*-splice PCR clones were

eliminated on the basis of this PCR-based screening process. Based on size, twenty-seven presumably unique clones were chosen for further analysis including eight from reaction one, nine from reaction two, three from reaction three, and seven from reaction four.

Each of the twenty-seven presumably unique clones were amplified with the -40 and reverse M13 primers in order to obtain the cloning sites within the sequencing data. Internal M13 primers, T7 and T3, were used to sequence these PCR products. The sequencing data was only reliable when the T3 primer was utilized. The location of the T7 primer within these PCR products required the cycle sequencing reaction to proceed through the *Alu* primer sequence prior to sequencing the *Alu*-splice insert. The unreliable data obtained by sequencing with the T7 primer indicates that the presence of the *Alu* primer site may affect the sequencing reaction. In addition, several of the inserts contained AT-rich regions in close proximity to the *Alu* end of the *Alu*-splice inserts. These regions likely disrupted the nucleotide balance early in the sequencing reactions, ultimately resulting in poor sequence data. In contrast, sequencing with the T3 primer, located at the splice end, allowed the cycle sequencing reaction to proceed through most of the *Alu*-splice insert before reaching these AT-rich regions. The nucleotide pool would have been disrupted late in the sequencing reactions, after which time most of the inserts would have already been sequenced.

Of the 27 inserts sequenced, 48% (13) contained only repetitive DNA, which is higher than the 20% found by Fuentes *et al.* (1997) in the 37 clones

located within 21q22.1-2.2. The splice primers are designed to hybridize to consensus sequences for splice sites. However, each consensus sequence may not represent the location of a functional acceptor or donor site and thus, an intron may not be present at all the sites where the primer hybridizes to the template. In addition, the *Alu* primers may hybridize to sites that are homologous to *Alu* consensus sequences, but are not actually located within an *Alu* repeat. These two types of "non-specific" hybridization could be the reason that some of the *Alu*-splice PCR products contain repetitive DNA.

Three of 14 *Alu*-splice PCR products containing unique DNA sequences were eliminated from further analysis. *Alu*-splice PCR product 2-7 was too small to be utilized as a probe, 2-12 contained cosmid vector DNA and 2-19 was identical to 2-16. The remaining 11 unique *Alu*-splice PCR products had to be edited, by eliminating the *Alu* and splice primer sites, prior to using them as probes. In addition, DNA repetitive elements which flanked the required unique DNA sequences in seven of the inserts (1-13, 2-2, 2-15, 3-2, 3-8, 4-4, and 4-12) had to be eliminated. Primers were designed for each insert that, when used to amplify the appropriate clone, accomplished this editing process and created PCR products which, based on homology searches with BLAST, presumably contained only unique DNA.

4.4 Localization of *Alu*-Splice Probes on Cosmid 25308

Prior to Southern blot analysis of human genomic DNA, the distribution pattern of the *Alu*-splice probes within cosmid 25308 was identified. This mapping analysis was performed to illustrate whether the *Alu*-splice probes were

representative of the entire cosmid. Based entirely on mathematical probabilities, one of the 11 probes should be located approximately every 2.8 kb. However, as illustrated in Figure 8, this distribution does not follow such a pattern. A higher proportion (55%) of the *Alu*-splice probes were localized to the 5.0 kb *Eco*RI fragment than would be expected by chance (16%). A lower proportion (27%) of the *Alu*-splice probes were localized to the 21 kb *Eco*RI fragment than would be expected by chance (69%). This distribution pattern may not actually be a misrepresentation of the cosmid, but may be the result of a non-uniform distribution pattern of *Alu* repeats and splice sites within cosmid 25308.

Mapping of the *Alu*-splice probes to the *Eco*RI fragments of cosmid 25308 serendipitously allowed for the localization of three of the *Bam*HI fragments originally sequenced. *Alu*-splice probe 2-15 contains a *Bam*HI site at nucleotide position 162. Sequence before this site was identified in the 1.8 kb *Bam*HI fragment and sequence after this site was identified in the 1.6 kb *Bam*HI fragment. In addition, the entire sequence of *Alu*-splice probe 4-12 was located within the 2.0 kb *Bam*HI fragment. Thus, the 1.6, 1.8 and 2.0 kb *Bam*HI fragments were indirectly localized to the 21 kb *Eco*RI fragment.

4.5 Mapping Analysis of *Alu*-Splice Probes to Cosmid 25308

The data obtained by localizing the *Alu*-splice probes to various restriction fragments of cosmid 25308, as summarized in Table 9, was utilized for several purposes. First, each *Alu*-splice probe hybridized to one or two distinct fragments of each of the seven restriction digestions of cosmid 25308. Thus, the

specificity and origin of the *Alu*-splice products was proven. Secondly, when all probes were compared, a distinct set of restriction fragments was detected by each *Alu*-splice probe, except for probes 1-9 and 3-2 which detected the same set of restriction fragments. This ensured that ten of the probes could be used to detect different restriction fragments of cosmid 25308 and, thus, maximize the amount of human genomic DNA analyzed. Thirdly, to be useful in searching for a DNA expansion, each probe had to detect a fragment that, if increased in size by very few base pairs, could identify two bands with similar size. For example, a normal *FRAXA* allele of $(CGG)_{30}$ would differ from a mutant *FRAXA* allele of $(CGG)_{230}$ by only 600 bp. Thus, for Southern blot analysis with each probe, a specific restriction endonuclease was chosen to digest human genomic DNA in order to detect small size differences between fragments using routine agarose gel electrophoresis. For this purpose, the optimum size of the detected restriction fragments would theoretically be as small as possible. However, with only 10 out of the 11 probes to use (only ten different regions since probes 1-9 and 3-2 appear to be detecting the same fragments) for the detection of an expansion in approximately 30 kb, it was necessary for each probe to detect as much DNA per fragment as possible without compromising the ability to detect an expansion. Thus, the optimum fragment size to be detected was chosen to be in the range of 1 to 4 kb. After the analyses of the "optimum" sized fragments, the remaining restriction fragments were not analyzed by other electrophoretic methods. Analysis of *FRA19B* using "non-optimum" sized

fragments was not pursued since the Lawrence Livermore Laboratory is currently sequencing cosmid 25308.

*Hind*III restriction fragments were generally too large to be able to identify any potential expansions by routine gel electrophoresis. Thus, no routine Southern blots containing genomic DNA digested with *Hind*III were used in the analysis. Analysis of the *Hind*III fragments could be performed by pulse field gel electrophoresis (PFGE). However, all 10 PCR probes were small (<620 bp) and PFGE analysis of the *Hind*III fragments was not efficient since the Lawrence Livermore Laboratory is currently sequencing cosmid 25308. Likewise, *Hinf*I restriction fragments were generally small and thus would not detect very much DNA per fragment. *Sma*I was very inefficient in the digestion of the human DNA samples and, therefore, was not utilized in further analyses.

4.6 Searching for a DNA Expansion

Five of the *Alu*-splice probes (1-9, 1-13, 2-16, 4-4, and 4-12) were eliminated as they did not detect a unique restriction fragment in the human genome. These fragments likely contain repetitive elements that were not detected by the homology searches. Thus, of the 11 probes, only six (2-2, 2-15, 3-2, 3-8, 4-2, and 4-7) hybridized to unique restriction fragments within the human genome. Interestingly, even though probes 1-9 and 3-2 detected the same set of restriction fragments of cosmid 25308, only 3-2 was useful for the detection of human genomic DNA since 1-9 did not provide a unique hybridization signal. In the end, the analysis of approximately 30 kb was performed by only six probes.

The position of an expansion cannot be eliminated from the 0.6 kb or 2.0 kb *Eco*RI fragment (Figure 4) since none of this DNA was analyzed. In addition, an expansion located in the 21 kb *Eco*RI fragment could have gone undetected since very little of this large fragment was analyzed. While a high proportion of the *Alu*-splice probes localized to the 5 kb *Eco*RI fragment, not all areas of the fragment were detected; therefore, the expansion could possibly have gone undetected within this region. However, the position of an expansion can be eliminated from the 1.9 kb *Eco*RI fragment, located at one end of cosmid 25308. Probe 3-2 detected this entire fragment and no size difference was visualized between *FRA19B* positive and negative individuals. This is consistent with FISH analysis using cosmid 16864. Cosmid 16864 also spans *FRA19B* (unpublished data). In addition, cosmid 16864 overlaps cosmid 25308 but does not contain the 1.9 kb *Eco*RI fragment. Thus, the area potentially harboring a DNA expansion is less than 28.6 kb.

4.7 Identification of Expressed Sequences

The *Alu*-splice PCR products were also analyzed for expressed DNA sequences. This was the original goal of *Alu*-splice PCR (Fuentes *et al.*, 1997) as *Alu* repeats are typically found within non-expressed regions associated with genes. From the six probes that contained unique sequence (2-2, 2-15, 3-2, 3-8, 4-2, and 4-7), only probe 4-2 identified an expressed sequence. Two transcripts, with approximate sizes of 4.5 and 7.0 kb, were detected in all tissues on the human MTN™ blot. Such a wide range of tissue expression suggests that the

Alu-splice probe 4-2, located within the 21 kb *Eco*RI fragment, contains a portion of a housekeeping gene with a genomic size larger than 7.0 kb.

4.8 Phenotypic Differences Among *FRA19B* Positive Individuals

Considerable variation exists between the phenotypes of the *FRA19B* positive individuals that have been identified to date. This variation is observed among individual members of one of the families, as well as between the two different families.

Individuals II-1, II-2, II-3, and II-4 of the Manitoban family all express *FRA19B*. However, the phenotypes of the brothers vary with II-1 and II-3 having schizophrenia, II-2 having autism and mental retardation, and II-4 having no obvious clinical phenotype (Chodirker *et al.*, 1987) (Figure 3). Nonetheless, the clinical phenotypes that do exist share a common theme of mental dysfunction. Variation between family members, common in disorders with a trinucleotide repeat as their molecular basis, is usually based on the number of trinucleotides within the repeat tract and/or on the methylation status of the promoter associated with the gene containing the expansion.

Furthermore, a polygenic mode of inheritance has been proposed for psychiatric disorders that may account for some of the phenotypic variation seen among the brothers of the Manitoban family (Schultz, 1999). On the other hand, *FRA19B* may cause effects on a specific allele of a neighboring gene, possibly accounting for the phenotypic variation observed between the Manitoban and Danish families (Côté, 1990).

5.0 SUMMARY AND CONCLUSIONS

The application of *Alu*-splice PCR to create functional probes from cosmid 25308 was successful and, thereby, alleviated the first major problem hindering the molecular analysis of *FRA19B*. However, the molecular basis for the expression of *FRA19B* has not yet been identified. Only six *Alu*-splice probes were useful in the analysis involving human genomic DNA; thus, the analysis of cosmid 25308 is incomplete in regard to the search for a DNA expansion.

Alu-splice PCR was also successful in identifying an expressed sequence within cosmid 25308. The expressed sequence is contained within two transcripts with approximate sizes of 4.5 and 7.0 kb. These transcripts appear to be biologically significant as they are both found to be expressed in a wide range of tissues including the heart, brain, placenta, lung, liver, skeletal muscle, kidney, and pancreas.

In conclusion, although no DNA expansion has been identified, it is important to complete the analysis of *FRA19B* for several reasons. *FRA19B* is located within a gene rich chromosome and, thus, the likelihood of a gene being present within this area is high (Therman and Susman, 1993). Such a gene is likely to be of importance since very few viable cytogenetic rearrangements have been identified within this region or within chromosome 19 (Schinzel, 1994). In addition, it has also been hypothesized that the normal sequence within *FRA19B* may be interspersed with various interrupting sequences. This would explain the difficulty identifying the molecular basis of fragility and, in addition, may account

for the stability of this region which results in the low frequency of *FRA19B* in the population.

6.0 FUTURE STUDIES

Currently, investigators from the Lawrence Livermore National Laboratory are sequencing the contig in which cosmid 25308 is located. When the sequence of cosmid 25308 becomes available, a CG-rich region, prone to expansion, may be identifiable. This would simplify the methodologies, possibly to PCR analysis, required to complete the search for an expansion within the *FRA19B* positive individuals. Typically, DNA within the *FRA19B* locus appears to be stable based on the low frequency of expression in the population. However, the two families studied must have undergone some type of mutation which rendered their alleles unstable and prone to expansion. For example, like the stabilizing AGG interspersions within the CGG repeat tract of *FRAXA* alleles, a specific interspersed DNA sequence, which typically provides stability to the *FRA19B* allele, may have been lost in certain members of these two families. Identification of the expansion and the underlying molecular structure within normal and expanded alleles could further the understanding of the dynamic mutation process involved in rare fragile site genesis.

The discovery of an expressed sequence located within cosmid 25308 should definitely be further investigated. To obtain the full length cDNA, *Alu*-splice probe 4-2 could be used to screen a human cDNA library. The cDNA could then be molecularly characterized for gene structure and potential function. The relationship between the gene and *FRA19B* could then be analyzed. Several relationships are possible based on the characterization of other fragile sites. *FRA19B* may not actually have an effect on the gene. However, *FRA19B*

could be similar to *FRAXA* or *FRAXF* and lie within a gene and be associated with a CpG island. When expanded, *FRA19B* could cause methylation of the CpG island and subsequent inactivation of the specific gene. In contrast, *FRA19B* could be similar to *FRA16B* in that it may exert *cis*-configuration effects on a neighboring gene possibly in an allele specific manner.

In conclusion, elucidating the effect of *FRA19B* may indicate its role, if any, in the clinical phenotypes observed in the Manitoba family. If *FRA19B* exerts effects on neighboring genes, this may further the understanding of the wide variation seen in the *FRA19B* positive individuals within and between the two families. Currently, the three autosomal rare fragile sites that have been cloned (*FRA11B*, *FRA16A* and *FRA16B*), unlike the X-linked rare fragile sites, are not associated with clinical phenotypes. However, identifying an effect of *FRA19B* that results in a clinical phenotype may be an important finding which leads to further understanding of other rare autosomal fragile sites and their, as yet, undetermined clinical effects.

7.0 REFERENCES

Ardeshir, F., E. Giulotto, J. Zieg, O. Brison, W.S. Liao, and G.R. Stark (1983) Structure of amplified DNA in different Syrian hamster cell lines resistant to N-(phosphonacetyl)-L-aspartate. *Molecular Cellular Biology* 3:2076-2088.

Ausubel, F.M., R. Brent, R.E. Kingston, D.D. Moore, J.G. Seidman, J.A. Smith, and K. Struhl (eds.) (1996) Current Protocols in Molecular Biology. John Wiley and Sons, Inc., USA.

Bell, M.V., M.C. Hirst, Y. Nakahori, R.N. MacKinnon, A. Roche, T.J. Flint, P.A. Jacobs, N. Tommerup, L. Tranebjaerg, U. Froster-Iskenius, B. Kerr, G. Turner, R.H. Lindenbaum, R. Winter, M. Pembry, S. Thibodeau, and K.E. Davies (1991) Physical mapping across the fragile X: hypermethylation and clinical expression of the fragile X syndrome. *Cell* 64:861-866.

Birnboim, H.C. and J. Doly (1979) A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Research* 7:1513-1517.

Blake, T.J., M. Shapiro, H.C. Morse III, and W.Y. Langdon (1991) The sequences of the human and mouse c-*cbl* proto-oncogenes show v-*cbl* was generated by a large truncation encompassing a proline-rich domain and a leucine zipper-like motif. *Oncogene* 6:653-657.

Boldog, F., R.M. Gemmill, J. West, M. Robinson, L. Robinson, E. Li, J. Roche, S. Todd, B. Waggoner, R. Lundstrom, J. Jacobson, M.R. Mullokandov, H. Klinger, and H.A. Drabkin (1997) Chromosome 3p14 homozygous deletions and sequence analysis of *FRA3B*. *Human Molecular Genetics* 6:193-203.

Brais, B., J.P. Bouchard, Y.G. Xie, D.L. Rochefort, N. Chrétien, F.M.S. Tomé, R.G. Lafrenière, J.M. Rommens, E. Uyama, O. Nohira, S. Blumen, A.D. Korcyn, P. Heutink, J. Mathieu, A. Duranceau, F. Codère, M. Fardeau, and G.A. Rouleau (1998) Sort GCG expansions in the *PABP2* gene cause oculopharyngeal muscular dystrophy. *Nature Genetics* 18:164-167.

Breslauer, K.J., R. Frank, H. Blocker, and L.A. Marky (1986) Predicting DNA duplex stability from the base sequence. *Proceeding of the National Academy of Science, USA* 83:3746-3750.

Brini, A.T., G.M. Lee and J.P. Kinet (1993) Involvement of *Alu* sequences in the cell-specific regulation of transcription of the γ chain of Fc and T cell receptors. *The Journal of Biological Chemistry* 268(2):1355-1361.

Burke, J.R., M.S. Wingfield, K.E. Lewis, A.D. Roses, J.E. Lee, C. Hulette, M.A. Pericak-Vance, and J.M. Vance (1994) The Haw River syndrome: dentatorubro-pallidoluysian atrophy (DRPLA) in an African-American family. *Nature Genetics* 7(4):521-524.

Chodirker, B., A. Chudley and M. Ray (1987) Fragile 19p13 in a family with mental illness. *Clinical Genetics* 31:1-6.

Côté, G.B. (1990) The *cis*-configuration effects of rare chromosomal fragile sites. *Clinical Genetics* 37:470-472.

Dekaban, A. (1965) Persisting clone of cells with an abnormal chromosome in a woman previously irradiated. *Journal of Nuclear Medicine* 6(10):740-746.

DiFiglia, M., E. Sapp, K.O. Chase, S.W. Davies, G.P. Bates, J.P. Vonsattel, and N. Aronin (1997) Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neurites in brain. *Science* 277:1990-1993.

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

Eichler, E.E., J.J.A. Holden, B.W. Popovich, A.L. Reiss, K. Snow, S.N. Thibodeau, C.S. Richards, P.A. Ward, and D.L. Nelson (1994) Length of uninterrupted CGG repeats determines instability in the *FMR1* gene. *Nature Genetics* 8:88-94.

Eichler, E.E., H.A. Hammond, J.N. Macpherson, P.A. Ward, and D.L. Nelson (1995) Population survey of the human *FMR1* CGG repeat substructure suggests biased polarity for the loss of AGG interruptions. *Human Molecular Genetics* 4:2199-2208.

Engelman, J.A., X.L. Zhang and M.P. Lisanti (1998) Genes encoding human caveolin-1 and -2 are co-localized to D7S522 locus (7q31.1), a known fragile site (*FRA7G*) that is frequently deleted in human cancers. *Federation of European Biochemical Societies Letters* 436:403-410.

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

Feinberg, A.P. and B. Vogelstein (1984) Addendum to: [Feinberg, A.P. and B. Vogelstein (1983) A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Annals of Biochemistry* 132:6-13.] *Anals of Biochemistry* 137:266-267.

Flynn, G.A., M.C. Hirst, S.J.L. Knight, J.N. Macperson, J.C.K. Barber, A.V Flannery, K.E. Davies, and V.J. Buckle (1993) Identification of the *FRAXE* fragile site in two families ascertained for X linked mental retardation. *Journal of Medical Genetics* 30:97-100.

Friedberg, E.C. (1985) DNA Repair. W.H. Freeman and Company, New York, NY, USA, pp 14-15.

Fu, Y.H., D.P.A. Kuhl, A. Pizutti, M. Pieretti, J.S. Sutcliffe, S. Richards, A.J.M.H. Verkerk, J.J.A. Holden, R.G. Fenwick, S.T. Warren, B.A. Oostra, D.L. Nelson, and C.T. Caskey (1991) Variation of the CGG repeat at the fragile site results in genetic instability: resolution of the Sherman paradox. *Cell* 67:1047-1058.

Fuentes, J.J., C. Pucharcos, M. Pritchard, and X. Estivill (1997) *Alu*-splice PCR: a simple method to isolate exon-containing fragments from cloned human genomic DNA. *Human Genetics* 101:346-350.

Gecz, J., A.K. Gedeon, G.R. Sutherlund, and J.C. Mulley (1996) Identification of the gene *FMR2*, associated with *FRAXE* mental retardation. *Nature Genetics* 13:105-108.

Gecz J., S. Bielby, G.R. Sutherland, and J.C. Mulley (1997) Gene structure and subcellular localization of *FMR2*, a member of a new family of putative transcription activators. *Genomics* 44(2):201-13.

Glover, T.W. (1981) FrdU induction of the X chromosome fragile site: evidence for the mechanism of folic acid and thymidine inhibition. *American Journal of Human Genetics* 33:234-242.

Glover, T.W. (1998a) Common fragile sites. In: Genetic Instabilities and Hereditary Neurological Diseases, R.D. Wells and S.T. Warren (eds), Academic Press, CA, USA, pp 75-83.

Glover, T.W. (1998b) Instability at chromosomal fragile sites. *Recent Results in Cancer Research* 154:185-199.

Godde, J.S. and A.P. Wolffe (1996) Nucleosome assembly on CTG triplet repeats. *The Journal of Biological Chemistry* 271(25):15222-15229.

Godde, J.S., S.U. Kass, M.C. Hirst, and A.P. Wolffe (1996) Nucleosome assembly on methylated CGG triplet repeats in the fragile X mental retardation gene 1 promoter. *The Journal of Biological Chemistry* 271(40):24325-24328.

Griffith, J.D. and Y.H. Wang (1998) Nucleosome analyses and diseases of chromatin structure. In: Genetic Instabilities and Hereditary Neurological Diseases, R.D. Wells and S.T. Warren (eds), Academic Press, CA, USA, pp 677-690.

Gu, Y., Y. Shen, R.A. Gibbs, and D.L. Nelson (1996) Identification of *FMR2*, a novel gene associated with the *FRAXE* CCG repeat and CpG island. *Nature Genetics* 13:109-113.

Gusella, J.F. and M.E. MacDonald (1996) Trinucleotide instability: a repeating theme in human inherited disorders. *Annual Review of Medicine* 47:201-209.

Hambor, J.E., J. Mennone, M.E. Coon, J.H. Hanke, and P. Kavathas (1993) Identification and characterization of an *Alu*-containing, T-cell-specific enhancer located in the last intron of the human *CD8α* gene. *Molecular and Cellular Biology* 13(11):7056-7070.

Hansen, R.S., C.T.K. Canfield, M.M. Lamb, S.M. Gartler, and C.D. Laird (1993) Association of fragile X syndrome with delayed replication of the *FMR1* gene. *Cell* 73(7):1403-1409.

Hewett, D.R., O. Handt, L. Hobson, M. Mangelsdorf, H.J. Eyre, E. Baker, G.R. Sutherland, S. Schuffenhauer, J. Mao, and R.I. Richards (1998) *FRA10B* structure reveals common elements in repeat expansion and chromosomal fragile site genesis. *Molecular Cell* 1:773-781.

Hewitt, S.M., G.C. Fraizer and G.F. Saunders (1995) Transcriptional silencer of the Wilms' Tumor gene *WT1* contains an *Alu* repeat. *The Journal of Biological Chemistry* 270(30):17908-17912.

Howeler, C.J., H.F.M Busch, J.P.M Geraedts, M.F. Niermeijer, and A. Staal (1989) Anticipation in myotonic dystrophy: fact or fiction? *Brain* 112:779-797.

Igarashi, S., R. Koide, T. Shimohata, M. Yamada, Y. Hayashi, H. Takano, H. Date, M. Oyake, T. Sato, A. Sato, S. Egawa, T. Ikeuchi, H. Tanaka, R. Nakano, K. Tanaka, I. Hozumi, T. Inuzuka, H. Takahashi, and S. Tsuji (1998) Suppression of aggregate formation and apoptosis by transglutaminase inhibitors in cells expressing truncated DRPLA protein with an expanded polyglutamine stretch. *Nature Genetics* 16:111-117.

Imbert, G., Y. Feng, D.L. Nelson, S.T. Warren, and J.L. Mandel (1998) *FMR1* and mutations in fragile X syndrome: molecular biology, biochemistry, and genetics. In: Genetic Instabilities and Hereditary Neurological Diseases, R.D. Wells and S.T. Warren (eds), Academic Press, CA, USA, pp 27-53.

Jennings, C. (1999) What causes schizophrenia? *Nature Neuroscience* 2(4):295.

Jones, C., P. Slijepcevic, S. Marsh, E. Baker, W. Langdon, R. Richards, and A. Tunnacliffe (1994) Physical linkage of the fragile site *FRA11B* and a Jacobsen syndrome chromosome deletion breakpoint in 11q23.3. *Human Molecular Genetics* 3:2123-2130.

Jones, C., L. Penny, T. Mattina, S. Yu, E. Baker, L. Voullaire, W. Langdon, G. Sutherlund, R. Richards, and A. Tunnacliffe (1995) Association of a chromosome deletion syndrome and a fragile site within the proto-oncogene *CBL2*. *Nature* 376:145-149.

Kang, S., K. Ohshima, M. Shimizu, S. Amirhaeri, and R.W. Wells (1995) Pausing of DNA synthesis *in vitro* at specific loci in CTG and CGG triplet repeats from human hereditary disease genes. *The Journal of Biological Chemistry* 270(45):27014-27021.

Knight, S.J.L., A.V. Flannery, M.C. Hirst, L. Campbell, Z. Christodoulou, S.R. Phelps, J. Pointon, H.R. Middleton-Prince, A. Barnicoat, M.E. Pembrey, J. Holland, B.A. Oostra, M. Bobrow, and K.E. Davies (1993) Trinucleotide repeat amplification and hypermethylation of a CpG island in FRAXE mental retardation. *Cell* 74:127-134.

Kochanek, S., D. Renz and W. Doerfler (1995) Transcriptional silencing of human *Alu* sequences and inhibition of protein binding in the box B regulatory elements by 5'-CG-3' methylation. *FEBS* 360:115-120.

Koob, M.D., M.L. Moseley, L.J. Schut, K.A. Benzow, T.D. Bird, J.W. Day, and L.P.W. Ranum (1999) An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). *Nature Genetics* 21:379-384.

Korneluk, R.G. and M.A. Narang (1997) Anticipating anticipation. *Nature Genetics* 15:119-120.

Kremer, E.J., R. Pritchard, M. Lynch, S. Yu, K. Holman, E. Baker, S.T. Warren, D. Schlessinger, G.R. Sutherland, R.I. Richards (1991) Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence $p(CCG)_n$. *Science* 252:1711-1714.

Kunst, C.B., and S.T. Warren (1994) Cryptic and polar variation of the fragile X repeat could result in predisposing normal alleles. *Cell* 77(6):853-61.

Le Beau, M.M., F.V. Rassool, M.E. Neilly, R. Espinosa III, T.W. Glover, D.I. Smith, and T.W. McKeithan (1998) Replication of a common fragile site, *FRA3B*, occurs late in S phase and is delayed further upon induction: implications for the mechanism of fragile site induction. *Human Molecular Genetics* 7:755-761.

Magenis, R.E., F. Hecht and E.W. Lovrien (1970) Heritable fragile site on chromosome 16: probable localization of haptoglobin locus in man. *Science* 170:85-86.

Martin, F.H. (1999) Molecular basis of the neurodegenerative disorders. *The New England Journal of Medicine* 340(25):1970-1980.

Matsuyama, Z., Y. Izumi, M. Kameyama, H. Kawakami, and S. Nakamura (1999) The effect of CAT trinucleotide interruptions on the age at onset of spinocerebellar ataxia type 1 (SCA1). *Journal of Medical Genetics* 36:546-548.

McInnis, M.G. and R.L. Margolis (1998) Anticipation and psychiatric disorders. In: Genetic Instabilities and Hereditary Neurological Diseases, R.D. Wells and S.T. Warren (eds), Academic Press, CA, USA, pp 401-411.

Meuth, M. (1984) The genetic consequences of nucleotide precursor pool imbalance in mammalian cells. *Mutation Research* 126:107-112.

Miller, J.F. (1988) Bacterial electroporation. *Molecular Biology Reports: BioRad Laboratories* 5:1-4.

Mishmar, D., A. Raht, S.W. Scherer, G. Nyakatura, B. Hinzmann, Y. Kohwi, Y. Mandel-Gutfroind, J.R. Lee, B. Drescher, D.E. Sas, H. Margalit, M. Platzer, A. Weiss, L.C. Tsui, A. Rosenthal, and B. Kerem (1998) Molecular characterization of a common fragile site (*FRA7H*) on human chromosome 7 by the cloning of a simian virus 40 integration site. *Proceedings of the National Academy of Sciences, USA* 95:8141-8146.

Mitas, M. (1997) Trinucleotide repeats associated with human disease. *Nucleic Acids Research* 25(12):2245-2253.

Mogk, R.L., N.L. Carson, A.E. Chudley, and A.J. Dawson (1998) Transmission of the *FRAXA* haplotype from three nonpenetrant brothers to their affected grandsons: an update with AGG interspersion analysis. *American Journal of Medical Genetics* 75:28-34.

Moore, H., P.W. Greenwell, C.P. Liu, M. Arnheim, and T.D. Petes (1999) Triplet repeats from secondary structures that escape DNA repair in yeast. *Proceedings of the National Academy of Science, USA* 96:1504-1509.

Nancarrow, J.K., E. Kremer, K. Holman, H. Eyre, N.A. Doggett, D. Le Paslier, D.F. Callen, G.R. Sutherlund, and R.I. Richards (1994) Implications of *FRA16A* structure for the mechanism of chromosomal fragile site genesis. *Science* 264:1938-1941.

Nelson, D.L. (1996) Allelic expansion underlies many genetic diseases. *Growth Genetics and Hormones* 12(1):1-4.

Nelson, D.L. (1998) FRAXE mental retardation and other folate sensitive fragile sites. In: Genetic Instabilities and Hereditary Neurological Diseases, R.D. Wells and S.T. Warren (eds), Academic Press, CA, USA, pp 65-74.

Nolin, S.L., F.A. Lewis, L.L. Ye, G.E. Houck, A.E. Glicksman, P. Limprasert, S.Y. Li, N. Zhong, A.E. Ashley, E. Feingold, S.L. Sherman, and W.T. Brown (1996) Familial transmission of the *FMR1* CGG repeat. *American Journal of Human Genetics* 59:1252-1261.

Norris, J., D. Fan, D. Aleman, J.R. Marks, P.A. Fureal, R.W. Wiseman, J.D. Iglehart, P.L. Deininger, and D.P. McDonnell (1995) Identification of a new subclass of *Alu* DNA repeats which can function as estrogen receptor-dependent transcriptional enhancers. *The Journal of Biological Chemistry* 270(39):22777-22782.

Orr, H.T. (1994) Unstable trinucleotide repeats and the diagnosis of neurodegenerative disease. *Human Pathology* 25:598-601.

Pearson, C.E., E.E. Eichler, D. Lorenzetti, S.F. Kramer, H.Y. Zoghbi, D.L. Nelson, and R.R. Sinden (1998) Interruptions in the triplet repeats of *SCA* and *FRAXA* reduce the propensity and complexity of slipped strand DNA (S-DNA) formation. *Biochemistry* 37:2701-2708.

Ritchie, R.J., S.J.L. Knight, M.C. Hirst, P.K. Grewal, M. Bobrow, G.S. Cross, and K.E. Davies (1994) The cloning of *FRAXF*: trinucleotide repeat expansion and methylation at a third fragile site in distal Xqter. *Human Molecular Genetics* 3:2115-2121.

Reidy, J.A. (1988) Role of deoxyuridine incorporation and DNA repair in the expression of human chromosomal fragile sites. *Mutation Research* 200:215-220.

Samadashwily, G.M., G. Raca and S.M. Mirkin (1997) Trinucleotide repeats affect DNA replication *in vivo*. *Nature Genetics* 17:298-304.

Sambrook, J., E.F. Fritsch and T. Maniatis (1989) Molecular Cloning: A Laboratory Manual, 2nd Edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Sarai, A., J. Mazur, R. Nussinov, and R.L. Jernigan (1989) Sequence dependence of DNA conformational flexibility. *Biochemistry* 28(19):7842-7849.

Scherzinger, E., R. Lurz, M. Turmaine, L. Mangiarini, B. Hollenback, R. Hasenbank, G.P. Bates, S.W. Davies, H. Lehrach, and E.E. Wanker (1997) Huntingtin-encoded polyglutamine expansions from amyloid-like protein aggregates *in vitro* and *in vivo*. *Cell* 90:549-558.

Schinzel, A. (1994) Human Cytogenetics Database. Oxford University Press, Great Britain.

Schultz, S.K. (1999) Schizophrenia. *The Lancet* 353:1425-1430.

Shastry, B.S. (1994) Heritable trinucleotide repeats and neurological disorders. *Experientia* 50(11-12):1099-1105.

Smith, M.A.C., M.A.V. de Souza, S.L.M. Payao, S. Pugliese, and J.J. Mari (1997) Schizophrenia and fragile sites. *Psychiatric Genetics* 7:175-177.

Strachan, T. and A.P. Read (1996) Human Molecular Genetics. BIOS Scientific Publishers Ltd., Oxford, USA, pp 150-201.

Sutherland, G.R. (1977) Fragile sites on human chromosomes: demonstration of their dependence on the type of tissue culture medium. *Science* 197:265-266.

Sutherland, G.R. (1979) Heritable fragile sites on human chromosomes I. Factors affecting expression in lymphocyte culture. *American Journal of Human Genetics* 31:125-135.

Sutherland, G.R. (1988) The role of nucleotides in human fragile site expression. *Mutation Research* 200:207-213.

Sutherland, G.R. and R.I. Richards (1995) The molecular basis of fragile sites in human chromosomes. *Current Opinion in Genetics and Development* 5:323-327.

Sutherland, G.R., E. Baker and R.I. Richards (1996) Fragile sites. *Encyclopedia of Molecular Biology and Molecular Medicine* 2:313-318.

Sutherland, G.R. and R.I. Richards (1999) Fragile sites - cytogenetic similarity with molecular diversity. *American Journal of Human Genetics* 64:354-359.

Therman, E., and M. Susman (1993) Human Chromosomes: structure, behavior and effects, 3rd edition, Springer-Verlag, New York, USA.

Timchenko, L.T. (1999) Myotonic dystrophy: the role of RNA CUG triple repeats. *American Journal of Human Genetics* 64:360-364.

Tommerup, N., J. Nielsen and M. Mikkelsen (1985) A folate sensitive heritable fragile site at 19p13. *Clinical Genetics* 27(5):510-4

Tunnacliffe, A., C. Jones, D. Le Paslier, R. Todd, D. Cherif, M. Birdsall, L. Devenish, C. Yousry, F.E. Cotter, and M.R. James (1999) Localization of Jacobsen syndrome breakpoints on a 40 Mb physical map of distal chromosome 11q. *Genome Research* 9(1):44-52.

Usdin, K. and K.J. Woodford (1995) CGG repeats associated with DNA instability and chromosome fragility form structures that block DNA synthesis *in vitro*. *Nucleic Acids Research* 23(20):4202-4209.

Vansant, G. and W.F. Reynolds (1995) The consensus sequence of a major *Alu* subfamily contains a functional retinoic acid response element. *Proceedings of the National Academy of Science, USA* 92:8229-8233.

Verkerk, A.J.M.H., M. Pieretti, J.S. Sutcliffe, Y.H. Fu, D.P.A. Kuhl, A. Pizutti, O. Reiner, S. Richards, M.F. Victoria, F. Zhang, B.E. Eussen, G.J.B. van Ommen, L.A.J. Blonden, G.J. Riggins, J.I. Chastain, C.B. Kunst, H. Galjaard, C.T. Caskey, D.L. Nelson, B.A. Oostra, and S.T. Warren (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* 65:905-914.

Vincent, A., D. Heitz, C. Petit, C. Kretz, I. Oberlé, and J.L. Mandel (1991) Abnormal pattern detected in fragile X patients by pulse-field gel electrophoresis. *Nature* 349:624-626.

Wang, L., W. Paradee, C. Mullins, R. Shridhar, R. Rosati, C.M. Wilke, T.W. Glover, and D.I. Smith (1997) Aphidicolin-induced *FRA3B* breakpoints cluster in two distinct regions. *Genomics* 41:485-488.

Wang, Y.H., R. Gellibolian, M. Shimizu, R.D. Wells, and J. Griffith (1996) Long CCG triplet repeat blocks exclude nucleosomes: a possible mechanism for the nature of fragile sites in chromosomes. *Journal of Molecular Biology* 263:511-516.

Wang, Y.H. and J. Griffith (1996) Methylation of expanded CCG triplet repeat DNA from fragile X syndrome patients enhances nucleosome exclusion. *The Journal of Biological Chemistry* 271(38):22937-22940.

Warren, S.T. and D.L. Nelson (1994) Advances in molecular analysis of fragile X syndrome. *Journal of American Medical Association* 271:536-542.

Warren, S.T. (1996) The expanding world of trinucleotide repeats. *Science* 271:1374-1375.

Westneat, D.F., W.A. Noon, H.K. Reeve, and C.F. Aquadro (1988) Improved hybridization conditions for DNA "fingerprints" probed with M13. *Nucleic Acids Research* 16:4161.

Wilke, C.M., B.K. Hall, A. Hoge, W. Paradee, D.I. Smith, and T.W. Glover (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. *Human Molecular Genetics* 5:187-195.

Wilmot, G.R. and S.T. Warren (1998) A new mutational basis for disease. In: Genetic Instabilities and Hereditary Neurological Diseases, R.D. Wells and S.T. Warren (eds), Academic Press, CA, USA, pp 3-12.

Wu, J., J. Grindlay, P. Bushel, L. Mendelsohn, and M. Allan (1990) Negative regulation of the human ε-globin gene by transcriptional interference: role of an *Alu* repetitive element. *Molecular and Cellular Biology* 10(3):1209-1216.

Yazawa, I., H. Nakase and H. Kurisaki (1999) Abnormal dentatorubral-pallidoluysian atrophy (DRPLA) protein complex is pathologically ubiquitinated in DRPLA brains. *Biochemical and Biophysical Research Communications* 260:133-138.

Yu, S., M. Mangelsdorf, D. Hewett, L. Hobson, E. Baker, H.J. Eyre, M. Lapsys, D. Le Paslier, N.A. Doggett, G.R. Sutherland, and R.I. Richards (1997) Human chromosomal fragile site *FRA16B* is an amplified AT-rich minisatellite repeat. *Cell* 88:367-374.

Zhong, N., W. Ju, J. Pietrofesa, D. Wang, C. Dobkin, and W.T. Brown (1996) Fragile X "gray zone" alleles: AGG patterns, expansion risks, and associated haplotypes. *American Journal of Medical Genetics* 64:261-265.

APPENDIX A:

Fluorescent *in situ* Hybridization of *FRA19B* With Cosmid 25308

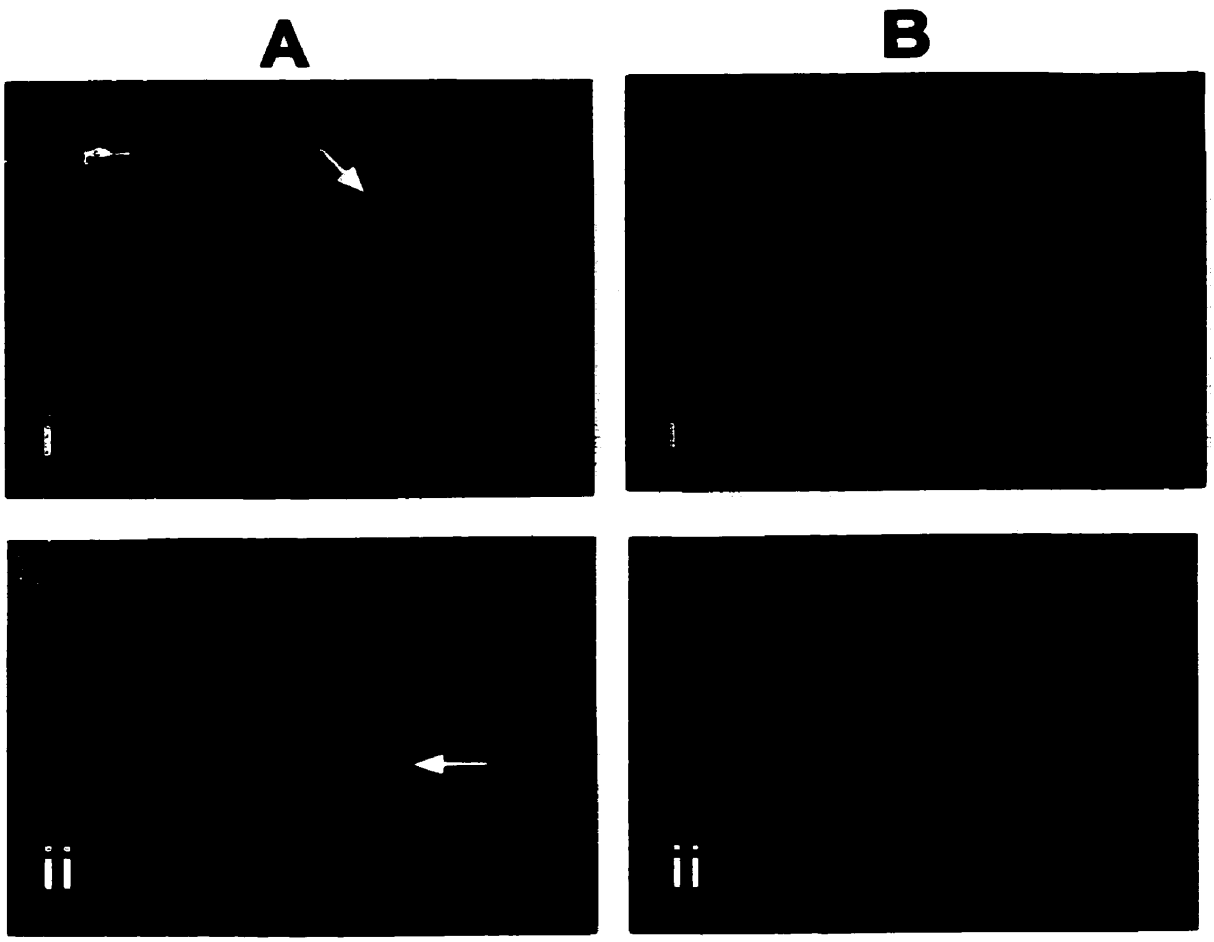

Figure 12: Fluorescent *in situ* hybridization analyses of two partial metaphase spreads illustrating cosmid 25308 spanning the *FRA19B* locus. A.) Metaphase chromosomes (blue) counter stained with 4,6-diamidino-2-phenylindole (DAPI) showing the fragile site on chromosome 19 (arrow). B.) Metaphase chromosomes (blue) counter stained with DAPI and cosmid 25308 (red) detected with Texas Red showing a split signal.

APPENDIX B:

DNA Sequences of *Bam*HI Fragments of Cosmid 25308

Bold type indicates homology to *Alu* repetitive elements.

Dashed lines (-----) indicate gaps in the sequence data.

2.0 kb *Bam*HI fragment of 25308

The underlined sequence indicates the position of the entire sequence of *Alu*-splice probe 4-12.

ATATTTCCTT CACTCTCTCC CTCCCTCCTA GATCACTGTC CCTGCCCTCT
CTTTCCTCCT TTCCACCCCC TTTGGGCCAG TCACTTTGCT **CTGCAACATC**
TGCCTTCCCG GTTTCAGAGC GATTCGCCGC CTCAGCCTCC GAGTAGCTGG
GAATTACAGG TGCGTGACCA CCACACCTGA CTCATTTTTG TATTTTTAGT
AGAGACAGGG TTTACATGTT GGCCAGGTGA CTCGAACTTC TGACCTCAAG
TGATCACCTG CCTCGGCTCC AAGGAGCGGT TAAGGGGAGA CCGG------
CTCGGCCTCC CAAAGTGCTG GGTTTACAGG CGTGACGACC GCGCCTGGCC
CATATTTATT TATTTGTTTG TAGAGATGGG CTCTCGCTCT GTTGCTTAGG
CTGGAGTGCA GTGGGGTGGT TATCACCCAC TGCAGCCTCA AACTCCTGGG
TTCAAGCGAT CCTCCCGAGT AGCTCTGGAT ACTCCGTGTG CACTACCGTG
CAAACATAAT TTTATTTTTT TCTGGACATG GGATCTTTCT GTGCTCCCCA
GGCTGGTCTC CCTAATACTT AAGCACTGCC AGACACGCGT GGTGCTCTGT
CTTCTGTCTG TTCTGTGTCT CCTCTGAAGG GCAGGGACAC TTGTGTGTTG
TGTTGCCTGC TGTATCCCCA CCACTAAGCA CACAGTAGGC AGTCTAAAGT
CCTTGCTGCT AGCTGATTAT AAGTTCAAGA CAAAAGAGGG AGGAGTCTAG
GATTGCCACC TGGAGCTCCA AGTGCAGATA GGGAACAGCC AGGGCTGCAG
GTGATACCCA GGCCTGGGGC TCAAGCGAGG TGTGGGGTGG AGGGACCACA
TCGGGCTCAT CTACATACGG GGCGGCACGA GAGAAAAAGG CTCT**CAAGCC**
GGGCACAGTG GCTCACACCT GTAATCCCAG CACTTTGGGA GGCCAAGGTG
GGTGGATCAC CTGAGGTCAG GAGTTCAAGA CCAGCCTGAC TAACATGGCG
AAACCCCATC TCTACTAAAA ATACAAAAAT TAGTCGGGCA TGGTGGCAGG
CGCCTGTAGT TCCAGCTACT CAGGAAGCTG AGGCAGGAGA ATTGCTTGAA
CCTGGGAGAC AGGGGTTGCA GTGAGCTGAG ATCTTGCCAC TGCCCTCCAG
CCTGGGTGAC AGAGTGAGAC TCCATGTCAA AAGAAAAG----------
AGGTGTGAGC CACTGTGCCG GCCTGAGAGC CTTTTTCTCC TGTGCCGCCC
CGTATGTACA TGAGCCCGAT GTGGTCCCTC CACCCCACAC CTCGCTTGAG
CCCCAGGCCT GGGTATCACC TGCAGCCCCT GGCTGTTCCC TATCTGCACT
TGGAGCTCCA GGTGGCAATC CTAGACTCCT CCCTCTTTTG TCTTGAACTT
ATAATCAGCT AGCAGCAAGG ACTTTAGACT GCCTACTGTG TGCTTAGTGG
TGGGGATACA GCAGGCGAAC ACAACACACA AGTGTCCCTG CCCTCAGAGG
GAGACACAGA ACAGACAGAA GACAGAGCAC CACAGCGTGT CTGGCAGTGC
TTAAGTATTA GGGAGACCAG CCTGGGGAGC ACAGAAAGAT CCCATGTCCA
GAAAAAAATA AAATTATCTT TGCACGGTAG TGCACACGGA GTATCCAGAG
CTACTC**GGGA GGATCGCTTG AACCCAGGAG TTTGAGGCTG CAGTGGGTGA**
TAACCACCCC ACTGCACTCC AGCCTAAGCA ACAGAGCGAG AGCCCATCTC
TACAAACAAA TAAATAAATA TGGGCCAGGC GCGGTCGTCA CGCCTGTAAA
CCCAGCACTT TGGGAGGCCG AGGCAGGTGG ATCACTTGAG GTCAGAAGTT
CGAGTCAGCC TGGCCAACAT GGTAAAACCC TGTCTCTACT AAAAATACAA
AAATGAGTCA GGTGTGGTGG CACGCAGGTG TAATCCCAGC TACTCGGGAG
GCTGAGGCGG GCGAATCGCT TGAACCGGGA AGGCAGATGT TGCAGTGAGC
AAAGTGACTG GCTGAAAGGG GGTGGAAAGG AGGAAAGAGA GGGCAGGGGA
CAGTGATCTA GGAGGGAGGG AGAGAGTGAA GGATCCCCGG GTACCGAGCT
CGAA

1.8 kb *Bam*HI fragment of 25308

The underlined sequence corresponds to that shared with *Alu*-splice insert 2-15.

CTGAGCACAG ACGGGGAGGT GGCAGGCTCT GGCCCCATCT CCAGGAGAAG
CGGTGAGGCA GGGAGGACAG ACAGGACCTT CTGAGGAAGC CCACGGAGCC
ACTCGGAGGG GCTGGGAGGG GCTGAGATGC TCTGGGTGGA TGAAGGGATA
CCTGTGGGAG GGGGGCTACT GGTGTCATCG AGGTCCCAGC ACTGGCCACG
GGCAAGGGGT GAGACCACAT AGCCTCAGGC ACGACGGGCA GGGAGAGGGC
AGGAACGGGC AGACAGGCAG GCACTCTGGC CTGGGGCACA GGCTGGTGGC
TCT-------- ----------
CCCTGTATCC AGCCTTGAAG GCGAGCACCA GTCCTGAGCA GAGTTCGAGA
CCAGCCTGCC CACATGTGAC CCGTCTCTAA TAAATCAAAT **AGCCAGGCAT**
GGTGGTGTGT GCCTGTAATC CCAGTCACTC AGGAGGTTGA GTCAGGAGAA
TTGCTTGAAC TCAGGACGGT GGAGGTTGCA GTGAGCTGAG TTTATGCCAT
TGCACTCCAG CCTGGGCGAC ACAGCAAGAC CCTGTCTCAA AATAAATAAA
TAAATAAATA AATAAATAAA TAAATAAATA AGATTAGACC AACACTAGCA
GAACAGCCAT GCCAGGTGCA CAAGGGGGCC AAACCCT

1.6 kb *Bam*HI fragment of 25308

The underlined sequence corresponds to that shared with *Alu*-splice insert 2-15.

GATCCCCTGC CTGGGGCTCT GTTCAGCAGC TGCCTGTACC CTCTGCAGCC
CTGGGCTTCC CCTGGGAGCC CCTGCCCCCT GTTGGCCTCA GGCCTGGGCA
GGCGTGTGTG TGTGTGGCAG GCACAGAGCA GAGATTGGGG TGGCAGTGGG
GTGTGGGGCG CGGAGCCCTC GATGAGGTCT CACATTAGGC CCTCCCTCTG
AGGGGTGGGG AACCTCTCCT GCTCCTCCCA CCCGACTGTT TTTATTTTGA
TGGGTATGTG GGGCCTGTTT GTTACCCCTC ACCCCTTTAA ACAATAATTT
ATTGAGGTGA AATTCATGTA ACATGAAACT AACCATTTAA AAGTGAAC--
GAGAGGGGCT GGTTGTATTT TAAGCATGTT GAGTGGTTCA GCAGGAACTC
GGAGGCCTTC TCTCCCTGTG GACGCTAGGC TGCCTTCCTA CAGGCCACAC
GGGCTTGGAA GCATGTGGCA CTGCCTGCCT GGGGCCCAGG AACTATTAAT
ATACCTGCAG CCACAGGCAG GCCCCGTGCC CGCCCAGGGC CCTGCTGTGC
TAACCCT

0.7 kb *Bam*HI fragment of 25308

GATCCAAATC CTTGAAAATA CGCTGGCAAC AACGCTGCTG AACCGCTCCG
CTCGCGCGTG GCGCTGACTG AAAGTGGCCA ACGTTGTTAC GAACATGCGT
TGGAAATCCT CACCCAATAT CAGCGACTGG TCGATGACGT CACGCAAATC
AAAACGCGCC CGGAGATGAT TCGTATTGGT TGCAGTTTTG GTTTTGGGCG
CCATATTGCG CCACTATTAC CGAACTGATG CGCAATTATC CTGAGTTACA
GGTGCATTTT GAATTGTTCG ATCGCAAATT GATTTAGTGC AGGATAATAT
TGATCTGGAT ATTCGATTAA TGACGAAATT CTGATTATTA TATTG-----
TATGCTGCGT TCCGAGTGGG ATGTGTTGCC GTTTCTGGAG AGTGGCAAAC
TGGTGCAAGT ATTGCCAGAG TATGCACAAA GCGCCAATAT CTGGGCTGTT
TACCGGGAGC CGCTCTATCG CAGCATGAAA TTACGTGTCT GCGTGGAATT
TCTGGCGGCA TGGTGCCAGC AACGGCTGGG CAAGCCCGAT GAAGGATATC
AGGTCATGTA GATCCACCAG A

APPENDIX C:

DNA Sequences of the *Alu*-Splice Inserts

Bold sequence indicates the primer regions
utilized to produce the "edited" *Alu*-splice probes.

1-9

GCGGTCACAG TGAAACCCAA ACCCTAACAA **TAGCGAGATA** **CTTTCGGTGC**
AGAGGCCGAG CTGACCCGTC ACCACCGGGA TGACTGGCTG CAAGCGGCGG
AGGCTGCGAC GCTCCCCACT GCCCCCTTGT GGCCGCACAG ACTTACTCCC
GCCTGCTGGA AGGGCACCCG CAGCAGCCGG CTCAGGAACA GTCCCAAAGC
CGATGTCAGG GACGGAGACA GCTCTGGTGC AGAATGGGAC CTTAACCCTG
GCCCTGCCAC CAGCCAAGAG AGCTCAGGGA GCCTGCATTC ATTTATCCAT
TCATTCATTC CTGACTCCCT CCTTCCCATA CCCCAGGCAC TGTTCCAGTT
CCTGGGATAT GGCAGTGAAA AAAACCAAGC CAGCCTCTGG GGTTCATGGG
ACCTGAATTC TAGCCTGTGC GTGCTGGGCA AGCAGGGGGT GGGGGCAGAT
AGCCATAGCA TGTTTTGTCA CCAAGAGTTG TGAACAAAAC TAGAAGGTTG
TGG**GGGAGGT** **CACAATGTGG** **TTT**TTTTTGT TTTTGTTTTT TTTTTGAGAG

1-13

GCGGTCCCAG GTTCAAGCAA TTCTCCTGTC TCAGCCTCCC AAGTAGCTGG
GATTACAGGC ACGCATCACC ACGTCTGGCT AATTTTTGTA TTTTGGTAGA
GACGGGTTCA CCATGTTGGC CAGGCTGGTC TCAAACTCCT GATTTCAAGT
GATCTGCTCA CTTTGGCCTC CCAAAGTGCT GTGATTACAG GCGGGAGCCA
CCGCATCCAA CCATCTCAGA CATCTCAGCA GCAGTGATAC CACTGTGGCA
GCCACAGATG ACATCCAGTG AGCATGGCGT GTTCAATACA ACTCTATTTA
TGGACACCAA AATTTGAATT TAATTTTCAC ATATCATGAA ATAGGTTTCT
TTGATTTTTT TTTT**CCATCC** **GCTTAAAAAG** **G**TAAAAACGT CATGTATTTT
TTAAAAGATA AAACAAAAGT TAAAGCCATT TTTAGTCGGT TCAAAAGCCG
GCAGTANACC AGAGTTGCCC TATCGGCTGT ANTTTGCTGA CCCATTTGTC
TAA**GCCATCA** **TTACCTTCAC** **AG**ANTTATTT CCACAACCTT TA

2-2

GCGGTCGCAG GT**TCAAGACT** **TGGCTTCCCC** **AGCTT**TACCT GCTTGCCACT
GCTGGGTGGA ACAGACTGGA GGAGTGAGGG CAGAAGCTGG GAGCAAAGAG
GGAGCTAT**AG** **AGAGGTGAAA** **A**GTGGGCTGG GCGCANTGGC TCACGCCTGG
AATCCCAGCA CTTTAGGAGG CTGAGACAGG ATGATCACTT AAGCCCAGGA
GTTCAAGACC ANCTTGGGCA AAATAGCAAG GCTCTGCCTC CACAAAAAGT
ATTTATAAAA ATTAGCTGTG TCCACGCATG GTGACTCACA CCTGTAATCC
ANGCACTTTG GGAGGCTGAG GTGGGCAGAT CACCTGAGGT CAGGAGTTCG
AGACCATCCT GGCCAACATG GTAAAACCCC GTGTAAAAAC ACAANAATTA
GCCAAGTGTG GTGGCTCACN CCTGTAATCC CCG

2-15

Italics sequence indicates a *Bam*HI restriction site (referred to on page 71).

GCGGTCGCAG GT**ATCCCTTC** **ATCCACCCAA** **GA**GCATCTCA GCCCCTCCCA
GCCCCTCCGA GTGGCTCCGT GGGCTTCCTC AGAAGGACCT GTCTGTCCTC
CCTGCCTCAC CGCTTCTCCT GGAGATGGGG CCAGAGCCTG CCACCTCCCC
GTCTGTGCTC A*GGATCC*CCT GCCTGGGGCT CTGTTCAGCA GCTGCCTGTA
CCCTCTGCAG CCCTGGGCTT CCCCTGGGAG CCCCTGCCCC CTGTTGGCCT
CAGGCCTGGG CAGGCGTGTG TGTGTGTGGC AGGCACAGAG CAGAGATTGG
GGTGGCAGTG GGGTGTGGGG CGCGGAGCCC TCGATGAGGT CTCACATTAG
GCCCTCCCTC TGAGGGGTGG GGAACCTCTC CTGCTCCTCC CACCCGACTG
TTTTTATTTT GATGGGTATG TGGGGCCTG**T** **TTGTTACCCC** **TCACCCCTTT**
AAACAATAAT TTATTGAGGT GAAATTCATG TAACATGAAA CTAACCATTT
AAAAGTGAAC

2-16

GCGGTCACAG GT**TCAAGACT** **TGGCTTCCCC** **AA**GCTTTACC TGCTTGCCAC
TGCTGGGTGG AACAGACTGG AGGAGTGAGG GCAGAAGCTG GGAGCAAAG**A**
GGGAGCTATA **GAGAGGTGAA** AAGTGGGCTG GGCGCAGTGG CTCACGCCTG
TAATCCCCG

3-2

GGCCGCTCTA GAACTAGTGG ATCCCCCGGG CTGCAGGAAT TCGATATCAA
GCTTATCGAT ACCGTCGACC ACTGCACTCC AGCCTGGGCG ACAGCGAGAC
TCTCTCAAAA AAAAAACAAA AACAAAAAAA ACCACATTGT GACCTCCCCC
ACAACCTTCT AGTTTTGTTC ACAACTCTTG GTGACAAAAC ATGCTATGGC
TATCTGCCCC CACCCCCTGC TTGCCCAGCA CGCACAGGCT ANAATTCAGG
TCCCATGAAC CCCAGAGGCT GGCTTGGTTT TTTTCACTGC CATATCCCAG
GAACTGGAAC AGTGCCTGGG GTATGGGAAG GAGGGAGTCA GGAATGAATG
AATGGATAAA TGAATGCAGG CTCCCTGAGC TCTCTTGGCT GGTGGCAGGG
CCAGGGTTAA GGTCCCATTC TGCACCANAG CTGTCTCCGT CCC**TGACATC**
GGCTTTGGGA **CTGT**TCCTGA GCCGGCTGCT GCGGGTGCCC TTCCAGCAGG
CGGGAGTAAG TCTGTGCGGC CACAAGGGGG CAGTGGGGAG CGTCNCAACC
TCCGCCGCTT GCAGCCAGTC ATCCCGGTGG TGACGGGTCA GCTCGGCCTC
TGCACCGAAA GTATCTCGCT ATTGTTAGGG TTTGGGTTCA CTGTGACCAN
CTGGGCGAGG CATCANTGCA CCAGCTCTCT GGCCCTGGNT TCCTGTGGGA
TGGGGCCCCT TCAAGAGCAT GGTCAGCGTC CCANGAGAAA GGGGCCTGTG
GGATCCTGCC **AAGGGTTTGG** **GCCCCTTTGT** GCACCTGGGC

3-8

GGCCGCTCTA GAACTAGTGG ATCCCCCGGG CTGCAGGAAT TCGATATCAA
GCTTATCGAT ACCGTCGACC ACTGCACTCC AGCCTGGGCG ACACAGTGAG
ACTCTGTCTC AAAAAAAAAA ATAAAAAAAA TA**AACTATAC** **ACCTGCCCCC**
TTCAACTGCC ACTCCCTGTC CCCTTTCCTT TCTTTCTTCT TTCCATGTCG
CACGTCACCT TTGAACATAT TACATATANC GATTTATTTC ATTGTCACNC
TTTCTTTGCC AAGATGTGAG TTCCATGCGA GCAGGGATTT TTGTCNCNCC
TTCTATAGCT GAACCCCACA GTCC**TTAACA** **GGGCCTGGCA** **TGTGT**CANTG
AANAAACNAG CCTCAGTTTC CTTACCTGCA AAAGAGGATG AATACATTGT
AAAGGANATG ATTCNTGTTG AGGATGANCC NCCGTGCCTG GCTGGTTGTG
ATTGGGCGGG TGGTGTTACC TGCCGCCGCG TGTATAAAGC CTGCCACGTA
AATCCCAGTC TAGGCTCTTT GGCAAGCTGA GCTCATGGGG GTGGACTAAC
CTGTCTGCTG GACTCGGGCA NTCGGGGGTT GGCCCGGATG AAGCTCCTCC
TGCTCCTCCA NTCANTTTGT NTCTGCAGGG CTCANCCTCT GGCCCCCTTC
CAGCTTCCTG GGTCCGAAAA

4-2

GGCCGCACTT ACCT**TGCTTT** **CCCCAAGATC** **AGCT**TTGCAA ACGCTAGCAA
CCTGCTGCCC TTCTCTCTTG CTTACTCTGG ATAAACTCAC CTCTTTCCCT
TCCTTGGTGC CTATCCCCGT GCAACTGAAC ATAATGGGGG GCGGGGAGAT
AAGAAAAAA**A** **CAGCTCTAAG** **CCAGGCAAAG** **TG**GCTCACGC CTGTAATCCC
CG

4-4

GGCCGCTCTA GAACTAGTGG ATCCCCCGGG CTGCAGGAAT TCGATATCAA
GCTTATCGAT ACCGTCGACG GGGATACAGG CGTGAGCCAC CGCATCCAAC
CATCTCAGAC ATCTCAGCAG CAGTGATACC ACTGTGGCAG CCACAGATGA
CATCCAGTGA GCATGGCGTG TTCAATACAA CTCTATTTAT GGACACCAAA
ATTTGAATTT AATTTT**CACA** **TATCATGAAA** **TAGG**TTTCTT TGATTTTTTT
TTTCCATCCG CTTAAAAAGG TAAAAACGTC ATGTATTTTT TAAAAGATAA
AACAAAAGTT AAAGCCATTT TTAGTCGGTT CAAAAGCCGG CAGTANACCA
GAGTTGCCCT ATCGGCTGTA GTTTGCTGAC CCATTTGTCT AAGCCATCAT
TACCTTCACA GAATTATTT**C** **CACAACCTTT** **AATTGGAC**TT TTTTTTTTTT
TTTTTTTGAA AGGGGGTCC

4-7

GGCCGCACAC ACCACCTTCC CCAGTGGCCT CAGTGTATGA CCTCGTTTCT
CAGTGGCACC CGGGGGCGGC TGCACCAGCC TCCCCGAGGC GCCACAGTGG
AAGCAGAATG TCACCCAGGG TGGGGAGGAA GGGCTGAGAG TGGCTGACCT
CCCACTCCCA GGTCCCTGGA CCGCTCTGCT CCTGTCTTCT ACCTGTGCTG
GCAAGACGAG AGGTGCCAGG GTCTCAATGC CAATGTCTCT GGGCTCTCGG
ATCCGAGGCC AGGGCCAGGT GCAGAGGCAG CGGTAGCCCG GGACCCGGAG
AAGCTGGAGC TGCTGTTCCC ATCCAGCAGA GCCCCGAGCG AGCAGGTAGC
AATCCAGCCT TGGAAGATGG TCATCGGACG CACGAGCTCC TGGCTTGGGC
GAGCAGCCAC AGCCATTCAT AGAGGATGAA TATGGCACAA GACAAACACA
GTGGGAAGTT GGAGAACCTG GGTCTCAGCC CTGGGCCTGC CCCCCATTCT
CAGTGGTACC CACAGCCTAC ACCCATCCCT GCGGTTGTGC CCTCANGGCC
TCCGAGGAGC ACTGGCCCAC CCAAANCGAC GCTGATCTCG CCCGTTTCAT
TCAAACGTTG ACATCTATCA NATGCTCACC CTGTACCTAA GACTATGCTN
AGACTGCCAA CCANGCCTCA ACCANGCCCT ATAANAAAAA NACTGGGCCG
GGCGCAATTG GCTCACNCCT GTTATCCCCG TTCAACCTCA AAGGGGGGGC
CCGTANCCAA ATTCCCCCNA AAATTAATT

4-12

GGCCGCTCTA GAACTAGTGG ATCCCCCGGG CTGCAGGAAT TCGATATCAA
GCTTATCGAT ACCGTCGACG GGGATTACAG GGTGAGCCAC TGCGCCTGGC
CCATATTTAT TTATTTGTTT GTAGAGATGG GCTCTCGCTC TGTTGCTTAG
GCTGGAGTGC AGTGGGGTGG TTATCACCCA CTGCAGCCTC AAACTCCTGG
GTTCAAGCGA TCCTCCCGAG TAGCTCTGGA TACTCCGTGT GCACTACCGT
GCAAAGATAA TTTTATTTTT TTCTGGACAT GGGATCTTTC TGTGCTCCCC
AGGCTGGTCT CCCTAATACT TAAGCACTGC CAGACACGCT GTGGTGCTCT
GTCTTCTGTC TGTTCTGTGT CTCCCTCTGA GGGCAGGGAC ACTTGTGTGT
TGTGTTGCCT GCTGTATCCC CACCACTAAG CACACAGTAG GCAGTCTAAA
GTCCTTGCTG CTANCTGATT ATAAGTTCAA GACAAAAGAG GGAGGAGTCT
ANGATTGCCA CCTGGAGCTC CAAGTGCANA TANGGAACAG CCANGGGCTG
CANGTGATAC CCAGGCCTGG GGCTCAAGCG AGGTGTGGGG TGGNGGGAAC
ACATCNGGCT CATCTACATA CNGGGCGGCA GCAGAGAAAA AGGCTCTCAA
GCCGGGCAAA AGTGGCTCAC GCCTGTTAAT CCCCGT