PHYSICAL MAPPING OF HUMAN CHROMOSOME 16

Sinoula Apostolou B.Sc. (Hons.)

A thesis submitted for the Degree of Doctor of Philosophy to The University of Adelaide

Department of Cytogenetics and Molecular Genetics,
Women's and Children's Hospital, North Adelaide, South Australia

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

August, 1997

CORRECTIONS

1. The symbol for the gene causing FMF has been changed from MEF to MEFV.

2. page 7, para 3: RFLPs are not always caused by single nucleotide changes.

3. page 8, para 3: The reference for Beckman and Weber is: Beckman, J.S., Weber, J.L. (1992). Survey of human and rat microsatellites. Genomics: 627-631.

4. page 68: the first series of FMF cases was reported by Siegal in 1945 (Ann. Intern. Med. 23: 1-21).

5. page 70: most markers used in linkage studies reported by Aksentijevich *et al* (1991) were RFLPs, not (AC)n markers.

6. table 3.4: '+' should be inserted in the cell for somatic cell hybrid CY107 and cosmid 60E2.

7. page 184, para 3: 75 cm^3 and 150 cm^3 should be 75 cm^2 and 150 cm^2, respectively.

8. figure 7.1: D16S253 should be D16S523.

9. figure 7.6 legend: should read D16S3070, a microsatellite marker, and D16S3275 were physically mapped.

STATEMENT

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

I give consent to this copy of my thesis, when deposited in the University Library, being available for loan and photocopying.

Sinoula Apostolou

CONTENTS

Statement I

Contents II

Abbreviations III

Summary V

Acknowledgements IX

Chapter 1 Literature Review 1

Chapter 2 Materials and Methods 74

Chapter 3 Identification of Cosmids Localised to the 16q24 Chromosomal Region 106

Chapter 4 Isolation of Transcripts Encoded by Cosmids Localised to 16q24 143

Chapter 5 Characterisation of Two Novel Transcripts Localised to 16q24.3-qter 178

Chapter 6 Mutation Analysis of Two Transcribed Sequences Localised to 16q24.3-qter 213

Chapter 7 Isolation and Identification of Transcripts in the Familial Mediterranean Fever Candidate Region 264

Chapter 8 Concluding Remarks 289

References 294

ABBREVIATIONS

(AC)n:	an AC repeat sequence, repeated *n* times
APRT:	adenosine phosphoribosyl transferase gene
bp:	base pairs
BACs:	bacterial artificial chromosomes
BLAST:	basic local alignment search tool
BSA:	bovine serum albumin
CDGE:	constant denaturant gel electrophoresis
cDNA:	complementary deoxyribonucleic acid
cM:	centimorgan
CMC:	chemical mismatch cleavage
DEPC:	diethyl pyrocarbonate
DGGE:	denaturing gradient gel electrophoresis
DNA:	deoxyribonucleic acid
dNTP:	deoxynucleoside triphosphate
EDTA:	ethylene diaminetetraacetic acid
ESTs:	expressed sequence tags
FA:	Fanconi anaemia
FAA:	Fanconi anaemia group A gene
FAB:	Fanconi Anaemia/Breast Cancer consortium
FISH:	fluorescence *in situ* hybridisation
FMF:	familial Mediterranean fever
GALNS:	N-acetyl galactosamine-6-sulfatase
HA:	heteroduplex analysis
HGP:	Human Genome Project
IPTG:	isopropyl-ß-thio-galactopyranoside
kb:	kilobase pairs
kD:	kilodalton

LOH:	loss of heterozygosity
Mb:	megabase pairs
MDE:	mutation detection enhancement
µg:	microgram
µl:	microlitre
mg:	milligram
ml:	millilitre
mRNA:	messenger ribonucleic acid
ORF:	open reading frame
PCR:	polymerase chain reaction
RFLP:	restriction fragment length polymorphism
RH:	radiation hybrid
RNA:	ribonucleic acid
RNase:	ribonuclease
RT-PCR:	reverse transcription polymerase chain reaction
SDS-PAGE:	sodium dodecyl sulphate polyacrylamide gel electrophoresis
SSCP:	single stranded conformation polymorphism
STSs:	sequence tagged sites
Tris:	tris (hydroxymethyl) aminomethane
TSG:	tumour suppressor gene
UTR:	untranslated region
VNTRs:	variable number of tandem repeats
YACs:	yeast artificial chromosomes
X-Gal:	gal-5-bromo-4-chloro-3-indoyl-B-D-galactopyranoside

SUMMARY

Physical Mapping of Human Chromosome 16

This thesis involved the construction of a detailed physical map of the distal band of the long arm of human chromosome 16, 16q24. Physical maps are composed of cloned DNA segments which allow chromosomes to be more amenable to detailed analysis. The 16q24 region has been demonstrated to possess a high gene density. These genes include the Fanconi anaemia group A gene (FAA) which has been localised to the 16q24.3 region by classic linkage analysis. Also, loss of heterozygosity (LOH) of the 16q24 region, further refined to 16q24.3-qter, has been demonstrated in sporadic breast tumours. This LOH indicates the presence of a tumour suppressor gene (TSG) in the region. The integration of this physical map with the chromosome 16 genetic map is of great importance as it can be used as a framework map which can benefit the positional cloning of candidate disease genes, including the TSG and FAA, mapped to this specific chromosomal region. During the course of this thesis, the International Fanconi Anaemia/Breast Cancer (FAB) consortium was established to expedite the positional cloning of the FAA and tumour suppressor genes localised to 16q24.3-qter.

The 16q24 region was deficient in markers, thus the first step toward constructing the physical map was the identification of cosmids from this chromosomal region. Cloned DNA segments, 35-45 kb in cosmid vectors, facilitate access to any chromosomal region for further analysis. An Alu PCR strategy was used to isolate DNA, as cosmids, in this region. This technique allows the specific PCR amplification of human DNA of unknown sequence, from complex mixtures of human and rodent DNA. It was applied to the isolation of human specific sequences directly from two human/rodent somatic cell hybrids, CY2 and CY18A, which contain the distal part of 16q as the only human chromosome 16 material. The Alu PCR amplified products from these hybrids were used as hybridisation probes for screening approximately 4000 clones of a gridded chromosome 16 specific cosmid library to enable the

identification of cloned DNAs from this predefined chromosomal region. Los Alamos National Laboratory (LANL) have assembled these cosmid clones into contigs, that is sets of clones which possess overlap, by repetitive sequence fingerprinting. They have been estimated to represent an 84% coverage of the euchromatin of chromosome 16. A total of 32 identified cosmids were confirmed to map to 16q24. An estimation of the amount of 16q24 represented by these cosmids, together with their contigs, is about 2 megabases.

The next step was the use of a selection of the cosmid clones localised to 16q24 to identify transcribed sequences encoded by these cosmids, using the approach of direct cDNA selection. In this study, the direct cDNA selection protocol was modified in an attempt to rapidly isolate longer cDNA fragments and to obtain greater transcribed sequence information. An enriched region specific cDNA library was generated from hybridisation of cDNA inserts to cloned genomic DNA localised to the region of interest. The PCR products from the second round of selection were used as hybridisation probes to screen 40,000 clones of a normalised infant brain cDNA library to identify homologous clones. Five cDNA clones were demonstrated to show specific homology to cosmids from which they were derived. Thus, the newly isolated cosmids were useful in the identification of transcribed sequences. These resources greatly benefit the positional cloning approach for the identification of disease genes assigned to this region of interest.

Subsequently, a collaborative effort led to the identification of additional cosmids mapping to the 16q24.3-qter region. Expressed sequences and microsatellite repeats already mapped to this region, were used as hybridisation probes for screening approximately 14,600 clones of a ten times coverage gridded chromosome 16 cosmid library. Cosmid walking was also performed by probing these filters with cosmid ends to extend the identified singleton cosmids and cosmid contigs in this region. These identified cosmids were assembled into contigs that extended over 650 kb of genomic DNA in the 16q24.3 region.

The cDNA clones localised to 16q24.3 are possible candidates for disease genes localised to this region, including the TSG and FAA. Thus, two novel cDNA clones, yc81e09 and yh09a04, were further characterised. This involved confirmation of their localisation to 16q24.3-qter using a panel of human/rodent somatic cell hybrids. Northern blots and reverse transcription polymerase chain reaction (RT-PCR) were used to determine the full-length sizes of the transcripts and expression patterns in various tissues. Clone yc81e09 demonstrated expression in peripheral blood lymphocytes (PBL), muscle, the frontal lobe and occipital lobe brain sections and a number of cell lines. This clone detected a transcript of approximately 3.7 kb in size. Clone yh09a04 was expressed in PBL, tonsil, muscle, the brain stem and frontal lobe brain sections and numerous cell lines. A transcript of approximately 2.5 kb was detected on a Northern blot. Subsequent Northern analysis from the FAB consortium detected transcripts of several sizes including 2, 3, 4.7, 5.5 and 7.5 kb, the most prominent of which was 4.7 kb in length.

Sequences of the two cDNA clones were obtained and compared to sequences in accessible nucleotide databases to identify overlapping sequences that may extend the sequence of the clones. The sequence of clone yh09a04 was extended with a homologous clone, yf14a03 but the sequence of clone yc81e09 was not extended. The remaining sequence of the 5.5 kb transcript, containing yh09a04, was finally identified by the FAB consortium. Additional sequence for yc81e09 has been obtained by the FAB consortium but more sequence is required to attain the full-length of this transcript. Homologies of two transcribed sequences to any known genes or protein motifs were also investigated to determine function, but no significant homologies were identified.

The identification of a candidate gene that may be responsible for a disease initiates a search for disease causing mutations, an essential step in the positional cloning of disease genes. Single stranded conformation polymorphism (SSCP) analysis of the two transcripts was conducted to determine if their sequence is altered in breast cancer patient samples displaying LOH at 16q24.3-qter, when compared to normal DNA sequence. The FAB consortium conducted SSCP analysis of the 5.5 kb transcript sequence in Fanconi anaemia (FA) patient samples. Three polymorphisms were identified for transcript yh09a04 and no differences were detected for yc81e09 in the segments that were investigated in the breast tumour samples. Four mutations were detected in the sequence of the 5.5 kb transcript, which included yh09a04/yf14a03, in FA samples. This transcript was therefore determined to be the FAA gene. Thus, the positional cloning strategy involving the physical mapping of cosmids, transcript identification and mutation analysis of candidate genes has been successful in the identification of the FAA gene. Additional work is required for the identification of the TSG.

The approach of direct cDNA selection was also applied to a second project involving the positional cloning of the MEF gene responsible for familial Mediterranean fever (FMF) localised to 16p13.3 by genetic linkage mapping. The international FMF consortium has constructed a YAC/cosmid contig encompassing the FMF candidate region. An enriched cDNA library was generated from hybridisation of cDNA inserts from a foetal brain cDNA library to seven of the cosmids localised to 16p13.3. Analysis of the cloned PCR products from one round of selection identified three transcripts which demonstrated homology to cosmids from which they were derived. These transcribed sequences contribute to the transcript map of the region and the FMF consortium is continuing in its efforts to identify the MEF gene by screening for disease specific mutations in the transcripts.

ACKNOWLEDGEMENTS

I would like to thank the Department of Cytogenetics and Molecular Genetics at the Women's and Children's Hospital for providing me with the opportunity to study for the Degree of Doctor of Philosophy. I would especially like to thank my supervisors Associate Professor David Callen, Professor Grant Sutherland and Professor Don Roberton for their help, advice and guidance during the course of the project.

Thanks to all my colleagues in the department who provided both a friendly working environment and much appreciated assistance with the project. In particular, thanks to Joanna Crawford and Scott Whitmore for their contributions to this project, as indicated in the text; Jozef Gecz for his helpful discussions and advice on various aspects of the project; Agi Gedeon, Ingrid Lensink, Lesley Ades, Andrew Donnelly, Oliva Handt, Gabriel Kremmidiotis, Hilary Phillips and Kathie Friend for their friendship throughout the years of my study. My most sincere thanks to my wonderful friend and colleague Bryone Kuss for all her help, support, encouragement and words of wisdom.

Thanks also to Professor Francis Collins and Dr. Danilo Tagle for the opportunity of working in their laboratory, particularly to Dan for teaching me the direct cDNA selection technique and for the ongoing discussions regarding the project.

Special thanks to my friend Katina Dimopoulos for her support and invaluable help with this thesis on numerous occasions, and to Kym Thalassoudis for all his help with the computing. Thanks to Konnie Tolios, Loula Tsitsis, Andrew Dimopoulos, Eleni Venetsanakos, Irene Kampouropoulos, and Cindy Wright for their friendship and encouragement throughout my studies.

Finally, to my family, George, Argiro and Nick Apostolou, I thank you for all for your love, support and for everything.

Two dejected assistants of Thomas Edison said:
"We've just completed our seven hundredth experiment and we still don't have the answer. We have failed."

"No, my friends, you haven't failed" replied Mr. Edison. "It's just that we know more about this subject than anyone else alive. And we're closer to finding the answer, because now we know seven hundred things not to do. Don't call it a mistake. Call it an education."

CHAPTER 1

Literature Review

1.1 INTRODUCTION 3

1.2 HUMAN GENOME PROJECT 6

1.2.1 Mapping the Human Genome 6

1.2.1.1 Genetic Mapping 7

1.2.1.2 Physical Mapping 10

1.3 SEARCHING FOR DISEASE GENES 20

1.3.1 Positional Cloning 20

1.3.2 Positional Candidate Approach 23

1.4 HUMAN CHROMOSOME 16 26

1.4.1 The Physical Map of Human Chromosome 16 26

1.4.2 The 16q24 Chromosomal Region 33

1.4.2.1 Gene-Rich Chromosomal Regions 33

1.5 THE GENETICS OF BREAST CANCER 36

1.5.1 Genetic Model for Tumour Progression 36

1.5.2 Breast Cancer 37

1.5.3 Tumour Suppressor Genes and Loss of Heterozygosity 38

1.5.4 Loss of Heterozygosity at 16q 41

1.5.5 Hereditary Breast Cancer 42

1.5.6 Identified Genes Involved in the Development of Breast Cancer 47

1.6 THE LONG ARM OF HUMAN CHROMOSOME 16 (16q) AND BREAST CANCER 50

1.6.1 Detailed Loss of Heterozygosity Studies of 16q in Breast Cancer 50

1.6.2 Correlation Between Clinicopathological Markers of Breast Cancer and LOH on 16q 55

1.6.3 Genetic Predisposition: Hereditary Breast Cancer and Chromosome 16 57

1.6.4 Candidate Tumour Suppressor Genes Localised to the 16q24 Chromosomal Region 58

1.7 DISEASES LOCALISED TO HUMAN CHROMOSOME 16 64

1.7.1 Fanconi Anaemia 64

1.7.2 Familial Mediterranean Fever 68

1.8 AIMS OF THE THESIS 72

1.1 INTRODUCTION

The Human Genome Project (HGP) is a co-ordinated international effort initiated in order to define the human genetic blueprint. The initial goals of the HGP include the construction of genetic, physical and transcript maps of the human genome for the identification and localisation of the estimated 50,000-100,000 genes thought to reside within the human genome, and ultimately to sequence the 3 billion base pairs of the human genome.

The genetic variation displayed in the human genome is a valuable resource for studies in molecular and medical genetics. The cloning and characterisation of genes underlying human inherited traits aid the elucidation of the molecular basis of physiological processes in higher organisms and contribute to the completion of the map of the human genome. The difficulty with cloning and characterising disease genes is that the vast majority of genes underlying a disease are known only by their phenotype, not their biochemical basis. If the biochemical basis of a disease is known, the gene can be cloned using the functional cloning approach which is based on information gained from the protein product. However, this approach cannot be extended to heritable traits and diseases where the biochemical basis is unknown. Therefore, the genes underlying these traits can only be identified using other approaches.

Since the inception of the HGP in 1990, significant progress has taken place in genetic and physical mapping which has led to the development of new technologies that have influenced new approaches for the identification of disease genes. These include the positional cloning approach, which initially requires the identification of the location of the disease gene in the genome through linkage analysis using genetic markers positioned on a genetic map. A physical map consisting of a contig of large genomic fragments across the relevant region is then constructed and can be used for the eventual identification of these genes. Once a candidate gene has been identified, its association with a heritable disorder requires additional evidence. This may include association of a gene mutation with the disease, or expression of the gene in an appropriate tissue and the location of an active protein. The study of the structure and function of a gene's protein product can help elucidate the pathophysiology of the disease. This protein product may also lead to the design of therapeutic drugs, or may be used for disease management or to improve the disease through gene therapy.

Both the scientific and medical communities can benefit from the information and materials arising from the HGP. The cloning and identification of disease genes may improve the management of a particular disorder, and the study of affected families with heritable disorders could provide valuable mapping information. The novel mechanism of trinucleotide repeat expansion (Bates and Lehrach, 1994), an example of the medical knowledge that has emerged from the HGP, is a mutation causing several genetic disorders, including fragile X mental retardation and myotonic dystrophy, and may also be responsible for other conditions such as ageing and cancer.

This thesis presents contributions toward the physical mapping of the 16q24 region of the long arm of human chromosome 16 (16q) and the positional cloning of two disease genes. These genes are the Fanconi anaemia group A gene localised to this region by genetic linkage mapping, and a tumour suppressor gene associated with a region of loss of heterozygosity at the 16q24 chromosomal region in sporadic breast tumours. Also, a gene responsible for familial Mediterranean fever (FMF) has been localised to 16p13.3 by genetic linkage mapping. The positional cloning strategy has been utilised for the construction of a cosmid contig spanning this candidate region. The work presented in this thesis contributes to the transcript map of this region which is being used for the identification of the FMF causative gene.

The literature review in this chapter is divided into two main sections. The first section outlines the goals of the Human Genome Project and focuses on the developments made toward the construction of the genetic, physical and transcript maps of the present day (1.2.1). The approach of positional cloning which has made a significant impact toward the ultimate HGP goal of identifying and isolating all genes in the human genome is also described (1.3.1). The second section describes the physical mapping of human chromosome 16 (1.4.1), particularly the properties and characteristics of the 16q24 chromosomal region (1.4.2 and 1.6). This chapter concludes with the main aims of this thesis (1.8).

1.2 HUMAN GENOME PROJECT

The Human Genome Project (HGP) which commenced in 1990, was organised in order to create detailed genetic and physical maps of the genome with the ultimate goal of obtaining the complete DNA sequence of the human genome. The goals of the 15 year project (Rossiter and Caskey, 1995) included:

- The construction of a high resolution genetic map of the human genome.
- The production of a variety of physical maps of all chromosomes with emphasis on maps that make DNA accessible to investigators for further analysis.
- The construction of detailed physical maps and genetic maps of selected organisms used exclusively in research laboratories as model systems.
- The identification and localisation of all human genes.
- The determination of the complete nucleotide sequence of the human genome.
- The development of new technologies necessary to achieve these goals.

1.2.1 Mapping the Human Genome

The human genome is the genetic material in the nucleus of a human cell which contains 3 x 10^9 base pairs (bp) of DNA (Bodmer, 1981) and is represented by a set of 46 chromosomes. Differences displayed in DNA sequences between individuals are the basis for genetic variation and aetiology of genetic disease. High resolution maps of all chromosomes are required for functional analysis of the genome as a whole, and for the identification and characterisation of its genes. Chromosome maps include genetic, physical and cytogenetic maps which are available at many different levels of resolution from chromosome bands to single base pairs. These maps can be integrated to produce a complete map of the human genome. By mapping the human genome, together with data on structural features of chromosomes, pseudogenes, repetitive sequences and regulatory elements, it should be possible to relate the structure of the genome to both its function and evolution.

1.2.1.1 Genetic Mapping

A genetic map specifies the relative order of genes and polymorphic marker loci along chromosomes. The more often two markers are inherited together in a family, the closer they are presumed to be in the genome. A genetic map constructed by linkage analysis may be utilised for the chromosomal assignment, regional localisation and positional cloning of new genes in the human genome. This mapping is dependent on the availability of polymorphic linkage markers on a high resolution map. In the 1970s, the lack of highly informative and evenly spaced markers made this type of mapping difficult, but the development of microsatellite markers along with large-scale semi-automated methods for marker isolation, typing and analysis (Gyapay *et al.* 1994; Buetow *et al.* 1994) some twenty years later, contributed to the rapid success of genetic mapping.

The first genetic maps were constructed using protein polymorphisms as genetic markers. These markers included cell surface antigens, blood group antigens, and isozymes (Race and Sanger, 1968; Harris *et al.* 1977). They required a diverse range of biochemical and immunological techniques for analysis and were relatively scarce when compared to the DNA markers which are presently available. Significant progress toward the generation of genetic maps of the human genome was made after the discovery of Restriction Fragment Length Polymorphisms.

In the 1980s, Restriction Fragment Length Polymorphisms (RFLPs) were used to define and follow transmission of loci in families with inherited disorders. RFLPs are single nucleotide changes affecting the presence or absence of a restriction enzyme site and consequently the size of the hybridising restriction fragment, which is detected by Southern analysis. In 1980, it was proposed that RFLPs could be used to construct a complete linkage map of the human genome (Botsein *et al.* 1980). By 1987, sufficient loci were identified to generate the first global human genetic map based on RFLPs (Donis-Keller *et al.* 1987). The average resolution between the markers was 10 cM and they covered approximately 95% of the human genome. RFLPs expedited the initial localisation of disease genes, for example Huntington disease (Gusella *et al.* 1983), cystic fibrosis (Tsui *et al.* 1985), and adult polycystic kidney disease (Reeders *et al.* 1985). Although RFLP markers were beneficial for linkage studies it was found that their polymorphic information content was often low and typing of large numbers of RFLP loci by Southern analysis was extremely labour intensive and used large amounts of DNA.

During the latter half of the 1980s, additional polymorphic markers were identified for use in linkage analysis. The advantage of these markers was that they were more polymorphic than RFLPs. They contained tandem repeats of short DNA sequences, variable number of tandem repeats (VNTRs) (Nakamura *et al.* 1987) or minisatellites (Jeffreys *et al.* 1985). VNTRs were located predominantly in the telomeric regions of chromosomes (Nakamura *et al.* 1988; Royle *et al.* 1988). Although such loci had the advantage of being multiallelic, thus were very informative, they were not randomly distributed throughout the genome, therefore left large regions of the genome unlinked to these loci. Linkage analysis with VNTRs was still based on Southern analysis as the fragment sizes were often too large for analysis by Polymerase Chain Reaction (PCR).

Further developments in molecular biology led to the isolation of a special type of marker, consisting of relatively short (less than 100 bp) tandemly repeated segments of DNA with repeat units of 2-6 bp (Beckmann and Weber, 1992). These markers, referred to as microsatellites, take advantage of the observation that the number of copies of short repeats can differ significantly among individuals at any one genomic site, ie. the microsatellite sequence can be highly polymorphic (Weber and May, 1989). The class of repeat most commonly detected is the (AC)n dinucleotide repeat. These highly polymorphic loci are abundant and ubiquitous throughout the genome. Genetic mapping with these markers is performed by using PCR and polyacrylamide gels (Weber and May, 1989). Each microsatellite is characterised by using PCR primers that anneal to single copy DNA flanking the repetitive element. This procedure permits high throughput typing of many samples and does not require cloning or Southern analysis. Microsatellites are also used as Sequence Tagged Sites (STSs) (see 1.2.1.2) in physical maps and provide points of alignment between the genetic and physical maps of human chromosomes.

Since their discovery, these highly polymorphic microsatellite markers have been critical for the construction of high resolution genetic maps and the localisation of disease genes. A milestone in human genetic mapping was published in 1992 by Weissenbach *et al*, who constructed a genetic map based on 813 (AC)n repeat PCR detectable markers that covered every chromosome (except the Y chromosome) (Weissenbach *et al.* 1992). Genetic linkage maps based on microsatellite markers at various resolutions were subsequently published for numerous human chromosomes which included chromosome 15 (Beckmann *et al.* 1993), chromosome 16 (Shen *et al.* 1994) and chromosome 21 (McInnis *et al.* 1993). These maps of chromosomes 15, 16 and 21 have an average interlocus resolution of 2.0, 3.2 and 2.5 cM respectively, and the highly informative markers on these maps were later used to construct a high resolution genetic map of the human genome.

In late 1994, the first of the major goals of the HGP to be reached was the generation of a genetic map of the human genome with markers spaced an average distance of 2-5 cM apart. More than 100 laboratories pooled their data to generate a comprehensive linkage map of the human genome (Murray *et al.* 1994). This map contains 5826 loci covering 4000 cM on a sex-averaged map, representing an average marker density of 0.7 cM. This map has limitations as only 908 of the markers are ordered with high confidence (odds ratio > 1000 : 1). These 908 markers constitute a framework map of about 4 cM resolution. Additional markers have been localised with respect to the framework map markers, with odds between 10 : 1 and 100 : 1. Subsequently, a genetic map composed entirely of microsatellite markers has been published (Dib *et al.* 1996). This map contains 5264 markers located to 2335 positions representing a marker density of 0.7 cM, or about one marker every 700 kb. This average marker density of 0.7 cM is well beyond the specified goal of 2-5 cM set in the HGP.

1.2.1.2 Physical Mapping

Physical maps of chromosomes are composed of cloned DNA segments which are the essential materials to commence large scale DNA sequencing of the human genome, the ultimate goal of the HGP. These maps specify the physical distances between markers on the chromosome, facilitate access to any chromosomal region and form the basis for the development of disease gene and transcript maps. Physical maps are necessary for exploring the functional significance of chromosome organisation and for increasing our understanding of many aspects of cellular and molecular biology.

Several technological developments from the HGP enabled the physical segmentation of the human genome. A range of vectors were developed for cloning DNA fragments of sizes ranging from approximately 20 kb up to 1 Mb, ie. plasmids, bacteriophage, cosmids, bacterial artificial chromosomes (BACs) and yeast artificial chromosomes (YACs). Construction of genome-wide physical maps was achieved when these genomic clones were ordered and arranged into contigs. Assembly of clones into contigs required sequence tagged sites (STSs) that acted as landmarks which overlapped clones along the physical map.

A number of physical mapping strategies with various levels of resolution were developed to determine the physical relationships between cloned DNA fragments. The physical map of lowest resolution is the cytogenetic map which is based on *in situ* hybridisation of cloned DNA fragments to metaphase chromosomes. Initially, this involved the use of tritium-labelled probes, then technically simpler fluorescently labelled probes were introduced to develop the Fluorescence *in situ* Hybridisation (FISH) technique (Landegent *et al.* 1987). Extended prophase chromosomes provide a resolution of about 3 Mb, and *in situ* hybridisation to interphase nuclei allows the ordering of cloned DNA segments with a resolution of several hundred kilobases (Trask, 1991). The resolution of *in situ* hybridisation for chromosome mapping was improved by using metaphase chromosomes with defined breakpoints, such as fragile sites or translocation breakpoints. The FISH technique also enabled the mapping of individual clones to chromosomes, the orientation of contigs along a chromosome, and determination of the chromosomal order of two probes as close together as 100 kb (Lawrence *et al.* 1990), but this was labour intensive.

Cytogenetic based physical maps were also constructed with human/rodent somatic cell hybrid panels. These panels are comprised of individual hybrids that contain segments of a particular human chromosome. Such hybrids can be derived from breakpoints ascertained in the human population and are beneficial for physical mapping as they contain a defined segment of a chromosome. Markers were localised to specific chromosomal regions by Southern hybridisation to DNA of the panel of somatic cell hybrids. The development of PCR allowed rapid screening of such panels of hybrids and required significantly lower amounts of DNA. Thus, the somatic cell hybrid panel is a valuable resource that allows the localisation of DNA fragments, cosmids, YACs, contigs, genes and STSs to individual chromosomes. A detailed cytogenetic based physical map of human chromosome 16 using a panel of somatic cell hybrids (Callen *et al.* 1995; Doggett *et al.* 1995) has been constructed and will be described in section 1.4.1.

Radiation hybrid (RH) mapping produces physical maps of increased resolution. Radiation hybrids are produced from the fragmentation of chromosomes in cultured cells with high doses of X-rays. These hybrid panels have been used to determine the physical location, orientation and long range order of contigs, DNA segments and genes (Green and Olson, 1990; Cox *et al.* 1990). Resolutions of less than one Mb can be achieved with these maps, thus they extend the resolution of genetic maps which were originally expected to be between 2-5 cM. They can be fully integrated with genetic maps to increase the overall density of ordered markers to refine physical maps. Physical distances between loci localised on RH maps can be determined using statistical procedures similar to genetic linkage mapping. The orientation and long range order of loci is based on the probability of cosegregation of markers on a chromosomal fragment contained in the hybrid. (Lawrence *et al.* 1991). Examples of RH maps which have been constructed include portions of chromosomes 5 (Warrington *et al.* 1991) and 16 (Ceccherini *et al.* 1992). Recently, radiation hybrid panels of the whole human genome were produced. One of the RH panels consists of 168 human/hamster cell lines each retaining about 32% of the human genome in random fragments of about 10 Mb (Gyapay *et al.* 1996). This panel was used to construct physical maps of 100 markers on chromosome 4 and 300 markers on chromosome 7 (review by Guyer and Collins, 1995).

The construction of contigs of overlapping cloned DNA fragments (Burke *et al.* 1987) provides even greater resolution for the physical map. A contig is an organised set of DNA clones that collectively cover a chromosomal region that is too long to clone in one piece (Olson *et al.* 1986). Contigs are overlapping cloned DNA fragments that are positioned relative to one another. Overlaps between the fragments can be detected using a fingerprinting technique which involves digestion of each clone with restriction enzymes and comparison of the lengths of the DNA fragments. In the late 1980s difficulties were encountered in the assembly of contigs of a large size for the construction of physical maps, as the size of the DNA fragments that could be cloned into vectors was a limiting factor. Cosmids and bacteriophages could accommodate DNA fragments up to approximately 40 kb. The development of a method which enabled cloning of large segments of DNA into Yeast Artificial Chromosome (YAC) vectors (Burke *et al.* 1987; Dawson *et al.* 1986) improved the prospect for construction of large scale physical maps of the human genome. The human genome comprises 3000 Mb and YACs revolutionised the study of the human genome, as DNA fragments of up to 1 Mb (Schlessinger, 1990; Chumakov *et al.* 1992) could be cloned in these vectors. The location of a YAC in the genome could be achieved either by FISH to G-banded chromosomes, or by hybridisation to a Southern blot containing DNA from a panel of somatic cell hybrids. Pulsed field gel electrophoresis (PFGE) (Smith *et al.* 1988) was introduced to construct long range physical maps of the human genome. This technique allowed the separation and analysis of high molecular weight human DNA fragments, up to 10 kb, and the construction of YAC contigs. PFGE was useful in generating restriction maps of chromosomal regions ranging in size from 250 kb to 5 Mb, using rare cutting restriction enzymes, eg. NotI, which preferentially cleaved within CpG islands (Brown and Bird, 1986).

The STS (Olson *et al.* 1989) was developed as a new type of marker for the construction of physical maps. This is a short (200-300 bp) unique DNA sequence that can be PCR amplified and can be used as a landmark during physical map construction, as it is easily detected by PCR assays. STSs alleviated the problems associated with cloned markers which were used as probes in physical mapping. Complex physical maps of the human genome based on restriction enzyme sites required the maintenance of tens of thousands of individual clones in bacterial strains that were used as probes to detect particular segments of DNA in Southern analysis. In contrast, sequences for STSs had advantages over the cloned markers which were previously utilised, as they could be made electronically available, and were easily distributed in the scientific community. The implementation of the STS only required the synthesis of the primers and the determination of appropriate PCR conditions. They could be used to carry out a large number of assays rapidly and in an automated fashion.

STSs have been used to identify pairs of overlapping genomic clones on the basis of their shared STS to obtain contiguous regions of several Mb. Contig closure to bridge the gaps between existing contigs could be achieved by a strategy involving the isolation of the ends of existing contigs for hybridisation to identify new clones that may link the contigs. The polymorphic microsatellite markers used for genetic mapping have an important role as STSs as they can be used to integrate the physical and genetic maps. Genetic markers flanking a disease gene also define the physical interval that contains the gene. In addition, the integration of contigs and genetic markers is the bridge between genetic mapping studies and disease gene identification.

Following these developments, multi-megabase contigs were constructed (Green and Olson, 1990; Anand *et al.* 1991; Silverman *et al.* 1991). STSs were used to identify overlapping YAC clones on the basis of their shared STS to obtain contiguous regions of several Mb. In 1992, a low resolution complete YAC based physical maps of human chromosome 21 (Chumakov *et al.* 1992b) and human Y chromosome (Foote *et al.* 1992) using STSs were published. In late 1992, the Centre D'Etude du Polymorphisme Humain (CEPH) MegaYACs were developed to further improve the construction of large scale physical maps. DNA fragments up to 1.4 million bases could be cloned in these vectors. These MegaYACs were used to construct the first generation map of the human genome (Cohen *et al.* 1993).

The early physical maps which were constructed using overlapping YACs were not entirely reliable, as errors were made in the ordering of clones. This was a result of the YAC libraries containing rearrangements relative to the genome from which the clones were derived (Trask *et al.* 1992). Approximately 1% of YACs containing human DNA were unstable which resulted in deletion derivatives, while 10% of yeast clones carried cotransformed YACs. Most total human genome YAC libraries contained 40-50% chimaeric clones (Bronson *et al.* 1991). Chromosome specific YAC libraries constructed from genomic DNA of somatic cell hybrids appeared to consist of fewer chimaeric clones (5-15%) (McCormick *et al.* 1993). Errors in the map were resolved by increasing the density of markers on the physical map, FISH analysis of the YACs and comparison to the results from genetic mapping efforts.

Recently, a variety of new vectors for the cloning of large sized DNA fragments, Bacterial Artificial Chromosomes (BACs) were developed. These included the bacterial-P1-derived clones, (Pierce *et al.* 1992), F factor-derived BACs (Shizuya *et al.* 1992) and a variation of the P1-derived clones termed PACs (Ioannou *et al.* 1994). BAC libraries consist of clones with an average size of 90-120 kb. Because of their large insert size, YACs are the system of choice for long range physical mapping. However, YACs are not ideal for fine structure analysis due to the rearrangements present in the clones. Low yields of pure human DNA from YACs also limits techniques which require large amounts of DNA, such as subcloning in the production of cosmid libraries, or the introduction of YACs into mammalian cells. In addition, YACs are not the ideal source of sequencing templates due to the lack of pure human DNA and the presence of chimaeric clones. Thus, BACs appear to be the most likely source for contig construction of chromosomal regions of interest and ultimately as template for large scale DNA sequencing. The frequency of chimaeric clones and rearrangements in BACs is significantly lower than in YACs. An additional feature of BACs is their low vector to insert ratio compared to cosmids, which is advantageous for sequencing the cloned DNA. Furthermore, regions which were unclonable in YACs, such as telomeres of chromosomes, can be cloned in BACs.

The great majority of STSs described thus far were derived from anonymous genomic DNA sequences which were successfully used for genome-wide and chromosome specific mapping. In the early days of the HGP, STS gene-based mapping was limited as there were not many human gene sequences available. This situation changed with large scale DNA sequencing projects and the emergence of Expressed Sequence Tags (ESTs) (Adams *et al.* 1991) which were established as sequences of cDNA clones (cloned DNA copies of mRNA molecules that direct protein synthesis) and were mapped to chromosomes (Polymeropoulos *et al.* 1992). ESTs serve the same features as STSs but also act as guideposts for mapping genes along chromosomes or as reference sequences for identifying genes along genomic DNA, which can help in the structural analysis of the human genome. Furthermore, the study of cDNAs is advantageous as it can lead to the isolation of full length sequences of cDNA and the eventual prediction of gene products and the functional analysis of the human genome.

Wilcox *et al* (1991) developed ESTs from the 3' untranslated region (UTR) of messenger RNAs (mRNAs) which was advantageous as this part of the mRNA sequence rarely contains introns, compared with the rest of the mRNA sequence. Therefore, a PCR product of the same size was produced for both genomic DNA and cDNA. In addition, sequences in the 3' UTR were useful as they are not as well conserved as those in the coding region making it easier to distinguish between individual genes and members of gene families that may be closely related in their coding sequences.

The ultimate goal of the HGP is to determine the nucleotide sequence of the human genome which will lead to the identification of all genes and their controlling regions. A major development which has improved transcript mapping and gene identification is the establishment of online databases, for example GenBank (Burks *et al.* 1985; Bilofsky and Burks, 1988) which includes dbEST and dbSTS, to permit the search and retrieval of sequences on the Internet. Also, the large scale sequencing projects and the subsequent entry of the ESTs in the dbEST database have improved the transcript map. In 1995, more than 125,000 human derived ESTs were located in dbEST. The cDNA sequencing initiative at Washington University, USA was expected to produce 5' and 3' sequences for 200,000 random cDNAs by 1997, which would extend the 65,000 human ESTs placed in the public arena by other groups.

As more gene-based markers accrue, homology information and links to the scientific literature greatly assists the value of these markers as they provide the EST framework for the sequence ready physical map. This data and the retrieval systems are invaluable for the isolation and identification of disease genes. The Basic Local Alignment Search Tool (BLAST) (Altschul *et al.* 1990) is a rapid database searching algorithm that searches for local areas of similarity between two sequences and then extends the alignments on the basis of defined match and mismatch criteria. Nucleotide sequences in the database demonstrating homology to a nucleotide sequence of interest can be identified using the BLAST-N program. Nucleotide sequences can be translated into amino acid sequences and database searches for homologous coding regions with BLAST-X can be performed, and the GRAIL program

(Uberbacher and Mural, 1991) can be used to identify coding sequences from genomic DNA sequence.

One of the major goals of the HGP is to produce a physical map containing 30,000 unique markers ordered with respect to each other and spaced on average every 100 kb. Various human genome-wide STS-based maps were completed recently. Chumakov *et al* (1995) constructed a YAC contig map covering approximately 75% of the human genome. This map comprised 225 contigs with an average size of 10 Mb, which were positioned with about 2600 STSs. A map based on 15,000 STSs at one Mb resolution from Hudson *et al* (1995) represents great progress in reaching the physical mapping goal. This map includes 7,000 STSs previously generated and mapped on meiotic linkage maps by other groups (Gyapay *et al.* 1994; Dib *et al.* 1996). An additional 3,000 STSs were developed from cDNA sequences in GenBank and ESTs from dbEST. Therefore, the locations of numerous human genes on physical maps have been integrated with linkage maps of the human genome. The remaining 5,000 STSs were developed from random genomic DNA sequencing.

Hudson *et al* (1995) also used the RH map as a tool for the independent estimation of DNA marker order. At the beginning of 1996, 6,000 STSs were placed on the RH map but only about 1,300 could be ordered with odds greater than 300 : 1. These 1,300 markers divided the RH map into "bins" of 2.5 Mb. The combination of the RH map and YAC-STS content map gave an estimated average resolution of about one Mb (Hudson *et al.* 1996). Gyapay *et al* (1996) also developed a whole genome radiation hybrid panel of 168 human/hamster cell lines. A framework map that spans all the autosomes and the X chromosomes was constructed using about 400 microsatellites of known genetic location. Approximately 370 ESTs mapped to chromosomes 1, 2, 14 and 16 were also localised on this map.

These maps provide frameworks for positioning additional markers. The current genetic and physical maps span essentially the whole human genome. Hudson *et al* (1996) estimated 94% physical coverage of the genome, whereas the YAC clone based physical maps

(Chumakov *et al.* 1995) provided about 75% coverage. However, few genes were localised to these framework maps. The number was limited to 3235 ESTs localised by Hudson *et al* (1995) and 318 cDNAs mapped to the CEPH megaYAC panel by Berry *et al* (1995). This work was significant as it represented a milestone and a link between the limited EST mapping of the past and the thousands of gene-based markers that would appear on maps in the future.

Nevertheless, these maps still fell short of the HGP goal of 30,000 precisely ordered STSs. Also, they did not provide an adequate scaffold for sequencing of the human genome. Further work was required to generate an additional 15,000 STSs and to order them. The HGP goal of STSs spaced at an average interval of 100 kb is expected to be achieved by 1998. To attain this goal, an EST mapping consortium was formed to map greater than 50,000 ESTs to 0.5 Mb intervals (perhaps 0.1 Mb intervals), utilising two radiation hybrid panels representing the whole genome, and the CEPH megaYAC panel. Gene-based STSs were mapped against one or more of these panels then localised relative to the common framework (Schuler *et al.* 1996). A total of 16,354 distinct loci were localised to these panels. A total of 15,284 markers were mapped to the RH panel only. Approximately 5000 human genes were mapped at the time this work was begun, thus the consortium increased the number of mapped human genes on the physical map by greater than three times. The effort is now being extended to localise the majority of human genes but this will not be complete until the entire sequence of the genome is obtained. The advantage of the increased number of ESTs mapped to specific chromosomal regions is that they may act as potential candidate genes for disease loci mapping to specific chromosomal regions. This information is critical for the isolation of disease genes using the positional candidate approach (see 1.3.2).

At present, the available physical maps are not optimal for large scale sequencing. Current maps of most chromosomes have a resolution of 100-200 kb and are based largely on YAC clones. This resolution is insufficient for obtaining desired clone coverage in any of the preferred BAC systems. Additionally there is a shortage of well characterised, highly redundant, large insert BAC libraries that are required initially to construct the maps. Thus, high resolution maps need to be constructed to begin genome sequencing.

The sequencing of cDNA clones is of use in the identification of new genes. However, it is important to characterise genomic DNA as it contains critical DNA sequences not included in mRNA sequences that regulate gene expression. Sequencing genomes of model organisms is also important as it provides information on the structure, function and evolution of the human genome. Comparative mapping provides information on the extent of synteny between different mammalian species and reflects the conservation and the diversification of the genome. Data from comparative gene mapping is also important in studies of human disease as genetic linkage analysis may locate human disease loci to a small syntenic region in the animal genome which can be investigated to determine whether any gene mapping in the syntenic region of the model organism is a possible candidate. It may also provide a better understanding of the principles of chromosomal evolution in mammals.

In summary, the genetic mapping goal of the HGP has been attained. The goal was to construct a map with a resolution of 2-5 cM, however, a genetic map with greater resolution of 0.7 cM (or 1 marker every 700 kb) was constructed (Dib *et al.* 1996). The physical mapping goal of 30,000 precisely ordered STSs (or 1 marker every 100 kb) is rapidly being approached with mapping efforts from Hudson *et al* (1995) who localised 15,000 STSs and Gyapay *et al* (1996) who mapped 16,354 distinct gene-based STSs.

1.3 SEARCHING FOR DISEASE GENES

The identification of the estimated 3% of the human genome that is expressed (Nowak, 1994) presents a difficult challenge. Various methods can be used for the identification of disease genes. Initially, the functional cloning approach was used to identify a gene causing a biochemical defect in a human disease, based on information from the protein product without reference to the chromosomal map position. Either the protein itself was isolated from normal tissue and used to identify the corresponding cDNA (via amino acid sequence or antibodies) or, an assay for protein activity was used to screen the products from the cDNA clones. The isolated cDNA was localised using somatic cell hybrids or by FISH. These were laborious procedures and only feasible in a small number of disorders, as knowledge of the protein products of most disease genes was either minimal or absent altogether. An example of functional cloning includes the isolation of the gene underlying phenylketonuria (Kwok *et al.* 1985).

Thus, it was necessary to develop new approaches for the identification of disease genes. In recent years DNA based techniques, the positional cloning and the positional candidate approaches, were developed for this purpose. Both strategies involve the utilisation of the resources made available from the HGP. Thus, the effectiveness of the advances made in the HGP could be tested by determining the extent to which these new technologies could be used in the identification of genes associated with diseases.

1.3.1 Positional Cloning

Positional cloning (originally called reverse genetics) was developed in the 1980s, and is the approach by which the identification of a gene is based on the map position through genetic analysis (Collins, 1992) and assumes no functional information. It is a multistep process which begins with the localisation of a disease gene to a particular region following genetic linkage analysis in a collection of families with multiple affected members. This preliminary localisation is followed by molecular analysis of the region, including finer genetic mapping, physical mapping, isolation of DNA, identification of transcripts, cDNA cloning and mutation analysis of candidate genes.

Instead of linkage analysis, the localisation of a disease gene to a particular region can be achieved with DNA from patients that possess visible cytogenetic structural abnormalities, such as deletions or translocations, which aid the low and high resolution mapping of the responsible gene. A large number (26/42) of positional cloning successes have depended on such structural abnormalities (reviewed by Collins, 1995). The first success in positional cloning of a disease gene was reported for the X-linked gene for chronic granulomatous disease (Royer-Pokora *et al.* 1986). Genes including dystrophin (Monaco *et al.* 1986) and the MYC oncogene (Taub *et al.* 1982) have been cloned by using data derived from cytogenetics. Many genes with somatically acquired mutations, most of which have a role in cancer, have also been identified using the positional cloning approach. In lymphomas and leukaemias, this has been achieved by cloning the breakpoints of the observed cytogenetic rearrangements and the subsequent identification of transcripts in the region (Stanbridge, 1992). Many genetic mutations have also been identified in solid tumours.

The expanded triplet repeat (Bates and Lehrach, 1994) is a special type of DNA rearrangement which has aided the localisation of disease genes and was shown to represent the mutational basis of various diseases. These genes included those for fragile X syndrome (Kremer *et al.* 1991) and myotonic dystrophy (Buxton *et al.* 1992; Aslanidis *et al.* 1992). Southern blot analysis or PCR amplification of cloned genomic DNA can be used to detect expanding trinucleotide repeat mutations.

A number of approaches have been used for the identification of genes from cloned genomic DNA. Overlapping YAC or cosmid clones isolated from the region of interest are used as substrates for the identification of transcribed sequences localised in the region. The strategies include the search for evolutionary conservation of DNA fragments by Southern hybridisation to genomic DNA isolated from different species on "zoo blots" (Monaco *et al.* 1986; Riordan *et al.* 1989). Rare cutting restriction enzymes which cleave preferentially in CpG islands (which recognise sites with one or more CpG dinucleotides) have also been used for gene identification. CpG residues are used as sign posts for the 5' ends of constitutively expressed genes (Bird, 1986; Larsen *et al.* 1992) since a large fraction (as high as 60%) of gene promoters are associated with high concentrations of unmethylated CpG dinucleotides, ie. CpG islands. Genomic DNA fragments and CpG island containing fragments can be used to directly probe Northern blots of RNA isolated from foetal and adult human tissues to determine whether the genomic fragment contains expressed sequence. If expression is detected, the genomic fragment can be hybridised directly to a cDNA library generated from the appropriate tissue. These strategies are feasible for a small number of genomic clones but are not feasible for the investigation of large portions of the chromosome as they are labour intensive.

Thus, it has been necessary to develop new and different approaches to rapidly clone genes and transcribed sequences. New transcripts can be isolated and localised to a specific chromosomal region using new techniques which include direct cDNA selection (Lovett *et al.* 1991; Tagle *et al.* 1993) and exon trapping (Duyk *et al.* 1990; Buckler *et al.* 1991). The direct cDNA selection approach is used to generate an enriched region-specific cDNA library from the hybridisation of cDNA inserts to cloned genomic DNA localised to the region of interest. The enriched cDNA fragments are subsequently cloned for further characterisation. In the exon trapping procedure, genomic DNA is cloned into a vector that contains functional sequences for RNA splicing. If exons with complementary splice signals are contained within the genomic DNA, they can be spliced into mature RNA after expression in mammalian cells *in vitro* and subsequently characterised. Techniques which involve the identification of genes from cloned genomic DNA including exon trapping and cDNA direct selection are described and discussed in more detail in chapter 4, sections 4.1.1 and 4.1.1.1.

The increasing success of the positional cloning approach for the isolation of disease genes is due to the expanding resources provided by the HGP. The generation of the HGP high resolution genetic map of the human genome (Murray *et al.* 1994; Dib *et al.* 1996) has enhanced the localisation of disease genes by genetic mapping studies. The identification of genes within candidate regions has been assisted by use of the physical maps of the human genome which have greater resolution than genetic maps. These maps include the genome-wide STS-based maps constructed from overlapping YAC clones (Chumakov *et al.* 1995), the radiation hybrid map (Gyapay *et al.* 1996), the integrated linkage and physical maps (Hudson *et al.* 1995) and the gene-based STSs mapped against RH and YAC panels (Schuler *et al.* 1996).

The list of genes isolated by positional cloning has grown rapidly since the inception of the HGP. In 1992, 13 inherited human disease genes, including those from cystic fibrosis (Riordan *et al.* 1989) and Wilms' tumour (Rose *et al.* 1990), were positionally cloned (Collins, 1992). In 1993, an additional 7 genes, which included the gene for Menkes disease (Vulpe *et al.* 1993), were cloned (Ballabio, 1993). By April, 1995, a total of 42 genes had been isolated (Collins, 1995). Genes isolated in this period included those for Huntington disease (The Huntington Disease Research Collaborative, 1993), early-onset breast/ovarian cancer (Miki *et al.* 1994) and spinal muscular atrophy (Lefebvre *et al.* 1995).

1.3.2 Positional Candidate Approach

The success of positional cloning of greater than 40 disease genes (Collins, 1995) confirms the progress in genetic mapping, physical mapping and gene isolation technologies arising from the HGP. However, the positional cloning approach is being superseded by a more streamlined and rapid approach called the positional candidate approach.

Genes isolated by positional cloning have generally been associated with relatively common diseases which are subject to ongoing research and have an availability of large pedigrees and samples for genetic mapping. With these pedigrees, it is possible to localise a disease gene to a region less than one Mb, based on the resolution of recently constructed genetic maps (Dib *et al.* 1996). Otherwise, the isolated genes have been associated with visible chromosomal rearrangements in patient DNA which allow precise localisation of the disease locus. Most genetic diseases do not fall into these categories as there are insufficient families for a precise genetic localisation, or there are no known patients with detectable chromosomal rearrangements. Usually, the size of the region to which a gene is assigned is less than 2 Mb, and the isolation and examination of all genes localised to that region is required. This procedure is expensive, laborious and not productive in many cases.

The positional candidate approach does not require the isolation of all genes in a region of interest but relies on information from previously isolated genes and transcribed sequences mapped to the region, such as the gene product and/or the function of the gene. The strategy involves a combination of mapping to the correct chromosomal subregion, generally by linkage analysis, followed by examination of the interval to determine whether suitable candidate genes are localised to the region of interest. Subsequently, mutations of that gene are investigated in affected individuals to determine whether the candidate gene is the disease gene. Marfan syndrome, for example, was mapped to chromosome 15q21.1 by genetic linkage analysis (Kainulainen *et al.* 1990) and a short time later the fibrillin gene (FBN1) was mapped to 15q21.1 (Magenis *et al.* 1991). Mutations were then identified in the fibrillin cDNA from patients with Marfan syndrome (Dietz *et al.* 1991). Other genes cloned using this positional candidate approach include those for Alzheimer's disease (Sherrington *et al.* 1995) and hereditary non-polyposis colon cancer (Fishel *et al.* 1993).

The success of the positional candidate approach will grow with the high resolution genetic map and an increasingly dense transcript map. The databases of transcribed sequences, eg. dbEST, may assist the pure candidate gene approach, but these sequences are not useful for the positional candidate approach without the associated mapping information. Since genetic mapping efforts usually result in candidate intervals of 0.5-5 Mb, the transcript map should also have this degree of resolution. Mapping of cDNAs to somatic cell hybrids or by FISH will not usually achieve this. The utilisation of YAC libraries or RHs is more appropriate. Thus, the recent mapping of gene-based STSs to RH and YAC panels by Schuler *et al* (1996) will be beneficial in increasing the success rate of the positional candidate approach.

The major challenge ahead for positional cloning and the positional candidate approaches will be the elucidation of genes responsible for predisposition to common polygenic disorders such as diabetes, asthma, hypertension, many forms of cancer and psychiatric disorders. Most diseases have a genetic component, and the determination of these genetic factors is important for modern medicine. The recently constructed high resolution physical maps, efficient isolation of transcribed sequences and an increasing number of ESTs in databases may enhance the identification and isolation of the genes for the polygenic diseases. Thus, gene identification is important as it is the main medical justification for the HGP, and it can allow the development of diagnostic and therapeutic advances of great potential medical benefit.

1.4 HUMAN CHROMOSOME 16

1.4.1 The Physical Map of Human Chromosome 16

Human chromosome 16 represents 3% of the human genome (Morton, 1991) and has been estimated to contain about 98 Mb of DNA. If the human genome contains 50,000-100,000 evenly distributed genes, it is expected that chromosome 16 would contain between 1,500-3,000 genes. This chromosome has been targeted for intensive genetic and physical mapping over the past ten years. YACs provide the most efficient means of ordering genomic fragments, however, the construction of the ordered clone physical map of chromosome 16 commenced before the development of YACs. In 1988, Los Alamos National Laboratory (LANL) proposed to use a library of cosmid clones to construct a map of overlapping cloned fragments spanning the entire length of chromosome 16. Longmire *et al* (1991) constructed a chromosome 16 specific cosmid library from flow sorted chromosome 16. The DNA was partially digested resulting in overlapping fragments that were beneficial for constructing physical maps of ordered overlapping clones.

Chromosome 16 was studied for a number of reasons. These included the availability of a hybrid cell line containing a single copy of human chromosome 16 which permitted accurate sorting of this chromosome from the mouse background and the production of pure chromosome 16 clones for use in map construction. A chromosome 16 satellite repeat sequence was also available for use as a probe to determine the purity of the sorted chromosomes. Furthermore, a panel of somatic cell hybrids containing portions of chromosome 16 was available for the localisation of probes into intervals along chromosome 16. This chromosome also contained gene loci for several human diseases including polycystic kidney disease (Reeders *et al.* 1985) and several types of cancer including leukaemia and breast cancer. The construction of a physical map comprising overlapping cosmid clones derived from chromosome 16 could facilitate the positional cloning of these genes.

The initial strategy used to construct a contig map for chromosome 16 was the fingerprinting of cosmid clones and the determination of the overlaps between the pairs of clones. Chromosome 16 cosmid contigs were assembled by repetitive sequence fingerprinting (Stallings *et al.* 1992a) of each cosmid clone, a modification of the fingerprinting technique. The repetitive sequence fingerprinting method relied on the interspersed repetitive sequences found in the genome. Each cosmid was fingerprinted using the restriction enzymes EcoRI and HindIII, individually and in double digests, then the digested DNA was hybridised using (GT)n repeat probes (Stallings *et al.* 1990; Stallings *et al.* 1992a). The (GT)n sequence was demonstrated to be scattered across most regions of the genome with an average spacing of 30 kb. Thus, it was assumed that one (GT)n sequence would be contained in each cosmid clone containing a 35 kb insert. Overlap between cosmid clones was based on the restriction fragment sizes and the distribution of the repetitive sequences within these fragments. A statistical analysis was used to determine the pairs of overlapping clones. Approximately 4000 cosmids were characterised and 576 cosmid contigs with an average size of 100 kb in length with 10-20% overlap were generated using this procedure. These contigs spanned 58% of chromosome 16. An additional 1171 single fingerprinted clones not contained within a contig (singletons) covered about 26% of the chromosome. In total, these fingerprinted cosmids represented 84% of the euchromatin of chromosome 16.

Although a substantial part of chromosome 16 was covered by fingerprinted cosmids, continuation of this process was no longer an efficient means to generate coverage of the whole chromosome. By this time, a YAC library from flow sorted chromosome 21 (McCormick *et al.* 1993) was constructed and STS markers had been developed. McCormick *et al* (1993) constructed a chromosome 16 YAC library with an average insert size of approximately 200 kb from flow sorted chromosome 16, for use in linking cosmid contigs along the length of the chromosome. STS markers were also developed from end clones of the cosmid contigs. These STSs were used in PCR to identify YAC clones that overlapped with the cosmid from which the STS was derived.

A human/rodent somatic cell hybrid panel was also generated for use in the physical mapping of chromosome 16 (Callen, 1986; Callen *et al.* 1988; Callen *et al.* 1989; Callen *et al.* 1990; Callen *et al.* 1992; Callen *et al.* 1995). In general, each hybrid contains an increasingly longer portion of chromosome 16 starting from the tip of the long arm of the chromosome. These hybrids are constructed by fusing mouse cells with human cells and growing them in a medium in which only the cells containing the adenine phosphoribosyltransferase (APRT) selectable marker, localised to 16q24.3, can survive. The hybrids are derived from human chromosomes with constitutional translocations and deletions that are ascertained during clinical cytogenetic studies.

The extension of the human/rodent somatic cell hybrid panel has resulted in the construction of cytogenetic based physical maps of human chromosome 16 with increasing resolution (Callen *et al.* 1988; Callen *et al.* 1989; Chen *et al.* 1991; Callen *et al.* 1992; Callen *et al.* 1995; Doggett *et al.* 1995). Initially, the relative order of the hybrid breakpoints was determined by *in situ* hybridisation to parental human cell lines and mapping gene probes and anonymous DNA fragments to a small panel of hybrids. In 1986, the first chromosome 16 hybrid cell lines were constructed. These comprised five cytogenetically defined hybrid cell lines containing portions of chromosome 16 (Callen, 1986). These hybrids allowed the localisation of DNA markers to one of six regions of chromosome 16. In 1988, an additional somatic cell hybrid was constructed (Callen *et al.* 1988). Eight DNA fragments were localised to human chromosome 16 by *in situ* hybridisation to metaphase chromosomes of the parental cell lines and/or Southern analysis using the somatic cell hybrids. By 1990 the somatic cell hybrid panel was extended to a total of 31 hybrids (Callen *et al.* 1990). Chromosome 16 was divided into 31 regions, together with the four chromosome 16 fragile sites, following the localisation of 45 probes to a panel of hybrids derived from human cell lines with breakpoints on the long arm of chromosome 16 (Chen *et al.* 1991).

Considerable progress in the cytogenetic based physical map of chromosome 16 was made in 1992 (Callen *et al.* 1992). The localisation of 235 DNA markers to a panel of 54 somatic cell hybrids containing various portions of chromosome 16 together with the four fragile sites and the centromere allowed the delineation of 52 intervals on chromosome 16. These markers, some of which were in STS format, included genes, cosmids and anonymous DNA probes. This resulted in a physical map with an average resolution of 1.6 Mb. Chromosome 16 cosmid contigs mapped to the panel represented about 11% of the euchromatin of chromosome 16. A total of 66 polymorphic markers including both RFLPs and (AC)n microsatellite repeats typed in the CEPH pedigrees were also included in this map. The physical location of these markers permitted a detailed correlation of the cytogenetic based physical map with the genetic map of chromosome 16.

Integration of the chromosome 16 genetic map with the transcript map, utilising the somatic cell hybrid panel as a common framework, allows easy and accurate access from the genetic map to candidate genes and their transcripts, thus facilitates the positional cloning of disease genes. In 1995, 141 genetic markers, 76 genes and 124 transcribed sequences were integrated into the physical map of chromosome 16 using the high resolution somatic cell hybrid panel (Callen *et al.* 1995). The map consisted of 93 somatic cell hybrids and four fragile sites that defined 82 intervals at an average resolution of 1.2 Mb. The integrated map is beneficial as it allows rapid selection of flanking microsatellite markers for detailed genetic localisation of a disease gene. In this map, the physical location of the genetic markers is known, thus the physical interval containing the disease gene is defined by somatic cell hybrid breakpoints, which is advantageous for the positional cloning of disease genes to chromosome 16. The genes and transcripts in each region then provide potential candidate genes for the mapped disease.

The most recent integrated physical, genetic and cytogenetic map of chromosome 16 was published in September, 1995 (Doggett *et al.* 1995). The high resolution somatic cell hybrid panel is the framework for this map. This map is comprised of 78 hybrid cell lines, which with the centromere and four chromosome 16 fragile sites divides chromosome 16 into 90 independently ascertained breakpoints in the euchromatin giving the map an average resolution of 1.1 Mb. A section of this physical map is shown in figure 1.1.

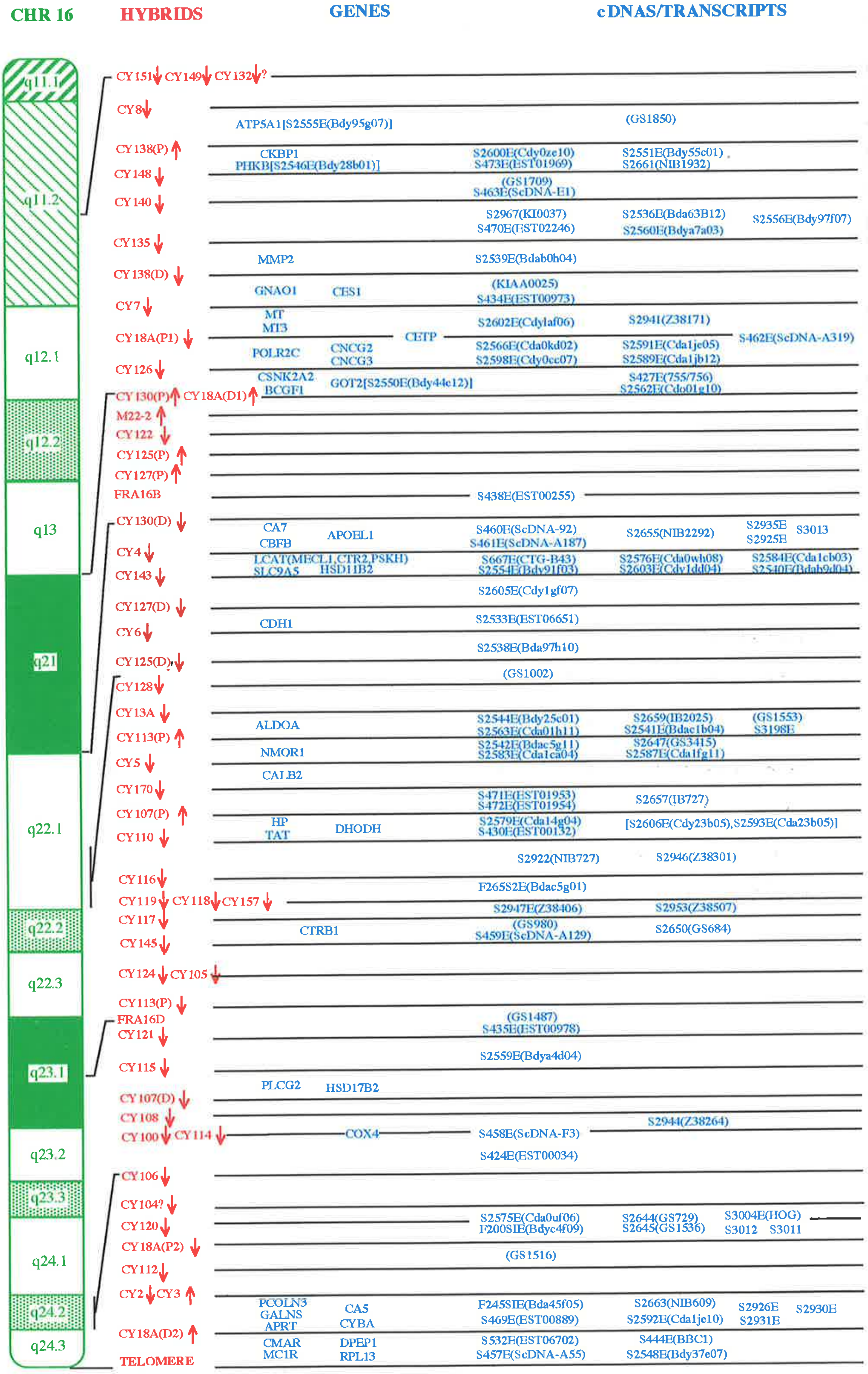
This physical map consists of a low resolution map and a high resolution map. The low resolution megaYAC map comprises 638 CEPH megaYACs that have been localised to and ordered within the somatic cell hybrid breakpoints with 418 STSs. The megaYAC map provides almost complete coverage of the euchromatin of chromosome 16. The exception is a 2.5 Mb portion of the 16p13.3 region of chromosome 16 which is not well represented in YACs. However, cosmid contigs and chromosome 16 specific YACs cover 1.14 Mb of this gap. The high resolution cosmid contig/miniYAC physical map consists of 320 contigs, 2,000 fingerprinted cosmids and 248 miniYACs, which are localised on the breakpoint map and integrated with the megaYAC map by clone hybridisation and STS screening. The cosmids in contigs cover more than 35% of the map. These sequence ready cosmids are beneficial for the large scale sequencing of the chromosome. The high resolution map covers more than 50% of the euchromatin.

The chromosome 16 cosmid contigs were generated by enriching for cosmids with (GT)n repetitive sequences which then formed the basis of the repetitive sequencing fingerprinting procedure (Stallings *et al.* 1990). This cosmid library does not contain randomly distributed cosmids as the (GT)n repeats are known to be deficient in gene-rich regions. Thus, it is likely that the euchromatin of chromosome 16 will be under represented by these fingerprinted cosmids. Furthermore, all regions of a chromosome cannot be cloned into vectors, thus a high resolution sequence ready map covering the whole of chromosome 16 may not be accomplished using these cosmid clones.

It is likely that the physical map contains errors which may be due to chimaeric clones, deleted clones, chromosome duplications and the presence of low abundance repeat sequences. Deletion derivatives and chimaeric clones have been identified in YAC libraries.

Figure 1.1

Ideogram of chromosome 16q showing an example of the physical mapping efforts for human chromosome 16 (Callen *et al.* 1995; Doggett *et al.* 1995). The regional localisation of genes, cDNA clones and transcripts is featured in this map. The breakpoints from hybrid cell lines containing different segments of chromosome 16 are represented by the horizontal lines. An arrowhead (pointing up or down) indicates which part of chromosome 16 is retained in each hybrid. The ideogram shows the precise order of breakpoints on chromosome 16, although the spacing between the breakpoints is arbitrary.


CHR 16
HYBRIDS
GENES
c DNAS/TRANSCRIPTS
q11.1
q11.2
q12.1
q12.2
q13
q21
q22.1
q22.2
q22.3
q23.1
q23.2
q23.3
q24.1
q24.2
q24.3
CY151 CY149 CY132?
CY8
ATP5A1[S2555E(Bdy95g07)]
(GS1850)
CY138(P)
CKBP1
PHKB[S2546E(Bdy28b01)]
S2600E(Cdy0ze10)
S473E(EST01969)
S2551E(Bdy55c01)
S2661(NIB1932)
CY148
(GS1709)
S463E(ScDNA-E1)
CY140
S2967(KI0037)
S470E(EST02246)
S2536E(Bda63B12)
S2560E(Bdya7a03)
S2556E(Bdy97f07)
CY135
MMP2
S2539E(Bdab0h04)
CY138(D)
GNAO1
CES1
(KIAA0025)
S434E(EST00973)
CY7
MT
MT3
S2602E(Cdylaf06)
S2941(Z38171)
CY18A(P1)
CETP
S462E(ScDNA-A319)
POLR2C
CNCG2
CNCG3
S2566E(Cda0kd02)
S2598E(Cdy0cc07)
S2591E(Cda1je05)
S2589E(Cda1jb12)
CY126
CSNK2A2
BCGF1
GOT2[S2550E(Bdy44e12)]
S427E(755/756)
S2562E(Cdo01g10)
CY130(P) CY18A(D1)
M22-2
CY122
CY125(P)
CY127(P)
FRA16B
S438E(EST00255)
CY130(D)
CA7
CBFB
APOEL1
S460E(ScDNA-92)
S461E(ScDNA-A187)
S2655(NIB2292)
S2935E
S2925E
S3013
CY4
LCAT(MECL1,CTR2,PSKH)
SLC9A5
HSD11B2
S667E(CTG-B43)
S2554E(Bdy91f03)
S2576E(Cda0wh08)
S2603E(Cdy1dd04)
S2584E(Cda1cb03)
S2540E(Bdab9d04)
CY143
S2605E(Cdy1gf07)
CY127(D)
CDH1
S2533E(EST06651)
CY6
S2538E(Bda97h10)
CY125(D)
(GS1002)
CY128
CY13A
ALDOA
S2544E(Bdy25c01)
S2563E(Cda01h11)
S2659(IB2025)
S2541E(Bdac1b04)
(GS1553)
S3198E
CY113(P)
NMOR1
S2542E(Bdac5g11)
S2583E(Cda1ca04)
S2647(GS3415)
S2587E(Cda1fg11)
CY5
CALB2
CY170
S471E(EST01953)
S472E(EST01954)
S2657(IB727)
CY107(P)
HP
TAT
DHODH
S2579E(Cda14g04)
S430E(EST00132)
[S2606E(Cdy23b05),S2593E(Cda23b05)]
CY110
S2922(NIB727)
S2946(Z38301)
CY116
F265S2E(Bdac5g01)
CY119 CY118 CY157
S2947E(Z38406)
S2953(Z38507)
CY117
CTRB1
(GS980)
S459E(ScDNA-A129)
S2650(GS684)
CY145
CY124 CY105
CY113(P)
FRA16D
CY121
(GS1487)
S435E(EST00978)
CY115
S2559E(Bdya4d04)
PLCG2
HSD17B2
CY107(D)
CY108
S2944(Z38264)
CY100 CY114
COX4
S458E(ScDNA-F3)
S424E(EST00034)
CY106
CY104?
CY120
S2575E(Cda0uf06)
F200SIE(Bdyc4f09)
S2644(GS729)
S2645(GS1536)
S3004E(HOG)
S3012
S3011
CY18A(P2)
(GS1516)
CY112
CY2 CY3
PCOLN3
GALNS
APRT
CA5
CYBA
F245SIE(Bda45f05)
S469E(EST00889)
S2663(NIB609)
S2592E(Cda1je10)
S2926E
S2931E
S2930E
CY18A(D2)
CMAR
MC1R
DPEP1
RPL13
S532E(EST06702)
S457E(ScDNA-A55)
S444E(BBC1)
S2548E(Bdy37e07)
TELOMERE

The chimaeric clones in the YAC library constructed with flow sorted DNA from human chromosome 16 comprise about 20% of the clones. The low abundance repeat sequences that are shared between chromosome 16 and other chromosomes may contribute to the identification of false positive clones. Errors in the map can be resolved by increasing the density of markers on the physical map, FISH analysis of the YACs and comparison to the results to the genetic map.

Integration of the chromosome 16 genetic map with the physical map was achieved by PCR screening of YAC and cosmid clones using microsatellite markers and assignment of these STS markers to the somatic cell hybrid breakpoint intervals (Doggett *et al.* 1995). A total of 570 STSs have an average spacing of 160 kb on the integrated map. The somatic cell hybrid panel also provided the framework for the integration of over 500 genes, ESTs, anonymous DNA markers and microsatellite repeats on the physical map. The chromosome 16 map of 90 Mb is 2-3 times the size of the first low resolution physical maps reported of chromosome 21 (Chumakov *et al.* 1992) and the Y chromosome (Foote *et al.* 1992) which is a significant achievement.

The integration of the genetic map with the physical map tests the integrity and consistency of the data between both maps and provides the opportunity of positional cloning and localisation of genes by linkage on chromosome 16. The cosmids, genes, ESTs and genetic markers integrated with the cytogenetic based physical map which has an average resolution of 1.1 Mb (Doggett *et al.* 1995), provide a resource for the positional cloning or the positional candidate approach for isolating any disease gene mapped to this chromosome. The construction of the chromosome 16 human/rodent somatic cell hybrid panel and the localisation of cosmid contigs and YACs provide a framework for the large scale sequencing of this chromosome, the final goal of the HGP.

1.4.2 The 16q24 Chromosomal Region

The proposed project for this thesis involved the detailed physical mapping of the 16q24 region of human chromosome 16 to aid the positional cloning of two disease genes localised to this region. Particularly, efforts were concentrated on this region of human chromosome 16 since the telomeric regions of human chromosomes were demonstrated to contain a high concentration of genes. In addition, the gene responsible for Fanconi anaemia group A was localised to 16q24 (see section 1.7.1). A region of loss of heterozygosity in sporadic breast cancer was localised to the 16q24 chromosomal region, which indicated the presence of a tumour suppressor gene. This will be described in more detail in sections 1.5.3 and 1.6.1.

1.4.2.1 Gene-Rich Chromosomal Regions

The human genome is composed of isochores, specifically, of large DNA regions (average size > 300 kb), which are homogeneous in base composition and belong to a small number of families characterised by different GC levels (Bernardi *et al.* 1985; Bernardi, 1989). There is a non-uniform distribution of genes between the isochore families. The H3 isochore family comprises 3% of the human genome that is GC-rich and gene-rich (Mouchiroud *et al.* 1991). The GC content of the H3 isochore family is at least eight times higher than the H1 and H2 isochore families (31% of genome), and at least 16 times higher than the L1 and L2 families (62% of genome).

This heterogeneity in gene distribution is exemplified by the chromosome 16 band 16q21, the most intensely staining positive G band of chromosome 16, where there is an absence of localised genes and cytogenetic breakpoints. Individuals who have a normal phenotype and are heterozygous for a complete deletion of this band have been described (Witt *et al.* 1988) which suggests that 16q21 contains little DNA of genetic importance. The other less intensely staining positive G bands of chromosome 16 do not show a lack of gene localisations and cytogenetic breakpoints. Several other reports of patients who have normal phenotypes and are heterozygous for deletions involving G bands have been documented. These include bands 13q21 (Couturier *et al.* 1985) and 11p12 (Barber *et al.* 1991).

To identify the chromosomal localisation of the H3 isochore family, Saccone *et al* (1992) conducted *in situ* hybridisation experiments to human metaphase chromosomes. The H3 isochore family is thought to contain one third of human genes, has the highest concentration of CpG islands and CpG doublets, the highest transcriptional and recombinational activities and a distinct open chromatin structure characterised by DNAase sensitivity, nucleosome free regions, scarcity of histone H1 and acetylation of histones H3 and H4 (Bernardi *et al.* 1985; Bernardi, 1989; Mouchiroud *et al.* 1991: Rynditch *et al.* 1991). The H3 isochore family was localised in telomeric bands, which are generally light staining G bands, and chromomycin bands that were mostly telomeric (Saccone *et al.* 1992). Saccone *et al* (1992) demonstrated signals at the telomeric bands on 1p, 2q, 4p, 5q, 7p, 8q, 9q, 10q, 11p, 12q, 16q, 17q, 19p, 20q, 21q and 22q. Results from the compositional mapping of the long arm of chromosome 21, also indicated that single copy sequences from the GC-richest isochores are located in a telomeric band (Dutrillaux, 1973) and a chromomycin band (Ambros *et al.* 1987). Data generated from a transcript map constructed for human chromosome 21 also confirmed that telomeric regions are gene-rich. The distribution of transcripts along the chromosome was demonstrated to increase toward the telomere of this chromosome (Yaspo *et al.* 1995).

The results of Saccone *et al* (1992) demonstrated that the long arm of chromosome 16, 16q, had strong signal located at the telomeric band, indicating a gene-rich region. An additional piece of evidence suggesting that the 16q24 region has a high concentration of genes is that the 16q24 region may be lethal when monosomic. Patients which are heterozygous for deletions in the 16q24.2-qter region have not been reported, in contrast to patients with deletions in the remainder of chromosome 16 who have a viable phenotype (Callen *et al.* 1993).

In summary, the H3 isochore family which contains a high gene density and has been localised to telomeric bands of chromosomes, has important properties which deserve special attention in the HGP. In view of the evidence for telomeric regions containing a high gene content and to maximise the biological information for the sequencing effort, the 16q24 chromosomal region was chosen for the construction of a detailed physical map consisting of cosmids and YACs. This map can eventually be used for the construction of a transcript map to identify the genes localised in this region and ultimately to determine the nucleotide sequence of this region.

1.5 THE GENETICS OF BREAST CANCER

1.5.1 Genetic Model for Tumour Progression

The development of cancer is a complex process associated with multiple genetic abnormalities in a single lineage resulting in alterations of the normal mechanisms controlling cellular growth and differentiation. It is likely that for each type of cancer either a different combination of progressive alterations are involved or, the same alterations can lead to differing phenotypes depending on the target tissue specificity. The human model for multistage carcinogenesis is based on studies in colorectal cancer (Vogelstein *et al.* 1988; Fearon and Vogelstein, 1990). These studies have provided proof for the multistep hypothesis of solid tumour development and demonstrated the involvement of accumulated genetic changes in a single lineage. These genetic changes include loss of Deleted in Colorectal Cancer (DCC) suppressor gene expression from loci on at least three chromosomes and the activation of the RAS protooncogene which lead to the progression of colorectal cancer. Thus, the molecular characterisation of these genetic alterations has revealed numerous genes required to transform a normal cell into a malignant cell. Moreover, a number of these genes have been demonstrated to be oncogenic in their own right confirming the close association between chromosomal aberrations and neoplasia.

DNA abnormalities affect two broad classes of cellular genes, protooncogenes and tumour suppressor genes (TSGs), both of which play critical roles in the control of cell growth and differentiation. Protooncogenes give rise to cancer only when altered by a mutational event, producing a version of the gene called an oncogene. These genes may be altered by two mechanisms. One way is through mutations leading to abnormalities in gene expression, including gene amplification, gene rearrangements and promoter mutations. The other way is through mutations in the coding region through point mutations, leading to the production of proteins with abnormal biochemical features. Tumour suppressor genes (TSGs) are oncogenic by virtue of their loss rather than activation. The normal function of these proteins is to restrain cellular proliferation. Both copies of these genes must be inactivated for oncogenesis to progress. TSGs will be discussed in more detail in section 1.5.3.

Thus, the involvement of a number of genes in the process of carcinogenesis indicates that there may be various routes in the multistage process of cancer evolution. The aetiology of cancer is complex and can be understood through the use of clinical, genetic, cytogenetic and molecular genetic approaches. Understanding the molecular basis of the behaviour of tumours may reveal why they fail to respond to therapy or metastasise. Moreover, insight in the molecular genetic evolution may yield new targets for prevention or therapy.

1.5.2 Breast Cancer

Breast cancer is the most common malignancy amongst women in Western countries with an incidence of about 1 in 9 (Swanson *et al.* 1993). Breast cancer accounts for approximately 20% of total female cancer mortality (Hirayama, 1989). Although the last 20 years have produced several advances in the area of early breast cancer detection and therapeutic management, the incidence of this disease has increased but the mortality has not decreased. It is likely that through a clearer understanding of the molecular mechanisms underlying breast cancer development, progression and metastasis, these numbers will be reduced significantly.

Mammary epithelial cells undergo complex alterations in response to hormones. These cells divide, differentiate, secrete milk proteins and die by apoptosis. Disease may emerge when these processes are disturbed. Breast cancers are all derived from ductal or alveolar epithelial cells. The aetiology of breast cancer is associated with multiple risk factors including genetic, hormonal and dietary factors, and environmental agents such as carcinogens and radiation (Hsu *et al.* 1994). It is known that the most important factor for the development of breast cancer involves the level of the hormone oestrogen (El-Ashry and Lippman, 1994) although it is not known whether oestrogen plays a role as an initiator and also as a promoter.

Oestrogen stimulates cell proliferation via the induction of growth regulatory genes. It is widely believed that breast cancer begins as an oestrogen-dependent and progresses to an oestrogen-independent state. This progression may involve the constitutive expression of genes that were once regulated by oestrogen, loss of expression of oestrogen receptors, lack of response to anti-oestrogen tamoxifen and a more aggressive disease pattern with poorer prognosis.

Knowledge of the molecular markers that are involved in the developmental pathway of the mammary epithelial cells and the effects that these markers may have on the interactions that occur between the mammary epithelial, stromal and adipose tissue is limited. These issues must be addressed by the identification and delineation of the changes that occur in the mammary tissue during the development of the mammary gland at the molecular level, as well as the genetic alterations that may contribute to the progression from normal growth through malignancy to metastasis.

1.5.3 Tumour Suppressor Genes and Loss of Heterozygosity

Tumour progression is a complex process involving numerous genetic alterations of oncogenes and tumour suppressor genes which affect cell proliferation (Fearon and Vogelstein, 1990). Tumour suppressor genes (TSGs) are believed to be involved in the control of normal suppression of cellular proliferation during development (Knudson, 1989) including cell cycle regulation, transcriptional repression, signal transduction modulation and DNA repair. The inactivation of TSGs through deletions of these genes (Nigro *et al.* 1989) is an important mechanism in the pathogenesis of neoplasia.

Three lines of evidence have supported the idea that genetic alterations underlying neoplastic transformation may involve the inactivation of these TSGs. Somatic cell hybrid experiments provided the first evidence for TSGs. Somatic cell hybrids generated between normal cells and cancer cells did not give rise to tumours when injected into a suitable host, unless the chromosomes from the somatic cell hybrids were lost to generate tumourigenic variants (reviewed in Harris, 1988). This suggested that tumourigenesis arose through specific chromosomal losses from the normal cells and was central to the idea that tumourigenicity was suppressed in the hybrid cells by a genetic mechanism. This observation was taken as evidence that recessive genetic changes were responsible for tumourigenesis and were complemented by chromosomes with normal alleles.

A second approach for the identification of inactivated genes in neoplastic transformation was the utilisation of family material demonstrating an inherited predisposition to the tumour. A proportion of most types of tumours occur in both familial and sporadic forms of the disease and TSGs were first identified in inherited cancers. Knudson (1971) proposed the "two hit" hypothesis for the development and progression of cancer which took into account the differences between inherited and sporadic cancer. This hypothesis stated the autosomal dominant form of familial cancers could be accounted for by recessive loss of function mutations at the cellular level. The "first hit" in heritable cancers is a specific germline mutation of a suppressor gene such that all cells carry this mutation. For sporadic tumours, the "first hit" mutation occurs somatically. The "second hit" in both groups of cancers is a mutation in, or loss of, the second copy of the same gene. This "second hit" always occurs somatically, to the effect that the cell becomes homozygous. Thus, this hypothesis suggests that while susceptibility to cancer is dominantly inherited, tumourigenesis itself is recessive. One normal allele prevents cancer, hence the term tumour suppressor gene.

Studies of the inherited forms of the cancer retinoblastoma revealed that the same gene, RB, localised to 13q14, is involved in both the inherited and sporadic forms of the disease. Diverse mutations of RB were found in the familial and sporadic forms of retinoblastoma. In the familial form, the affected individual inherits a non-functional mutant allele from an affected parent that predisposes individuals to retinoblastoma through somatic recessive mutations such as chromosome loss or nucleotide substitutions, small deletions, mitotic recombination or gene conversion in the remaining normal allele (Dunn *et al.* 1989). The frequency of somatic mutational events is sufficiently high so that most individuals that inherit a mutant allele will develop one or more tumours depending on how many somatic mutations occur. This leads to apparently dominant mode of inheritance. In contrast, sporadic forms of retinoblastoma involve two independent somatic mutational events, the second of which must occur in the same cell or descendants of the cell that received the first mutation (Cavenee *et al.* 1983). The inactivation of both alleles is rare and sporadic tumours thus appear later in life.

The idea that the molecular basis of retinoblastoma involves recessive loss of function mutations was supported by demonstrating that in some cases of multiple tumours there was a constitutional deletion of chromosomal band 13q14.1 (Francke, 1978) in all somatic cells from familial tumours, and occasional deletions of 13q14 in sporadic tumours (Balaban *et al.* 1982). In the sporadic form of retinoblastoma, loss of heterozygosity of markers was demonstrated on one chromosome 13 due to either non-disjunction, mitotic recombination or gene conversion (Cavenee *et al.* 1983). However, the normal allele on chromosome 13 inherited from the non-affected parent was lost in the tumour, showing that LOH was acting to uncover a germline mutation (Cavenee *et al.* 1985). Based on the Knudson model of two recessive mutations leading to loss of function, it was demonstrated that separate genetic events were responsible for inactivation of each allele within a tumour.

The events from the "second hit" which result in the absence of genetic material could be detected by loss of heterozygosity (LOH) for markers flanking the locus. This LOH in tumours was used as an indication of the presence of a TSG. Thus, the third line of evidence for TSGs was provided by loss of heterozygosity (LOH). LOH involved the identification of regions of the genome where a genetic marker which was heterozygous in normal somatic tissue was homozygous in tumour tissue due to loss of one allele. The closer any linked marker is to the putative locus, the higher the frequency with which LOH is scored in tumours. It has been estimated that the frequency of LOH in tumour cells is at least two orders of magnitude higher than that of point mutation (Weinberg, 1991; Harwood *et al.*

1993). LOH is therefore greatly favoured over other mechanisms to eliminate the wild-type allele. As a consequence, non-random LOH is widely assumed to imply the presence of a TSG at that locus.

1.5.4 Loss of Heterozygosity at 16q

Many solid tumours have been demonstrated to display LOH on the long arm of chromosome 16, 16q, by analysis of (AC)n repeats and RFLPs. In hepatocellular carcinoma (Tsuda *et al.* 1990), 52% of informative cases demonstrated a common region of allele loss between the haptoglobin (HP) gene (16q22.1) and the chymotrypsinogen B (CTRB) gene (16q22.3-16q23.2). However, greater than 60% of cases with LOH were suggested to be due to monosomy of the entire chromosome 16. Allele loss was also demonstrated in 60% of prostate cancer tumours at 16q (Bergerheim *et al.* 1991; Kunimi *et al.* 1991), with a commonly deleted region between D16S4 (16q22.1) and 16qter. Other tumours with LOH at 16q include Wilms' tumour (Maw *et al.* 1992), and primitive neuroectodermal tumours (Thomas *et al.* 1991). These observations suggest that a gene or group of genes on this chromosome arm are suppressing the development of tumours from different histological origin. However, in these studies detailed localisation of the region(s) of chromosome 16q showing LOH is yet to be determined in order to identify the gene(s) involved in tumour suppression.

Studies on LOH in breast cancer have implied that TSGs may exist on chromosomes 1q, 3p, 8q, 9q, 11p, 13q, 16q, 17p, 17q and 18q (Chen *et al.* 1989; Devilee *et al.* 1989; Lundberg *et al.* 1987; Ali *et al.* 1987; Mackay *et al.* 1988; Sato *et al.* 1990; Devilee *et al.* 1991). The cytogenetic study of breast tumours showed that the most frequently lost chromosomes were 8, 13, and 16 and the most common region of karyotypic change was on chromosome 1 at q21-qter (Rodgers *et al.* 1984). Dutrillaux *et al* (1990) demonstrated rearrangements of chromosome 1 and/or 16 to be the most frequent aberration leading to gain of 1q and loss of

16q. Chen *et al* (1989) suggested that a TSG may be located at 1q23-q32 from RFLP studies of 48 breast tumours.

A total of 39 RFLP markers covering all the autosomal chromosomes were used to determine how many chromosomes or parts of chromosomes were involved in the development of primary breast carcinoma (Sato *et al.* 1990). Frequent LOH was observed in pairs of normal and breast tumour tissue with the RFLP markers on 17p and 16q, D17S74 and D16S7 respectively. Allele loss of chromosome 17p was shown in 33 of 59 (56%) informative cases and of 16q in 19 of 42 (45%) informative cases. This incidence of LOH on 16q in breast cancer had not previously been reported, although chromosome 16 had been implicated in breast cancer in the cytogenetic studies of Rodgers *et al* (1984).

Another LOH study was conducted in which all non-acrocentric chromosome arms were tested with at least one RFLP marker for allelic imbalance in 86 breast tumours (Devilee *et al.* 1991). They also introduced the term 'allelic imbalance' to refer not only to the loss of an allele but also to an increase of one allele with respect to the other in the tumour DNA. Several chromosome arms, in addition to chromosome arms containing known TSGs (13q, 17p, 18q), showed allelic imbalance in more than 30% of the informative cases. These results included 1q (50% allelic imbalance), 3p (34%), 6q (50%), 8q (40%), 9q (36%), 11q (315), 15q (37%), 16q (40%) and 17q (34%). Allelotype studies by Larsson *et al* (1990) and Harada *et al* (1994) have reported comparable frequencies of LOH which indicates that these chromosomal arms are candidates for containing TSGs. These results are shown in figure 1.2.

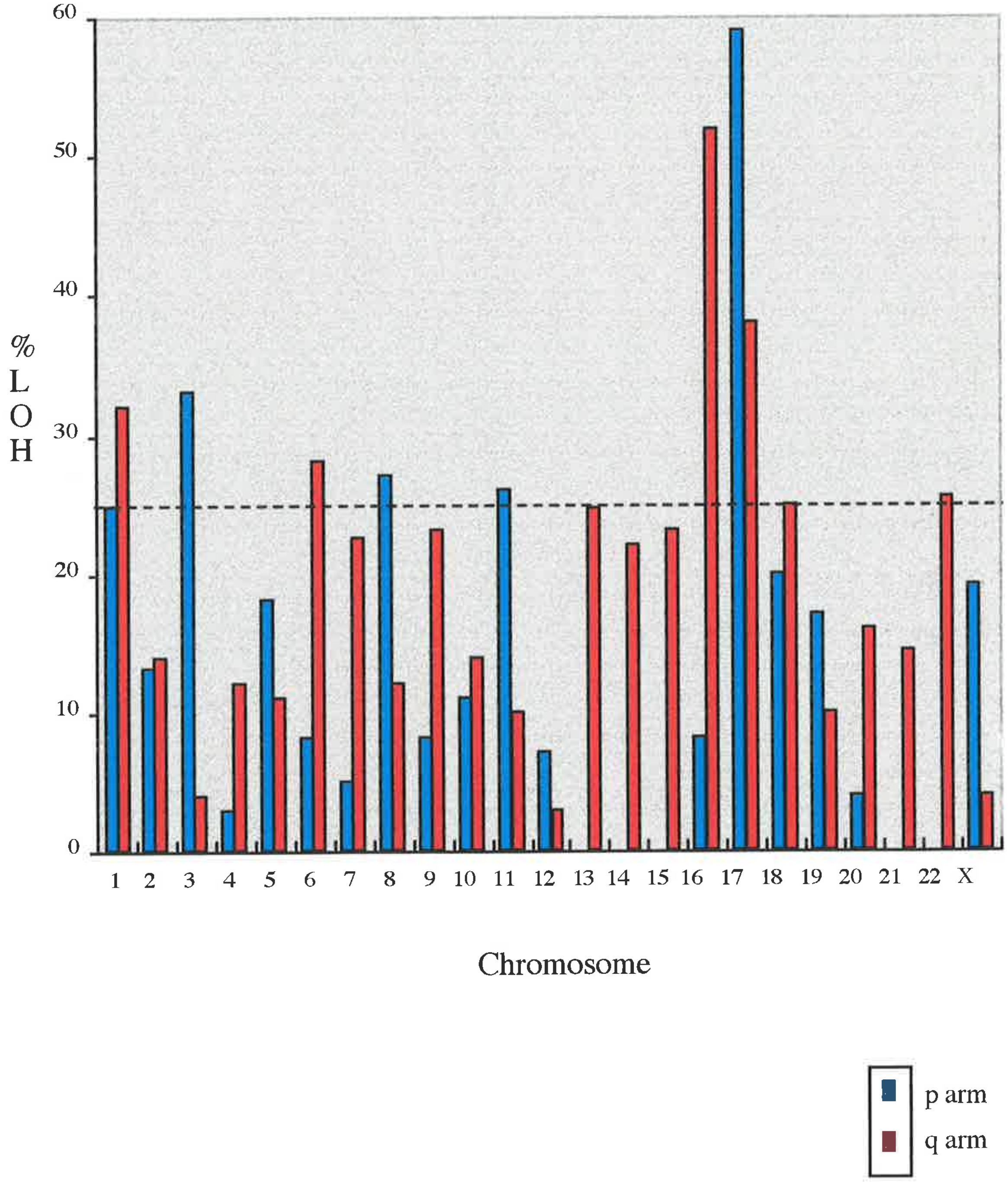
1.5.5 Hereditary Breast Cancer

Genes involved in the pathway of tumourigenesis have been identified using family material demonstrating an inherited predisposition to early onset breast cancer. It has been estimated that 5-10% of breast cancer cases in Western Countries are due to a highly

Figure 1.2

Frequency of allele losses on individual chromosomal arms summarised from various allelotype studies in breast cancer (Ali *et al.* 1987; Mackay *et al.* 1988; Chen *et al.* 1989; Devilee *et al.* 1989; Larsson *et al.* 1990; Sato *et al.* 1990; Devilee *et al.* 1991; Harada *et al.* 1994). The most frequent LOH was observed for chromosomal arms 17p and 16q. The red columns indicate the q arm and the blue columns indicate the p arm.

Allelotype of Breast Cancer



penetrant, autosomal dominant inherited predisposition (Newman *et al.* 1988) with patients developing breast (or ovarian) cancer relatively early in life. Studies of pedigrees with clustered breast cancer cases have identified at least four susceptibility genes which include BRCA1, BRCA2, TP53 (or p53) and the androgen receptor. In 1990, an autosomal dominant susceptibility gene predisposing to breast and ovarian cancer, BRCA1, was mapped to 17q21 by linkage analysis (Hall *et al.* 1990). LOH studies were performed with four (AC)n repeat markers at the 17q12-21 chromosomal region by PCR using lymphocyte DNA and DNA extracted from familial breast and ovarian tumours (Smith *et al.* 1992). Of the two breast tumours and eleven ovarian tumours examined in four affected families, one breast tumour and eight ovarian tumours showed allele loss for these markers suggesting the putative breast-ovarian cancer gene, BRCA1, to be a TSG.

BRCA1 was recently isolated by positional cloning (Miki *et al.* 1994; Futreal *et al.* 1994) and distinct BRCA1 mutations have been identified (Castilla *et al.* 1994; Friedman *et al.* 1994; Hogervorst *et al.* 1995; Shattuck-Eidens *et al.* 1995). Germline BRCA1 mutations have been detected in more than 100 families and tumours with about 70% of mutations being frameshifts and 10% nonsense mutations (Castilla *et al.* 1994; Friedman *et al.* 1994; Hogervorst *et al.* 1995; Shattuck-Eidens *et al.* 1995) which lead to truncation of the mutant protein. Mutations have been scattered throughout the BRCA1 gene and no hot-spot has been identified. Germline mutations displayed in BRCA1 which were demonstrated to uniformly involve loss of the wild type BRCA1 allele (Smith *et al.* 1992), appear to account for greater than 75% of breast and ovarian cancer in families that exhibit a high incidence of both types of tumour and about 40-50% in families displaying only inherited breast cancer (Narod *et al.* 1995). The study of approximately 200 families with at least four cases of breast cancer, has revealed that about 50% of families have mutations and/or linkage to BRCA1 and about 30% appear linked to BRCA1 (reviewed in Szabo and King, 1995).

The characterisation of BRCA1 is important in generating insight into the cellular basis of familial breast and ovarian cancer as well as providing clues for sporadic breast cancer. BRCA1 encodes a protein of about 190 kD (Miki *et al.* 1994) with a DNA binding zinc finger motif near the amino terminus which suggests that it may be a transcription factor. The expression of BRCA1 is induced by oestrogen but its function is unknown (Marquis *et al.* 1995). Jensen *et al* (1996) conducted a search on the functional domains of BRCA1 and demonstrated a granin consensus sequence. Granins are secreted proteins triggered by the activation of cyclic AMP. Expression of some members of the granin family is regulated by oestrogen (Thompson *et al.* 1992). The function of granins is not clear but they may be protein precursors to biologically active peptides which may assist in packaging peptide hormones and/or modulating the processing of peptide hormones (Huttner *et al.* 1991).

Genetic changes on chromosome 17 are frequently described in breast tumours (Sato *et al.* 1990; Devilee *et al.* 1991; Smith *et al.* 1992). Nagai *et al* (1994; 1995) identified five distinct regions of LOH on chromosome 17 in sporadic primary breast tumours that may be responsible for growth suppression. Two of these regions are located on 17p, one spanning the TP53 locus and the other in 17p13.3 distal to TP53. The other three regions are on 17q and include the region encompassing BRCA1 and between 17q12-qter. LOH of different regions of chromosome 17 have also been described in oesophageal tumours (Swift *et al.* 1995), cervical carcinoma (Park *et al.* 1995) and non-small cell lung cancer (Fong *et al.* 1995) suggesting that there are additional genes to BRCA1 and TP53 on chromosome 17 that suppress the development of tumours of different histological origin.

LOH around the BRCA1 locus has been observed in 30-70% of sporadic breast and ovarian tumours and somatic mutations should be detected in sporadic tumours if BRCA1 is a TSG. However, somatic mutations in BRCA1 have only been described in four out of 47 cases of sporadic ovarian (not breast) tumours (Merajver *et al.* 1995). This is the first finding to implicate BRCA1 directly in sporadic tumours and it is surprising that somatic mutations have not been detected in sporadic breast tumours. Hence, further work is required to identify somatic mutations in BRCA1 in sporadic breast tumours and to determine the genetic effect of the frequent LOH occurring in the region encompassing BRCA1 in sporadic breast cancers.

Observations of LOH at flanking markers considerable distances from BRCA1 has led to suggestions that other TSGs lie proximal and/or distal to BRCA1 on chromosome 17. Since mutations in BRCA1 have been rarely described in sporadic ovarian cancer and the relationship of the BRCA1 gene to ovarian tumours of borderline malignancy is unclear, a detailed deletion map of polymorphic markers in a 650 kb region of 17q21 was constructed (Tangir *et al.* 1996). This study of 26 sporadic ovarian cancers of borderline malignancy revealed a common region of deletion approximately 60 kb centromeric to BRCA1. This suggested that inactivation of the BRCA1 gene may not be responsible for the development of borderline ovarian tumours and that another TSG may play a role in sporadic ovarian cancer development and possibly in other tumours.

Although BRCA1 accounts for breast and ovarian cancer in more than 75% of families with genetic predisposition to the diseases mutations in a second breast cancer susceptibility gene, BRCA2, cause approximately 25% of familial breast (but not ovarian) cancer. Families with breast cancer linked to BRCA2 also suffer from a higher incidence of male breast cancer (Wooster *et al.* 1994) which suggests that BRCA2 is also an important breast cancer gene. BRCA2 was localised to 13q12-q13 (Wooster *et al.* 1994) and subsequently isolated by positional cloning (Wooster *et al.* 1995) with the identification of germline mutations in this gene using tissues from familial breast cancer patients. LOH on chromosome 13q has also been observed in sporadic breast and other cancers which suggests a somatically mutated TSG is localised in the region. BRCA2 is a strong candidate for this TSG and analysis of a large series of tumours to investigate whether BRCA2 is somatically mutated during oncogenesis has commenced.

The BRCA2 breast cancer susceptibility gene has been demonstrated to encode a 10-12 kb transcript by Northern analysis (Wooster *et al.* 1995). There are no clues to its functions as its sequence was not demonstrated to have strong homologies to sequences in the publicly available DNA or protein databases. For many other genes involved in breast cancer, such as CCND1, EMS1, EGF, DCC, prohibitin, and NM23, the evidence is still largely circumstantial or has been obtained only from *in vitro* studies on breast cancer cell lines (Van der Vijver, 1993).

The third locus, TP53, localised at 17p13, is a gene mutated in the rare Li-Fraumeni Syndrome pedigrees (Malkin *et al.* 1990) (see section 1.5.6). Mutations in a fourth gene, the X-linked androgen receptor, lead to breast cancer among men with the rare Reifenstein syndrome (Wooster *et al.* 1992).

1.5.6 Identified Genes Involved in the Development of Breast Cancer

Breast cancer is a heterogeneous disease with differences in recurrence rates, ploidy, proliferation rates, steroid and growth factor receptor expression and alterations in TSGs and oncogenes. Extensive studies have been conducted into the cellular and molecular pathways for oestrogen and other steroid hormones. Many of the target genes of oestrogen and their role in cellular proliferation have been identified. A limited number of oncogenes and tumour suppressor genes have also been examined for their presence and role in breast cancer and will be described in this section.

The number of oncogenes and tumour suppressor genes which have been examined is basically restricted to TP53, MYC, NEU and RB. The TP53 gene is a nuclear transcription factor (Levine *et al.* 1991) localised to 17p13.1 and is the single most mutated gene in cancer. Point mutations are the most common type of mutation found in this TSG. Germline loss of TP53 accounts for less than 1% of breast cancers and about 50% of sporadic breast cancers contain an altered TP53 gene (Coles *et al.* 1992). Germline mutations have been identified in

Li-Fraumeni patients, a familial cancer with a high incidence of mesenchymal and epithelial neoplasms at multiple sites including breast carcinomas, osteosarcomas and brain tumours. Somatic mutations in TP53 are common in breast and other tumour types, particularly higher grade intraductal and invasive carcinomas.

TP53 plays a crucial role in the cell cycle acting as a checkpoint monitor to levels of DNA damage during the transition between the G_1 and S phases of the cell cycle. In the event of DNA damage or deregulated growth, wild type TP53 is induced and leads to either cell cycle arrest or programmed cell death (apoptosis). Mutated TP53 does not do this, thus allowing for the selection of unstable populations (Lane, 1992; Vogelstein and Kinzler, 1992). Loss of TP53 is a two step process in which a mutation occurring in the first allele decreases the ability of TP53 to regulate cell growth. Alteration in the second allele completely knocks out normal TP53 expression. The frequency of mutations of the TP53 gene in a wide spectrum of human cancers suggest that TP53 inactivation may be a crucial and perhaps obligatory step in oncogenesis. This contrasts with other TSGs whose role in suppressing neoplasia appears to be restricted to certain tissues or cell lineages.

Gene amplification of the NEU and MYC protooncogenes is common in breast cancer and is associated with poor clinical course. The NEU gene is located on 17q and is found to be amplified or overexpressed in about 20% of breast tumours. The protooncogene MYC is a nuclear phosphoprotein localised to 8q24. It has been implicated in a large number of malignancies. The MYC protooncogene is important in breast cancer as about 20-30% of breast tumours (Escot *et al.* 1986) have abnormalities in MYC expression including gene amplification, rearrangement and overexpression. Overexpression has been shown to cause mammary tumours in transgenic mice (Stewart *et al.* 1984). It plays a role in cell proliferation as mitogenic stimulation of cells causes an increase in MYC. Decreased MYC is associated with terminal differentiation and loss of proliferative potential (Erisman and Astrin, 1988).

The retinoblastoma (RB) gene product is a nuclear phosphoprotein which undergoes cell-cycle dependent phosphorylation and acts as a negative regulator of the cell cycle. LOH for in the region of RB has been found in about 36% of sporadic breast tumours (Cox *et al.* 1994). The RB gene has been found to be mutated in breast cancer (Horowitz *et al.* 1989). In primary tumours where loss of RB expression has been observed, a proportion of cells still express RB suggesting that a mutation in this gene is a later event in breast cancer.

1.6 THE LONG ARM OF HUMAN CHROMOSOME 16 (16q) AND BREAST CANCER

1.6.1 Detailed Loss of Heterozygosity Studies of 16q in Breast Cancer

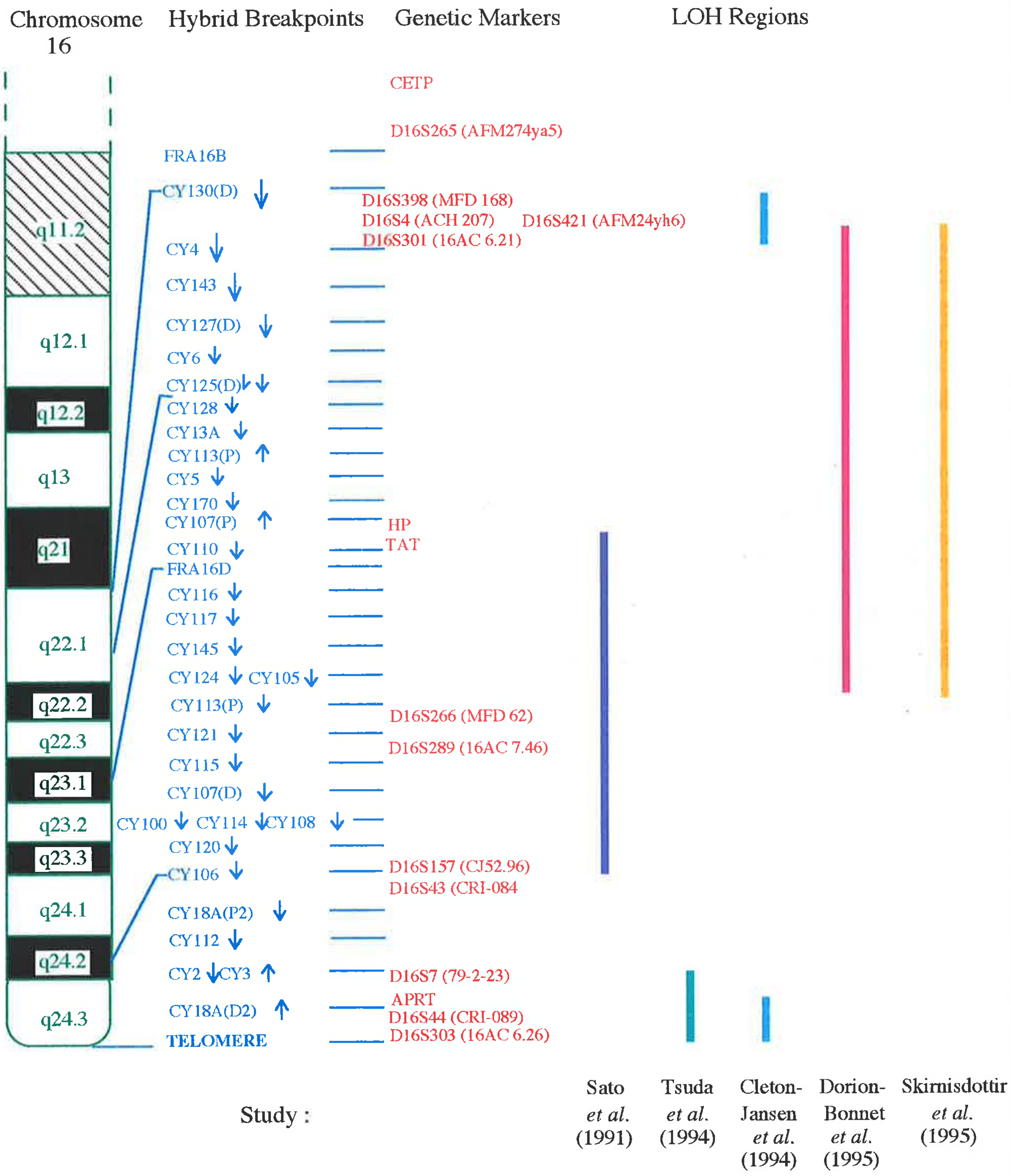
After the long arm of chromosome 16 was demonstrated to be involved in LOH in breast tumours, several studies were conducted to determine the more precise localisation of a TSG(s) on 16q using polymorphic markers mapped to chromosome 16. Sato *et al* (1991) showed frequent LOH of 16q in 51% of breast tumours (78 of 153). Also, a common region of deletion of chromosome 16q in breast cancer was identified in 57% (12 of 21) of cases at the HP locus (16q22.1) and between the HP and D16S157, probe CJ52.96, at 16q22-23 loci. This more accurate localisation of a region of LOH on 16q, shown in figure 1.3, was the first step toward the isolation of a putative TSG.

As the HGP progressed, the physical and genetic maps of chromosome 16 became more refined through the utilisation and localisation of a larger number of polymorphic markers. These polymorphic markers assisted in further refining the commonly deleted regions of chromosome 16 in breast tumours. Tsuda *et al* (1994) screened 78 paired samples of primary breast tumour and normal tissue with 27 polymorphic markers from chromosome 16 by RFLP analysis. An additional 152 tumours were screened with 6 markers CETP, D16S4, HP, TAT, D16S7 and APRT, spanning 16q13-q24.3, by RFLP analysis. LOH on 16q was detected in 38 of 78 (49%) tumours implicating this region as the most commonly deleted in breast cancer. The entire long arm of chromosome 16 was suggested to be lost in 15 of these tumours while the other 23 were observed to have partial allele loss on 16q. In comparison, LOH of the short arm of chromosome 16 was detected in 17 of 73 tumours (23%) but 16 of these tumours demonstrated allele loss on 16q.

Figure 1.3

Summary of location of regions identified from LOH studies of 16q in breast cancer. Sato *et al* (1991) demonstrated LOH between markers HP and D16S157, Tsuda *et al* (1994) demonstrated LOH between markers D16S43 and qter and Cleton-Jansen *et al* (1994) demonstrated two regions of LOH from APRT to D16S303 and between the markers D16S398 and D16S301. Recent studies demonstrate similar regions of LOH at 16q22.1 and 16q22.3-qter (Dorion-Bonnet *et al.* 1995) and 16q22-23 (Skirnisdottir *et al.* 1995).

LOH Studies on Chromosome 16 in Breast Cancer



The incidence of LOH was calculated to be 36% at the MT2 locus (16q13) and greater than or equal to 38% in the loci distal to MT2. In particular, the region deleted most frequently was the 16q24.2-qter region, between D16S43 and D16S155 and the telomere of the long arm. Included in this region were the markers D16S7, APRT and D16S44. In total, 127 of 230 (55%) tumours showed LOH on 16q and LOH on 16q24.2-qter was demonstrated in 118 of 225 (52%) tumours at the loci D16S7, APRT and/or D16S44, figure 1.3. Thus, a tumour suppressor gene was strongly suggested to be present in this region. This common region of deletion, 16q24.2-qter, differs from the regions reported by Tsuda *et al* (1990) for hepatocellular carcinoma and Sato *et al* (1991) for breast carcinoma. They observed the commonly deleted region to lie on 16q22-q23. Tsuda *et al* (1994) also demonstrated LOH distal to 16q12 and between 16cen-q22.1 in a number of their breast tumour samples. The collective results from these studies suggest that multiple TSGs, other than the putative one localised to 16q24.2-qter, may exist on 16q.

In another study, the region of 16q involved in LOH in sporadic breast cancer was extensively investigated using 79 paired tumour and normal DNA samples with an additional 20 polymorphic markers mapped to 16q (Cleton-Jansen *et al.* 1994). The percentage of tumours showing allelic imbalance with at least one informative marker on 16q was 63%. Tumours with a breakpoint, defined as a switch on the chromosome arm from LOH to retention of heterozygosity, can determine a small region of overlap when the breakpoints of a series of tumours are compared, and may provide information on the location of the putative TSG. The results demonstrated that at least two distinct regions of 16q were involved in allelic imbalance, suggesting that more than one TSG was located on this chromosome arm. LOH region I near the telomere at 16q24.3 from APRT to D16S303 which is retained (the genetic distance between the two bordering markers is 3 cM) and LOH region II on band 16q22.1 flanked by D16S398 and D16S301 (1.3 cM). These regions are shown in figure 1.3. The percentage of tumours showing LOH at region I was 52% and region II was 47%. These results are in agreement with those found in previous LOH studies of 16q in breast tumours (Larsson *et al.* 1990; Devilee *et al.* 1991; Sato *et al.* 1991; Tsuda *et* *al.* 1994). Cleton-Jansen *et al* (1994) observed concurrent allelic imbalance in different regions of 16q in individual tumours. The frequency of tumours displaying these complex allele losses was 6%. The authors did not determine whether the allelic imbalance in the two separate regions of 16q involved the same parental copy of chromosome 16. Recent allelic imbalance studies conducted to gain further insight into the smallest region of overlap on 16q in breast carcinomas have demonstrated the most recurrent LOH region to be at 16q22.1 and 16q22.3-qter (Dorion-Bonnet *et al.* 1995) and 16q22-23 (Skirnisdottir *et al.* 1995) which is in agreement with the previously published results (figure 1.3). Figure 1.4 shows a summary of studies involving breast tumours which have LOH involving different regions of 16q only and the frequency of LOH in the various regions.

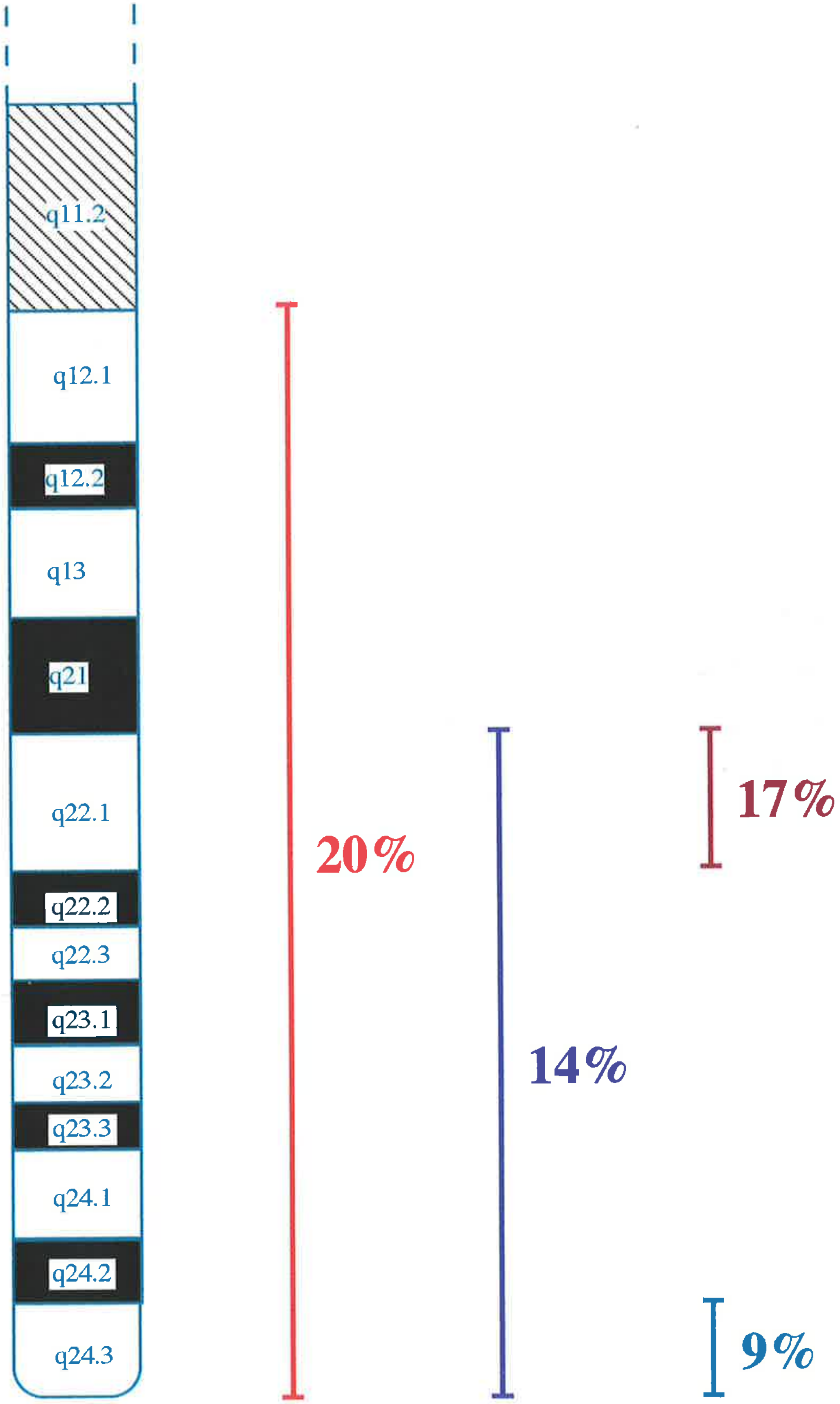
A cytogenetic study of breast tumours also demonstrated rearrangements on 16q involving bands 16q22-24 (Hainsworth *et al.* 1991) which are the same bands which contain the two regions of LOH on 16q. Loss of 16q has been observed in breast tumours with few or even no other cytogenetic abnormalities and has therefore been considered to represent an early event in tumourigenesis (Rodgers *et al.* 1984; Dutrillaux *et al.* 1990). These results are verified by the results from a study in which allelic imbalance of 16q in breast tumours did not correlate with allelic imbalance on other chromosome arms (Cornelisse *et al.* 1992).

In summary, extensive LOH studies suggest that there are at least two TSGs involved in breast cancer on the long arm of chromosome 16 at 16q22.1 and at 16q24.3. It is possible that these genes are also pleiotropically involved in suppressing the development of tumours of different histological origin, although for these tumours detailed localisation studies of the regions of chromosome 16q showing LOH has not yet been reported. As mentioned, allelotype studies of breast tumours and other tumours, such as hepatocellular carcinoma and Wilms' tumour, have demonstrated that LOH can involve more than one chromosome arm and also multiple sites on the same chromosomal arm. Since loss of 16q was proposed to be an early event in tumourigenesis, cloning a gene located on 16q may provide a clue for the elucidation of an early step in the multistep process of tumourigenesis.

Figure 1.4

Summary of various studies on LOH in different regions of 16q in breast cancer (Cleton-Jansen *et al.* 1994; Cleton-Jansen, personal communication). The frequencies of LOH have been determined by calculation of the number of breast tumours demonstrating LOH at 16q only out of the total number of tumours studied.

Loss of heterozygosity on 16q in primary breast tumours



The precise delineation of the smallest region of LOH is instrumental in any attempt to positionally clone a TSG localised in a region of interest. The distal region of LOH of 16q has been identified from APRT to the microsatellite marker D16S303 in the 16q24.3 chromosomal region. It will be possible to clone the TSG using either the positional cloning approach or the positional candidate approach. This will involve the construction of a detailed physical map across the 16q24.3-qter region and the isolation of genes or transcripts localised to the region. Candidate genes or transcripts already located in the region may also be characterised. They can subsequently be studied to determine whether they are altered in tumour DNA compared to normal DNA from the same patient.

The discovery of TSGs will have an immediate impact on the diagnosis of high risk individuals and the development of new therapeutic strategies. Most of the currently known TSGs have been identified by studies of inherited cancer, but the familial forms of tumours tend to be rare. Numerous regions of LOH have been identified in both familial and sporadic breast cancer, but all the TSGs presumed to be located in the regions of LOH have not been identified. Further progress in understanding the molecular mechanisms of breast cancer depends on the knowledge of the basic biology of the mammary epithelial cells as well as the functions of BRCA1, BRCA2 and other TSGs in the growth control of these cells.

1.6.2 Correlation Between Clinicopathological Markers of Breast Cancer and LOH on 16q

Various studies have attempted to correlate molecular genetic findings with the clinicopathological features of breast tumours. These studies may provide a clue to the biological role of LOH in breast cancer development. Frequent allele losses were detected with marker D16S7 (79-2-23) localised to 16q24.3. It was demonstrated that loss of chromosome 16q frequently coincided with lymph node metastasis in 67% of tumours (8 of 12) compared to 40% of tumours which retained chromosome 16q (8 of 20) (Sato *et al.* 1990). In contrast, the frequency of tumours with LOH on 17p and lymph node metastasis was not different to the group that retained 17p. Although the correlation between LOH on 16q and lymph node metastasis was not statistically significant in this study (Sato *et al.* 1990) or in another study from Harada *et al* (1994) it was suggested that this weak trend may serve as prognostic marker for metastatic potential in primary breast cancer.

In another study, Skirnisdottir *et al* (1995) revealed a slight but significant correlation between LOH on 16q and high progesterone receptor content. Comparison of tumours with and without LOH on 16q at the D16S421 locus at 16q22.1 revealed a slight correlation with high progesterone receptor content. The data showed no significant difference in survival between patients with LOH on 16q and those with a normal allele pattern on the long arm of chromosome 16.

However, further studies have failed to demonstrate any correlation of LOH on 16q with the clinicopathological features of breast tumours. Tsuda *et al* (1994) demonstrated the incidence of LOH on 16q to be high irrespective of differences in lymph node status, clinical stage, tumour size, histological grade and type, or oestrogen receptor status. LOH on 16q was frequent even in the group where histological grades and types indicated a favourable clinical outcome and in cases without metastasis. LOH on 16q was strongly suggested to occur commonly in breast cancer at a very early developmental stage among both highly aggressive and relatively low-grade tumours.

Thus, alteration of chromosome 16 has been suggested to play a role in the genesis of breast cancer irrespective of the grade of biological aggressiveness, clinical stage, tumour size, or oestrogen and progesterone receptor status. Other gene alterations such as NEU oncogene amplification, TP53 mutation, and LOH on other chromosomal arms have been related to the development of various aggressive biological and morphological characteristics such as high proliferation rate, marked nuclear atypia and/or dedifferentiation (Tsuda *et al.* 1990b; Sato *et al.* 1991; Iwaya *et al.* 1991).

1.6.3 Genetic Predisposition: Hereditary Breast Cancer and Chromosome 16

Hereditary breast cancer accounts for less than 10% of all breast cancer cases but the identification of breast cancer susceptibility genes, described in section 1.5.5, is a major achievement as these genes may provide a better understanding of all forms of these cancers in addition to the potential development of more effective therapies. There is evidence that additional genes involved in breast cancer exist in the genome as regions of LOH have been identified on numerous chromosomal arms in breast tumours.

In a study of LOH in tumours from patients with familial breast cancer (Lindblom *et al.* 1993), LOH was found not only on 17p (TP53), 17q (BRCA1) and on the X-chromosome (androgen receptor gene), but also on chromosome arms 8p, 19p and 16q. Frequently occurring allele losses were correlated with oestrogen receptor status, lymph node status, tumour size and distant metastases at follow up 2-15 years later. LOH at 16q was the most common genetic event detected. A highly significant correlation between LOH for at least one marker on 16q (19 of 67 cases) and the presence of distant metastases at follow up was demonstrated. Most of the tumours analysed retained heterozygosity at 16p suggesting that LOH at 16q is a specific event. The most frequently lost marker was D16S7 (79-2-23) in 18 of 57 informative cases. LOH at 16q was not correlated with oestrogen receptor status, lymph node status, tumour size or low differentiation grade. Similarly, the presence of distant metastases at follow up did not correlate with these parameters, suggesting that this could be an independent prognostic parameter. Survival data showed that patients with LOH at 16q were more likely to have a recurrence with distant metastasis compared to patients whose tumours did not show LOH at 16q. LOH at 17q was the second most common (18 of 75 cases) genetic event detected (Lindblom *et al.* 1993). A weak association between lymph node positivity and LOH at 17q was demonstrated. There was no significant correlation between LOH at 17q and oestrogen receptor status, lymph node status, tumour size or distant metastases.

In conclusion, frequent allele losses have been demonstrated with marker D16S7, mapped to 16q24.3, in both familial (Lindblom *et al.* 1993) and sporadic (Sato *et al.* 1990) breast tumours. Sato *et al* demonstrated a weak association between LOH at 16q in sporadic breast tumours and lymph node metastasis. The identification of a potential genetic marker, LOH at 16q, leading to future recurrence with distant metastases, may have clinical implications. The characteristic course of breast cancer is the initial spread of tumour cells to the lymphatic vessels, termed lymphogenic spread. The disease may progress by the spread of tumour cells via the blood stream to the rest of the body, referred to as haematogenic spread, and in some cases, breast cancer relapses after a long disease free interval. This indicates that the process of haematogenic spread is a separate event that is independent of lymphogenic spread, in addition to aggressive behaviour of the tumour. The finding of a separate event, LOH at 16q, which correlates with distant metastasis, but not with lymph node positivity or tumour size, suggests that a TSG is involved in haematogenic spread in familial breast cancer.

Distant organ metastases originate from selected cell clones of primary tumours and are assumed to carry all genetic features essential for the formation and maintenance of initial tumour growth (Fidler and Hart, 1982). In addition, they presumably harbour alterations for the metastatic spread. The examination of primary tumours and metastases may allow the recognition of features common to and changes in both tumours and may help elucidate the mechanisms involved in metastatic formation.

1.6.4 Candidate Tumour Suppressor Genes Localised to the 16q24 Chromosomal Region

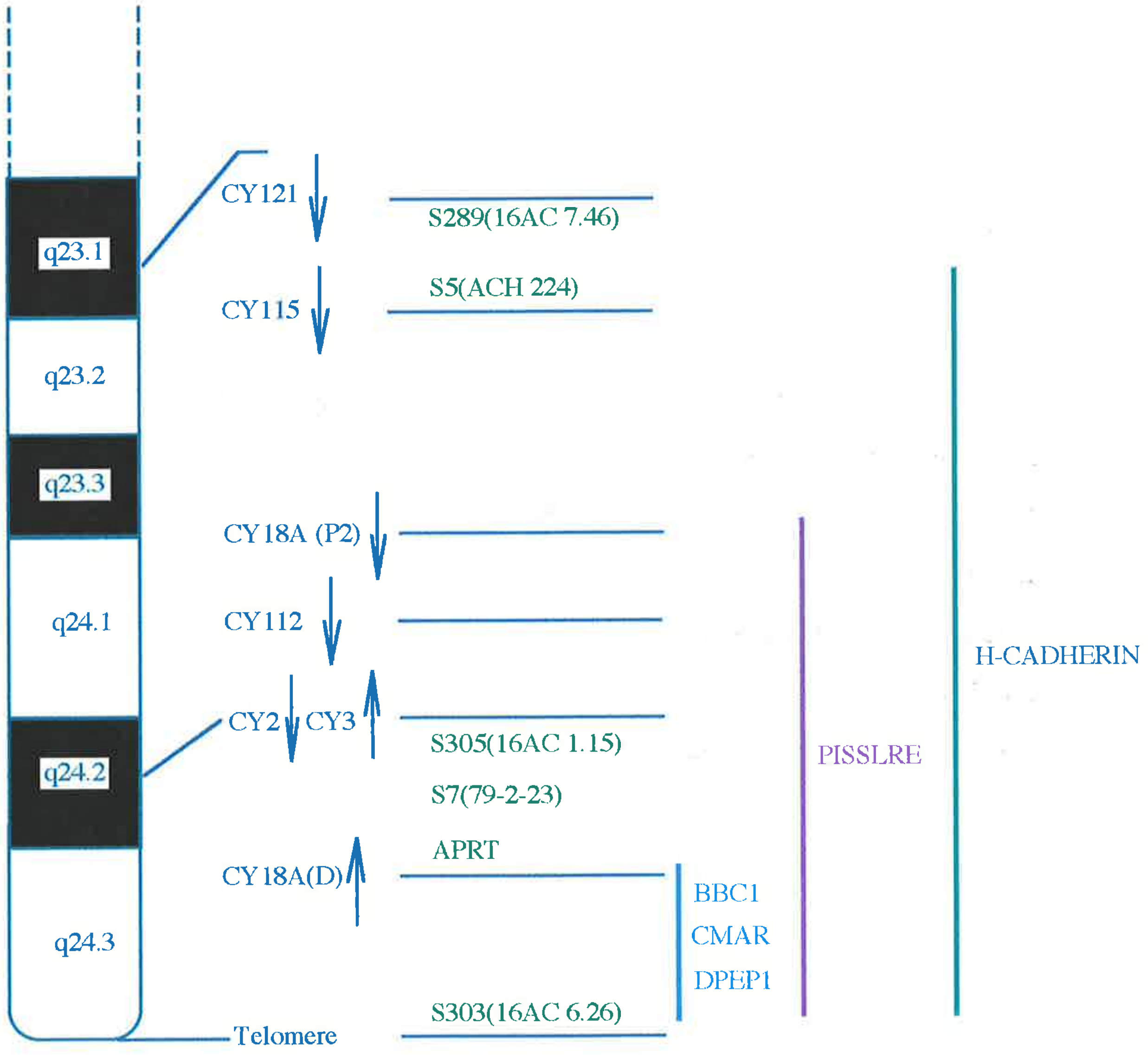
Positional cloning for the isolation of tumour suppressor genes can be complemented by the investigation of already cloned genes in the chromosomal region of interest as possible candidates. There are several genes on the long arm of chromosome 16, shown in figure 1.5, that are candidate tumour suppressor genes. The most promising of these candidate genes

Figure 1.5

Genes localised on the physical map of the 16q24 chromosomal region which may be possible candidates for the tumour suppressor gene involved in sporadic breast cancer.

CMAR	(Pullman and Bodmer, 1992; Koyama *et al.* 1993)
DPEP1	(Austruy *et al.* 1993)
BBC1	(Cleton-Jansen *et al.* 1995)
PISSLRE	(Bullrich *et al.* 1995)
H-cadherin	(Lee, 1996)

Candidate Tumour Suppressor Genes Localised to 16q24



code for proteins involved in cell adhesion. These molecules are members of the cadherin and integrin families are involved in cell-cell and cell-matrix adhesion and have been implicated in epithelial differentiation, carcinogenesis and metastasis (Takeichi, 1991; Behrens, 1993; Takeichi, 1993).

Cadherin dysfunction has been implicated in tumour development (Takeichi, 1991). Cell-cell association has been demonstrated to be disorganised in tumours and it is believed that this is due to the unregulated behaviour of tumour cells including invasion and metastasis (Behrens, 1993; Takeichi, 1993). A decrease in cellular adhesion or interactions through loss of cadherin molecules may enhance neoplastic progression as well as tumour cell invasion (Oka *et al.* 1993).

H-cadherin, or cadherin-13, localised to 16q24 by FISH to human metaphase chromosomes (Lee, 1996) is a favourable candidate TSG. This gene encodes a protein of 713 amino acids and is a member of the cadherin family that is involved in normal tissue cell-cell adhesion (Takeichi, 1991). A significant decrease in the expression of H-cadherin was shown in human breast carcinoma cell lines and breast tumours. The introduction and overexpression of H-cadherin in human breast cancer cells was shown to decrease the growth rates of the cells. These results implied that H-cadherin is one of the genes responsible for maintaining normal cellular phenotypes. The expression patterns, ability to inhibit tumour growth *in vitro,* chromosome localisation and cell adhesion properties of H-cadherin, suggest it has an important role in growth regulation in normal human mammary cells and is a favourable candidate TSG.

An association between the altered expression of E-cadherin (UVO), localised to the region of LOH on chromosome 16 at 16q22.1 (Mansouri *et al.* 1988) in various tumours including breast tumours, and advanced tumourigenic stage has been demonstrated (Takeichi, 1993; Behrens, 1993; Oka *et al.* 1993). Mutations were detected throughout the entire coding region of E-cadherin in 27 of 48 infiltrating lobular breast tumours, but none in breast tumours of other histopathological subtypes (Berx *et al.* 1995, 1996). The majority of these frameshift and non-sense mutations were found in combination with LOH of the wild type E-cadherin locus. These mutations were predicted to generate a secreted E-cadherin fragment instead of a transmembrane protein with cell adhesion activity. Additionally, they were suggested to be involved in the particular aetiology of sporadic breast cancers.

The gene encoding CMAR, cell matrix adhesion regulator, has been localised to 16q24.3 by linkage analysis (Koyama *et al.* 1993) and by physical mapping using chromosome 16 somatic cell hybrids (Callen *et al.* 1995). This gene encodes a protein of 142 amino acids and is a member of the integrin family, which are molecules that provide a functional bridge between extracellular matrix molecules and cytoskeletal components within the cell. The CMAR cDNA clone was isolated after it was observed to increase cell adhesion to components of the extracellular matrix in a colon cancer cell line (Pullman and Bodmer, 1992). Similarities of the CMAR protein sequence to known α and β integrin chain sequences were not identified. A protein motif search identified an N-terminal myristoylation motif, suggesting a cytoplasmic sub-membrane location for the protein, and a C-terminal consensus tyrosine kinase phosphorylation site, which is involved in the control of CMAR function. Removal of the tyrosine residue by site directed mutagenesis abolished the enhancement of cell-matrix adhesion. This suggested that CMAR is most likely activated by tyrosine phosphorylation and may have a role in signal transduction, coupling the cytoplasmic portion of adhesion molecules to the cytoskeleton and enabling the external environment to influence cell behaviour. It was proposed that loss of CMAR activity could be an early step in tumour invasion and metastasis, causing loss of differentiation induction or tumour cell release from cell or extracellular matrix attachments (Pullman and Bodmer, 1992).

As mentioned in section 1.6.2, regions of LOH on chromosome 16 have been suggested to play a role in the genesis of breast cancer irrespective of the grade of biological aggressiveness, clinical stage, tumour size, or oestrogen and progesterone receptor status.

The data suggests that the existence of a TSG on 16q24 is involved in the haematogenic spread of breast cancer cells. Cadherins and integrins have been implicated in breast cancer invasive capabilities, and loss of adhesion in tumour cells through LOH on chromosome 16 agrees with the results of Lindblom *et al* (1993) who described that LOH on chromosome 16 is associated with the development of distant metastases. Thus, both the H-cadherin and CMAR genes are candidate tumour suppressor genes .

Recently, a positive cell cycle regulator, PISSLRE, was localised to 16q24 (Bullrich *et al.* 1995). This gene was cloned through homology to positive cell cycle regulators (the cdc2 family). The PISSLRE protein is a novel member of the cyclin dependent kinase (cdk) family of protein/serine threonine kinases. It has a molecular weight of 33 kD and contains many conserved motifs of cdks that are required for cyclin binding and stabilisation. The progression of the cell cycle is regulated by the sequential activation of cdks through interactions with cyclically expressed proteins called cyclins. These complexes are inactivated by cyclin-cdk inhibitors. The first growth phase of the cell, G_1, has been implicated in the cellular events that lead to oncogenesis (Hunter and Pines, 1994). G_1 progression has been found to be highly regulated by cdk inhibitors that act in G_1 as a result of different cellular processes.

The cyclin-dependent kinases, cyclins and inhibitors of cyclin-cdk complexes have been reported to be involved in human neoplasia, and the localisation of PISSLRE to the region of LOH at 16q24 suggests that this gene may be involved in tumourigenesis. The predicted amino acid sequence of this protein shows 55% homology to p58/GTA which has been shown to inhibit entry into S phase when overexpressed in Chinese hamster ovary cells (Grana *et al.* 1994). Hence, although PISSLRE is a cdc2-related protein kinase, it may function as a negative regulator of cell cycle progression. However, Li *et al* (1995) showed that overexpression of both antisense and mutant constructs of PISSLRE in the U2OS tumour cell line suppressed cell growth when compared to cell growth with wild-type PISSLRE. The mutant forms of PISSLRE also stopped cell cycle progression in G_2-M. This indicated that PISSLRE is essential for cellular proliferation, and its effect is exerted in G_2-M. In summary, PISSLRE has sequence characteristics of a TSG and expression patterns of an oncogene, therefore, further studies are required to elucidate the properties of the PISSLRE gene.

Two distinct genes localised to 16q24 have been isolated by differential screening of tumour and benign or normal tissue. One gene designated breast basic conserved gene (BBC1), has been mapped to 16q24.3 using chromosome 16 somatic cell hybrids (Cleton-Jansen *et al.* 1995). BBC1 was demonstrated to have decreased mRNA expression in malignant breast tumours compared to benign tumours (Adams *et al.* 1992b). BBC1 has homology to rat (Olvera and Wool, 1994) and human ribosomal protein L13. The function of this protein is possibly as a transcription regulator since a DNA binding hydrophobic leucine zipper motif characteristic of transcription factors, was demonstrated in the sequence of ribosomal protein L13 (Tsurugi and Mitsui, 1991). Thus, it is likely that BBC1 is not a suitable candidate TSG.

The renal dipeptidase gene, DPEP1, selected as a potential TSG, was isolated from a cDNA library constructed from mature kidney cDNA which was subtracted with an excess of Wilms' tumour mRNA (Austruy *et al.* 1993a). The cDNA clones were selected according to a differential pattern of expression, ie. positive with RNA from mature kidney and negative with RNA from Wilms' tumour RNA). This gene has been localised to 16q24.3 by physical mapping with the use of somatic cell hybrids (Austruy *et al.* 1993b). DPEP1 is a zinc metalloproteinase located in the kidney membrane and hydrolyses dipeptides. Although the DPEP1 cDNA clone was isolated as a potential TSG from the subtracted kidney cDNA/Wilms' tumour mRNA library, its effective role in tumourigenesis has yet to be demonstrated. Thus, DPEP1 is not likely to be the TSG localised at 16q24.3.

1.7 DISEASES LOCALISED TO HUMAN CHROMOSOME 16

Two diseases localised to human chromosome 16, Fanconi anaemia and familial Mediterranean fever, will be discussed in this section.

1.7.1 Fanconi Anaemia

Fanconi anaemia (FA) is a rare autosomal recessive disorder with an estimated frequency of homozygotes in the order of 1 : 350,000, and heterozygotes 1 : 300 (Digweed, 1993). FA is usually a fatal disease with a mean survival of 16 years. Manifestation of the disease is extremely pleomorphic and may include congenital malformations, abnormal skin pigmentation and skeletal (radius and thumb dysplasia) and renal anomalies (Auerbach and Allen, 1991; Strathdee and Buchwald, 1992; Alter, 1993; Liu *et al.* 1994). Cells derived from FA patients exhibit an increase in chromosome instability and hypersensitivity to bifunctional DNA cross-linking agents, such as diepoxybutane and mitomycin C (Auerbach, 1993). FA is commonly thought of as a disorder of DNA repair, but the critical molecular target of the cross linking agents is still unknown. This disorder is associated with progressive bone marrow failure and a 15,000 fold increased risk of developing acute myeloid leukaemia, although a variety of solid tumours have also been observed (Auerbach *et al.* 1989; Auerbach and Allen, 1991; Liu *et al.* 1994). Thus, FA is a model disease to study for insight into the causes of leukaemia and cancer.

FA has diverse clinical symptoms and is considered to be genetically heterogeneous. Patients are classified into complementation groups which are defined by cell fusion studies that take advantage of the cross-linking agent sensitivity of the FA cells. Hybrids where the cross-linker sensitivity is corrected are assumed to result from the combination of cells from different complementation groups. Hybrids that still behave like the patients cells are produced by the fusion of cells from the same complementation group. Complementation analysis of lymphoblastoid cell fusion hybrids has revealed the existence of four different complementation groups (FA-A to FA-D) (Strathdee *et al.* 1992b) in seven lymphoblastoid cell lines from unrelated FA patients. In 1995, FA lymphoblastoid cell lines representing each of the four known complementation groups were used to classify 13 unrelated FA patients through cell fusion and complementation analysis (Joenje *et al.* 1995). One cell line was demonstrated to complement all four reference cell lines and therefore represents a new complementation group, designated FA-E. This result now implies that there are at least five complementation group genes associated with FA.

Complementation groups are thought to reflect genetic heterogeneity with each group related to a distinct gene playing a part in a biochemical pathway that, when interrupted, leads to a specific type of disease. The genetic heterogeneity of the disease has been a major obstacle in the positional cloning of FA genes by classical linkage analysis. The FAC gene was identified by isolation of its cDNA from an episomal expression library that complemented the cross-linker hypersensitivity of the transfected FA-C lymphoblastoid cell line HSC536 (Strathdee *et al.* 1992c). However, this approach of functional complementation is yet to identify genes for the other complementation groups.

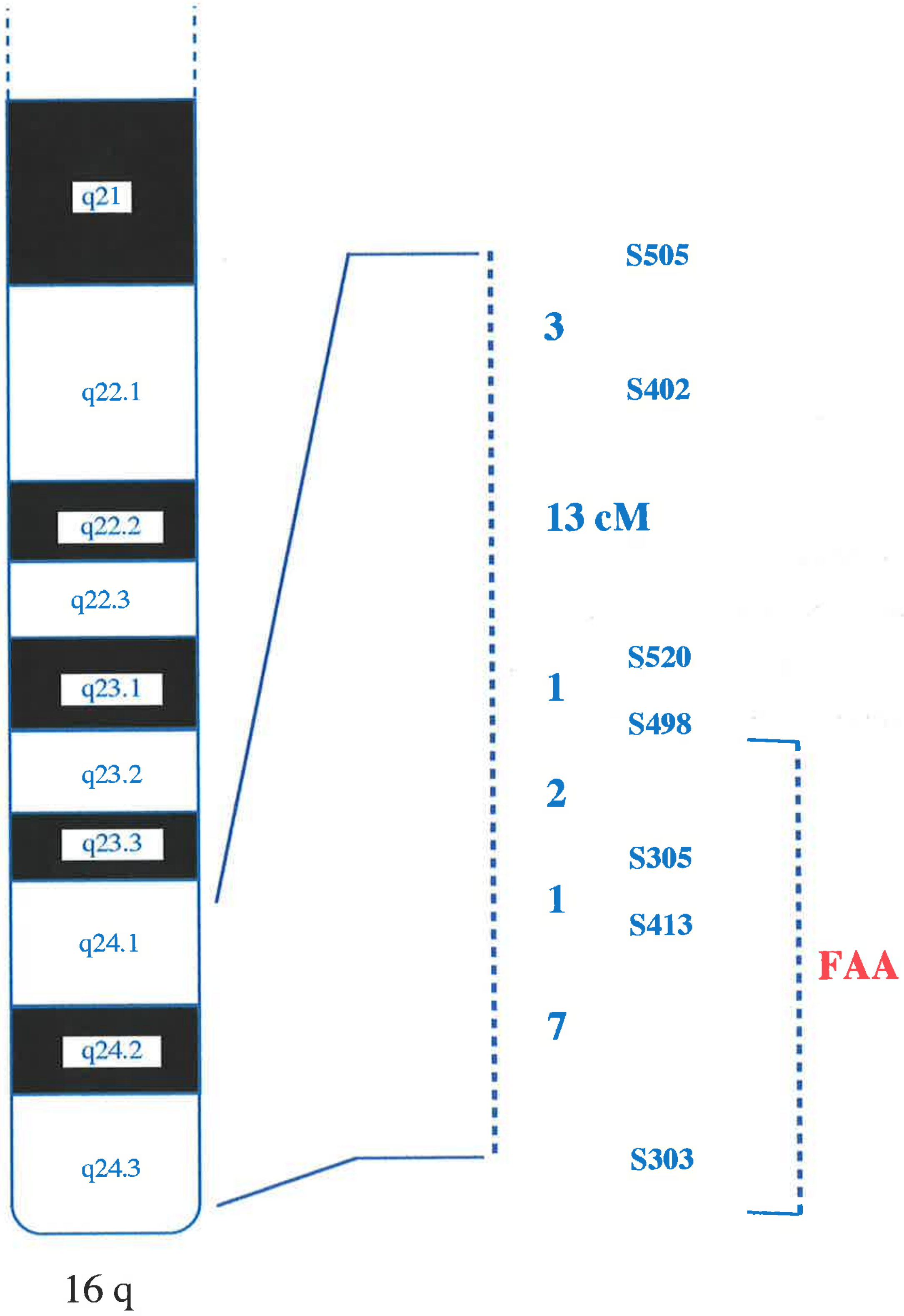
The FAC gene was mapped to chromosome 9q22.3 by *in situ* hybridisation (Strathdee *et al.* 1992c). The genomic structure of the 1674 nucleotide coding region of FAC was elucidated and is composed of 14 exons (Gibson *et al.* 1993a) encoding a 63 kD cytosolic protein of unknown function. The FAC gene was also screened for mutations in FA patients by chemical cleavage of FAC cDNA (Gibson *et al.* 1993b) and single strand conformation polymorphism analysis of individual exons (Verlander *et al.* 1994). Among the six pathogenic mutations identified, loss of exon 4 from the FAC transcript (Verlander *et al.* 1994) accounted for the majority of FA in non-Ashkenazi Jewish families. It was proposed that this mutation was associated with a severe phenotype of multiple congenital malformations and early onset of haematological disease (Verlander *et al.* 1994). However, the number of FA patients with specific FA mutations is still too limited to allow meaningful correlations of genotype with clinical phenotype.

The most prevalent complementation group is FA-A which has been observed in two thirds of classified FA patients (Buchwald, 1995). In November 1995, Pronk and coworkers published a study in which FA-A families, classified by complementation analysis, were used to localise the FAA gene by classical linkage analysis (Pronk *et al.* 1995). A genome-wide search was undertaken in 9 families and evidence for linkage of FAA to microsatellite marker D16S520 at 16q24.3 was obtained. An additional six markers from the region were then typed in the families. The linkage map and the markers used are shown in figure 1.6 (Gyapay *et al.* 1994; Doggett *et al.* 1995). The highest lod score, 8.01 at $\theta = 0.00$, was obtained with the microsatellite marker D16S305 but no recombinants were detected with any of the three most distal markers on the map. Recombination data indicated that the location of FAA is distal to D16S498 which is a genetic distance of 10.5 cM from D16S303, but the distance from D16S303 to the telomere is unknown. Thus, the FAA gene was localised to the 16q24.3 chromosomal region.

Linkage disequilibrium analysis provided further evidence for the location of FAA to this chromosomal region. The highest reported incidence of FA, 1 in 22000, is in the Afrikaner population of South Africa and is most likely due to a founder effect (Rosendorff *et al.* 1987). Twenty-one families which were genotyped showed no recombination with the most distal marker, D16S303. This locus also exhibited strong allelic association in 67% of the disease chromosomes. Significant allelic association was also observed with marker D16S305 in 62% of the disease chromosomes. Moreover, marker D16S413 was present in 51% of the disease chromosomes. The disequilibrium coefficient was lower for the proximal markers D16S520 and D16S498. In the 31 disease bearing chromosomes where the marker haplotype was determined, 15 carried the particular haplotype for the three most distal markers D16S305-D16S413-D16S303, compared with none of the 29 non-disease bearing chromosomes. These data supported the suggestion that a founder mutation had contributed to the high incidence of FA in the Afrikaner population of South Africa.

Figure 1.6

Linkage map of microsatellite markers from chromosome 16q24 used to localise the FAA gene (Pronk *et al.* 1995). Genetic distances are indicated in centiMorgans (cM). Somatic cell hybrid analysis places the three most distal markers in band 16q24.3. The FAA gene is localised distally to D16S498 from recombination data.


q21
q22.1
q22.2
q22.3
q23.1
q23.2
q23.3
q24.1
q24.2
q24.3
16 q
S505
3
S402
13 cM
S520
1
S498
2
S305
1
S413
FAA
7
S303

Thus, the Fanconi anaemia complementation group A gene has been localised to the 16q24.3-qter chromosomal region. This region does not contain any obvious candidate genes and the localisation of FAA is the first step toward the cloning of the gene using the approach of positional cloning. Genomic DNA clones localised in the detailed physical map of the 16q24 chromosomal region will be of use for the isolation and identification of transcripts which may be candidates for the FAA gene. The identification of the gene may provide an insight into the pathway associated with the prevention or repair of DNA damage.

1.7.2 Familial Mediterranean Fever

Familial Mediterranean fever (FMF) is an autosomal recessive disorder of unknown pathogenesis. McKusick reported that the first cases were described in 1954, were from Lebanon, and were mostly of Armenian background (McKusick, 1992, No. 249100). This inflammatory disease is characterised by recurring attacks of fever with sterile peritonitis, pleurisy, and/or synovitis (Sohar *et al.* 1967). Patients with FMF may also develop systemic amyloidosis, characterised by extracellular deposits of an abnormal degradation-resistant protein called amyloid, which makes it a potentially fatal illness. Treatment of patients with colchicine is known to prevent the recurring attacks of fever in 90% of FMF patients. It also prevents renal amyloidosis in 67% of patients who are treated early in the course of the disease (Zemer *et al.* 1986; Livneh *et al.* 1994).

This disease occurs primarily in individuals of Mediterranean descent, especially in non-Ashkenazi Jews (Sohar *et al.* 1967), Armenians (Schwabe and Peters, 1974), Anatolian Turks (Ozdemir and Sokmen, 1969) and Middle Eastern Arabs (Barakat *et al.* 1986). The frequency of the disease gene among these populations is extraordinarily high, reaching 1 in 22 among Jews in North Africa and 1 in 14 among immigrant Armenians in Los Angeles (Rogers *et al.* 1989). Although the biochemical basis of this disorder is unknown, the clinical manifestations of FMF suggest a lesion in a molecule important to the understanding of inflammation in general (Aksentijevich *et al.* 1993b).

For several years there has been an ongoing effort to map the FMF susceptibility gene [designated the gene symbol MEF (Aksentijevich *et al.* 1993b)] by molecular genetic techniques. The initial studies involved the identification of protein polymorphisms. The serum amyloid A gene was initially suspected to be abnormal in FMF patients (Sack *et al.* 1988). To test the role of serum amyloid A (SAA) and P (APCS) in FMF patients, Shohat *et al* (1990a) studied 17 informative families (15 Armenians and 2 non-Ashkenazi Jews) and 8 FMF patients with amyloidosis using a candidate gene approach. Their data showed no evidence for any FMF-associated polymorphism in any of 41 Armenian and Jewish FMF patients. The results of Shohat *et al* (1990a) excluded close linkage between SAA and MEF, and between APCS and MEF. Thus SAA and APCS were not the genes causing FMF and defects in them did not seem to contribute to the pathogenesis of FMF amyloidosis. Sack *et al* (1991) also studied the relationship between FMF and the SAA gene (localised to chromosome 11p) by typing alleles of an (AC)n repeat and a RFLP in the SAA gene cluster in Israeli FMF kindreds. The results from linkage analysis eliminated a minimum of 10.4 cM including and surrounding the SAA gene cluster as the site of the MEF gene.

Subsequently, general linkage studies using RFLPs were undertaken. Shohat *et al* (1990b) screened 14 Armenian families from California for 19 polymorphic markers (4 red cell antigens, 10 red cell enzymes and 5 serum proteins). These markers were located on 12 different human chromosomes. The results showed no significant association of FMF with any of the markers studied. The distribution in patients and controls was very similar for most markers. Linkage was rejected with 14 markers which excluded the MEF gene from those parts of the human genome. Linkage could not be excluded with 5 other markers (one of which was phosphoglycolate phosphatase located on chromosome 16pl3) which was probably due to the small number of families being informative for these markers.

Thus, studies using protein polymorphism and RFLP analysis had been conducted, but localisation of the FMF susceptibility gene had not been achieved. With the development of polymorphic microsatellite genetic markers and the construction of detailed genetic maps, the localisation of the MEF gene was more amenable. Genetic linkage studies using these microsatellite markers were the first step potentially leading to the positional cloning of the gene responsible for FMF.

Aksentijevich *et al* (1991) performed linkage analysis on 18 FMF families of North African and Iraqi descent, from Israel. Linkage analysis with 95 (AC)n DNA markers excluded about 30% of the human genome, with the majority of chromosomes 1, 9, 10 and 15 also being excluded. Four markers (D17S74, D17S40, D17S35 and GH) on the long arm of chromosome 17 appeared to be linked to the FMF susceptibility gene. Multipoint linkage analysis yielded a lod score of 3.54 at a 15 cM map distance from D17S40. In this report, Aksentijevich *et al* (1991) presented a most likely map order: D17S74 - GH-Dl7S351-D17S40 - MEF.

After demonstrating linkage of the MEF gene to the markers on chromosome 17q, Pras *et al* (1992) and Shohat *et al* (1992) independently reported the MEF gene to be linked to markers on chromosome 16 at band pl3.3. Multipoint linkage analysis placed MEF centromeric to D16S84, with a maximal lod score greater than 19, but did not establish flanking markers for MEF.

Clinical differences were exhibited for at least one of the major signs of FMF which included abdominal pain, joint pain, chest pain and/or skin eruptions, and one of the minor signs of increased erythrocyte sedimentation rate, leucocytosis and/or elevated serum fibrinogen levels (Eliakim *et al.* 1981; Sohar *et al.* 1967) in the Armenian FMF patients and the non-Ashkenazi Jewish patients. Because of the clinical differences in these two populations, the possibility of genetic locus heterogeneity was examined. For this purpose, Shohat *et al* (1992) typed FMF families from both Armenian and non-Ashkenazi Jewish populations with the markers on chromosomes 16p and 17q. Their results indicated that the gene for FMF was linked to the α-globin complex at 16pl3.3 in both ethnic groups. The present data support the view that a single locus causes FMF. Although there is no evidence for locus heterogeneity in FMF, allelic heterogeneity is possible as there are consistent clinical differences among the various affected populations (Sohar *et al.* 1967; Ozdemir and Sokmen, 1969; Schwabe and Peters, 1974; Barakat *et al.* 1986).

In the report of Aksentijevich *et al* (1993a), the previous localisation of MEF on chromosome 17q (Aksentijevich *et al*, 1991) was proved to be a "false positive" (type I) error. They indicated that chromosome 17q did not encode a major FMF susceptibility gene for some of the families, nor did it encode a disease modifying gene (second locus). Subsequent analysis of 31 non-Ashkenazi Jewish families with newly isolated (AC)n repeat and RFLP markers localised the MEF gene between VK5 (D16S94) on the telomeric side, and 24.1 (D16S80) on the centromeric side, in a genetic interval of approximately 9 cM (Aksentijevich *et al.* 1993b). Subsequent studies by Fischel-Ghodsian *et al* (1993) further refined this interval between 24.1 (D16S80) and SM7 (D16S283), at a distance of approximately 3 cM centromeric from D16S283.

1.8 AIMS OF THE THESIS

The aim of this project is to construct a detailed physical map of the 16q24 chromosomal region. The focus to map the 16q24 chromosomal region is in view of the high gene-density and the localisation of two disease genes to this region. The gene for Fanconi anaemia group A (FAA) has been localised to 16q24.3-qter by linkage analysis. In addition, a tumour suppressor gene (TSG) is located in this region, which has been demonstrated by loss of heterozygosity (LOH) studies of the 16q24 region, further refined to 16q24.3-qter, in sporadic breast tumours. A detailed physical map forms the basis for attempts to positionally clone candidate genes for disease genes, including FAA and the TSG, mapped to specific chromosomal regions.

The 16q24 chromosomal region is poorly represented by cloned DNA fragments, thus, the first aim of the project involves the identification and mapping of genomic cosmid clones to this chromosomal region. An Alu PCR approach will be utilised in the first instance to identify and map cosmids to the 16q24 chromosomal region.

The next aim of the physical mapping is the utilisation of cosmid clones localised to the 16q24 region to identify transcribed sequences encoded by the cosmids, using the approach of direct cDNA selection. The purpose of isolating transcribed sequences is for the construction of a transcript map of the region which will benefit the positional cloning approach for the identification of disease genes, including FAA and the TSG, localised to this region of interest.

Since the TSG and FAA genes are located in the 16q24.3-qter region, a collaborative effort will focus on the construction of a cosmid contig of this region to assist the positional cloning of these genes. STSs and cosmid ends isolated from cosmids already localised in this region will be used to identify additional cosmids for this contig.

The final aim of the 16q24 project is the characterisation of novel transcripts localised to the 16q24.3-qter region, which are possible candidates for disease genes localised to this region, including the TSG and FAA.. This will involve the determination of the full-length sizes of the transcripts by Northern analysis, and expression patterns in various tissues by reverse transcriptase-PCR. The isolation of full-length clones of the transcripts will also be performed. The cDNA clones will be sequenced and compared to sequences in nucleotide databases to identify overlapping sequences that may extend the sequence of the clones. Homologies of the transcribed sequences to any known genes or protein motifs will also be investigated to determine whether they possess any relevant functions.

The transcripts will also be investigated for disease causing mutations, a necessary step in the positional cloning of the tumour suppressor and FAA genes. Single stranded conformation polymorphism analysis of the transcripts in breast tumours displaying LOH at 16q24.3-qter will be performed. Members of the Fanconi Anaemia/Breast Cancer consortium will perform SSCP analysis of the transcripts in Fanconi anaemia patient samples.

The approach of direct cDNA selection will also be applied to a second project which involves the positional cloning of the MEF gene, localised to the 16p13.3 region by linkage analysis, which is responsible for familial Mediterranean fever (FMF). The FMF consortium has constructed a YAC/cosmid contig encompassing the FMF candidate region. The aim of my work is to identify and isolate transcripts mapping to this region which will aid the construction of a transcript map of the region. This transcript map will benefit the positional cloning of the MEF gene localised to this region. This positional cloning approach may lead to the eventual identification of the gene and mutations causing familial Mediterranean fever.

CHAPTER 2

Materials and Methods

2.1 INTRODUCTION 77

2.2 MATERIALS 78

2.2.1 Enzymes 78

2.2.2 Electrophoresis 78

2.2.3 Radiochemicals 79

2.2.4 Buffers and Solutions 79

2.2.5 Bacterial Media 82

2.2.5.1 Liquid Media 82

2.2.5.2 Solid Media 82

2.2.6 Antibiotics 83

2.2.7 Bacterial Strains 83

2.2.8 Vectors 83

2.2.9 Miscellaneous Materials, Kits and Fine Chemicals 84

2.3 METHODS 85

2.3.1 DNA Isolation 85

2.3.1.1 Large Scale Isolation of Plasmid/Cosmid DNA 85

2.3.1.1.1 Protocol I 85

2.3.1.1.2 Protocol II 86

2.3.1.2 Small Scale Isolation of Plasmid/Cosmid DNA 87

2.3.1.2.1 Protocol I 87

2.3.1.2.2 Protocol II 87

2.3.1.3 Isolation of Peripheral Lymphocyte DNA 88

2.3.1.4 Purification of DNA 89

2.3.1.4.1 Phenol/Chloroform Extraction of DNA 89

2.3.1.4.2 Ethanol Precipitation of DNA 90

2.3.1.4.3 Spectrophotometric Estimation of Nucleic Acid Concentration 90

2.4 SUBCLONING OF HUMAN DNA SEQUENCES 91

2.4.1 Preparation of Vector DNA and Human DNA Inserts 91

2.4.2 Dephosphorylation of Vector DNA 91

2.4.3 Ligation Reactions 92

2.4.4 Competent Cells and Transformation 92

2.5 ENZYME DIGESTION, GEL ELECTROPHORESIS AND SOUTHERN BLOT ANALYSIS 94

2.5.1 Restriction Endonuclease Digestion of DNA 94

2.5.2 Agarose Gel Electrophoresis of DNA 94

2.5.3 Molecular Weight Markers 95

2.5.4 ^{32}P Radio-isotope Labelling of DNA 95

2.5.4.1 Primer Extension 95

2.5.4.2 Pre-reassociation of Repetitive DNA 95

2.5.5 Transfer of DNA to Nylon Membranes 96

2.5.5.1 Plaque/Colony Lifting 96

2.5.5.2 Southern Blotting 96

2.5.5.3 Transfer of DNA of Low Molecular Weight 96

2.5.5.4 Transfer of DNA of High Molecular Weight 97

2.5.6 Prehybridisation, Hybridisation and Washing 97

2.5.6.1 Using Probes Labelled by Primer Extension 97

2.6 POLYMERASE CHAIN REACTION (PCR) 98

2.6.1 PCR Primers 98

2.6.2 Oligonucleotide Deprotection and Cleavage 99

2.6.2.1 Oligonucleotide Purification n-butanol Method 99

2.6.3 Subcloning PCR Products 99

2.7 DNA SEQUENCING 101

2.7.1 Automated DNA Sequencing 101

2.7.2 Double stranded DNA Sequencing 101

2.7.3 PCR Product Sequencing 101

2.7.4 Loading and Running the Sequencing Gel 102

2.8 COSMID VECTORS 103

2.8.1 sCos Vectors 103

2.8.2 Construction of a Chromosome 16 Specific Ordered Cosmid Library 104

2.8.2.1 Repetitive Sequence Fingerprinting and Assembly of Contigs 104

2.8.2.2 Construction and Screening of High Density Cosmid Grids 105

2.1 INTRODUCTION

Most of the methods described in this chapter were well established and used routinely in the Department of Cytogenetics and Molecular Genetics at the Women's and Children's Hospital (Adelaide, Australia). In this chapter the materials and methods used in common throughout the whole project will be described. The specialist technologies and approaches will be described in their corresponding chapters.

All enzymes were obtained from commercial sources and were used in accordance with the manufacturer's specifications. All chemicals and solvents were of analytical grade.

2.2 MATERIALS

2.2.1 Enzymes

All restriction endonucleases were obtained from New England Biolabs (Beverley, Massachusetts, USA)

Other enzymes were obtained from the following sources:

E.coli DNA Polymerase I	Amersham, Australia
(Klenow Fragment)	
Heat killed^{TM} phosphatase	Epicentre Technologies, Wisconsin, USA
Lysozyme	Sigma Chemical Co. St. Louis, Missouri, USA
Proteinase K	Sigma Chemical Co. St. Louis, Missouri, USA
RNase A	Boehringer, Mannheim, Germany

The stock solution (10 mg/ml) was incubated at 100°C for 10 min to inactivate any DNAase activity.

T4 DNA Ligase	Promega, USA
T4 Polynucleotide Kinase	Pharmacia Biotech, Uppsala, Sweden
Taq Polymerase	Boehringer Mannheim, Germany

2.2.2 Electrophoresis

The reagents were obtained from the following companies:

Acrylamide		Biorad, California, USA
Agarose	- Type N.A	Pharmacia Biotech, Uppsala, Sweden
	- Low Gelling Temperature	Progen, Queensland, Australia
Ammonium Persulphate (APS)		Acros Organics, New Jersey, USA
Bisacrylamide		Biorad, California, USA
(N,N'-methylene-bis-acrylamide)		
Bromophenol Blue		B.D.H. Chemicals LTD, Poole, Dorset England
Ethidium Bromide		Boehringer Mannheim, Germany

Molecular weight markers:	Bresatec, Adelaide, Australia
Spp-1 Bacteriophage restricted with EcoRI	
puc19 DNA restricted with HpaII	
N, N, N, N-tetramethylethylene diamine (TEMED)	Biorad, California USA
Urea	Ajax, NSW, Australia
Xylene Cyanol	Tokyo Kasei, Tokyo, Japan

2.2.3 Radiochemicals

alpha ^{32}P-dCTP, 3000 Ci/mmole	Radiochemical Centre, Amersham
gamma ^{32}P-ATP, 5000 Ci/mmole	Radiochemical Centre, Amersham

2.2.4 Buffers and Solutions

Solution and Equipment Sterilisation

All solutions were made with distilled water and sterilised by autoclaving at 120°C for 15-30 minutes. Microcentrifuge tubes and disposable pipette tips were also autoclave-sterilised.

Buffers and solutions routinely used in this study were as follows:

Denhardt's solution	0.1% (w/v) Ficoll
	0.1% (w/v) polyvinyl pyrrolidine
	0.1% (w/v) BSA
Formamide Loading Buffer	92.5 % (v/v) formamide
	20 mM EDTA
	0.1% (w/v) xylene cyanol
	0.1% (w/v) bromophenol blue

10 x Loading Buffer	50% (v/v) glycerol
	1% (w/v) SDS
	100 mM EDTA
	0.1% xylene cyanol
	0.1% (w/v) bromophenol blue
10 x Ligation Buffer	0.5 M Tris-HCl, pH7.4
	0.1 M $MgCl_2$
	0.1 M dithiothreitol
	10 mM spermidine
	10 mM ATP
	1 mg/ml bovine serum albumin
10 x MOPS Buffer	0.2 M MOPS
	50 mM sodium acetate
	1 mM EDTA
Phosphate Buffered Saline	130 mM NaCl
(PBS)	10 mM $NaHP0_4$
	10 mM NaH_2P0_4 pH7.2
10 x PCR mix (Boehringer, Mannheim)	20 mM Tris-HCl pH8.0
	0.1 mM EDTA
	1 mM DTT
	100 mM KCl
	50% (v/v) glycerol
	0.5% (v/v) Tween 20
	0.5% (v/v) Nonidet P40
	plus 2 mM dATP, dGTP, dTTP, dCTP.

20 x SSC	3 M NaCl
	0.3 M tri-sodium citrate pH7.0
20 x SSPE	3 M NaCl
	0.18 M $\mathrm{NaH_2PO_4.2H_2O}$
	20 mM EDTA pH7.4
TBE	89 mM Tris-base
	89 mM boric acid
	2.5 mM EDTA pH8.3
TE	10 mM Tris-HCl pH7.5
	0.1 mM EDTA
TE + glucose	50 mM Tris-HCl pH8.0
	20 mM EDTA
	50 mM glucose
TES	25 mM Tris-HCl pH8.0
	10 mM EDTA
	15% (w/v) sucrose

2.2.5 Bacterial Media

2.2.5.1 Liquid Media

All liquid media were prepared in millipore water and sterilised by autoclaving, antibiotics and other labile chemicals were added after the solution had cooled to 50°C.

The compositions of the various media were as follows:

Luria (L) - Broth	1% (w/v) Bacto-tryptone (Difco) 0.5% (w/v) Bacto yeast extract (Difco) 1% (w/v) NaCl, pH to 7.5 with NaOH.
2 x YT	1.6% (w/v) Bactotryptone 1% (w/v) Bacto yeast extract 0.5% (w/v) NaCl, pH to 7.5 with NaOH.

2.2.5.2 Solid Media

L-Agar	L-broth 1% (w/v) Bacto agar
L-Amp	L-broth 1% (w/v) Bacto agar ampicillin (100 g/ml)
L-Kanamycin	L-broth 1% (w/v) Bacto agar kanamycin (50 g/ml)

2.2.6 Antibiotics

Ampicillin	Sigma Chemical Co.
Kanamycin	Boehringer Mannheim

2.2.7 Bacterial Strains

The bacterial strains used in this project are listed below:

E.coli XL1-Blue : *sup*E44 *hsd*R17 *rec*A1 *end*A1 *gyr*A46 *thi rel*A1 *lac*$^-$ F'[*pro*AB$^+$ *lac*I^q lacZΔM15 Tn10 (tet^r)] host for recombinant plasmids and M13 bacteriophage, purchased from Stratagene.

E.coli DH5α: *sup* 44 Δ*lac* U169 (p80 *lac* ZΔM15) *hsd* R17 *rec* A1 *end* A1 *gyr* A96 *thi*-1 relA1 host for recombinant plasmids, obtained from the *E. coli* Genetic Stock Centre, Yale University, New Haven.

Stock cultures of these (and plasmid transformed bacteria) were prepared by dilution of an overnight culture with an equal volume of 80% glycerol and stored at either -20°C, or -80°C for long term storage. Single colonies of bacteria, obtained by streaking the glycerol stock onto agar plates of suitable medium were used to inoculate liquid growth medium, and the bacterial cultures were grown at 37°C with continuous shaking to provide adequate aeration.

2.2.8 Vectors

The vectors used in this study are the following:

Phagemid Vectors:

pBLUESCRIPT SKII +	Stratagene, La Jolla, California, USA

Plasmid Vectors:

puc19	Bresatec, Yanisch-Perron *et al.* (1985)

2.2.9 Miscellaneous Materials, Kits and Fine Chemicals

ABI Prism^{TM} Dye Terminator Cycle Sequencing Kit	Perkin Elmer
ABI Prism^{TM} Dye Primer Cycle Sequencing Kit	Perkin Elmer
Chemicals for Oligonucleotide Synthesis	Applied Biosystems
Deoxynucleotides (dNTPs)	Boehringer Mannheim
Dideoxynucleotides (ddNTPs)	Boehringer Mannheim
Dideoxysequencing Kits	Perkin Elmer
Guanidinium isothiocyanate	Sigma Chemical Co.
Human placental DNA	Sigma Chemical Co.
Hybond N+^{TM} Nylon Membrane	Amersham
Isopropyl thio-B-D-galactoside (IPTG)	Boehringer Mannheim
Mega Priming Oligolabelling Kits	Amersham
Phenol	Wako
Photographic film	Polaroid
pGEM-T cloning kit	Promega
Salmon sperm DNA	Sigma Chemical Co.
SDS	Sigma Chemical Co.
Spermidine	Sigma Chemical Co.
Streptavidin Coated Magnetic Beads	Dynal
TRIzol	Gibco BRL
Wizard prep columns	Promega
X-ray film	Kodak

2.3 METHODS

2.3.1 DNA Isolation

2.3.1.1 Large Scale Isolation of Plasmid DNA and Cosmid DNA

2.3.1.1.1 Protocol I:

(modification of Sambrook *et al.* 1989)

10 ml Luria Bertoni (LB) medium containing ampicillin (50 μg/ml) was inoculated with a single fresh bacterial colony. The culture was incubated at 37°C for 5-7 hours with vigorous shaking, and then transferred to 400 ml LB containing ampicillin (50 g/ml). After overnight incubation at 37°C with vigorous shaking, the culture was transferred to 50 ml centrifuge tubes. The tubes were left on ice for 15 minutes and then spun at 3000 rpm for 15 minutes in a Jouan CR3000 centrifuge at 4°C. The supernatant was discarded and the cell pellets were gently resuspended in a total of 1.2 ml TE and glucose containing 60 μl of 80 mg/ml lysozyme. The cell suspension was left at room temperature for 20 minutes and on ice for 1 minute. 1.2 mls of 0.2 M NaOH/1% SDS was added to the cell suspension, gently mixed and incubated on ice for a further 5 minutes. 1.8 ml of ice cold 3 M potassium acetate (pH4.3) was then added, and mixed by inversion. After resting on ice for 10 minutes, the lysed cell debris was cleared by centrifugation in a Beckman J2-21M/E centrifuge with a JA20 rotor at 15,000 rpm for 15 minutes.

The supernatant was mixed with 2 volumes of ethanol and after 5 minutes at room temperature, precipitated nucleic acids were pelleted by centrifugation at 15,000 rpm for 15 minutes. The DNA pellet was washed twice in 2 ml of 70% ethanol, air-dried and resuspended in 200 μl TE. To eliminate RNA in the DNA preparation, 10 μl of 1 mg/ml RNase was added to the DNA solution and incubated at 37°C for 1 hour. To eliminate proteins in the DNA preparation, 100 µl of 3 x proteinase K buffer, 10 µl of 10% SDS and 2 µl of 10 mg/ml proteinase K were added to the DNA solution and incubated for 1 hour at 37°C. Following incubation, the DNA was phenol extracted, ethanol precipitated and dissolved in 200 µl of TE.

2.3.1.1.2 Protocol II:

(QIAGEN plasmid/cosmid purification)

A single bacterial colony was inoculated into 500 ml conical flasks containing 200 mls of LB and 200 µl of thawed ampicillin (50 mg/ml), and incubated overnight at 37°C with shaking at 225 rpm. DNA templates were prepared according to the manufacturer's instructions using the plasmid/cosmid purification protocol and Qiagen tip-20 columns (Qiagen Plasmid Handbook for Plasmid Mini Kit, 1993). Overnight bacterial cultures were centrifuged at 3,000 rpm for 10 minutes at 4°C. The supernatant was removed and the tubes inverted for 5 minutes prior to returning right-side up. The bacterial pellet was resuspended in a total of 1.5 mls of resuspension buffer (stored at 4°C), the pellet gently dislodged with a tip, and the contents equally divided into 1.5 ml tubes. A total of 1.5 mls of lysis buffer was added and the contents were mixed gently by inversion and incubated at room temperature for 5 minutes. To this a total of 1.5 mls of neutralisation buffer (stored at 4°C) was added, the contents mixed gently by inversion, and incubated on ice for 5-10 minutes. The tubes were microcentrifuged at 13,200 rpm for 15 minutes. The supernatant was recovered into an eppendorf tube, and then microcentrifuged as before.

Qiagen columns were equilibrated with 1 ml of equilibration buffer, and each supernatant loaded onto a Qiagen column. The columns were washed 4 times with 1 ml wash buffer. The DNA was eluted with 800 µl elution buffer and collected into eppendorf tubes placed beneath the columns. The DNA was precipitated with 700 µl isopropanol, mixed well and centrifuged at 13,200 rpm for 15 minutes. The pellet was washed with 1 ml 70% ethanol and centrifuged as before for 10 minutes. The supernatant was discarded, the pellet air-dried, and resuspended in 50 µl water.

2.3.1.2 Small Scale Isolation of Plasmid DNA

2.3.1.2.1 Protocol I:

(modification of Birnboim and Doly, 1979)

A single bacterial colony was inoculated to 1.5 ml of LB medium containing ampicillin (50 g/ml) in a 10 ml tube. The culture was incubated at 37°C overnight with vigorous shaking. The culture was transferred to an eppendorf tube and spun in an eppendorf centrifuge at 14,000 rpm for 2 minutes. After discarding the supernatant, the cell pellet was well resuspended in 100 µl of cold fresh TES medium and 0.25 ml of 100 mg/ml lysozyme. The cell suspension was left at room temperature for 5 minutes before 200 µl of 0.2 N NaOH/1% sodium dodecyl sulphate (SDS) was added and mixed well. The mixture was incubated on ice for 5 minutes and then 150 µl of cold 3 M sodium acetate (pH4.6) was added. After 5 minutes on ice the mixture was spun in an eppendorf centrifuge for 10 minutes. The supernatant was carefully drawn off and the nucleic acids precipitated with 2 volumes of ethanol. The DNA pellet was finally washed twice with 70% ethanol, air-dried and resuspended in 50 µl sterile H_2O.

2.3.1.2.2 Protocol II:

(QIAGEN plasmid purification)

A single bacterial colony was inoculated into 50 ml centrifuge tubes containing 10 mls of LB and 10 µl of thawed ampicillin (50 mg/ml), and incubated overnight at 37°C with shaking at 225 rpm. DNA templates were prepared according to the manufacturer's instructions using the plasmid purification protocol and Qiagen tip-5 columns (Qiagen Plasmid Handbook for Plasmid Mini Kit, 1993). Overnight bacterial cultures were centrifuged at 3,000 rpm for 10 minutes at 4°C. The supernatant was removed and the tubes inverted for 5 minutes prior to returning right-side up. The bacterial pellet was resuspended in 300 µl of resuspension buffer (stored at 4°C), the pellet gently dislodged with a tip, and the contents added to 1.5 ml tubes containing 300 µl lysis buffer. The contents were mixed gently by inversion and incubated at room temperature for 5 minutes. To this, 300 µl of neutralisation buffer (stored at 4°C) was added, the contents mixed gently by inversion, and incubated on ice for 5-10 minutes. The tubes were microcentrifuged at 13,200 rpm for 15 minutes. The supernatant was recovered into an eppendorf tube, and then microcentrifuged as before.

Qiagen columns were equilibrated with 1 ml of equilibration buffer, and each supernatant loaded onto a Qiagen column. The columns were washed twice with 1 ml wash buffer. The DNA was eluted with 800 µl elution buffer and collected into eppendorf tubes placed beneath the columns. The DNA was precipitated with 700 µl isopropanol, mixed well and centrifuged at 13,200 rpm for 15 minutes. The pellet was washed with 1 ml 70% ethanol and centrifuged as before for 10 minutes. The supernatant was discarded, the pellet air-dried, and resuspended in 20 µl water.

2.3.1.3 Isolation of Peripheral Lymphocyte DNA

(modification of Wyman and White, 1980)

Blood samples were collected in 10 ml tubes containing EDTA and were allowed to cool to room temperature before being stored at -20°C. For DNA lymphocyte isolation, the frozen blood sample was thawed and transferred to a 50 ml centrifuge tube. Cell lysis buffer was added to 30 mls. After mixing, the tube was left on ice for 30 minutes. The cell suspension was spun in the Jouan centrifuge at 3500 rpm for 15 minutes at 4°C. The supernatant was aspirated down to 5 mls, then cell lysis buffer was added again to 30 mls. Centrifugation was repeated.

The supernatant was carefully aspirated and 3.25 ml Proteinase K buffer, 0.5 ml of 10% SDS and 0.2 ml of Proteinase K (10 mg/ml) were added and well mixed with the cell pellet. The tube containing the cell suspension was sealed with parafilm, secured on a rotating wheel (10 rpm) and incubated overnight at 37°C. DNA extraction was performed twice with phenol and twice with phenol/chloroform. Following ethanol precipitation the DNA pellet was dissolved in 0.1 ml of TE.

2.3.1.4 Purification of DNA

2.3.1.4.1 Phenol/Chloroform Extraction of DNA

Solutions of DNA were extracted with phenol/chloroform to remove proteins and other contaminants. Equal volume of phenol (TE saturated) was added to the DNA solution and vortexed vigorously for 1 minute. After vortexing, the mixture was centrifuged for 5 minutes at full speed in an eppendorf centrifuge. The upper aqueous phase which contained the DNA was removed leaving a white interface of denatured protein and the lower organic phase. When small quantities of DNA were being handled, the organic-phase was re-extracted with TE and the aqueous phases pooled.

For better phase separation and optimal purification of DNA by this method, phenol extraction was sometimes followed by a phenol/chloroform extraction. Here, 0.5 volume of phenol and 0.5 volume of chloroform were added to the DNA solution, vigorously mixed and centrifuged for five minutes. The aqueous phase was removed and added to an equal volume of chloroform : isoamyl alcohol (24:1). The vortex, centrifugation and aqueous phase removal steps were repeated once again. Following either extraction procedure, the DNA was then ethanol precipitated.

2.3.1.4.2 Ethanol Precipitation of DNA

The DNA sample was made 300 mM with respect to sodium acetate using a 3 M stock solution at pH5.2. 2.5 volumes of cold ethanol were added and the tube mixed well. The mixture was incubated at -20°C for one hour or longer. Precipitated DNA was pelleted at 14,000 rpm for ten minutes and washed once in 70% ethanol. After drying in air or under vacuum, the DNA was redissolved in sterile water.

2.3.1.4.3 Spectrophotometric Estimation of Nucleic Acid Concentration

The absorbance at 260 nm was used to estimate the concentration of nucleic acids (Sambrook *et al.* 1989). The following extinction coefficients were used : 0.05 for double stranded DNA, 0.04 for single stranded DNA and RNA, and 0.03 for oligonucleotides (units are in optical density/μg^{-1}/cm). An OD_{260} : OD_{280} ratio of 1.8 or greater implied minimal contamination with protein (Sambrook *et al.* 1989). Spectrophotometry was performed using a Cecil Model CE2010 spectrophotometer.

2.4 SUBCLONING OF HUMAN DNA SEQUENCES

2.4.1 Preparation of Plasmid Vector DNA and Human DNA Inserts

This protocol is a modification of Sambrook *et al* (1989)

500 ng of vector DNA (Bluescript phagemid vector SK II)was digested with a restriction endonuclease which cleaves the polylinker, in a total volume of 20 µl at the required temperature, specified for each restriction endonuclease, for one hour. Digestion was tested by running 1 µl of digested and undigested vector DNA samples side by side on a minigel which was stained with EtBr and visualised under UV light. Human DNA (cloned in cosmid, or a PCR product) was digested with the same restriction enzyme that cleaved the vector, or with an enzyme which generated compatible ends. This DNA sample was subsequently checked on a minigel for complete digestion then was extracted once with an equal volume of phenol/chloroform followed by ethanol precipitation. Final DNA concentration of insert DNA was adjusted to 20 ng/µl with sterile water.

2.4.2 Dephosphorylation of Vector DNA

(product handbook of Epicentre Technologies)

To prevent self-ligation of vector digested with a single restriction enzyme, the 5' terminal phosphate group was removed with HK^{TM} phosphatase. The vector DNA was digested to completion with an appropriate restriction enzyme in a total volume of 50 µl. 6 µl of 1 unit/µl phosphatase, 7 µl of 10 x buffer and 7 µl of 50 mM $CaCl_2$ were added into the digestion. The reaction was carried out at 30°C for 1 hour and then followed by heating at 65°C for 20 minutes to inactivate the phosphatase. The dephosphorylated vector DNA was directly used for ligation.

To test the efficiency of dephosphorylation, 1 µl of dephosphorylated vector DNA was ligated and transformed into *E. coli* strain XL1-Blue. If the 5' terminal phosphate group was removed, the vector could not recircularise, therefore only a few colonies were seen on the plate because the linear form DNA is very inefficient in transformation. The vector was then ready for use.

2.4.3 Ligation Reactions

(modification of the method from product handbook of Promega, Wisconsin, USA)

Ligation reactions were carried out with a vector : insert molar ratio of approximately 1:3 to maximise intermolecular ligation rather than intramolecular ligation. Usually, for 20 ng of linearised and phosphatased vector, 2 µl of 10 x ligation buffer, 1-2 units of T4 DNA ligase and insert DNA (5-50 ng) were added and the reaction mixture (in a total volume of 20 µl) was incubated at 12-16°C overnight. The efficiency of the ligation reaction was normally checked by religation of the EcoRI digested Spp-1 DNA under the same conditions as the sample DNA. The religated and non-religated DNA samples were separated on an agarose minigel. The disappearance of low molecular weight bands and increasing intensity of the large molecular weight bands indicated a high efficiency of the ligation reaction.

2.4.4 Competent Cells and Transformation

(modification of Sambrook *et al.* 1989)

A single colony of the *E.coli* host strain was inoculated into 5 ml of L-broth (where appropriate the L-broth was supplemented with an antibiotic) and the culture incubated overnight at 37°C with continuous shaking. The overnight culture was then diluted 100 fold into 50 ml of L-broth (plus antibiotic) and the incubation continued at 37°C, with shaking, until the culture reached an absorbance at A_{600} of 0.6-0.8. The cells were then pelleted by centrifugation at 2,000 x g for 5 min, resuspended in 2.5 ml of an ice cold solution of 0.2 M $MgCl_2$ and 0.5 M $CaCl_2$ and left on ice for 60 min.

200 µl of this cell suspension was mixed with 2-5 µl of the DNA ligation reaction mix and left on ice for 30 min. The cells were then heat shocked at 42°C for 1 min, cooled on ice for 1 minute and LB plus 20 mM glucose added. Following incubation at 37°C for 20-30 min, the cells were pelleted by gentle centrifugation and resuspended in 100 µl of LB. The resuspended cells together with 30 µl of 2% 5-bromo-4-chloro-3-indolyl D-galactoside (X-gal) and 30 µl 0.1 M isopropylthio D-galactoside (IPTG) were spread onto appropriate antibiotic plates and incubated overnight at 37°C. Recombinant plasmids were detected as white colonies due to the disruption of the coding region of the lacZ gene fragment by the insert.

2.5 ENZYME DIGESTION, GEL ELECTROPHORESIS AND SOUTHERN BLOT ANALYSIS

2.5.1 Restriction Endonuclease Digestion of DNA

Restriction endonuclease digestion of DNA was carried out using the buffer systems provided by New England Biolabs (see product handbook of Biolabs). Generally, four units of enzyme was added for each microgram of DNA to be digested and the reaction mix was incubated for at least 12 hours for genomic DNAs (plasmid and cosmid DNA digests were incubated for 2-4 hours) to ensure complete digestion. To ensure that the enzymic activity was not affected by glycerol, the volume of restriction enzyme(s) did not exceed 1/10 of the final volume of reaction mix, especially when two or more different enzymes were used simultaneously. Reactions were terminated by the addition of 0.1 volume of 10 x Loading Buffer (2.2.4)

2.5.2 Agarose Gel Electrophoresis of DNA

Electrophoresis of DNA to be used for Southern blot analysis was carried out using agarose (0.8%-1.2%) dissolved in 0.5 or 1 x TBE and cast on 14 cm x 11 cm x 0.3 cm perspex horizontal casts. Electrophoresis was performed in BRL horizontal tanks containing 0.5-1 x TBE buffer at 15-100 mA, until the bromophenol blue dye front had migrated an appropriate distance to ensure that adequate separation of the DNA fragments had taken place. Analytical agarose minigels (for checking digestions etc) were electrophoresed for one hour at 100 volts in a Biorad Mini-sub^{TM} DNA cell. Low gelling temperature agarose gels were used for purification of DNA fragments required for subsequent analyses. DNA was visualised under UV light after staining the gel in 0.02% ethidium bromide solution for 10-30 minutes.

2.5.3 Molecular Weight Markers

Depending on the sizes of the DNA fragments to be analysed on analytical or low gelling temperature gels, EcoRI digested Spp-1 phage or puc19 DNA digested with HpaII were used as molecular weight markers. EcoRI digested Spp-1 phage were used as molecular weight markers in Southern blot analysis.

2.5.4 ^{32}P Radio-Isotope Labelling of DNA

2.5.4.1 Primer Extension

Labelling of double stranded DNA was performed by primer extension of random oligonucleotides (Feinberg and Vogelstein, 1983) using the Amersham Megaprime DNA labelling systems kit. In brief, a small quantity of DNA insert (25-50 ng) was denatured at 100°C for two minutes and added to a solution containing random nonamers, dATP, dGTP, dTTP, alpha ^{32}P-dCTP, Klenow enzyme and buffer. The mixture was incubated at 37°C for 30-60 minutes.

2.5.4.2 Pre-reassociation of Repetitive DNA
(Sealy *et al.* 1985)

^{32}P-labelled DNA thought to contain repetitive DNA sequences was pre-reassociated prior to DNA hybridisation, to enable suppression of the repetitive sequences. The labelled DNA was mixed with an equal volume of 5 x SSC and a 2000 fold excess of sonicated human placental DNA (Sigma). The mixture was denatured at 100°C for 10 minutes, cooled on ice for one minute and then incubated at 65°C in a water bath for 1-2 hours. This allowed the preferential hybridisation of high copy repetitive elements of the radiolabelled and human placental DNAs. The mixture was then added to prewarmed hybridisation mix, and applied to the Southern blot filters.

2.5.5 Transfer of DNA to Nylon Membranes

2.5.5.1 Plaque/Colony Lifting

(Grunstein and Hogness, 1975; Benton and Davis, 1977)

Bacteria harbouring a recombinant vector were plated out as described in 2.4.4. After growth overnight at 37°C a replica was made by gently laying a nylon membrane disc onto the plate surface. Plates containing top-agar were first cooled to 4°C. The filter was keyed to the plate and then transferred to 0.5 M NaCl/0.5 M NaOH for 10 minutes to lyse host cells and denature DNA. After neutralisation for 10 minutes in 2 M NaCl/0.5 M Tris HCl pH8.0, the filter was rinsed in 2 x SSC and baked in the microwave at high heat for 30 seconds.

2.5.5.2 Southern Blotting

(Sambrook *et al.* 1989; Reed and Mann, 1985)

DNA was digested with the appropriate restriction enzymes and electrophoresed on a 0.8-2.0% (depending on the size of the DNA fragments being separated) agarose gel in 1 x TBE buffer. Following staining with ethidium bromide, the gel was visualised under UV light to determine the extent of the digestion of the DNA sample.

2.5.5.3 Transfer of DNA of Low Molecular Weight

The DNA fragments were denatured by soaking the gel in 0.4 N NaOH for 10 minutes. The Hybond N+^{TM} (Amersham) nylon membrane was cut to the size of the gel and placed in the 0.4 N NaOH transfer solution briefly before transfer. DNA in the agarose gel was transferred to the prepared filter by capillary action (using absorbent paper) for 1-16 hours. Following transfer the filters were baked for 45 seconds in a microwave oven.

2.5.5.4 Transfer of DNA of High Molecular Weight

The DNA fragments were denatured by soaking the gel in 0.5 N NaOH, 2.5 M NaCl for 1 hour. The gel was then neutralised in 1.5 M NaCl, 0.5 M Tris-HCl pH7.5 for 1 hour. The Hybond N+^{TM} (Amersham) nylon membrane was cut to the size of the gel and placed in the 10 x SSC transfer solution briefly before transfer. DNA in the agarose gel was transferred to the prepared filter by capillary action for 1-16 hours. Following transfer the filters were baked for 45 seconds in a microwave oven.

2.5.6 Prehybridisation, Hybridisation and Washing

2.5.6.1 Using Probes Labelled by Primer Extension

Prior to hybridisation, nylon filters were prehybridised at 42°C for 1 hour in a solution consisting of 50% (v/v) deionised formamide, 10% (w/v) dextran sulphate, 5 x SSPE, 2% SDS, 1 x Denhardt's and 100 g/ml salmon sperm DNA. Hybridisations were incubated overnight at 42°C with between 1-10 ng/ml of probes labelled by primer extension (2.5.4.1). The filters were washed twice for 10 minutes in Wash A (2 x SSPE, 1% SDS) at 42°C. They were then washed twice for 10 minutes in Wash A at 65°C. Finally the filters were washed twice for 15 minutes in Wash B (0.1 x SSPE, 0.1% SDS) at 65°C Autoradiography was carried out at -80°C in the presence of tungsten intensifying screens.

2.6 POLYMERASE CHAIN REACTION (PCR)

All PCRs were performed in a Perkin Elmer-Cetus thermal cycler (Norwalk, CT, USA). Incubations for the majority of reactions were performed in a 20 µl final volume comprising of 2 µl 10 x PCR mix, 1.5 mM $MgCl_2$, 150 ng of each primer, 100 ng template DNA, 1 unit of *Taq* DNA polymerase and sterile water to 20 µl. The solution was mixed well and overlaid with one drop of paraffin oil. The reaction conditions used for individual PCR reactions are described in the relevant results chapters. Positive and negative controls were always included in each set of reactions. The various controls which were utilised for each experiment are described in the results chapters.

Following amplification, aliquots of the PCR reactions were analysed by electrophoresis on 1-2% agarose gels in TBE buffer (2.5.2). The reaction products were visualised by illumination under UV light.

2.6.1 PCR Primers

Oligonucleotides for PCR were designed to contain a similar proportion of purine and pyrimidine bases and there were usually no stretches of more than four consecutive purines and no repeat sequence DNA. In addition, primer pairs were carefully checked at their 3' ends to avoid the possibility of primer-dimer formation. The standard length of oligonucleotides used in this study was 24 nt. All oligonucleotides were synthesised using an Applied Biosystems 391 DNA synthesiser in the Department of Cytogenetics and Molecular Genetics at the Women's and Children's Hospital.

2.6.2 Oligonucleotide Deprotection and Cleavage

A 1 ml syringe was attached to the synthesis column containing the synthesised oligonucleotide. 1 ml of ammonium hydroxide was pipetted into a 1.5 ml tube. The ammonium hydroxide was drawn into the column/syringe, minimising bubbling in the syringe which was allowed to stand for 1 hour at room temperature. The syringe was removed and the ammonia solution was collected in a 1.5 ml centrifuge tube. The ammonia/oligonucleotide solution was incubated at 55°C overnight.

2.6.2.1 Oligonucleotide Purification n-Butanol Method

(from Sawadogo and Van Dyke, 1991)

The cleaved and deprotected oligonucleotide in ammonium hydroxide solution was cooled to room temperature. The oligonucleotide was then added to 10 mls of n-butanol in a 50 ml centrifuge tube. The solution was vortexed for 15 seconds, then centrifuged at 3,000 rpm for 45 minutes. The supernatant was discarded and the pellet was resuspended in 1 ml of water. An additional 10 mls of n-butanol was added, then vortexed, centrifuged and supernatant discarded as before. The pellet was air-dried and resuspended in 100 µl of water. The DNA was quantitated and the concentration of oligonucleotide was adjusted to 1 mg/ml.

2.6.3 Subcloning PCR Products

The pGEM-T Vector system (Promega) takes advantage of the single deoxyadenosine nucleotides that are added to the 3' end of all duplex molecules by the thermostable polymerases used in PCR. These A-overhangs are used to insert the PCR product into the specifically designed vector pGEM-T which provides single 3' T overhangs at the insertion site.

One µl of PCR product was used in a ligation reaction for cloning into a 3 kb pGEM-T vector. Two separate ligation mixes (mix A and B) were used with each PCR product. Mix A and B both consisted of 1 µl of T4 DNA 10 x ligase buffer (50 ng/µl) of T vector and 1 µl of T4 DNA ligase enzyme (3 U/µl). In addition, mix A contained 1 µl of amplified PCR product and 6 µl of water, whereas mix B contained 0.5 µl of PCR product and 6.5 µl of water. Ligation reactions were carried out with a vector: insert molar ratio of 1:1 to maximise intermolecular as opposed to intramolecular ligation. The reactions were allowed to proceed for 3 hours at 15°C and then overnight at 4°C. XL1-Blue cells were transformed according to section 2.4.4.

2.7 DNA SEQUENCING

2.7.1 Automated DNA Sequencing

The automated sequencing was performed with an Applied Biosystems 373A DNA Automated Sequencer and *Taq* Dye Deoxy^{TM} Terminator Cycle Sequencing Kit or the *Taq* Dye Primer (M13 reverse and M13 forward) Sequencing Kit and protocols. The automated sequencing was used to sequence double stranded plasmid DNA or PCR amplified product DNA.

2.7.2 Double Stranded DNA Sequencing

Double stranded DNA templates were diluted to concentrations of 200-250 ng/µl in preparation for sequencing. The ABI Prism^{TM} Ready Reaction Dye Primer Cycle Sequencing Kit protocol was followed (Perkin Elmer), using primers adjacent to the vector cloning site in either the forward (-21M13) or reverse (M13) direction. Templates were also sequenced using the ABI Prism^{TM} Dye Terminator Cycle Sequencing Kit (Perkin Elmer) using the procedure described in the Perkin Elmer Cycle Sequencing Handbook.

2.7.3 PCR Product Sequencing

PCR products of double stranded DNA were amplified by using conditions dependent on the primer pairs being used. The PCR products were purified from the oligonucleotide primers using Qiaquick columns (Qiagen) according to the manufacturer's instructions.

Dye terminator sequencing of double stranded PCR product comprised a total of 30-90 ng of PCR amplified DNA, 1 µl of 20 ng/µl primer, reaction mix and H_20 to 20 µl The reactions were concentrated with ethanol. Primers used in the dye terminator reactions were specific to each template and are listed in the relevant results chapters.

Dye primer sequencing comprised a total of 60-180 ng of PCR product, divided into 4 separate tubes with reaction buffer/primer (M13 forward or M13 reverse) mix. The reactions were concentrated with ethanol.

2.7.4 Loading and Running the Sequencing Gel

Denaturing acrylamide gels, used for resolving fluorescently labelled dideoxy terminator or dye primer PCR fragments in conjunction with the software of the ABI 373A DNA sequencer, were made, according to the protocol supplied. The sequencing gel mix contained the following reagents: 7.5 M urea (Biorad), 6% acrylamide: bisacrylamide solution (19:1; Biorad), and 1 x TBE (pH 8.3). The contents were placed in a clean parafilm-sealed conical flask under hot water, then made up to an 80 ml volume with distilled water, vacuum-filtered through a Nalgene 0.2 M filter unit, and de-gassed. To this solution, 350 µl of freshly made 10% potassium persulphate was added and mixed, followed by 45 µl of TEMED. This solution was syringe-poured between the two pre- prepared gel plates which had been cleaned with a non-fluorescent detergent, rinsed with water, separated by 0.4 mm spacers, taped around the edges, and clamped. After pouring, a plastic strip was applied to the top edge of the plates (where the shark's teeth were to be placed), and the gel left to polymerise over at least 2 hours. Then the clamps and tape were removed, the plates cleaned again, and the teeth inserted.

The gel was pre-electrophoresed at 35 watts (constant) and 45°C for 1 hour prior to loading of the samples. To each sample was added 4 µl of formamide solution (deionised formamide: 50 mM EDTA pH 8.0 = 5: 1) and mixed well. The solution was centrifuged briefly to collect all the liquid at the bottom of the tube. Before loading, the sample was heated at 90°C for 2 minutes. 3 µl of each sample was loaded on the gel. The gel was run for 12 hours under the conditions stated above, using 1 x TBE, diluted from a 3 MM Whatman-filtered 10 x TBE stock. The sequence data was visualised as a 4-colour chromatogram.

2.8 COSMID VECTORS

Cosmid vectors were originally designed to clone and propagate large segments of genomic DNA. In their simplest form, cosmid vectors are modified plasmids that carry the DNA sequences (cos sequences) required for packaging DNA into bacteriophage lambda particles. Because cosmids carry an origin of replication and a drug resistance marker, cosmid vectors can be introduced into *E. coli* by standard transformation procedures and propagated as plasmids. Cosmid vectors have been constructed to contain a large variety of structural elements designed to improve sequence representation, to simplify or expedite the structural or functional analysis of cloned DNA, or to simplify the construction of high quality representative libraries.

2.8.1 sCos Vectors

sCos vectors are a family of vectors that have been designed specifically for the mapping and functional analysis of human chromosomes. The construction of these vectors, described elsewhere (Evans *et al.* 1989), included the presence of two cos sites so that packaging could be carried out with high efficiency and without requiring size selection of the insert DNA. The vectors also include the presence of T7 and T3 bacteriophage promoters for the synthesis of "walking" probes. Unique restriction sites were incorporated for the removal of the insert from vector, for example; NotI sites for sCos1, NotI and SacII sites for sCos2 and SfiI sites for sCos4. A plasmid origin of replication (ori) was also included for giving high yields of cosmid DNA when preparing templates. In this project sCos1 cosmid vector was used.

2.8.2 Construction of a Chromosome 16 Specific Ordered Cosmid Library

The chromosome 16 specific cosmid library was constructed at the Los Alamos National Laboratory (LANL), New Mexico and has been described (Stallings *et al.* 1990). In general, human chromosomes 16 were isolated from a somatic cell hybrid CY18. A single chromosome 16 was the only human chromosome present in this hybrid. After partial digestion with Sau3A and dephosphorylation with calf intestinal alkaline phosphatase, the chromosomal DNA was ligated to the cloning arms from the cosmid vector sCos1. *In vitro* packaging and infection of *E. coli* yielded 1.75 x 10^5 independent recombinants, giving a 67 fold statistical representation.

2.8.2.1 Repetitive Sequence Fingerprinting and Assembly of Contigs

(performed at Los Alamos National Laboratories (LANL),
New Mexico, USA)

An approach has been developed for the identification of overlapping cosmid clones by exploiting the high density of repetitive sequences in complex genomes as described by Stallings *et al.* (1990). By coupling restriction digestion mapping with oligo probes targeting abundant interspersed repetitive sequences such as Alu (Jelenik and Schmid, 1982), Ll (Scott *et al.* 1987) and $(GT)_n$ (Rich *et al.* 1984; Weber and May, 1989), a "fingerprint" is obtained. The initial analysis of fingerprint data are the pairwise comparison of the fingerprint between cosmid clones. Clones that overlap will share restriction fragments of similar size with similar repetitive sequences. A probability of overlap is assigned to each pair of clones based on the number of shared restriction fragments with the same repetitive sequences.

The analysis of the treatment of fingerprint data and the algorithm used to detect pairwise overlap has been described (Balding and Torney, 1991). Once overlapping pairs have been identified, the clones are ordered and assembled into contigs based upon the information in overlapping pairs. For contig construction a computer program was used based on a genetic algorithm - genetic contig assembly algorithm, GCAA. GCAA represents possible maps as strings of numbers encoding the lengths and positions of the clones. The quality of any possible map is measured by a fitness function that takes into account the overlap likelihoods, overlap extents and clone lengths that the map is intended to fill.

By the repetitive sequence fingerprinting of approximately 4000 cosmid clones obtained from the chromosome 16 specific library, a cosmid contig map was developed (Stallings *et al.* 1992a). The clones were organised into 576 contigs and 1171 cosmid clones not contained within a contig.

2.8.2.2 Construction and Screening of High Density Cosmid Grids

High density filters of arrayed chromosome 16 specific cosmid clones were produced by LANL. A Beckman Biomek 1000 was used to stamp bacterial colonies onto Biodyne nylon hybridisation membranes (1536 clones/membrane) as already described (Longmire *et al.* 1991).

Membranes were hybridised overnight with probes labelled by primer extension in 6 x SSC, 10 mM EDTA pH8.0, 10 X Denhardt's, 1% SDS, 0.1 mg/ml denatured sonicated salmon sperm DNA, at 65°C. Following hybridisation membranes were washed in 2 x SSC, 0.1% SDS at room temperature (quick rinse); twice in 2 x SSC, 0.1% SDS at room temperature for 15 minutes; and twice in 0.1% SDS at 50°C for 30 minutes. Double stranded DNA probes were labelled with ^{32}P to a specific activity of 10^8 cpm/µg by primer extension labelling (2.5.4.1) and pre-reassociated (2.5.4.2) prior to hybridisation of the probe to the filters.

CHAPTER 3

Identification of Cosmids Localised to the 16q24 Chromosomal Region

3.1 INTRODUCTION 107

3.1.1 Isolation of Cosmids From Defined Subregions of Human Chromosomes 108

3.1.2 Alu PCR Strategy for the Isolation of Cosmids Localised to the 16q24 Chromosomal Region 109

3.1.3 Isolation of Cosmids Localised to 16q24.3 Using STSs and Cosmid Ends 111

3.2 MATERIALS AND METHODS 113

3.2.1 Hybrid Cell Line Panel 113

3.2.2 Alu PCR Amplification 113

3.2.3 Screening the Chromosome 16 Cosmid Library 114

3.2.4 Localisation of Cosmids to the 16q24 Region 115

3.3 RESULTS 117

3.3.1 Generation of Alu PCR Amplification Products and Identification of Cosmids From the 16q24 Chromosomal Region 117

3.3.2 Localisation of Identified Cosmids 120

3.3.3 Identification of Cosmids Using STSs and Cosmid Ends Localised to 16q24.3 129

3.4 DISCUSSION 137

3.1 INTRODUCTION

Upon commencement of the project for this thesis in October 1992, the initial aim was to construct a detailed physical map of the 16q24 chromosomal region. This was in view of the recognised characteristics of 16q24. These included the localisation of the H3 isochore family to chromosomal telomeric bands, including the telomeric band of 16q (Saccone *et al.* 1992) (see 1.4.2.1). Thus, it was postulated that the 16q24 region was gene-rich, as the H3 isochore family contains one third of all human genes. Also, loss of heterozygosity (LOH) of the chromosome 16 long arm, 16q, was demonstrated by (AC)n repeat and RFLP analysis in many solid tumours including hepatocellular carcinoma (Tsuda *et al.* 1990) and prostate cancer (Bergerheim *et al.* 1991; Kunimi *et al.* 1991) (see 1.5.4). LOH of 16q was also demonstrated in sporadic primary breast cancer cases at a frequency of 45% - 51% (Sato *et al.* 1990; Sato *et al.* 1991) and allele losses in 42.8% of tumours were observed with marker D16S7, mapped to 16q22-q24 (Sato *et al.* 1991). These observations indicated that a putative tumour suppressor gene (TSG) (or genes) was localised on this chromosome arm. Detailed localisation of the region(s) of chromosome 16q showing LOH had not been determined in 1992, but the high frequency of allele loss observed with D16S7 in sporadic breast tumours implied that a TSG may be localised in the 16q24 chromosomal region. The results of the various studies indicated that the TSG may also be pleiotropically involved in suppressing the development of tumours of different histological origin.

Thus, the 16q24 region was chosen for the construction of a detailed physical map consisting of YACs and cosmids. Following finer genetic mapping of the region of LOH, this physical map can assist in the positional cloning of the TSG. In addition, the physical map will make a contribution to the construction of transcript maps to ultimately identify all the genes in this region.

3.1.1 Isolation of Cosmids From Defined Subregions of Human Chromosomes

Progress in mapping the human genome requires the rapid and efficient isolation of large numbers of DNA probes from predefined subregions of specific chromosomes. Somatic cell hybrids containing intact human chromosomes or subchromosomal fragments, isolated in rodent cell backgrounds, have been powerful reagents for human genome analysis including gene mapping and isolation of chromosome specific sequences (Ruddle, 1981). Techniques for the isolation of human sequences from a heterologous rodent background have included cloning the DNA of a somatic cell hybrid and the subsequent screening of this recombinant library with human specific repeat sequence probes to identify human clones (Gusella *et al.* 1980). Isolation of human sequences from predefined subregions has also been achieved by amplification of DNA fragments microdissected from chromosomal bands of metaphase chromosomes (Ludecke *et al.* 1989), but this procedure generates very short probes that are difficult to clone.

Nelson *et al* (1989) developed a technique termed Alu PCR which involved the amplification of human DNA of unknown sequence from complex mixtures of DNA from humans and other species, and applied it to the isolation of human specific sequences directly from human/rodent hybrid DNA. The Alu PCR method is based on PCR amplification of human sequences using primers directed at the human specific portion of Alu, a short interspersed repeat element (SINE). These 300 bp repeat elements, found ubiquitously in human DNA, number approximately 900,000 in the haploid human genome and occur on average every 4 kb throughout it (Britten, 1988). The distance between any pair of Alu repeats may vary considerably since Alu sequences appear to be enriched in Giemsa light staining chromosome bands, which are GC and gene-rich regions, and are deficient in other regions (Korenberg and Rykowski, 1988). A consensus Alu sequence has been established and regions of the repeat that are reasonably well conserved have been identified (Kariya, 1987). These regions include those that are conserved among primates and rodents, and others that have evolved to be primate specific. There is sufficient sequence divergence of the human and rodent repetitive elements to reduce cross hybridisation of human specific primers to rodent Alu repeats. A single PCR primer designed to recognise these conserved regions allows specific amplification of human sequences from rodent backgrounds in hybrid cells when two Alu elements in either convergent or divergent orientation are less than 5 kb apart.

The Alu PCR technique has been adapted for a number of applications in addition to the isolation of human specific sequences directly from human/rodent hybrid DNA. These include the rapid isolation of human genomic DNA from a variety of cloned sources including YACs, cosmids and bacteriophages, thus providing a useful tool for analysis of the human genome. Alu PCR products have also been utilised directly as probes for *in situ* hybridisation experiments on metaphase spreads and have been useful in characterising the human content of human/hamster hybrid cell lines (Lichter *et al.* 1990).

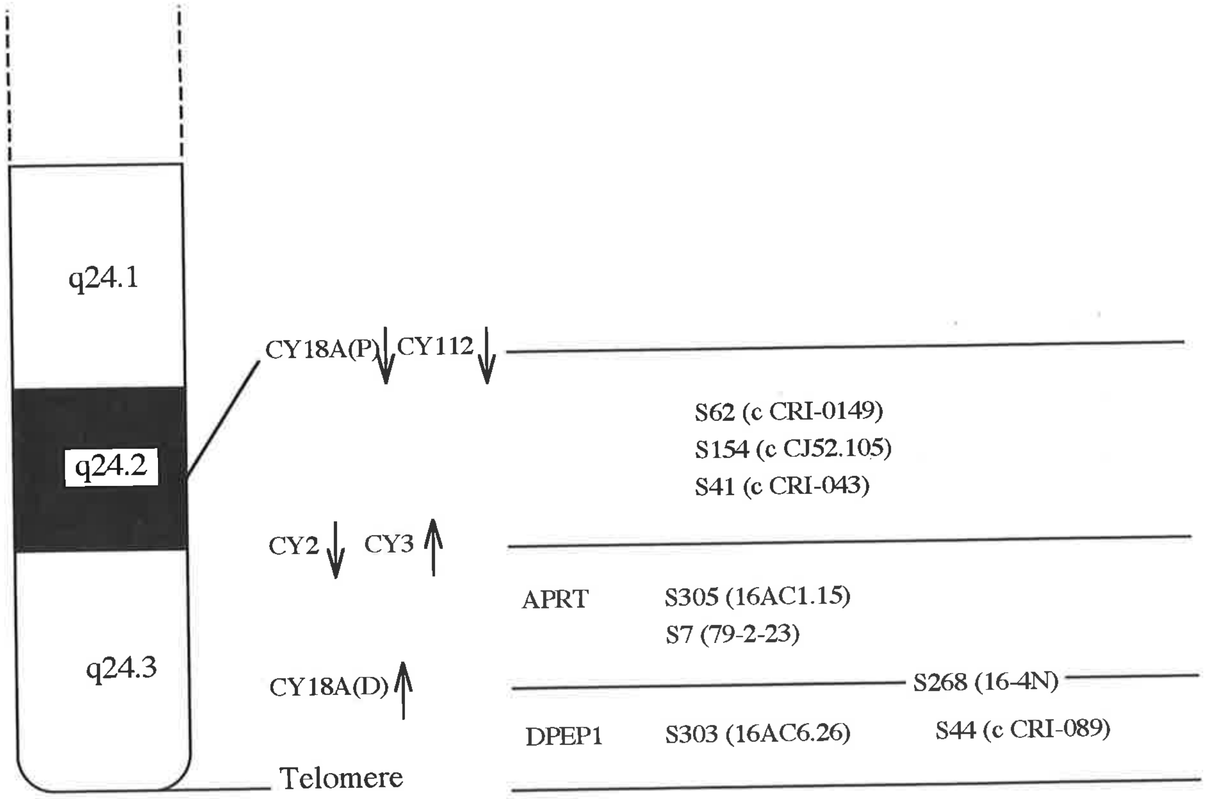
3.1.2 Alu PCR Strategy for the Isolation of Cosmids Localised to the 16q24 Chromosomal Region

The cytogenetic based physical map of chromosome 16 constructed with human/mouse somatic cell hybrids (Callen *et al.* 1995; Doggett *et al.* 1995) has made this chromosome more amenable to detailed analysis. The hybrids possess a defined segment of chromosome 16 and they are derived mostly from naturally occurring rearrangements, but some have arisen from spontaneous rearrangements during construction. The physical map of the 16q24 chromosomal region published in 1992 (Callen *et al.* 1992) is shown in figure 3.1. At that time, this region was not well represented by genes or cloned DNA, thus it was decided to utilise the Alu PCR strategy to increase the density of cosmids in the 16q24 region and proceed toward the construction of the physical map of this region. Two somatic cell hybrids, CY2 and CY18A, in which the only human chromosome 16 material was the distal part of the long arm of chromosome 16, shown in figure 3.2 part A, were used as templates in the Alu PCR reaction in order to specifically amplify human sequences. These amplified

Figure 3.1

Physical map of the 16q24 chromosomal region published in 1992 (Callen *et al.* 1992). This region is represented by two genes APRT and DPEP1, four cosmids, indicated by a 'c' and four polymorphic markers.

16q24 chromosomal region



products were then used as hybridisation probes for screening the chromosome 16 specific gridded cosmid library (see 2.8.2). Since many of these cosmids were assembled in contigs, this hybridisation enabled the identification of cosmids and members of their contigs, from this predefined chromosomal region.

3.1.3 Isolation of Cosmids Localised to 16q24.3 Using STSs and Cosmid Ends

Over the time engaged in experiments involving the Alu PCR strategy and the localisation of cosmids to the 16q24 chromosomal region, developments in the Human Genome Project (HGP) resulted in the refinement of genetic and physical maps of human chromosome 16. New somatic cell hybrids were constructed and a large number of polymorphic markers, genes and transcribed sequences (Adams *et al.* 1992; Koyama *et al.* 1993; Austruy *et al.* 1993; Whitmore *et al.* 1994; Callen *et al.* 1995; Scott *et al.* 1996) were assigned to chromosome 16, including 16q24. The polymorphic markers assisted in refining the commonly deleted regions of chromosome 16 in breast tumours. Tsuda *et al* (1994) demonstrated LOH at 16q24.2-qter in 52% of breast tumour samples. Cleton-Jansen *et al* (1994) identified LOH from APRT to D16S303 at 16q24.3 at a similar frequency in breast tumours. Also, the Fanconi anaemia group A gene (FAA) was localised to the 16q24.3 region by classical linkage analysis (Pronk *et al.* 1995). Based on this refined mapping it was decided to restrict the construction of the detailed physical map to a smaller part of 16q24, namely 16q24.3 between APRT and D16S303, the smallest region defined to contain the TSG and FAA. A detailed physical map consisting of cosmid contigs spanning this region will form the basis of the positional cloning or candidate gene approaches for the identification of these genes. Based on linkage data, the accurate physical distance of this region was estimated to be less than 2 to 3 Mb.

STSs can be used to link and integrate genomic clones into contigs. Thus, a strategy involving STSs was utilised to identify new cosmids mapping to the 16q24.3 region. The expressed sequences and microsatellite repeats already mapped to this region of chromosome 16 were utilised as hybridisation probes for screening filters containing a gridded chromosome 16 specific cosmid library with a redundancy of ten from the Los Alamos National Laboratory (LANL) (2.8.2). The approach of cosmid walking was also utilised to extend the singleton cosmids and cosmid contigs mapped to this region. This involved the generation of cosmid ends to probe the ten times coverage cosmid filters and cosmid overlaps were confirmed by a restriction enzyme digest protocol. The construction of this physical map was a collaborative effort involving myself, Mr. Scott Whitmore and Ms. Joanna Crawford, Department of Cytogenetics and Molecular Genetics, Women's and Children's Hospital, South Australia.

3.2 MATERIALS AND METHODS

3.2.1 Hybrid Cell Line Panel

The DNA from a number of somatic cell hybrids containing either a cytogenetically normal chromosome 16 or portion thereof, and human parental cell line DNA (ie. the cell line from which the somatic cell hybrid was derived) were used for Alu PCR and Southern analysis in this study. The portion of chromosome 16 present in each cell line is described in table 3.1. The CY18A and CY2 hybrids were used as templates in Alu PCR reactions. The CY18A hybrid arose spontaneously from the CY18 hybrid and was thought to contain only chromosome 16 material from 16q24.2 to q24.3. CY2 contains a der (X) t(X:16)(q26;24) such that the most telomeric region of 16q is attached to the long arm of the X chromosome. The construction of somatic cell hybrids and details of the human parental cell lines have been previously described (Callen, 1986; Callen, 1990).

3.2.2 Alu PCR Amplification

The 517 (Nelson *et al.* 1989) and A-33 (Chumakov *et al.* 1992a) oligonucleotides were used separately as amplimers for PCR amplification. Primer A-33 was located between bases 229-213, and primer 517 between bases 241 and 263, of the human specific consensus at the 3' end of the Alu repeat sequence. These oligonucleotides were synthesised and deprotected as described in sections 2.6.1 and 2.6.2. The oligonucleotide sequences were as follows: 517, 5'- CGACCTCGAGATCTYRGCTCACTGCAA -3', where Y = T or C and R = A or G, and A-33, 5'- CACTGCACTCCAGCCTGGGCGAC -3'. The 517 primer (Nelson *et al.* 1989; Ledbetter *et al.* 1990) and the A-33 primer (Chumakov *et al.* 1992a) were previously shown to be completely human specific when DNAs from human/rodent cell hybrids were used as template.

The PCR reaction (see 2.6) was performed on a Perkin Elmer DNA thermal cycler under the following conditions (modified from Zucman *et al.* 1992). 200 ng of hybrid DNA and 500 ng of oligonucleotide in a final volume of 100 µl were used per amplification reaction. After an initial denaturation at 94°C for 10 minutes, 35 cycles were carried out with the following parameters: denaturation at 94°C for 2 minutes, annealing at 55°C for 2 minutes and extension at 72°C for 4 minutes. A 10 minute extension was performed at the end of the last cycle.

PCR products generated from the Alu primers were purified and concentrated using Wizard PCR prep minicolumns (Promega) according to the manufacturer's instructions and resuspended in sterile water. The Alu PCR products were analysed by electrophoresis on a 1.5% agarose gel (see 2.5.2).

3.2.3 Screening the Chromosome 16 Cosmid Library

Los Alamos National Laboratory (LANL) has constructed a chromosome 16 specific cosmid library of which some 4000 cosmids were assembled into contigs by use of repetitive sequence fingerprinting (Stallings *et al.* 1990) (see section 2.8.2). Filters containing these 4000 cosmids arrayed in a high density format represent approximately one times coverage of the chromosome. 100 ng of Alu PCR product labelled by primer extension, as in 2.5.4.1, with repetitive DNA sequences pre-reassociated using human placental DNA (2.5.4.2), was added to these filters and hybridised at 65°C for 16 hours (2.8.2.2). Filters were washed, as described in 2.8.2.2, and autoradiographed for one day at -70°C with one intensifying screen.

3.2.4 Localisation of Cosmids to the 16q24 Region

Cosmids were grown overnight in LB supplemented with kanamycin at 50 µg/ml. Cosmid DNA was extracted by the alkaline lysis procedure (Sambrook *et al.* 1989) described in 2.3.1.1.1. Gel electrophoretic patterns of EcoRI digested cosmid DNA (2.5.1 and 2.5.2) were analysed after ethidium bromide staining to determine the integrity of the DNA.

Human DNA, DNA from the human/mouse hybrid cell lines including CY18, CY18A, CY112, CY2, CY3, CY104, CY115, CY120, CY107 (see table 3.1), and the mouse cell line A9, were digested with HindIII. After electrophoresis of this DNA (2.5.2), Southern blot filters were made from agarose gels by the methods described in 2.5.5.2 and 2.5.5.4. The filters were prehybridised then probed with labelled, pre-reassociated cosmid DNA (2.5.4.2) in order to localise the cosmids. Hybridisation was carried out at 42°C for 16 hours and filters were washed following the procedure outlined in 2.5.6.1. Probes were stripped from filters before hybridisation with the next probe.

FISH analysis with a selection of cosmids was performed by Mrs. Helen Eyre. The location of fluorescence-labelled probes was determined by hybridisation to metaphase chromosomes *in situ* as described by Callen *et al* (1988).

Table 3.1

Panel of human/mouse hybrid cell lines containing DNA from the long arm of chromosome 16 used in Alu PCR and physical mapping studies of the 16q24 chromosomal region.

Hybrid Cell Line	Mouse Parent	Human Parent Rearrangement	Portion of 16 Present	Reference
CY18	A9		complete 16	Callen (1986)
CY115	A9	t(9;16)(p22;q24.1)	q24.1-qter	Callen *et al.* (1992)
CY107	A9	del(16)(q22.3q24.3)	pter-q22.3 q24.3-qter	Shen *et al.* (1994)
CY120	A9	t(4;16)(q25;q24.2)	q23 - qter	Callen *et al.* (1990)
CY104	A9	t(?;16)(?;q24.3)	q24.3-qter	Callen *et al.* (1995)
CY18A	A9	-	q24.2 - q24.3	Shen *et al.* (1994)
CY112	A9	r(16)(p13.3q24.2)	q24.2 - qter	Shen *et al.* (1994)
CY3	A9	t(X;16)(q26;q24.2)	pter - q24	Callen *et al.* (1986)
CY2	A9	t(X;16)(q26;q24.2)	q24 - qter	Callen *et al.* (1986)

3.3 RESULTS

3.3.1 Generation of Alu PCR Amplification Products and Identification of Cosmids From the 16q24 Chromosomal Region

In this study, primers 517 and A-33 were utilised to promote the specific amplification of inter-Alu sequences from two human/mouse somatic cell hybrids, CY2 and CY18A. These two hybrids contain unique regions of 16q24 together with a common region of overlap. The primers were first tested for their ability to specifically amplify human DNA from the human/mouse DNA mixture, using a range of annealing temperatures in the PCR reaction. Under the conditions described, primers were found to yield amplification products only from DNA of the hybrid cells and not from DNA of the parental mouse cell line, A9 (data not shown). Thus, the primers were demonstrated to be human specific. No products were observed in control reactions in which no template was added (data not shown). Figure 3.2 part A shows an ideogram of the portion of chromosome 16 present in the two human/mouse hybrids and figure 3.2 part B shows the gel electrophoretic pattern of the Alu PCR products obtained. The Alu PCR amplified products from the cell hybrids contain a series of bands ranging in size from 0.36 kb to 1.86 kb. Some PCR products of each hybrid were of unique size, in addition to others that were of equal size which presumably corresponded to the overlapping region.

Purified Alu PCR products obtained from the CY18A template were pooled in one tube as were the Alu PCR products from the CY2 template. Each Alu PCR product was labelled and hybridised to the filters of the arrayed chromosome 16 cosmid library. Figure 3.3 shows representative autoradiographs. All cosmids that provided a positive signal were grouped into three classes according to their pattern of hybridisation with the two Alu PCR probes. These patterns were consistent with the cosmids being derived from one of three regions defined by the somatic cell hybrids. Specifically, these are regions I, II and III corresponding respectively to CY18A only, the CY2 / CY18A overlapping region and CY2 only, indicated

Figure 3.2

A. Ideogram of the portion of the long arm of chromosome 16 present in the two human/mouse hybrids CY2 and CY18A. The horizontal lines indicate the breakpoints of the human/mouse somatic cell hybrids.

Cosmids positive from Alu PCR screening were grouped into one of three regions defined by the somatic cell hybrids. Region I corresponds to CY18A only; Region II corresponds to CY18A / CY2 overlap; and Region III corresponds to CY2 only.

B. Gel electrophoretic pattern of the Alu PCR products obtained using the 517 and A-33 primers.

Lane 1:	Spp-1 markers
Lane 2:	517 primer, CY2
Lane 3:	517 primer, CY18A
Lane 4:	A-33 primer, CY2
Lane 5:	A-33 primer, CY18A

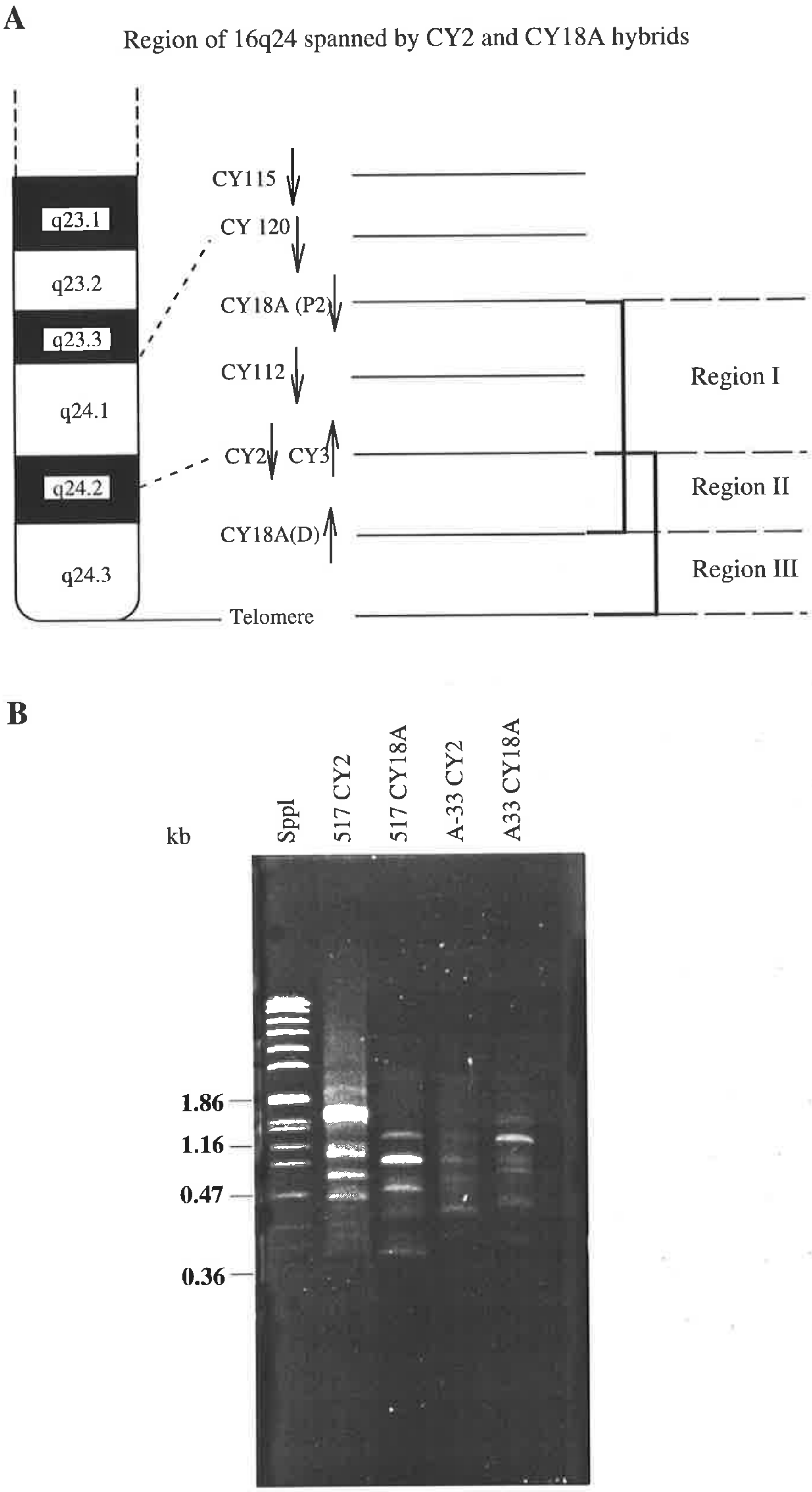
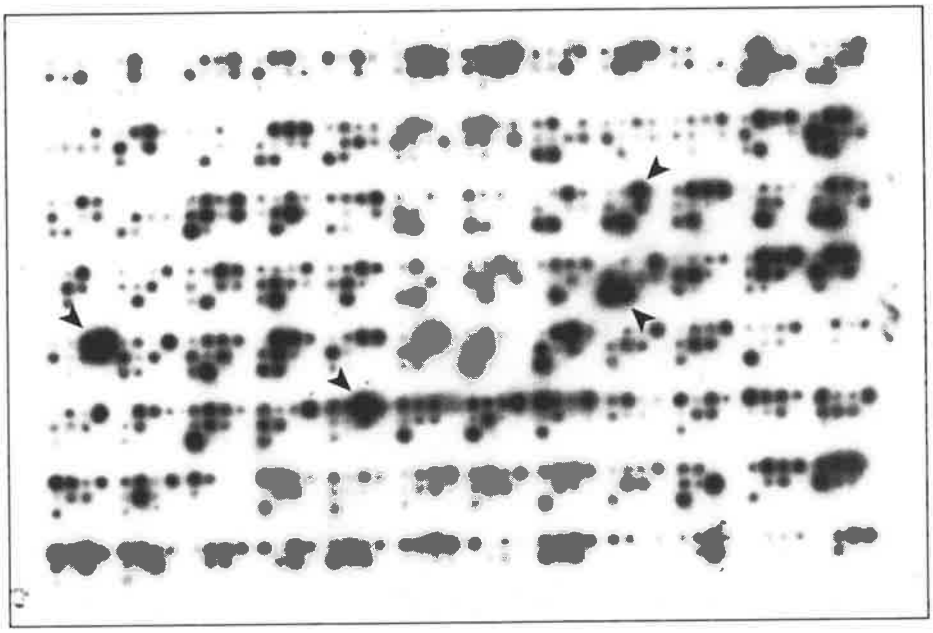

Region of 16q24 spanned by CY2 and CY18A hybrids

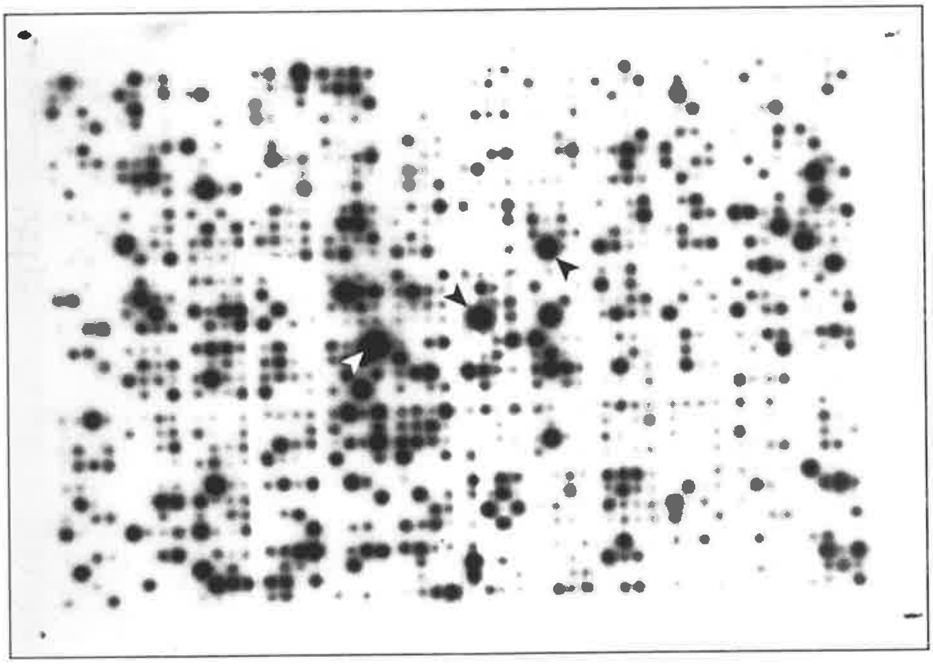
Figure 3.3

Representative autoradiographs of pooled 517 and A-33 primed Alu PCR products from the CY2 somatic cell hybrid (top) and the CY18A (bottom) somatic cell hybrid, hybridised to the gridded array of chromosome 16 cosmid clones. Positive signals are indicated by the arrows. One membrane of the cosmid library filters, from a set of three, is shown.

CY2



CY18A



in figure 3.2 part A. This analysis identified a total of 52 cosmids potentially located in regions I, II and III. Region I contained 23 cosmids, Region II contained 25 cosmids and Region III contained 4 cosmids. These data are listed in table 3.2.

3.3.2 Localisation of Identified Cosmids

The initial step in confirming the localisation of the cosmids identified from screening the chromosome 16 cosmid library with the Alu PCR products was to determine whether any of them had been mapped previously by Southern analysis or PCR. This was achieved by accessing mapping data from a database of chromosome 16 cosmid localisations (LANL). Several cosmids mapped to 16p11.2, 16p13 and 16q22, suggesting there could be common repetitive elements shared between 16q24 and these three regions of chromosome 16. Cosmids which were previously mapped in regions other than 16q24 were eliminated from subsequent analysis.

The location of the remaining cosmids was determined by Southern analysis using a panel of somatic cell hybrids containing DNA of the most distal part of chromosome 16q, indicated in table 3.1. A selection of representative autoradiographs are shown in figure 3.4. The localisation of a number of cosmids was also confirmed by FISH analysis. Nine cosmids, mostly members of contigs 228, 188 and 236.1, were positive for CY3 and CY18A, but not the other hybrids containing DNA from 16q24. These results are shown in table 3.3. Further characterisation of the CY18A hybrid, performed by Ms. Sharon Lane, Department of Cytogenetics and Molecular Genetics, Women's and Children's Hospital, indicated that it contained a region of 16q13 which had not been identified previously. Two additional members of contigs 228 and 236.1, identified from screening the chromosome 16 cosmid library, gave a total of eleven cosmids likely to be localised to the 16q13 region of chromosome 16.

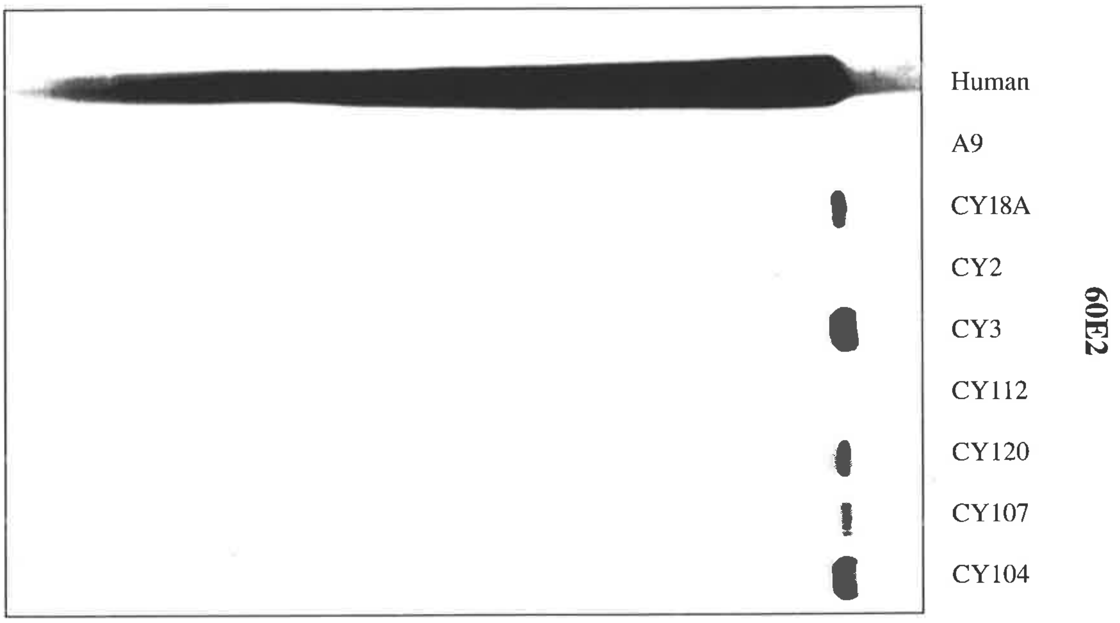
Table 3.2

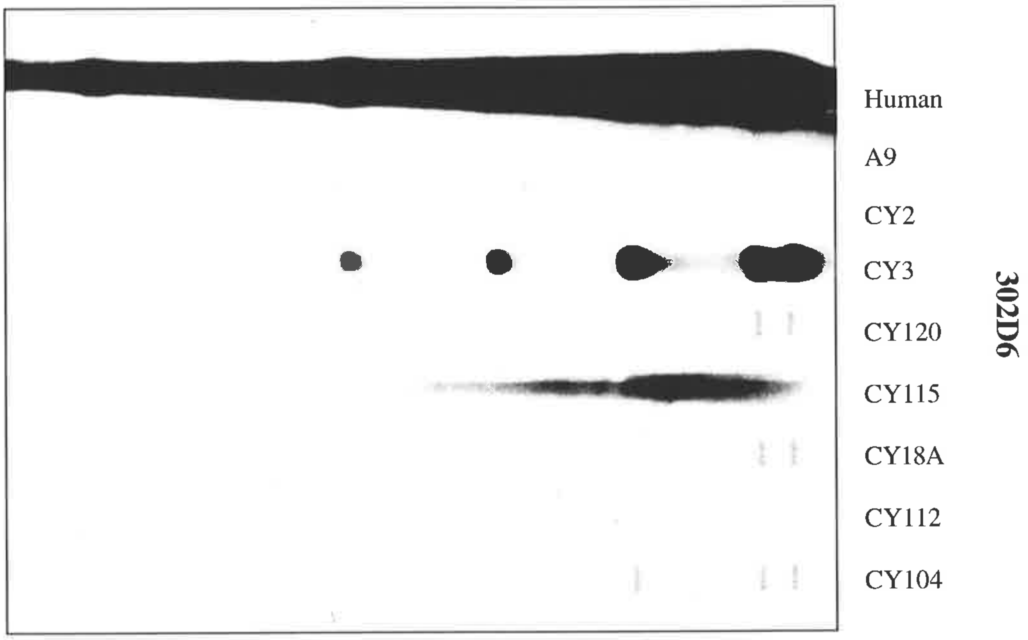
This table indicates the number of positive cosmids and contigs resulting from the chromosome 16 cosmid library screening with Alu PCR products from somatic cell hybrids CY2 and CY18A. The number of cosmids and contigs have been grouped according to their potential localisation in Regions I, II or III of the 16q24 chromosomal region.

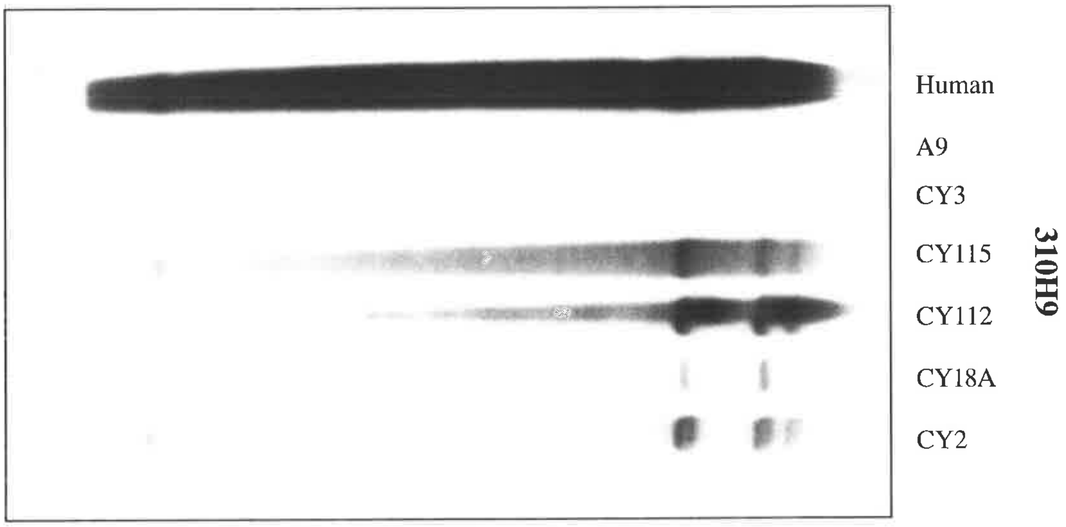
	Number of contigs	Number of cosmids in contigs	Number of singleton cosmids	Total number of cosmids
Region I (CY18A)	10	13	10	23
Region II (CY18A / CY2 overlap)	10	22	3	25
Region III (CY2)	2	2	2	4

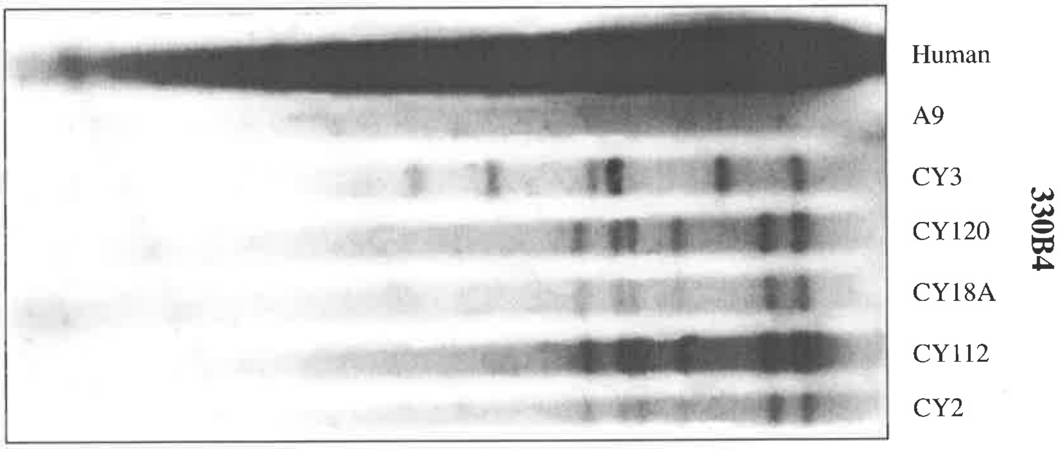
Figure 3.4

Representative autoradiographs of cosmids 60E2, 302D6, 310H9, 330B4 and 303B4, localised to 16q24 by Southern analysis. Each filter contained HindIII digested DNA from human, mouse cell line A9, and somatic cell hybrids containing the most distal portion of the chromosome 16 long arm.


Human
A9
CY18A
CY2
CY3
CY112
CY120
CY107
CY104
60E2


Human
A9
CY2
CY3
CY120
CY115
CY18A
CY112
CY104
302D6


Human
A9
CY3
CY115
CY112
CY18A
CY2
310H9


Human
A9
CY3
CY120
CY18A
CY112
CY2
330B4

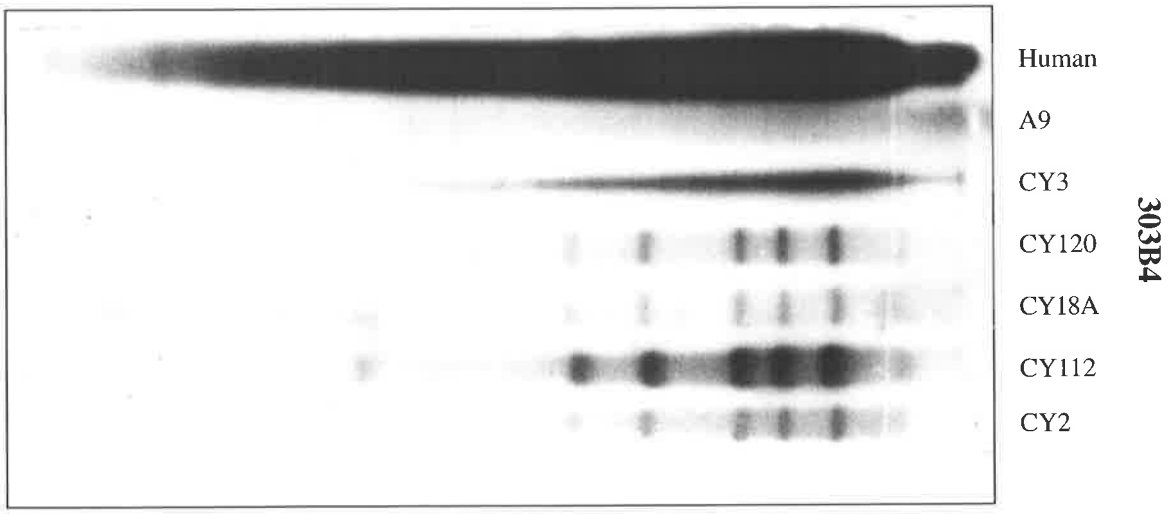

Human
A9
CY3
CY120
CY18A
CY112
CY2
303B4

Table 3.3

Cosmids positive for Alu PCR products from CY18A localised to 16p13 by Southern analysis. The '+' and '-' indicate the somatic cell hybrids which were homologous to these cosmids.

Cosmid	Contig no.	Human	CY18	A9	CY3	CY2	CY120	CY115	CY18A	CY112	Database
7D3	228	+		+	+	-	-		+	-	
5G6	228	+		-	+	-	-		+	-	
48C7	228	+		+	+	-	-		+	-	
71H10	188	+		-		-	-	-	+	-	
319A11	188	+		-	+	-	-	-	+	-	CY12-CY180A (16p11)
329A8	236.1	+		-	+	-		-	+	-	
314H12	236.1	+		-	+	-		-	+	-	
41A10	55.5	+		+	+	-	-	-	+	-	
319B5	S	+	+	-	+	-	-		+	-	

The cosmids mapped to the 16q24 chromosomal region by Southern analysis are listed in table 3.4 and their locations with respect to somatic cell hybrid breakpoints are shown in figure 3.5. A total of 32 of the 52 originally identified cosmids map to their expected locations at 16q24. A summary of the Alu PCR positive cosmids, displayed according to their contig status is shown in table 3.5. The results indicated that 12 of 23 possible cosmids mapped to Region I (CY18A), 18 of 25 cosmids mapped to Region II (CY2 / CY18A overlap) and 2 of 4 cosmids mapped to Region III (CY2). The results from the Southern analyses were compared to current chromosome 16 mapping data recently by accessing the LANL World Wide Web site, http://www-ls.lanl.gov. A number of cosmids localised to 16q24 were members of cosmid contigs from which more than one cosmid member was identified as being positive following screening of the chromosome 16 cosmid library. Table 3.6 shows the number of cosmids which were positive in each of these contigs, termed multiple hit contigs. Several contigs, such as 67.2, had a cosmid positive for CY18A but not for CY2, and vice versa, but the same cosmid was not positive for both hybrids. The localisation of more than one cosmid per contig provided further verification of the contig assembly which was based on repetitive sequence fingerprinting performed at LANL.

An estimation of the minimum amount of 16q24 represented by these cosmids is 2.04 megabases. This figure is based upon addition of the lengths of cosmid contigs, assembled by repetitive sequence fingerprinting (described in 2.8.2.1), and singleton clones. The data regarding the lengths of each cosmid contig and singleton cosmid was obtained from the LANL World Wide Web site.

Table 3.4

Cosmids localised to 16q24 by Southern analysis. The Alu PCR products and somatic cell hybrids which showed homology to each cosmid are indicated by '+' and '-'.

Cosmid	Contig no.	Alu CY2	Alu CY18A	Human	CY18	A9	CY3	CY2	CY120	CY115	CY104	CY107	CY18A	CY112	Database	FISH
307A2	665		(+ 348B2)	+		+	+	-	+		+		+	-	321D1, q24 CY18A(P2)-CY112	
305D7	682.4.2		+	+	+	+	+	-	+	+	+		+	-	21B7, p12.3 CY13-CY15	
60E2	682.3		+	+		-	+	-	+		+		+	-		
302D6	198.1		+	+		-	+	-	+		+		+	-		confirmed
50E12	198.1		+ (+ 50H12)	+		-	+	-	+				+	-	50E12, q13 CY18A(P1)-CY126	
41E11	S		+	+		-	+	-		+			+	-		
50C12	S		+	+		-	+	-	+				+	-		
302H6	S		+	+	+	-	+	-			+		+	-		
312E9	S		+	+		-	+	-	+		+		+	-		
307F8	S		+	+		-	+	-		+			+	-		
45F2	S		+	+		-	+	-	+		+		+	-		
302G7	231.1		(+ 34F10)	+	+	-	+	+	+				+	-	34F10, 308G6, 38H11, q24 CY2/3-CY18A(D2)	confirmed
310H9	76.1	(+ 317G1)	(+ 51A1, 317G1)	+		-	-	+		+			+	+		confirmed
14G10	76.1		+	+		-	-	+		+			+			confirmed
301F4	639	+	+	+		-	-	+		+			+	+		confirmed

Cosmid	Contig no.	Alu CY2	Alu CY18A	Human	CY18	A9	CY3	CY2	CY120	CY115	CY104	CY107	CY18A	CY112	Database	FISH
304F5	67.2	(+ 311F9)	(+ 73H6)	+	+	-	-	+	+		+		+	+		
330B4	549	+	+	+	+	-	+ (different bands)	+	+				+	+		confirmed + 16p11.2
305C6	549	+	+	+		-	+ (different bands)	+	+				+	+		
318A10	549	+	+	+		-	+ (different bands)	+		+			+	+		
340E8	540	+	+	+		-	-	+	+				+	+		confirmed
380C10	540	+	+	+		-	-	+	+				+	+		
304A7	658	+	+	+		-	-	-	+	+	+		+	+		confirmed + 16p11.2
329E4	658	+	+	+		+	+ (different bands)	+		+			+	+		
310D1	10.2	(+ 307A10, 303G7)	(+ 307A10)	+	+	-	-	+	+				+	+	303G7 CY2/3CY18A(D2)	confirmed
303B4	S	+	+	+		-	-	+	+				+	+		confirmed
313G3	S	+	+	+		-	+ (different bands)	+	+				+	+		confirmed + 16p11
309A3	S	+	+	+		-	-	+		+			+	+		confirmed
317E5	S	+		+	+	-	-	+	+				-	+		confirmed
37B2	S	+ (+ yh09a04)		+	+	-	-	+	+		+		-	+		

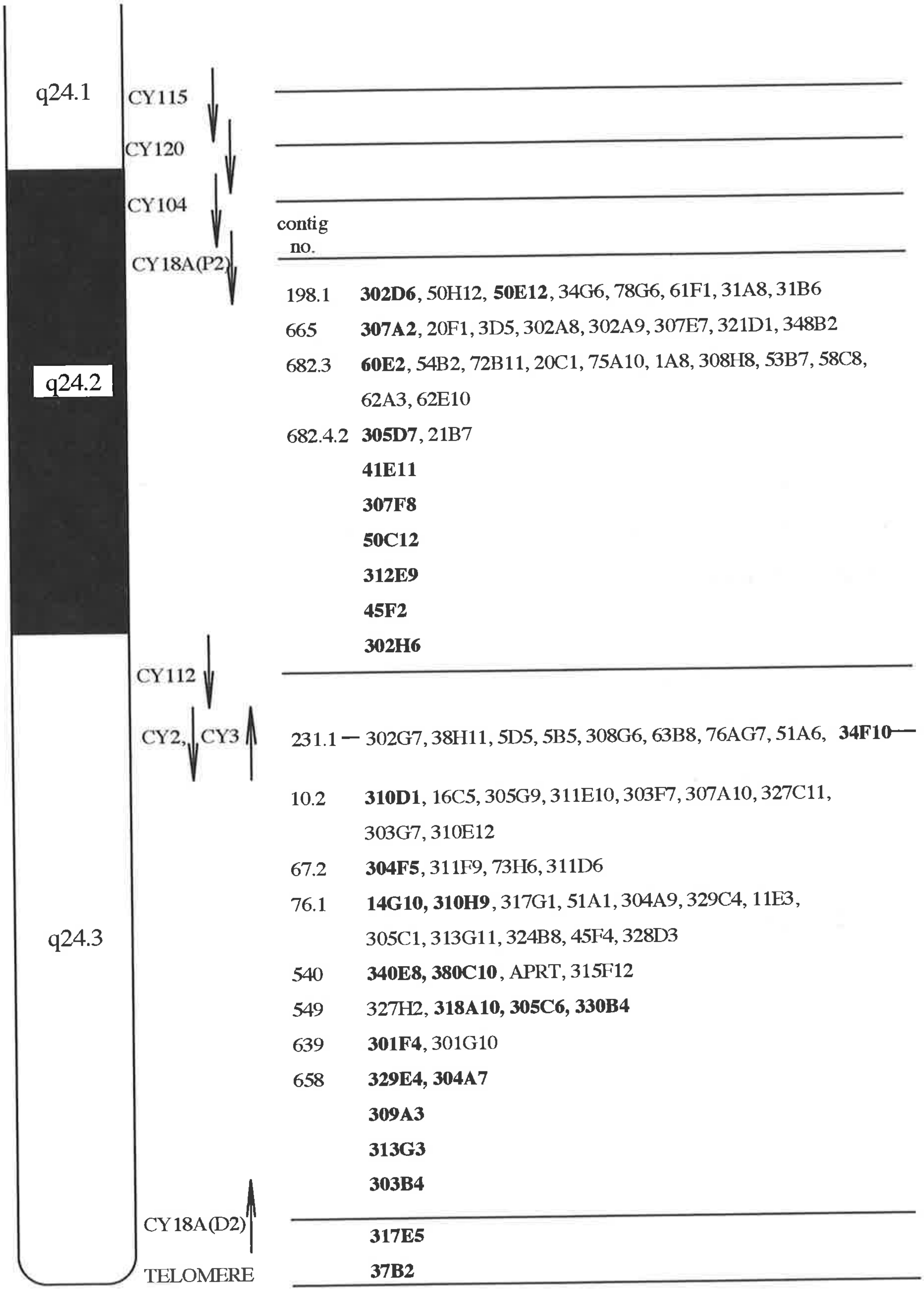
Figure 3.5

Cosmids indicated in bold lettering were localised by Southern analysis to the 16q24 chromosomal region. The additional cosmids shown are members of the contig containing the localised cosmid.

The APRT singleton cosmid, containing the APRT gene, was demonstrated to be homologous to cosmids in contig 540 and was included as a member of this contig. An additional gene, GALNS (N-acetylgalactosamine 6-sulfatase), was previously demonstrated to be homologous to contig 540 (Morris *et al.* 1994).

Contig 231.1 has been localised to the CY3 / CY2 hybrid breakpoint.

16q24 cosmids and their contigs



contig no.	
198.1	**302D6**, 50H12, **50E12**, 34G6, 78G6, 61F1, 31A8, 31B6
665	**307A2**, 20F1, 3D5, 302A8, 302A9, 307E7, 321D1, 348B2
682.3	**60E2**, 54B2, 72B11, 20C1, 75A10, 1A8, 308H8, 53B7, 58C8, 62A3, 62E10
682.4.2	**305D7**, 21B7
	41E11
	307F8
	50C12
	312E9
	45F2
	302H6
231.1 —	302G7, 38H11, 5D5, 5B5, 308G6, 63B8, 76AG7, 51A6, **34F10**—
10.2	**310D1**, 16C5, 305G9, 311E10, 303F7, 307A10, 327C11, 303G7, 310E12
67.2	**304F5**, 311F9, 73H6, 311D6
76.1	**14G10, 310H9**, 317G1, 51A1, 304A9, 329C4, 11E3, 305C1, 313G11, 324B8, 45F4, 328D3
540	**340E8, 380C10**, APRT, 315F12
549	327H2, **318A10, 305C6, 330B4**
639	**301F4**, 301G10
658	**329E4, 304A7**
	309A3
	313G3
	303B4
	317E5
	37B2

Table 3.5

Number of positive cosmids and contigs located in Regions I, II or III of 16q24.

	Number of contigs	Number of cosmids in contigs	Number of singleton cosmids	Total number of cosmids
Region I (CY18A)	5	6	6	12
Region II (CY18A / CY2 overlap)	8	15	3	18
Region III (CY2)	0	0	2	2

Table 3.6

Number of Alu PCR positive cosmids which are members of multiple hit contigs. The number of cosmids localised to each region (I, II or III) is indicated.

Region I (CY18A)	**Region II** (CY18A / CY2)	**Region III** (CY2)	No. cosmids in contig	Contig No.
	1	1	9	10.2
1		1	4	67.2
2	1		12	76.1
3			8	198.1
	2		4	540
	2		4	549
	2		2	658

3.3.3 Identification of Cosmids Using STSs and Cosmid Ends Localised to 16q24.3

As stated, it was decided to restrict the construction of the detailed physical map to a smaller part of 16q24 based on the localisation of the TSG and the FAA genes to the 16q24.3 region. Contributions to the construction of this physical map were made by myself, Mr. Scott Whitmore and Ms. Joanna Crawford, Department of Cytogenetics and Molecular Genetics, Women's and Children's Hospital, South Australia.

The strategy of STS mapping was used to identify new cosmids mapping to the 16q24.3 region. The STSs already localised in the 16q24.3 region from ongoing physical mapping efforts of chromosome 16 were utilised as hybridisation probes for screening filters of the gridded chromosome 16 specific cosmid library with a ten times coverage of chromosome 16 (LANL) (2.8.2). These markers included four novel cDNAs, c113, EST06702 (D16S532E), yc81e09 and yh09a04 and three genes, cell adhesion regulator (CMAR), renal dipeptidase (DPEP1) and melanocyte stimulating hormone receptor (MC1R). The cosmid end from 317E5, which is homologous to the BBC1 gene, and the microsatellite marker 16AC 6.26 (D16S303) were also used.

The approach of cosmid walking was then utilised to extend the singleton cosmids and cosmid contigs mapped to this region. This involved the generation of cosmid ends to probe the ten times coverage chromosome 16 cosmid filters. The isolation of cosmid ends was performed by S. Whitmore and J. Crawford. Cosmid DNA was digested with NotI, to release the insert, together with a range of restriction enzymes including BamHI, HincII and XbaI. The digests were electrophoresed and nylon filters were prepared by Southern blot transfer (2.5.5.2). DNA fragments representing the terminal sequences of the genomic insert, or cosmid ends, were identified by the successive hybridisation of T7 and T3 bacteriophage promoter sequences to the filters. These sequences flank the sCos1 NotI cloning site (2.8.1). Preparative gels were made and the DNA fragments that hybridised to the T7 and T3 oligomers were excised in agarose.

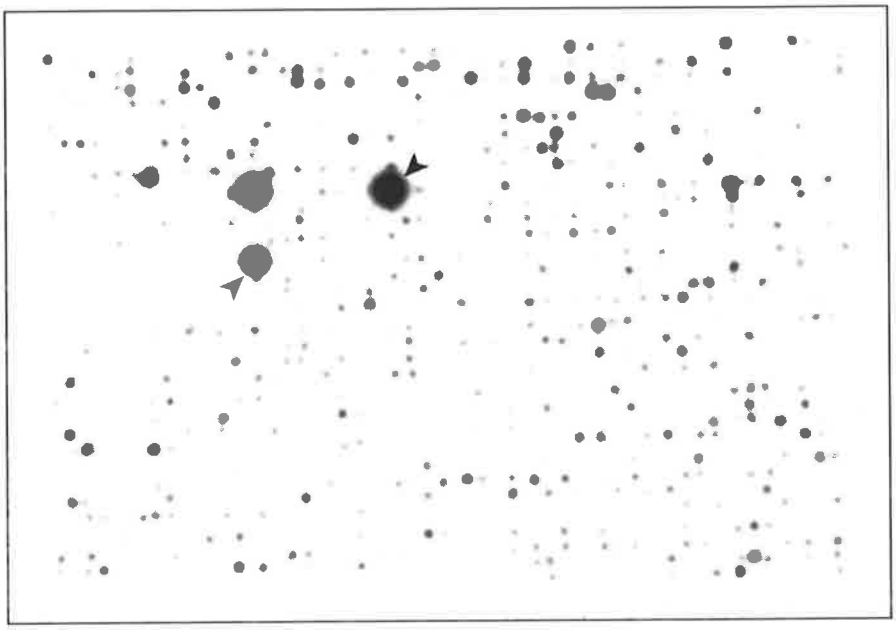
My contribution to this work involved screening the ten times coverage high density cosmid grids with probes of the STSs and cosmid ends (2.8.2.2) to identify positive cosmids. Examples of these autoradiographs are shown in figure 3.6. In order to confirm whether the identified cosmids were true positives, dot blots of cosmids (2.5.5.1) were screened with their corresponding probe to detect a positive signal. Cosmid overlaps were confirmed by a restriction enzyme fingerprinting protocol performed by S. Whitmore and J. Crawford.

Table 3.7 shows cosmids which were positive for the STSs and also displays cosmids which were homologous to various cosmid ends which were generated during contig construction. Figures 3.7 - 3.10 exhibit the cosmids assembled in their respective contigs. Markers yc81e09, MC1R and EST06702 shared homology with several cosmids, therefore were linked in the same contig. Cosmids homologous to the CMAR marker, 317E5 cosmid ends and the DPEP1 marker were linked in the same contig following the identification of cosmids using cosmid ends, and their organisation into contigs. The CMAR/317E5/DPEP1 contig linked up with the yh09a04 contig by cosmid 378G9. The yc81e09/MSH-R/EST06702, CMAR/317E5/DPEP1/yh09a04, c113 and 16AC 6.26 (D16S303) contigs are estimated to extend over 650 kb of genomic DNA in the 16q24.3 region.

Figure 3.6

Representative autoradiographs of the gridded array of chromosome 16 cosmid clones, from the ten times coverage library, hybridised with radiolabelled ends of cosmids 348F3 (top) and 317E5 (bottom). Positive signals are indicated by the arrows. One membrane of the cosmid library filters, from a set of ten, is shown.

348F3
end



317E5
end

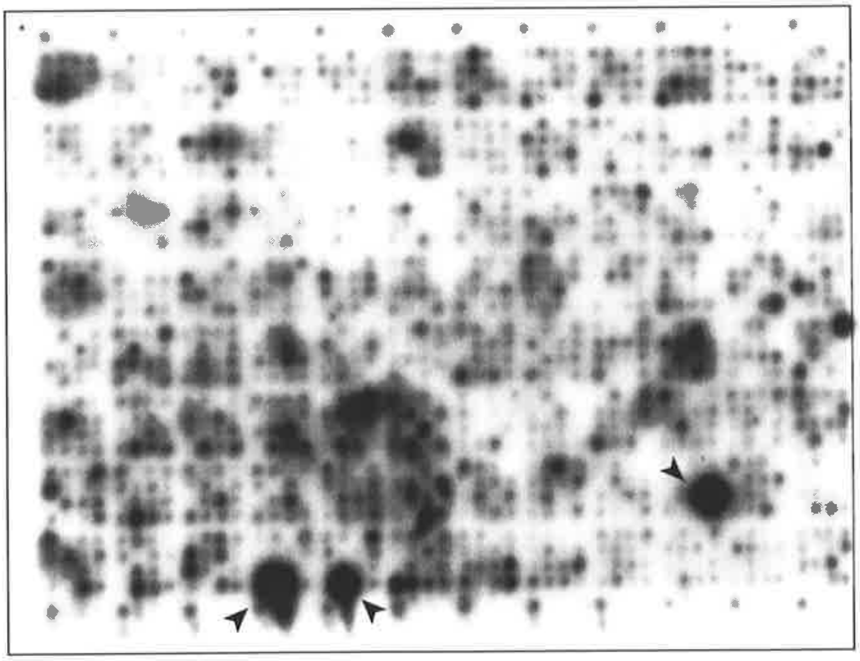


Table 3.7

Cosmids identified by hybridisation of the chromosome 16 cosmid library with STSs mapped to the 16q24.3-qter region. Cosmids which were homologous to various cosmid ends which were generated during contig construction are also displayed.

STS or Cosmid End	Cosmids Identified
yc81e09 (ScDNA-A55)	377F1
	336A8
	416C1
	337G8
	374E6
	360D7
360D7 end	354E12
354E12 end	354E12
	360D7
MC1R gene	374E6
	337G8
	416C1
	336A8
	377F1
377F1 end	377B2
	318C5
318C5 end	393B2
	403B8
403B8 end	403B8
	393B2
EST06702	377F1
	336A8
	416C1
	337G8
	374E6

STS or Cosmid End	Cosmids Identified
yh09a04 cDNA	352A12
	309E9
	361G2
	361H2
361H2 end	378G9
378G9 end	378G9
352A12 end	431F1
431F1 end	387F9
	372A3
	436G3
	444B9
444B9 end	444B9
c113 cDNA	434D8
	396B4
	361E4
	372C1
	348F3
348F3 end	434E1
	412H12
	408G1
	365A7
365A7 end	408G1
	365A7
434D8 end	396B4
	434D8
16AC 6.26 STS	425E4
425E4 T7 end	369E1
369E1 end	369E1
425E4 T3 end	425E4

STS or Cosmid End	Cosmids Identified
CMAR gene	432B2
	384F2
317E5 end (BBC1)	408G10
	317E5
	432B2
	376F9
	410H4
	342G5
	325G2
	384F2
	383H6
383H6 end	447A5
	427B11
427B11 end	427B11
408G10 end	383G9
	422A5
	365H1
DPEP1 gene	444C11
	444D11
	324F5
	435H5
	435H6
444C11 end	340B6
340B6 end	365H1
	383G9
	317E5
	410H4
435H6 end	421E4
	371D5
371D5 end	344H1
344H1 end	378G9

Figure 3.7

Three cosmid contigs constructed after initial hybridisations of the chromosome 16 cosmid library with probes of the CMAR, BBC1 (317E5) and DPEP1 genes, and subsequent hybridisations with ends of newly identified cosmids, were linked to form one contig of approximately 220 kb in size.

CMAR / 317E5 / DPEP1 contig

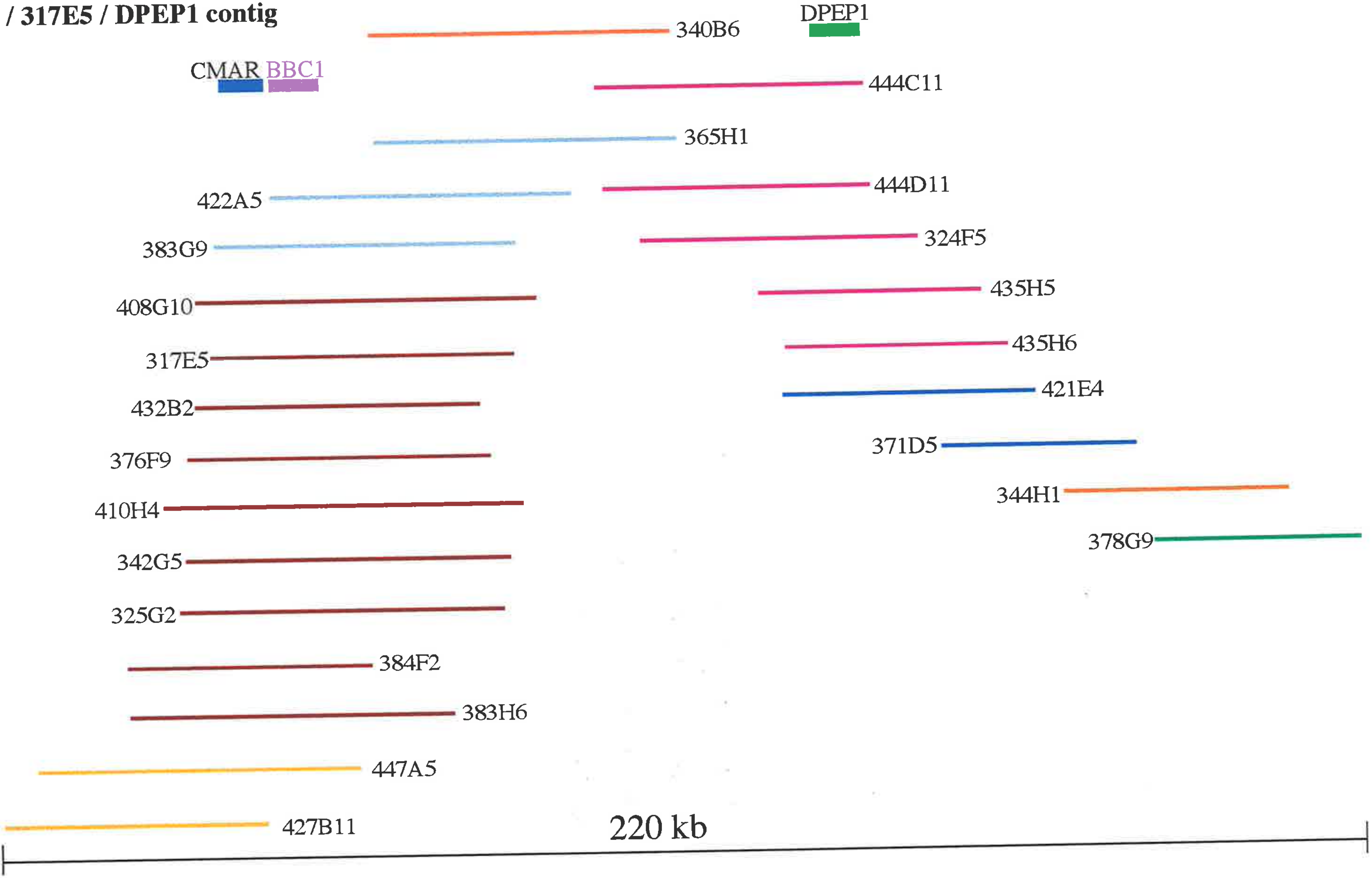


Figure 3.8

Three cosmid contigs constructed after initial hybridisations of the chromosome 16 cosmid library with probes of ScDNA-A55, EST06702 and MC1R, and subsequent hybridisations with ends of newly identified cosmids, were linked to form one contig of approximately 100 kb in size.

yc81e09 / MC1R contig

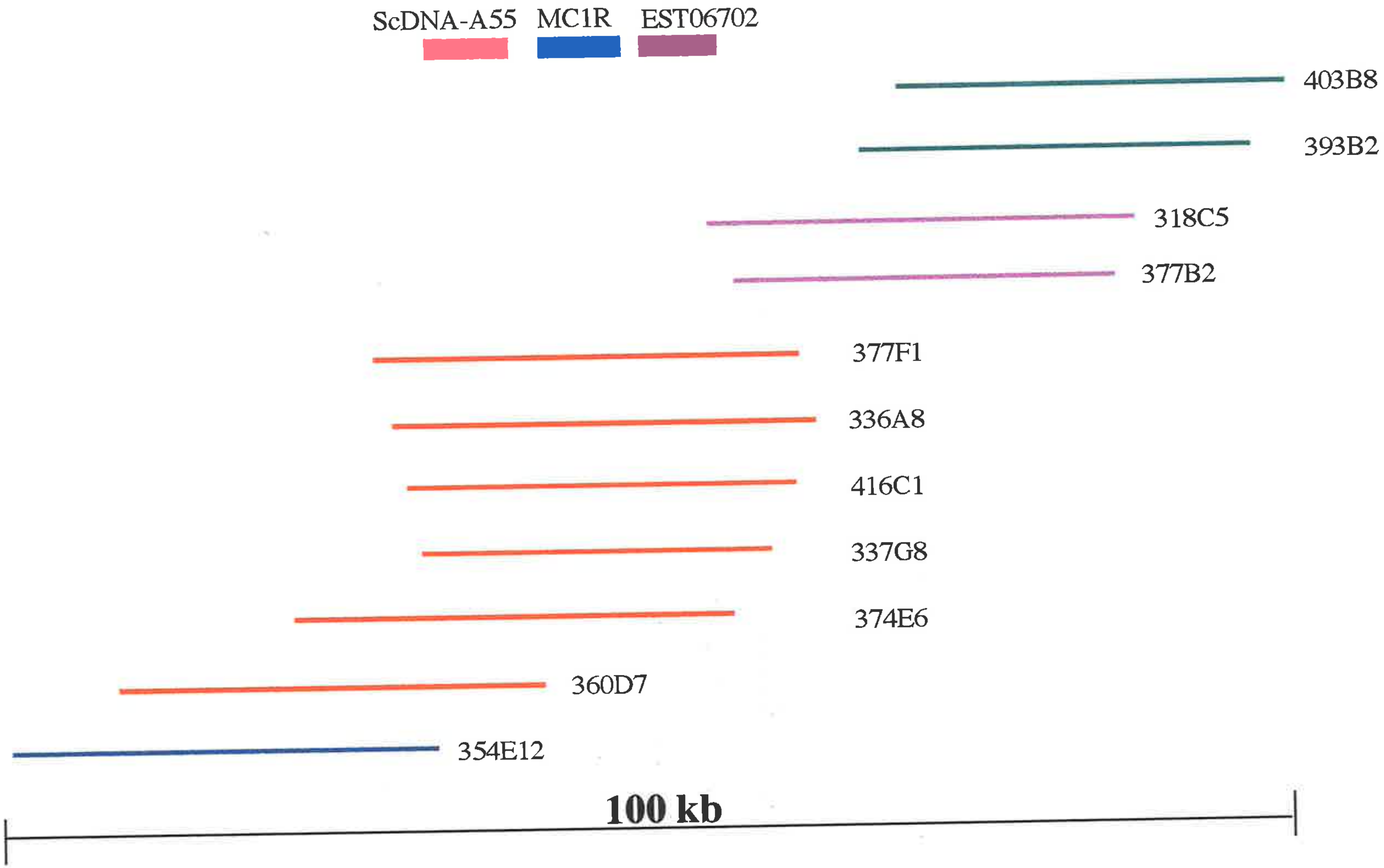


Figure 3.9

Cosmid contig constructed after initial hybridisation of the chromosome 16 cosmid library with the yh09a04 cDNA and subsequent hybridisations with ends of newly identified cosmids generates one contig of approximately 140 kb in size.

yh09a04 contig

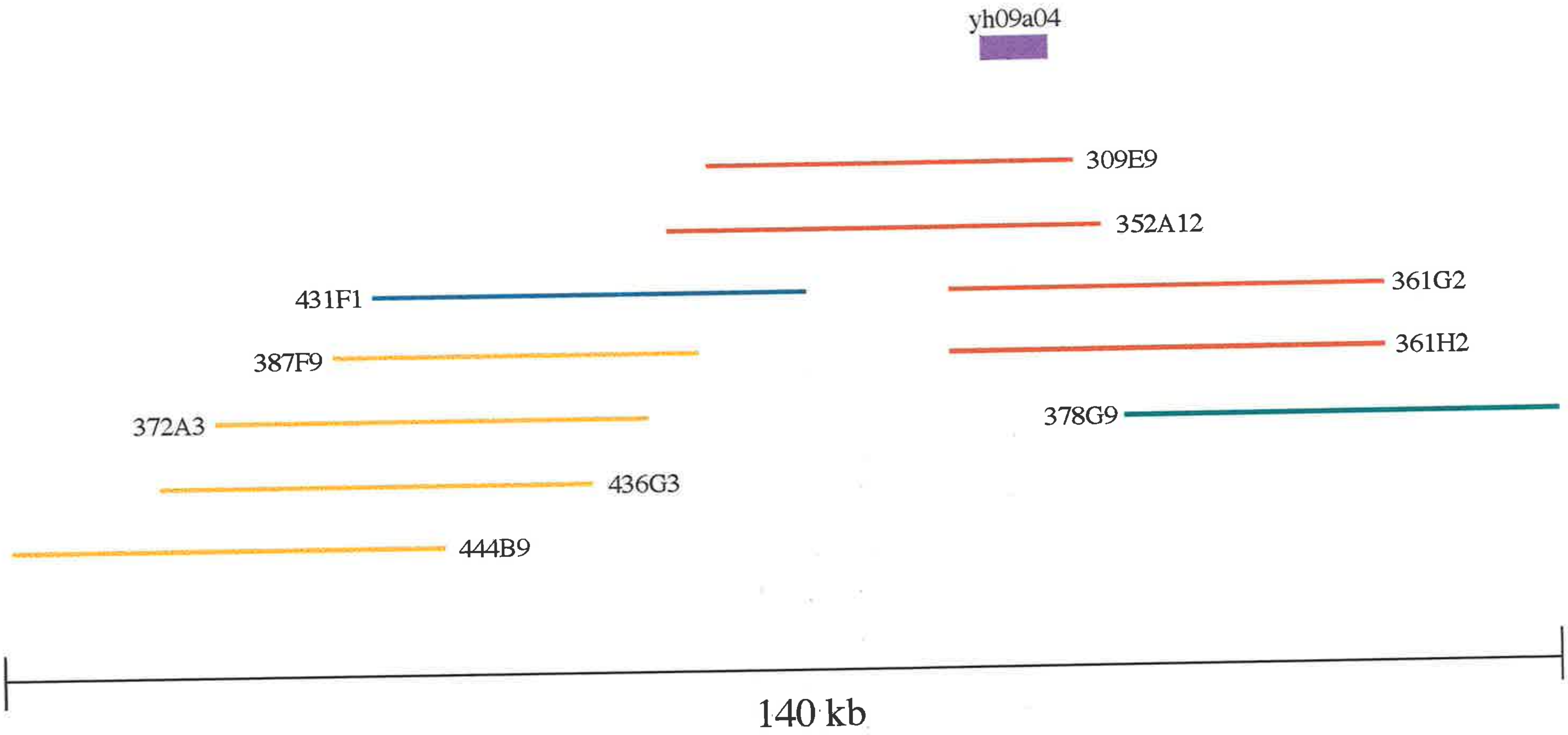
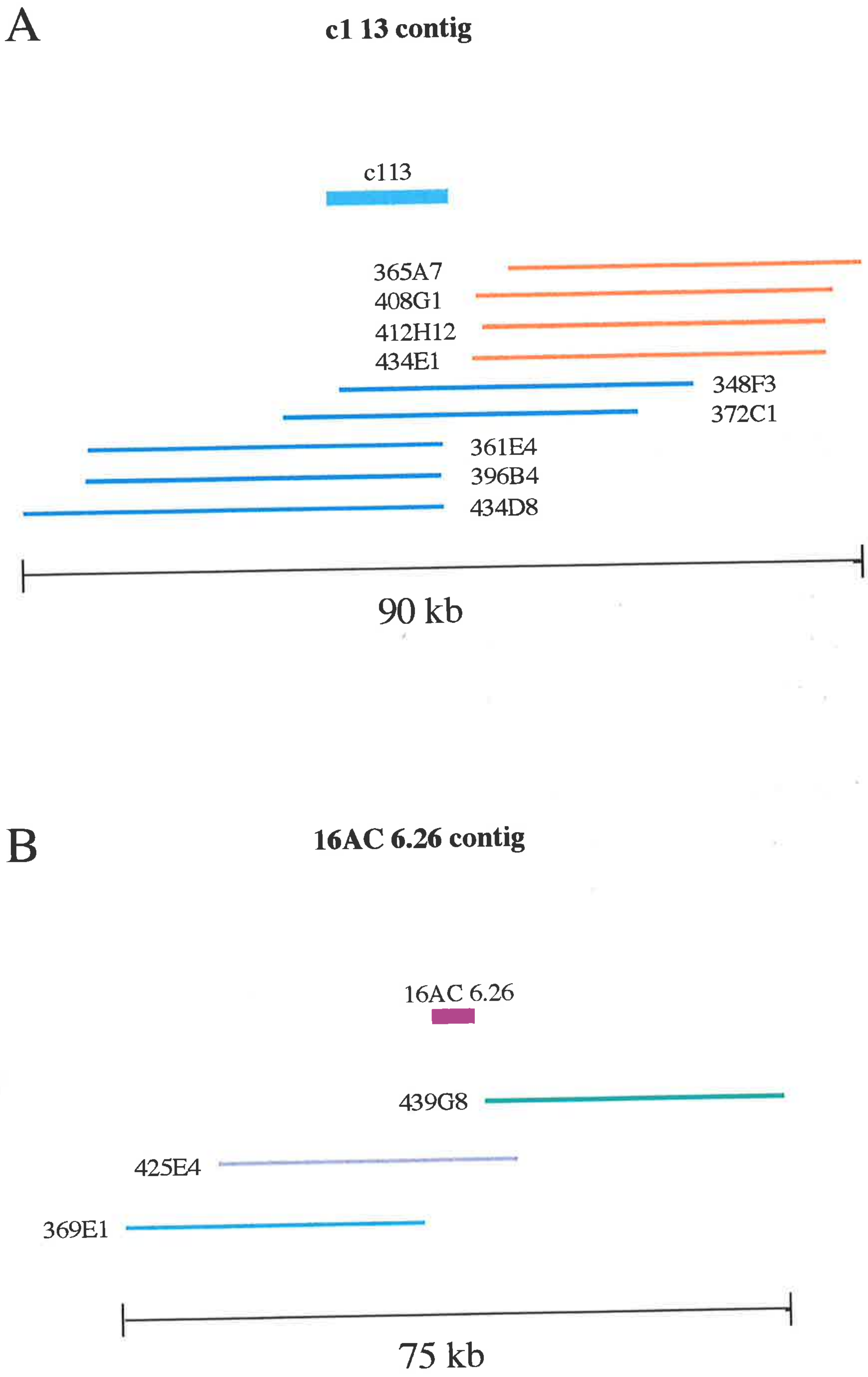


Figure 3.10

A. A cosmid contig of approximately 90 kb in size was constructed after initial hybridisation of the chromosome 16 cosmid library with the c113 probe and subsequent hybridisations with ends of newly identified cosmids.

B. A cosmid contig of approximately 75 kb in size was generated after initial hybridisation of the chromosome 16 cosmid library with the 16AC 6.26 microsatellite and subsequent hybridisations with ends of newly identified cosmids.


A
c1 13 contig
c113
365A7
408G1
412H12
434E1
348F3
372C1
361E4
396B4
434D8
90 kb
B
16AC 6.26 contig
16AC 6.26
439G8
425E4
369E1
75 kb

3.4 DISCUSSION

A combination of 517 and A-33 primed Alu PCR products from the CY2 and CY18A somatic cell hybrids and hybridisations to approximately 4000 cosmids of the chromosome 16 specific cosmid library, has enabled the targeted isolation of 32 cosmids from the predefined chromosomal subregion, 16q24. The Alu PCR procedure allows the selective amplification of human specific inter-Alu sequences from a mouse background. The efficiency of the method was evaluated by investigating the number of individual cosmids mapping to somatic cell hybrids, containing DNA from the most distal portion of chromosome 16, by Southern analysis. From a total of 52 cosmids placed in one of three regions in which they could potentially map, 32 cosmids were identified as being placed in the correct chromosomal subregion. The results indicated that a higher percentage of cosmids, 72%, mapped to their expected location for Region II, CY2 / CY18A overlap, compared with 52% mapped to Region I, CY18A only. This may be due to a better interpretation of the signal obtained from the two sets of autoradiographs for cosmids mapping to Region II, in contrast to interpretation of positive signal from one autoradiograph (CY18A) for Region I. Also, cosmids which belonged to multiple hit contigs, of which there was a large number for Region II, had an increased chance of being mapped to their correct location.

The Alu PCR technique has been used in a variety of applications including the hybridisation of Alu PCR products from human chromosome 21 somatic cell hybrids to identify YAC clones from a megabase insert-size YAC library (Chumakov *et al.* 1992). These YACs were subsequently assembled into large contigs covering most of the long arm of chromosome 21. The Alu PCR method has also been successfully applied in the isolation of cosmids from chromosomal subregions by numerous authors. These include Nahmias *et al* (1995) who identified a total of 1400 cosmids from a chromosome 9 cosmid library following hybridisation of Alu PCR products derived from irradiation hybrids containing the 9q34 region. These cosmids were assembled into contigs by a fingerprinting procedure and are currently being used as a resource for the isolation of expressed sequences that may assist in the positional cloning of the tuberous sclerosis disease gene (TSC1) which is localised to 9q34. Zucman *et al* (1992) identified 125 independent loci from a chromosome 22 cosmid library, following hybridisation of Alu PCR products derived from four chromosome 22 somatic cell hybrids. These hybrids contained unique regions of chromosome 22q together with common regions of overlap. Eighty-five percent of the cosmids analysed were demonstrated to map to their corresponding locations. Rohme *et al* (1994) demonstrated a similar frequency of cosmids mapping to their corresponding region. Alu PCR products isolated from two X chromosome radiation hybrids containing region Xq13.1 were used as hybridisation probes to identify a total of 39 cosmids from an X chromosome cosmid library. A total of 11 of 13 analysed cosmids (85%) were demonstrated to map to the correct region.

An explanation for the higher percentage of erroneously localised cosmids observed in this study compared with the percentage observed in the published results (Zucman *et al.* 1992; Rohme *et al.* 1994) may be the inability to differentiate reliably weak positive signals from background. A number of false positive cosmids identified were members of contig 55 or mapped to regions of the chromosome 16 specific low abundance minisatellite repeats, 16p11.2, 16p12, 16p13.11, and 16q22.1 (Stallings *et al.* 1992b; Stallings *et al.* 1993). These false positive cosmids may contain repetitive elements common to the 16q24 and low abundance minisatellite repeat regions. Stallings *et al* (1992b) initially detected these repeats when difficulty was experienced in ordering cosmid clones into contigs for some regions of chromosome 16. Contig assembly was achieved by identification of pairs of cosmid clones that shared similar sized restriction fragments containing similar repetitive sequences (Stallings *et al.* 1990). A set of 78 clones containing the chromosome 16 specific low abundance repeats resulted in one large contig, number 55, comprising 2% of all cosmids. These cosmids could not be ordered in a linear array to form a contig, as the large number of overlapping clones either displayed branching in many different directions, or formed stacks that extended in an upward direction. This was due to the low abundance repeat sequences causing false overlaps between non-contiguous clones.

Southern analysis of several cosmids, including 330B4 and 304A7, demonstrated these cosmids to be localised between hybrid breakpoints CY112 and CY18A(D). Their autoradiographs demonstrated a series of human bands in CY3 which were distinct from the human bands demonstrated in CY2 and the mouse bands in A9. The Southern analysis of cosmid 330B4 is shown in figure 3.4. This result implies that these cosmids map at 16q24.3 and also another region of chromosome 16. FISH results demonstrated signal in two regions, 16qter and 16p11.2, the latter being the region known to contain the chromosome 16 specific low abundance repeats. Thus, the 16q24 and 16p11.2 regions share common repetitive elements.

Results from Southern analysis also led to the identification of a number of discrepancies in contig assembly and inconsistencies with mapping data available from the LANL chromosome 16 cosmid database. One example is cosmid 37B2, mapped between CY18A(D) and the telomere of 16q using the yh09a04 cDNA which is homologous to this cosmid. Originally, this cosmid was a member of contig 87.2 which was first localised to 16q24.2, but is now mapped to 16q13 using STSs from cosmids 4D5 and 26E3. Cosmid 305D7 from contig 682.4.2 mapped to 16q24 but the database suggested that 21B7, another member of this contig, maps to 16p13. Cosmids 302D6 and 50E12 from contig 198.1 were localised to 16q24, but the database suggests that 50E12 maps to 16q13. These discrepancies may be due to the presence of the chromosome 16 low abundance repeats leading to false overlaps in the contig map, chimaeric clones, deleted clones or chromosome duplications. Another reason may be due to the inadequate stringency of the algorithm being used for the identification of overlaps. Thus, the independent hybridisations performed in this study were beneficial in assessing the accuracy of the chromosome 16 contig map.

In this Alu PCR study, products from hybrids CY2 and CY18A led to the successful isolation of 32 cosmids in the 16q24 subregion. Other methods used for the isolation of human sequences from hybrid cells require direct cloning to produce a large recombinant library and the identification of clones based upon hybridisation to human repetitive sequences. The Alu PCR technique circumvents the construction of libraries from hybrids and allows amplification of human sequences directly from hybrid DNA. Alu PCR material from chromosome 16 cell hybrids used for screening the chromosome 16 specific cosmid library specifically focuses the cosmid screening only on this chromosome. Thus, even hybrids that contain non chromosome 16 human material, in addition to chromosome 16 DNA, can be used to identify cosmids derived from chromosome 16. This is the situation for the CY2 hybrid which contains some of the X chromosome. Also, because the various inter-Alu sequences are amplified to differing extents, their direct cloning would have resulted in recurrent isolation of the same probes. Instead, when these amplified sequences are used for cosmid library screening, the different levels of amplification are evidenced by differing intensities of hybridisation signals observed for the corresponding cosmids.

Several contigs were identified in which more than one cosmid member was positive following screening of the chromosome 16 cosmid library. An example of these contigs, termed multiple hit contigs, is contig 67.2 which possessed a cosmid positive for CY18A but not for CY2, and vice versa. This may have been due to variable amplification of DNA by Alu PCR. Alternatively, certain contig members may not have been identified because of low signal intensity. The identification of more than one cosmid per contig using Alu PCR products also provided further verification of the contig assembly which was based on repetitive sequence fingerprinting (LANL).

Alu PCR has significant limitations as it will only occur between pairs of Alu sequences that are positioned at an appropriate distance and orientation with respect to each other. The utility of this technique also has restrictions based on the apparent asymmetric distribution of the repeat sequences within the genome, leaving some regions of the genome significantly under represented. The Alu-bubble PCR technique (Munroe *et al.* 1994) is an alternative method which can be used on any human fragment that contains a single Alu sequence. This technique has been demonstrated to generate ten times more products in YACs than Alu PCR. The representation and distribution of amplification products is also increased. FISH analysis using a probe of products generated from a hybrid containing whole chromosome 11 has been demonstrated to highlight the whole of chromosome 11 (Munroe *et al.* 1994). Thus, the more complex probe generated through Alu-bubble PCR could be utilised to screen the chromosome 16 library to possibly identify more cosmids mapping to the 16q24 chromosomal region.

It is likely that a greater number of cosmids map to the gene-rich and GC-rich 16q24 region, but these were not identified in this study. This may have been due to the limitations of the gridded (one times coverage) chromosome 16 cosmid library which was used. The chromosome 16 cosmid contigs were generated by enriching for cosmids with (GT)n repetitive sequences which then formed the basis of the repetitive sequencing fingerprinting procedure (Stallings *et al.* 1990). This chromosome 16 cosmid library does not contain randomly distributed cosmids as these (GT)n repeat regions are known to be deficient in gene-rich regions (Stallings *et al.* 1991). It is therefore expected that the 16q24 region is not well represented in the filters of the gridded chromosome 16 cosmid library with one times coverage. Since the development of this cosmid library comprising 4000 clones, additional cosmids have been included in a set of filters containing approximately 14,600 gridded chromosome 16 specific cosmids. Thus, screening this cosmid library which has a ten times coverage of chromosome 16, with the Alu PCR products generated from hybrids CY2 and CY18A may possibly identify more cosmids mapping to the 16q24 region.

An alternative strategy which can be utilised to rapidly eliminate false positive clones, instead of Southern analysis of all the identified cosmids, is the initial hybridisation with Alu PCR probes from the somatic cell hybrids to dot blots of the identified cosmids. Cosmids displaying a negative signal for this hybridisation can be excluded and the remaining cosmids localised by Southern analysis.

The Alu PCR, STS landmark mapping and cosmid walking procedures are rapid, inexpensive and easy to perform. As illustrated, they have tremendous potential for the isolation and analysis of small regions of human DNA. A small number of cosmids were localised to the 16q24 region prior to the utilisation of the Alu PCR approach. Cosmids identified from screening the chromosome 16 cosmid library with Alu PCR products from CY2 and CY18A hybrids were estimated to represent an estimated 2.04 Mb of the 16q24 region. The STS landmark mapping and cosmid walking study, was successful in generating five contigs originating from nine STSs. These contigs have now been linked to each other, with the exception of contig c113, using cosmid ends which were isolated following my contribution to this project. These contigs are estimated to extend over 650 kb of the 16q24.3 region.

YACs were also utilised in an attempt to order cloned DNA within the 16q24.3 interval by PCR based STS mapping, but they only provided limited information. The megaYACs containing this region are apparently unstable and undergo rearrangement. If necessary, mapping can be extended by use of STSs to identify BACs which may link and integrate contig c113 to the large CMAR - 16AC 6.26 (D16S303) contig and enable a complete map of cloned DNA to be constructed.

Thus, the isolation of new loci in the 16q24 chromosomal region assists the construction of a physical map and contributes to the positional cloning of disease genes localised to this region, including the TSG involved in LOH and the FAA gene at 16q24.3. The new cosmids mapping to this region can also be screened for additional polymorphic repeat markers which can utilised to narrow the region of LOH in primary breast cancer. The development of ordered cosmid contig maps for this region will also provide rapid access to cloned DNA for the large scale sequencing goal of the HGP.

CHAPTER 4

Isolation of Transcripts Encoded by Cosmids Localised to 16q24

4.1 INTRODUCTION 144

4.1.1 Approaches for the Identification of Genes in Genomic DNA 145

4.1.1.1 Exon Trapping 147

4.1.1.2 Direct cDNA Selection Strategy for the Isolation of Transcribed Sequences 149

4.1.2 Comparison of the Exon Trapping and Direct cDNA Selection Techniques 152

4.2 MATERIALS AND METHODS 154

4.2.1 Preparation of cDNA and Cosmid DNA 156

4.2.2 Hybridisation and Isolation of Specific cDNAs 157

4.2.3 Analysis of cDNA Clones 158

4.2.4 Sequencing of cDNA Clones 159

4.3 RESULTS 160

4.3.1 Direct cDNA Selection System 160

4.3.2 Selection of cDNAs for the GALNS Gene 160

4.3.3 Isolation of Transcribed Sequences Using Cosmids Localised to 16q24 161

4.3.4 Localisation of cDNAs to Cosmid DNA 163

4.3.5 Sequence Analysis of Positive cDNA Clones 167

4.4 DISCUSSION 169

4.5 CONCLUSION 176

4.1 INTRODUCTION

The development of a transcript map of the human genome integrating the physical and cytogenetic maps is an important goal of the Human Genome Project (HGP). Physical maps comprising YACs and cosmids allow immediate access to large regions of the human genome and provide easily accessible materials for the analysis and characterisation of the coding potential of specific regions. Transcribed segments isolated from specific genomic regions are utilised in the positional cloning procedure for the identification of genes for human inherited traits assigned to those regions. Transcripts can be used as anchor points on the human genome map to facilitate ordering of large genomic clones and to confirm contig integrity. They can aid in the construction of a human gene map and can also serve as initiation points for the continuous sequencing of large genomic fragments, hence, they contribute to the ultimate goal of the HGP. Region-specific transcribed segments are essential reagents for studies involving human genome organisation including gene (and pseudogene) density and distribution, transcriptional activity and CpG island association and how these features relate to the cytogenetically observable features of G and R bands. Furthermore, once a disease gene in a particular region is cloned, the development of a detailed transcription map of this region will allow further assessment of possible regulatory relationships between the genes in that region.

The aim of the work presented in this chapter was to use a selection of cosmids localised to the 16q24 chromosomal region to identify transcribed sequences encoded by these cosmids, using the approach of direct cDNA selection. The cosmids chosen for this procedure were identified by the Alu PCR approach and were localised between the CY2 somatic cell hybrid breakpoint and the telomere of the long arm of chromosome 16, 16q (see chapter 3, figure 3.5). At the commencement of this part of the project, the 16q24 chromosomal region was known to possess a high concentration of genes (Saccone *et al.* 1992) and loss of heterozygosity (LOH) had been observed in sporadic breast tumours with marker D16S7, mapped to 16q22-q24 (Sato *et al.* 1991) distal to the CY2 somatic cell hybrid breakpoint, which indicated the presence of a tumour suppressor gene (TSG).

4.1.1 Approaches for the Identification of Genes in Genomic DNA

The construction of physical maps of large genomic segments has been achieved but the identification and recovery of expressed sequences from the cloned DNA remains a cumbersome task. A number of diverse techniques have been used to identify genes in cloned genomic DNA. These techniques identify genes by one of several features, which can be found in some but not all genes. For example, individual segments of genomic DNA can be examined for cross-species homology to detect evolutionarily conserved coding sequences by Southern analysis of 'zoo blots' (Monaco *et al.* 1986; Riordan *et al.* 1989). These genomic DNA segments can then be used as hybridisation probes to screen cDNA libraries to obtain more coding sequence. This technique is limited because not all genes are sufficiently conserved to show cross-species hybridisation. Also, this application can only examine short genomic segments at a time, making the systematic investigation of large portions of the chromosome difficult.

Another method is based on the observation that a large fraction (as high as 60%) of gene promoters, especially those of housekeeping genes, are associated with high concentrations of unmethylated CpG dinucleotides, ie. CpG islands (Bird, 1986). Thus, DNA around CpG islands is highly enriched for the first exons of genes. In this method, gene identification is performed by digestion of genomic DNA with a rare cutting restriction enzyme which cleaves preferentially in CpG islands (one which recognises a site with one or more CpG dinucleotides). These enzymes include NotI, BssHII, EagI and SacII (Larsen *et al.* 1992b). The identified sequences flanking the CpG islands are subcloned and used as hybridisation probes to screen cDNA libraries. This method is limited as the identification and subcloning steps can be laborious and, depending on the source of the cDNA library, the gene may or may not be represented. Additionally, all genes which do not have CpG islands cannot be identified.

cDNA libraries containing human transcripts derived from human/rodent somatic cell hybrids can also be used to isolate human transcribed sequences. In this method, heteronuclear RNA (hnRNA) is isolated from somatic cell hybrids (Liu *et al.* 1989) using hexamers reverse complementary to splice donor sites for first strand cDNA synthesis. The human transcripts can be identified by screening the resulting heteronuclear-cDNA library with total human DNA, since the original hnRNA still retains introns containing species-specific repetitive elements. This method has the advantage of scanning entire chromosomes or subchromosomal regions that are present in the hybrids. The limitations of this technique are that only genes with introns can be isolated, and these introns must contain repetitive elements. In addition, only those genes expressed in the hybrid cell line can be isolated.

The direct screening of cDNA libraries with radiolabelled genomic clones generated from YACs (Denizot *et al.* 1992) or microdissected regions of a chromosome (Kao and Yu, 1991) can also be used to identify genes. This technique has a number of difficulties, including low signal to noise ratio when screening cDNA libraries with YAC probes, resulting in low reproducibility. This is due to the sequence complexity of the labelled genomic DNA requiring a large amount of unlabelled total human DNA to block out high copy repeats in order to avoid the identification of erroneous repeat containing cDNAs.

With the recent development of automated sequencing techniques it has become feasible to partially sequence large numbers of random cDNA clones (Adams *et al.* 1993). Nevertheless, the chromosomal assignment of the ESTs that are generated requires considerable additional effort (Polymeropoulos *et al.* 1992). However, the recently published gene map of the human genome (Schuler *et al.* 1996) is evidence that this approach is rapidly progressing and is becoming an increasingly important resource for gene identification.

The identification and mapping of genes in large genomic regions requires sensitive methods with a high throughput. The methods so far described are not likely to rapidly detect the majority of genes in large genomic regions of, for example, a megabase or greater.

4.1.1.1 Exon Trapping

New approaches have been developed in recent years to rapidly identify expressed sequences in large chromosomal regions. These approaches include exon trapping and direct cDNA selection. Exon trapping (Buckler *et al.* 1991; Duyk *et al.* 1990; Church *et al.* 1994) is targeted at the capture of expressed sequences directly from large genomic regions. The basis of this procedure is the recognition of exon splice site sequences in genomic DNA by the cellular splicing apparatus. Intron sequences are removed from premessenger RNA molecules leading to conversion to mRNA and the rescue of transcriptional units.

The exon trapping procedure involves the digestion of genomic DNA at particular restriction sites (eg. BamHI) and the subcloning of fragments into a trapping vector, eg. pSPL3, at the same restriction site or a restriction site producing compatible ends. For example, if a single exon is contained within a BamHI genomic fragment and is flanked by a splice acceptor and splice donor site, this exon will potentially be "trapped". The constructs are transfected into mammalian cos-7 cells. After transient expression, RNA is isolated and reverse transcribed using a vector specific primer to yield first strand cDNA. Synthesis of the second cDNA strand is followed by digestion with BstXI. The splice acceptor and splice donor sites of the pSPL3 vector contain BstXI half sites. These sequences form a complete BstXI site if an exon is not trapped, thus are used to eliminate vector background. The resulting PCR products are sequenced to identify the trapped exon and can also be used to screen cDNA libraries to obtain a cDNA clone of the gene.

This procedure does not require prior knowledge of a gene's structure and is independent of the expression of the gene in a particular cell line. Exon trapping is a useful procedure in the positional cloning strategy for isolating candidate disease genes and has been successful for the isolation of genes including the gene responsible for Huntington disease (Huntingtons Disease Research Collaborative, 1993), the Menkes disease gene (Vulpe *et al.* 1993) and the glycerol kinase gene (Walker *et al.* 1993). Protocols have been designed to trap both internal exons (Buckler *et al.* 1991; Duyk *et al.* 1990; Church *et al.* 1994) and 3' terminal exons (Krizman and Berget, 1993). The 3' terminal exon trapping procedure has a number of advantages over internal exon trapping. These include larger sized 3' terminal exons, 200-400 bp on average, compared to the average 120 bp size of internal exons. Thus, 3' terminal exons yield greater sequence information per trapped exon. The vast majority of 3' untranslated regions (UTRs) of vertebrate genes are comprised of a single exon (Niwa and Berget, 1991), which consequently minimises multiple analysis of the same gene. Also, 3' terminal exons contain a highly conserved poly A signal which is a defining hallmark of a last exon (Moore *et al.* 1992) and can be used to recognise these trapped exons.

The exon trapping method is potentially more efficient than the methods described in section 4.1.2. However, this technique has a number of disadvantages. Exon trapping has limited use in that it cannot find genes that do not contain introns. The success of the exon trapping procedure depends on the positions of the exons relative to the restriction sites of the enzymes used to digest the genomic DNA in the initial step. The exon must lie within the DNA spanned by the restriction enzyme sites in order to be trapped. Also, exon trapping is incapable of specifically detecting genes with fewer than three exons (Brennan and Hochgeschwender, 1995). Cryptic splice sites in the human genome can cause false positives to occur. Splicing events may be tissue specific or temporally regulated and may not occur in the packaging cell lines used in the exon trapping system.

4.1.1.2 Direct cDNA Selection Strategy for the Isolation of Transcribed Sequences

The direct cDNA selection technique was developed to allow rapid and complete identification of transcripts in large genomic regions and is complementary to the exon trapping procedure. The original protocols of direct cDNA selection involved the hybridisation of PCR amplified inserts from cDNA libraries to YAC DNA immobilised on nylon membranes (Lovett *et al.* 1991; Parimoo *et al.* 1991). Transcripts that were enriched through specific hybridisation to YAC DNA were eluted, amplified by PCR, then either evaluated or used for further rounds of hybridisation. The highly repetitive elements present in the DNAs were suppressed by blocking either the YAC DNA (Parimoo *et al.* 1991) or the PCR amplified cDNA inserts (Lovett *et al.* 1991) with sheared total human DNA. The extent of cDNA enrichment was determined by hybridising a gene probe to clones from the original and enriched cDNA libraries, then comparing the number of clones detected by the probe. The enrichment of specific cDNA sequences has been shown to vary by a factor of less than one thousand to several thousand fold depending on the number of selection cycles and the complexity of the genomic DNA.

This filter hybridisation procedure was neither easily quantitated nor sensitive, thus, magnetic bead capture was introduced to modify the technique. This involved hybridising biotin labelled genomic DNA to cDNA inserts in solution and subsequently immobilising the hybrid molecules on magnetic beads coated with streptavidin (Korn *et al.* 1992; Morgan *et al.* 1992; Tagle *et al.* 1992). The beads were washed to remove non-specific cDNAs, then bound cDNAs were eluted from the beads and PCR amplified using nested vector primers. The pool of cDNAs from this "first enrichment" was sometimes subjected to another cycle of hybridisation to genomic DNA to yield "second enrichment" cDNAs. The PCR amplified and enriched cDNAs were finally cloned. The resulting cDNA libraries are enriched for sequences that are homologous to the immobilised genomic DNA. The modified procedure allows better control of hybridisation compared with filter hybridisation, since this is conducted in solution (Britten *et al.* 1985). The stability and strength of the high affinity biotin-streptavidin coupling allow DNA manipulations at any desired stringency, including thermal denaturation and elution of annealed cDNAs, and simple and efficient pre-reassociation, hybridisation and washing.

The direct cDNA selection technique is amenable to the analysis of genomic segments cloned in a variety of vectors including lambda phages, YACs, P1 phages and cosmids (Korn *et al.* 1992; Morgan *et al.* 1992; Baens *et al.* 1995). However, enrichment with cosmid or P1 DNA is more efficient than with lambda clones due to the higher insert-to-vector ratio. YAC DNA as a source of genomic DNA is not as favourable as cosmid DNA as the isolation of YAC DNA of high purity is difficult. When such DNA has been used for direct cDNA selection, artefacts resulting from contaminating yeast ribosomal sequences have been reported to account for 30-60% of enriched clones (Lovett *et al.* 1991; Morgan *et al.* 1992). This problem can be overcome, to some extent, by pre-reassociation of the YAC DNA with total yeast DNA prior to hybridisation with cDNAs.

Several studies have documented the possible levels of enrichment of cDNAs which can be achieved by direct cDNA selection. Enrichments of cDNA by several thousand fold were demonstrated in the hybridisation of cDNAs to YAC DNA immobilised on nylon membranes after two rounds of selection (Lovett *et al.* 1991; Parimoo *et al.* 1991). High levels of enrichment of cDNA were also achieved using the magnetic bead capture procedure after two rounds of selection. Morgan *et al* (1992) showed > 100,000 fold enrichment with YAC DNA and Korn *et al* (1992) demonstrated an 80,000 fold enrichment with cosmid DNA.

The direct cDNA selection technique can be applied to the analysis of expressed sequences that are rare, developmentally expressed or tissue specific using a variety of genomic DNA and cDNA library sources. Various cDNA libraries including an oligo-dT primed, MboI digested library (Morgan *et al.* 1992), a normalised short insert random primed library (Forster and Rabbitts, 1993) and a combination of oligo-dT and random primed library

(Korn *et al.* 1992) have been used in this procedure. There are advantages and disadvantages in using different libraries but random primed libraries, which presumably retain overlap between cDNA clones, are the cDNA sources of choice.

One problem with direct cDNA selection is the presence of repetitive sequences in cDNA and genomic clones which can hybridise to each other. These hybridisations can be suppressed by pre-reassociating either one, or both, of the DNA sources with an excess of human cot-1 DNA or total human placental DNA. However, an unsolved problem is the presence of low copy repeats, or repeats that are only present in specific subchromosomal regions (Tagle *et al.* 1993; Tagle *et al.* 1994). These low copy repetitive elements can be present in up to 10% of the cDNAs enriched after the selection process and such clones are generally false positives.

Direct cDNA selection has been demonstrated to be a powerful and straightforward technique in numerous applications. For example, Korn *et al* (1992) was successful in isolating 81 cDNAs encompassing a region of 900 kb in Xq28. Morgan *et al* (1992) evaluated 24 unique cDNA clones isolated from a 425 kb YAC at 5q23-31. This YAC was used for two rounds of selection with eight different tissue sources. Rommens *et al* (1993) generated a transcript map of the 1 Mb Huntington disease gene region. YACs localised to 4p16.3 were used to obtain a total of 58 different cDNA clones. Peterson *et al* (1994) showed that 76 selected cDNA clones were derived from a 1.2 Mb region surrounding marker D21S55 on chromosome 21. Baens *et al* (1995) characterised 117 cDNAs isolated by direct cDNA selection using pools of cosmids localised to chromosome 12p. They demonstrated the majority of clones to map to 12p13 which is similar to the distribution of the known 12p genes. Thus, the direct cDNA selection technique has been successfully applied in the rapid identification and isolation of cDNA sequences encoded by large genomic segments of the human genome.

4.1.2 Comparison of the Exon Trapping and Direct cDNA Selection Techniques

The performances of the exon trapping and direct cDNA selection methods have been compared by Yaspo *et al* (1995) using a group of 81 cosmids localised to chromosome 21, representing approximately 2 Mb of non-contiguous DNA. High density grids were prepared with exon and cDNA libraries constructed from this group of cosmids. A total of 79 non-overlapping coding sequences were identified. Comparison of the two methods by cross hybridisation of clones from each library revealed that the individual libraries contained different or overlapping cDNAs, not multiple copies of the same clone. In addition, the number of potential coding elements recovered in the exon and cDNA resources was determined. Seventy percent of the tested exons detected selected cDNAs. The remaining exons may have contained sequences corresponding to rare transcripts. Alternatively, the absence of homologous selected cDNAs may be a result of the cDNA sources used for cDNA selection not containing these transcripts, or the size of the exon contained in the cosmid may have been too short to be detected by this procedure. In contrast, only 44% of the selected cDNAs detected exons in the exon libraries. These cDNAs generally detected only one individual exon, indicating that the exon libraries may not contain all the exons contained within the selected cDNA libraries. Thus, the cDNA enrichment generated clones which were not present in the exon library.

Exon trapping has been demonstrated to show a dramatic loss in exon recovery proportional to an increase in the input DNA complexity (Yaspo *et al.* 1995). Two strategies were used to trap exons from the cosmids. The first strategy involved division of the 81 cosmids into seven pools, while the second involved the division of these cosmids into three pools. The first strategy produced a less redundant and more complex library, compared to the library from the second strategy. The isolation of new independent exons was more efficient from the first library and may be explained by a greater initial genomic complexity. The entire exon library was not characterised, however, 29 different exon sequences were identified to be distributed among 19 cosmids. In comparison, 36 independent transcripts were identified by direct cDNA selection with the same genomic source. Analysis of the selected cDNA clones indicated that this procedure generates overlapping cDNAs, rather than multiple copies of the same clone, which is advantageous for the construction of longer cDNA sequences.

The distribution of the exons and selected cDNAs in the group of 81 cosmids was also investigated (Yaspo *et al.* 1995). Both exons and cDNAs were isolated from approximately half of the total cosmids. Only cDNAs were isolated from 3.7% of the cosmids. However, only exons were trapped from greater than 13% of the cosmids. Thus, this data indicated that exon trapping performed better for the identification of a large number of potentially transcribed loci. Nevertheless, cDNA selection provided greater coverage in terms of the coding sequence length. The sizes of the expressed sequences identified by exon trapping and direct cDNA selection were also demonstrated to differ. Trapped exons ranged from 50-250 bp (S. Whitmore, personal communication) while selected cDNAs typically averaged 500-600 bp in size (Cheng *et al.* 1994; Yaspo *et al.* 1995). Thus, direct cDNA selection exhibits more sequence information than exon trapping.

It was found that direct cDNA selection and exon trapping are complementary, thus using a combination of both procedures is advantageous for completeness and accuracy. To summarise, direct cDNA selection is less technically demanding than exon trapping and the number of redundant clones generated is more limited. Direct cDNA selection is more complete in terms of sequence length but exon trapping identifies a greater number of potential transcripts. Exons allow the orientation of corresponding cDNAs and can be used directly in experiments, for example on genomic DNA for mapping or gene dosage. The ease of isolation of processed pseudogenes is a disadvantage of cDNA selection. However, unprocessed pseudogenes which retain exon-intron structure are also targets for exon trapping. Thus, sequence analysis is required for the identification of spurious cDNAs or exons. Further, the occurrence of both exons and cDNAs at a given locus provides strong evidence for a true transcribed sequence.

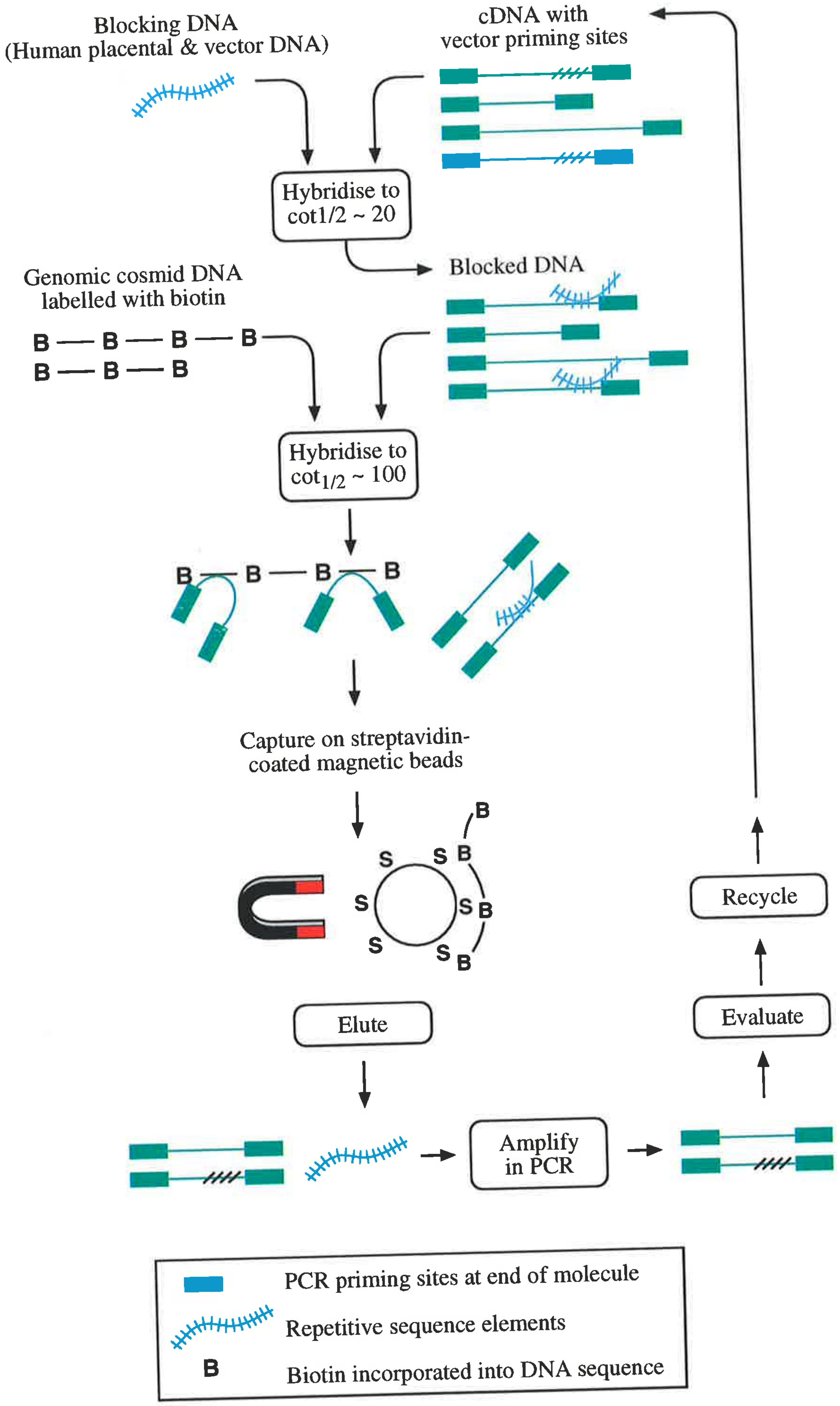
4.2 MATERIALS AND METHODS

The direct cDNA selection procedure was performed under the guidance of Dr. Danilo Tagle (Laboratory of Gene Transfer, National Center for Human Genome Research, National Institutes of Health, Bethesda, MD, USA). Analysis of identified cDNA clones was undertaken in the Department of Cytogenetics and Molecular Genetics, Women's and Children's Hospital, South Australia.

The isolation of cDNA sequences from a foetal brain cDNA library using cosmids from the 16q24 chromosomal region was performed by the approach shown schematically in figure 4.1. The technique involved the digestion of DNA from cosmids localised to the 16q24 chromosomal region and subsequent labelling of the DNA with biotin using the nick translation procedure. PCR amplified inserts of the cDNA library were depleted of repetitive sequences and vector sequences by hybridisation to total human DNA and sCos1 DNA. The repeat free material was hybridised in solution with biotinylated cosmid DNA and bound to streptavidin-coated magnetic beads. After a series of stringent washes, the cDNAs hybridised to the cosmid DNA were eluted from the beads and amplified by PCR. The selection-amplification procedure was repeated, resulting in cDNA populations highly enriched in sequences hybridising to the biotinylated cosmid DNA.

Figure 4.1

Schematic description of the biotin-streptavidin magnetic bead capture system for the direct cDNA selection protocol.


Blocking DNA
(Human placental & vector DNA)
cDNA with
vector priming sites
Hybridise to
cot1/2 ~ 20
Blocked DNA
Genomic cosmid DNA
labelled with biotin
B — B — B — B
B — B — B
Hybridise to
cot$_{1/2}$ ~ 100
B — B — B — B
Capture on streptavidin-
coated magnetic beads
S S S S S S
B B B B
Elute
Amplify
in PCR
Evaluate
Recycle
PCR priming sites at end of molecule
Repetitive sequence elements
B Biotin incorporated into DNA sequence

4.2.1 Preparation of cDNA and Cosmid DNA

A modification of the methods published by Parimoo *et al* (1991), Lovett *et al* (1991), Korn *et al* (1992), and Tagle *et al* (1993) was utilised. Inserts from approximately 2x10^6 clones of an oligo-dT and random primed human foetal brain cDNA library of 10^6 recombinants (Stratagene) in the λ-ZAP vector were amplified by PCR (see 2.6) using T3 and T7 vector primers in a final volume of 100 µl. The primer sequences were as follows: T3, 5'-GCTCGAAATTAACCCTCACTAAAG -3', and T7, 5'-GAATTGTAATACGACTCACCTATAG -3'. Primers were synthesised and deprotected as described in sections 2.6.1 and 2.6.2. After an initial denaturation at 94°C for 2 minutes, thermal cycling was performed for 30 cycles under the following conditions: 94°C denaturation for 30 seconds, 60°C annealing for 30 seconds and 72°C extension for 1 minute and 30 seconds plus an additional 5 seconds extension per cycle using the automatic segment extension option on the 9600 Perkin Elmer DNA thermal cycler. A 10 minute extension was performed at the end of the last cycle.

To scale up the amount of cDNA produced, four independent PCR amplified aliquots of the library were pooled and a 10 µl aliquot was analysed by electrophoresis on a 1.2% agarose gel (2.5.2) to determine the extent of the amplification. The PCR products were cleaned of excess primers, dNTPs and *Taq* enzyme using Wizard PCR prep minicolumns (Promega) according to the manufacturer's instructions, except that the columns were eluted with 50 µl TE at 60°C. The DNA was ethanol precipitated, lyophilised (2.3.1.4.2) and resuspended in TE. The DNA concentration was determined (2.3.1.4.3) then adjusted to 1 µg/µl.

Cosmids to be used in the direct cDNA selection procedure were grown overnight in LB supplemented with kanamycin at 50 µg/ml. Cosmid DNA was extracted by the alkaline lysis procedure (Sambrook *et al.* 1989) described in 2.3.1.1.1. Gel electrophoretic patterns of EcoRI digested cosmid DNAs were analysed after ethidium bromide staining to determine the integrity of the DNA.

One μg aliquots of DNA from each cosmid were digested (2.5.1) with either EcoRI or HindIII restriction enzyme for 2 hours at 37°C. The EcoRI and HindIII digests of each cosmid were pooled and 500 ng of this DNA was biotinylated by nick translation (BRL Bionick labelling system). The final concentration of the biotinylated cosmid DNA was 20 ng/μl.

4.2.2 Hybridisation and Isolation of Specific cDNAs

Biotinylated DNA preparations of several cosmids were pooled to form two groups. The human high copy repeat and vector sequences of the PCR amplified cDNA library inserts were blocked by pre-reassociation with an excess of unlabelled human placental DNA, added to a cot value of 20, and sCos1 vector DNA. Two μg of sheared human placental DNA and 200 ng of sCos1 DNA were mixed with 2 μg of amplified cDNA. Hybridisation solution consisting of 0.12 M sodium phosphate pH7.0 was added to give a total volume of 10 μl. The mixture of DNAs was denatured for 5 minutes at 95°C then hybridised for 4 hours at 60°C. Then, 200 ng of the pooled biotinylated cosmid DNA (20 ng/μl) was denatured for 5 minutes at 95°C, added to the pre-reassociated amplified cDNA (10 μl) and hybridised in a total volume of 20 μl at 60°C for 24 hours.

Specific cDNAs bound to biotinylated cosmid DNAs were captured using streptavidin-coated magnetic beads (Dynal). About 1 mg (10 μg/μl) of these beads was used for each hybridisation performed. These beads were pre-washed three times in 6 x SSC then resuspended in two volumes (200 μl) of TE (see 2.2.4)/1 M NaCl. The beads were added to each 20 μl hybridisation mixture and suspended by rotation at room temperature for 15 minutes. The beads with captured cosmid DNA/cDNA hybrids were collected with a magnet for 1 minute, resuspended in 100 μl of 1 x PCR buffer then washed two times at room temperature for 15 minutes with rotation. This was followed by three washes at 60°C for 15 minutes with 100 μl of 1 x PCR buffer. The beads were resuspended in 100 μl of 1 x PCR buffer and the captured cDNAs were eluted by boiling the sample for 5 minutes. The beads were collected with the magnet and the eluted cDNAs pipetted to a fresh tube.

Ten µl of the eluted cDNAs were PCR amplified in 100 µl reactions using nested vector primers SK, 5'- CGCTCTAGAACTAGTGGATCC -3', and KS, 5'- CGAGGTCGACGGTATCGATAAG -3' under the conditions described in 4.2.1. These eluted and amplified "first enrichment" PCR products, together with the amplified PCR products from the original starting cDNA library, were electrophoresed on a 1% agarose gel (2.5.2) to compare the sequence complexities of the two cDNA samples.

Further enrichment was achieved by recycling this "first enrichment" amplified cDNA in a further round of magnetic bead capture, using the procedure described above. The bound cDNA from the second round of enrichment was then eluted, PCR amplified using the nested vector primers SK and KS, and purified using Wizard prep minicolumns (Promega) according to the manufacturer's instructions. This "second enrichment" cDNA was labelled by primer extension (2.5.4.1) and hybridised (2.5.6.1) to filters containing 40,000 cDNA clones arrayed in a high density format, to identify homologous cDNA clones. The cDNA clones are from a 3' oligo-dT primed normalised infant brain cDNA library of 2.5 x 10^6 recombinants in the phagemid L-BA vector (Soares *et al.* 1994). This step was performed by Christa Prange at Lawrence Livermore National Laboratory, USA. Positive cDNA clones were subsequently sent to us as bacterial stabs.

4.2.3 Analysis of cDNA Clones

The inserts of the normalised cDNA library clones which were positive from the screening were PCR amplified directly from colonies. Colonies of each cDNA clone were gridded on ampicillin plates and cells from each colony were transferred to the PCR reaction mixture using a sterile pipette tip. Inserts were amplified by PCR (2.6) using puc forward, 5'- TGTGAGCGGATAACAATTTCACACAGGA -3', and puc reverse, 5'-

CACGACGTTGTAAAACGACGGCCAGT -3', vector primers in 50 µl reactions. After an initial denaturation at 94°C for 4 minutes, thermal cycling was performed for 30 cycles under the following conditions: 94°C denaturation for 1 minute, 55°C annealing for 1 minute and 72°C extension for 3 minutes. A 10 minute extension was performed at the end of the last cycle.

Fifteen µl of each PCR product were electrophoresed on 1% agarose gels (2.5.2) to assess the sizes of the inserts. The gels were subsequently Southern blotted and the membranes were used to screen the cDNA clones with radiolabelled total human DNA (2.5.4.1 and 2.5.6.1) to determine which clones contained high copy repetitive elements.

Cosmid fragments displaying homology to each cDNA clone were identified. Aliquots of PCR amplified inserts of the cDNA clones were purified by electrophoresis through 1% low melting point agarose gels (2.5.2). Individual cDNA inserts were excised from the gel and subsequently labelled by primer extension (2.5.4.1). Filters containing 500 ng of cosmid DNA restricted with either EcoRI, HindIII, HpaII or HaeIII were hybridised with radiolabelled total human DNA (2.5.4.1 and 2.5.6.1) to determine which cosmid fragments contained high copy repeats. The filters were subsequently probed with individual labelled cDNA inserts (2.5.6.1).

4.2.4 Sequencing of cDNA Clones

DNA of cDNA clones was prepared using Qiagen columns (2.3.1.2.2) and sequenced using dye primer kits (Applied Biosystems Inc.) (2.7.2). The sequences were analysed using an automated DNA sequencer (ABI 373A) (2.7.4) and compared to nucleotide databases using the BLAST-N program. Statistically significant sequence similarities were considered to be of P < 1 e -0.05.

4.3 RESULTS

4.3.1 Direct cDNA Selection System

The method of direct cDNA selection generally involves the cloning of selected cDNAs from the first or second round of selection and the subsequent analysis of the library of selected cDNA clones. The sizes of the inserts of these clones are generally small and the subsequent screening of a cDNA library is necessary to obtain longer clones. To overcome this shortcoming, the "second enrichment" PCR products generated from the second round of selection in this study, were radiolabelled and used directly to identify homologous oligo-dT primed normalised infant brain cDNA library clones by hybridisation to filters of 40,000 arrayed clones. The identified cDNA clones were then screened to eliminate those clones containing high copy repetitive elements. The remainder of the clones were analysed for specific homology to cosmid DNA. Thus, the direct cDNA selection protocol was modified in an attempt to efficiently isolate longer cDNA fragments and to obtain greater transcribed sequence information. This modification is also advantageous as the steps involved in cloning the PCR products and screening cDNA libraries for longer clones are eliminated.

4.3.2 Selection of cDNAs for the GALNS Gene

The N-acetyl galactosamine-6-sulfatase (GALNS) cDNA was demonstrated to show homology to four overlapping chromosome 16 cosmids, APRT, 340E8, 380C10, and 315F12, mapped to 16q24.3 (Morris *et al.* 1994). Cosmids 340E8, 380C10, and 315F12 were used as a positive control and test system for the direct cDNA selection experiment to identify coding sequences for the GALNS gene. After two rounds of selection, the eluted and amplified "second enrichment" cDNA PCR products were labelled and hybridised to the arrayed normalised infant brain cDNA library. The PCR amplified inserts of the nineteen identified clones were electrophoresed on agarose gels and transferred to nylon membranes. Two clones containing high copy repetitive elements were identified by screening with total human DNA and eliminated from subsequent analysis. Probing these clones with a full length single copy GALNS cDNA clone (provided by Dr. C.P. Morris, Department of Chemical Pathology, Women's and Children's Hospital, South Australia) identified five positive clones, or 30% of the non-repetitive clones, which demonstrated that enrichment of the GALNS cDNA from a total cDNA library was successful. The autoradiograph of the positive cDNA clones hybridised to GALNS cDNA is shown in figure 4.2. Two of the positive clones were randomly chosen and DNA sequence from both ends of each cDNA was obtained. The sequence showed homology to the sequence of GALNS. Thus, the test system for direct cDNA selection demonstrated that specific GALNS cDNA sequences were enriched after two rounds of selection and the frequency of high copy repetitive clones was 10%.

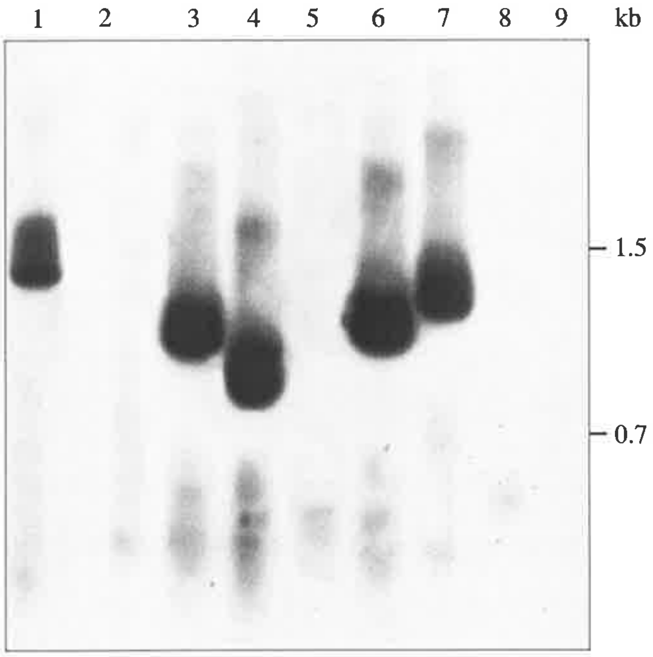
4.3.3 Isolation of Transcribed Sequences Using Cosmids Localised to 16q24

Using the direct cDNA selection approach, two libraries of selected cDNA fragments were generated using two independent cosmid pools. Cosmid pool 1 consisted of 38H11 and 51A6 from contig 231.1 and the singleton cosmid 37B2; and cosmid pool 2 consisted of 14G10, 317G1, 51A1, 304A9 from contig 76.1 and 318A10, 305C6, 330B4 from contig 549. Visualisation of the "first enrichment" and "second enrichment" PCR products on an agarose gel demonstrated a smear of DNA fragments ranging in size between 300 to 800 base pairs. The sizes of these cDNAs were smaller than the cDNA inserts of the starting library whose sizes were up to 1.2 kb. Thus, the sequence complexity in the selected library was reduced when compared to the starting library. Bands corresponding to any contaminating vector sequences in the starting cDNA library were not observed. PCR products were not observed in control reactions in which template was not added (data not shown).

PCR products from the "second enrichment" were used to directly identify homologous normalised foetal brain cDNA library clones by hybridisation to filters of arrayed clones. Inserts of the 72 identified clones were sized on 1% agarose gels that were then Southern

Figure 4.2

Normalised infant brain cDNA clones were identified by hybridisation of homologous "second enrichment" PCR products, generated from cosmids containing coding sequences for the GALNS gene after two rounds of direct cDNA selection. This autoradiograph demonstrates the identification of true positive normalised cDNA clones by hybridisation with a single copy GALNS cDNA. The five homologous cDNA clones range in size from 0.8 kb to 1.4 kb.


1
2
3
4
5
6
7
8
9
kb
1.5
0.7

blotted. The sizes of the inserts of the cDNA clones ranged from 0.8 kb to 1.8 kb. An example of the gels with PCR amplified cDNA inserts are shown in figure 4.3. Clones containing high copy repeat sequences were identified by hybridisation with total human DNA. These clones were eliminated from further analysis. A total of 29 clones, 49% of the cDNA clones in pool 1 and 30% in pool 2, contained high copy repetitive elements. These results are summarised in table 4.1.

4.3.4 Localisation of cDNAs to Cosmid DNA

To determine whether the cDNAs were homologous to the cosmid DNA from which they were derived, each non-repetitive purified cDNA insert was labelled and used as a probe on Southern blots containing EcoRI, HindIII, HaeIII or HpaII digests of the cosmids. Examples of these hybridisations are shown in figure 4.4. After a small number of cDNAs were analysed in this manner, similar hybridisation patterns were observed. The similarity in hybridisation patterns may represent the repeated isolation of overlapping cDNA fragments. To address this issue and to avoid analysing the same cDNA fragment repeatedly, membranes of the 20 cDNA inserts identified with "second enrichment" PCR products from pool 1 and the 23 cDNA inserts identified with "second enrichment" PCR products from pool 2 were hybridised, along with the filters containing the restricted cosmid DNAs. The results of the cDNA insert membrane hybridisation were used to pick new cDNAs for each round of hybridisation.

Of these cDNAs, five had a hybridisation pattern consistent with the recognition of specific cosmid fragments. Many of the cDNA clones were demonstrated to have homology to other cDNA clones which were selected. For example cDNA yh09a04 showed homology to nine other clones, and cDNA clone ym06c10 was homologous to six other clones. Most of the other cDNA clones were homologous to one other cDNA clone. Nine cDNA clones derived from pool 2 cosmids had hybridisation patterns suggesting that they recognised human low

Table 4.1

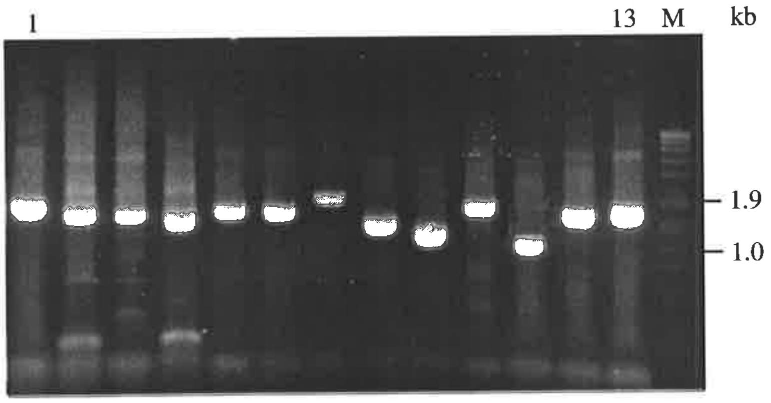
A summary of the results from the direct cDNA selection experiment. Normalised infant brain cDNA library clones were hybridised with "second enrichment" PCR products, from two rounds of selection, generated from three independent cosmid pools. The total number of clones identified and the number containing high copy repeats are indicated.

	Cosmids positive control	Cosmid pool 1	Cosmid pool 2
Total number of positive cDNA clones	19	39	33
Number of cDNA clones containing high copy repeats	2	19	10
Number of non-repetitive cDNA clones	17	20	23

Figure 4.3

An example of agarose gels displaying the PCR amplified inserts of 25 normalised infant brain cDNA clones homologous to "second enrichment" PCR products, generated from two rounds of selection, using cosmids from pools 1 and 2. The sizes of the inserts shown range from 1 kb to 1.9 kb.

Lane M corresponds to Spp-1 markers.


1
13
M
kb
1.9
1.0

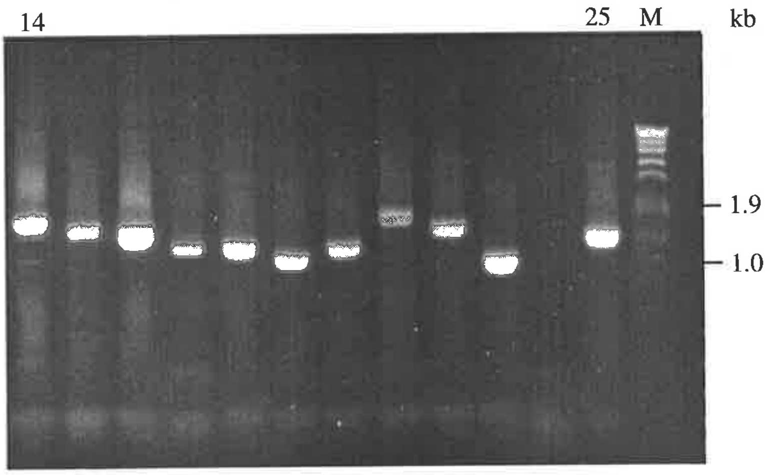

14
25
M
kb
1.9
1.0

Figure 4.4

Hybridisation of selected cDNA clones to cosmid DNAs. The autoradiographs represent hybridisations of cDNA clones to restricted cosmid DNA. These are as follows:

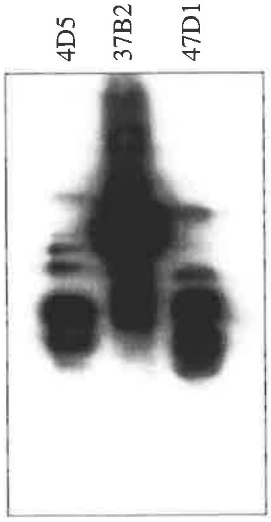
1. cDNA clone yh09a04 hybridised to HpaII digested cosmid DNA

2. cDNA clone ym06c10 hybridised to HindIII (H) and EcoRI (E) digested cosmid DNAs (2a) and HpaII digested cosmid DNAs (2b)

3. cDNA clone yg81f09 hybridised to EcoRI digested cosmid DNAs

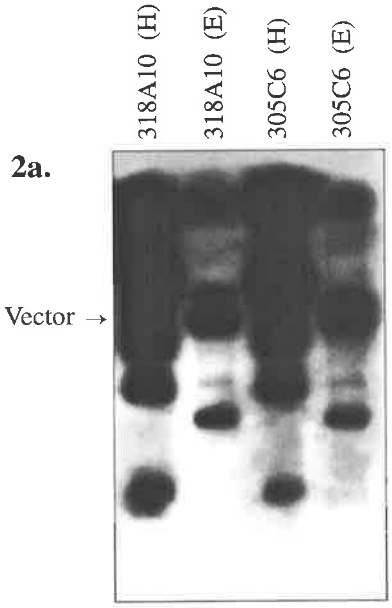
1.

yh09a04

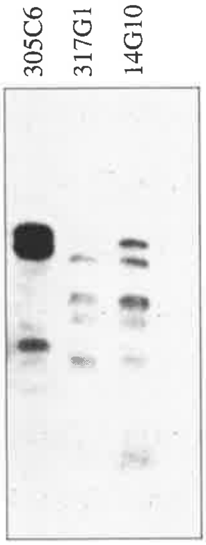


ym06c10

2a.

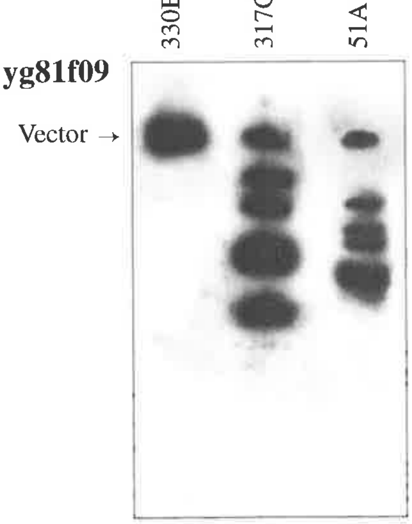


2b.



3.

yg81f09



copy repetitive elements. The remainder of the clones from pools 1 and 2 had homology to vector sequence only.

A list of the cDNA clones demonstrating homology to unique cosmid fragments is presented in table 4.2. Two clones from pool 1 and three clones in pool 2 showed homology to cosmids. Refined localisation of the cDNAs was not attempted, but preliminary localisation to 16q24 was defined by the mapping information of the corresponding cosmid with which the cDNA showed homology.

4.3.5 Sequence Analysis of Positive cDNA Clones

Approximately 300 base pairs of DNA sequence from both ends of each cDNA was obtained and compared to the accessible nucleotide databases using the BLAST-N program. Results are also shown in table 4.2. The sequences identified by this analysis were either homologous to randomly sequenced cDNA clones or did not show any similarity to any known sequence.

Table 4.2

A list of cDNA clones selected from two rounds of enrichment which show homology to specific bands of the indicated cosmids. GenBank Accession numbers of the cDNA clones are shown together with any homologous entries found upon searching the accessible nucleotide databases using the BLAST-N program. Statistically significant sequence similarities were considered to be of $P < 1$ e -0.05.

Hybrid Breakpoint	cDNA	Homologous Cosmids	GenBank Accession No.	Homologous cDNA clones (Accession No.)
CY18A(P)				
CY112				
CY2 / CY3				
1.	ym06c10	318A10, 305C6	H10940	
2.	yg81f09	317G1, 51A1	R54581	
3.	c-1ztb04	318A10, 305C6	F07214, F03493	
CY18A(D)				
4.	yh09a04	37B2	R61618, R60950	R07162
5.	yh01f01	37B2	R52867	
TELOMERE				

4.4 DISCUSSION

The application of the direct cDNA selection procedure has been described for the identification of transcribed sequences in the 16q24 chromosomal region. These isolated cDNA clones may be candidates for genes localised to this region including the gene for Fanconi anaemia group A (FAA) and the tumour suppressor gene associated with LOH in sporadic breast tumours. A limitation of the original direct cDNA selection procedure is the production of small insert (200-500 bp) (Gecz *et al.* 1993) sublibraries. cDNA libraries must be screened if longer clones are desired. In this study, an attempt was made to overcome this limitation by screening 40,000 clones of an arrayed normalised cDNA library with pools of selected cDNA PCR products. This oligo-dT primed library is derived from infant brain and is directionally cloned in the phagemid L-BA vector (Soares *et al.* 1994). This library has the following features. It contains long clones with an average size of 1.7 kb; the length of the segment representing the mRNA poly(A) tail is short, allowing an increase in useful sequence information from the 3' end of the clone; it has a broad representation of transcripts and has a very low frequency of non-recombinants (0.1%), mitochondrial and ribosomal transcripts (Adams *et al.* 1993). Two pools of cosmids covering approximately 300 kb of non-contiguous 16q24 genomic DNA were used to identify five normalised cDNA library clones ranging in size from 0.9 kb to 1.6 kb, which hybridised to distinct fragments of cosmid DNA.

The cosmids used in the cDNA selection were generated from the monochromosomal 16 somatic cell hybrid, CY18 (Stallings *et al.* 1990), hence they should be free of contamination by any other human chromosome. However, the methodology of direct cDNA selection is based on hybridisation by sequence homology and may also enrich pseudogenes and members of multigene families. The cDNAs that were demonstrated to hybridise to distinct cosmid fragments in this experiment may include pseudogenes, since this level of analysis offers no differentiation between genes and pseudogenes. However, further mapping, expression studies and sequencing can determine whether the cDNAs are actively transcribed genes or are pseudogenes.

The results of the positive control describing the recovery of GALNS cDNA clones from the cDNA library using three cosmids carrying GALNS sequences, indicated that the selection conditions utilised were successful in isolating homologous GALNS clones. The level of enrichment in this study was not estimated as the filters containing the arrayed normalised cDNA clones were not readily accessible to determine the number of copies of GALNS present in these filters. The number of clones containing high copy repeats was 10%, which is similar to frequencies found in other studies. These frequencies have been described to vary from 1% (Lovett *et al.* 1991), 7% (Fan *et al.* 1993) and 10% (Baens *et al.* 1995) to 20-50% (Cheng *et al.* 1994). Thus, pre-reassociation of repeat sequences fails to block all the high copy repeat sequences. Unexpectedly, the blocking of the repetitive elements seems to have been less successful in subsequent experiments using other cosmids, since a high frequency of selected cDNA clones containing high copy repeats from cosmid pools 1 and 2, 49% and 30% respectively, was demonstrated.

The repetitive sequences of the cDNA inserts were suppressed in this study, as a significant number of cDNA libraries have been demonstrated to contain repetitive sequences (Adams *et al.* 1992). The hybridisation of total human DNA to a normalised cDNA library revealed that approximately 10% of the clones contained high copy repeats (Crampton *et al.* 1981). This, in part, is due to clones possessing unspliced or partially spliced mRNA which contain introns with high copy repeats. However, high copy repeats can also be included in the 3' UTR of cDNA clones. For example, in a recent study, Yulig *et al* (1995) extracted over 1600 entries described as human complete cDNAs from the GenBank database and assessed the frequency with which the Alu repeat sequence occurred in these sequences. They demonstrated that 5% of the fully spliced human cDNAs contained Alu sequences, generally in the 3' UTR.

The high proportion of cDNA clones containing high copy repeats selected from cosmids in pools 1 and 2 in this study may be a result of insufficient blocking of repeat sequences before hybridising the cDNAs to genomic DNA, or may be due to these cosmids containing different concentrations or types of repeats compared to the cosmids used in the positive control experiment. A higher concentration of repeats may alter the effectiveness of suppression of repetitive sequences in the selection steps. The enrichment of repeat containing cDNAs and the subsequent identification of homologous cDNAs from the normalised library may have been partly avoided by suppressing repetitive sequences in both the genomic DNA and the cDNA.

The presence of Alu repeats in the 3' UTRs of the cDNA clones from the oligo-dT primed normalised library may have contributed further to the number of identified clones containing high copy repetitive elements. The "second enrichment" cDNAs may have consisted only of unique sequence, but their subsequent hybridisation to the normalised cDNA library may have identified normalised cDNA clones containing unique sequences homologous to the probes, together with Alu sequences in the 3' UTR of the clones.

Therefore, not all cDNAs that contain repetitive elements are necessarily non-specific background. Analysis of these cDNA clones by Southern hybridisation to membranes containing restricted cosmid DNA requires suppression of these repetitive sequences or isolation of the single copy sequences from the clones to identify cosmid restriction fragments containing sequences homologous to the cDNAs. An alternative method for investigation of these clones, which was performed in this study, is by Southern hybridisation to membranes containing genomic DNA restricted with frequent cutting endonucleases, such as HpaII or HaeIII. These enzymes generate small DNA fragments that are more likely to contain only single copy sequence, compared to enzymes such as EcoRI that produce larger DNA fragments which contain single copy and repetitive sequences.

Some non-repetitive clones from pool 2 did not display homology to a specific cosmid, but to a number of bands in unrelated cosmids. This implied that the cDNA contained low copy repeat sequences which were most likely enriched from region-specific low copy repeats contained in the genomic clones. FISH analysis of cosmids in contig 549 (table 3.4) demonstrated signal at 16q24 and the chromosome 16 specific low abundance repeat region, 16p11.2. The presence of these repeat sequences in the cosmids used from contig 549 may account for some of the false positive clones which were identified. These low copy repeat sequences will not be adequately suppressed with human placental DNA, but the characterisation of this type of cDNA clone can be avoided in future studies by using these clones to screen selected libraries or as blocking agents in future direct cDNA selection experiments.

False positive cDNA clones from both pools were also demonstrated to hybridise only to cosmid vector bands indicating possible contamination of the foetal brain cDNA starting library with vector sequences. Suppression of vector sequences in the PCR amplified inserts of the starting cDNA library and the "first enrichment" cDNAs was attempted with sCos1 DNA prior to hybridisation of these cDNAs to genomic DNAs. The isolation of cDNAs homologous to vector sequences may have been avoided by suppressing vector sequences in the genomic DNA, or in both sources of DNA.

The identified normalised cDNA library clones were screened to detect redundant clones. This was achieved by hybridising the labelled insert of each cDNA clone to filters containing PCR amplified inserts of the identified cDNA clones. cDNA inserts displaying positive signals were eliminated from further analysis. Clones which were of identical size were considered to contain identical sequence, but longer clones were likely to contain additional sequence at the 5' end, thus were further analysed. Soares *et al* (1994) compared the frequencies of particular cDNAs in the original and normalised infant brain libraries. They observed that the most abundant species of the original cDNA library are drastically diminished after normalisation and all the frequencies are brought within the range of one order of magnitude. Also, sequence analysis of 187 randomly chosen cDNA clones demonstrated all but four of these to contain unique sequence. However, the results obtained in this study suggest that this normalisation procedure is not entirely successful.

The number of transcripts expected to be isolated from the cosmids in this study can be estimated using results from previous studies. The overall density of selected cDNA fragments, with an average size of 500 bp, has been demonstrated to be one in every 20 kb in cosmids (Petersen *et al.* 1994; Cheng *et al.* 1994). Morgan *et al* (1992) also identified nine cDNA clones with an average size of 0.8 kb from a 425 kb YAC. Thus, one normalised cDNA clone with an average size of 1.2 kb may be represented by approximately 50 kb of genomic DNA and about five or six normalised cDNA clones may be expected to be isolated from the cosmids used in this study which cover 300 kb of genomic DNA.

Five cDNA clones which hybridised to different cosmid fragments were identified in this study, but it is not known whether these clones are real transcribed sequences or pseudogenes. Thus, it is possible that the number of true transcripts identified in this study is less than five. If this is the situation, the small number of identified clones may be explained by the limited complexity of the original amplified commercial cDNA library used for the initial selection. Alternatively, this may be a reflection of the limited complexity of the normalised cDNA clones present in the set of filters which were probed with the "second enrichment" cDNAs.

PCR amplification of the inserts of the cDNA library using vector primers results in a bias against the larger cDNAs since the longer clones are not present in the selected "first enrichment" PCR products. This limitation may be overcome in two ways. The large cDNAs may be cut into smaller pieces by random shearing or restriction enzyme digestion. Alternatively, the synthesis of double stranded cDNA can be achieved by random priming to build a collection of overlapping molecules of various sizes. Linkers or adapters are then ligated to the cDNA fragments and PCR amplification to generate cDNA inserts for the direct cDNA selection procedure can be achieved using primers homologous to the linker sequences. This procedure avoids the cloning step. In such a population of cDNAs the amplification product of every molecule can be potentially selected, and a greater number of cDNAs encoded by these cosmids can be identified.

There were no cDNA clones recovered for cosmids 38H11 and 51A6 from pool 1. These cosmids may not contain any coding sequences. Alternatively, transcripts in these cosmids may not be represented in the foetal brain cDNA library that was used as the starting library in this study. The highly complex commercial foetal brain cDNA library was chosen, as the direct cDNA selection technique is dependent on the use of cDNA libraries generated from a particular tissue source. It has been demonstrated that the sequence complexity of non-neural tissues, such as liver or kidney, is always two to three-fold lower than the sequence complexity of the brain (Van Ness and Hahn, 1980) and only 35% of brain mRNA species have been estimated to be shared with non-neural tissues (Chaudhari and Hahn, 1983). It is unclear to what extent the foetal brain mRNA represents the entire human mRNA population, therefore, in order to select all encoded cDNAs from a given region, the ideal would be to use a cDNA library representing all possible transcribed sequences. An alternative approach to attempt to increase the complexity of the cDNA library is to pool libraries from multiple tissue sources and from various developmental stages, with each library "tagged" with different linkers.

During the course of this study, the region of LOH at 16q24 in breast cancer was refined to 16q24.3-qter between markers APRT and D16S303 (Cleton-Jansen *et al.* 1994) thus, finer mapping of all the identified cDNA clones from the direct cDNA selection experiment to human DNA and somatic cell hybrid DNAs was not performed. The genetic map of the 16q24 chromosomal region shows APRT to be 2.2 cM distal to the D16S7 genetic marker (Kozman *et al.* 1993) and close to the CY18A(D) somatic cell hybrid breakpoint. Also, the FAA gene was localised to 16q24.3-qter by linkage analysis (Pronk *et al.* 1995). In 1995, the Fanconi Anaemia/Breast Cancer (FAB) consortium was established for greater input and effort into the positional cloning of the tumour suppressor and FAA genes. Thus, attention was focussed on the construction of a detailed physical map for 16q24.3-qter to expedite the positional cloning approach for the identification of these genes. The majority of cosmids, except 37B2, used in the direct cDNA selection experiment were not localised to this region, thus were not applicable for the positional cloning of these genes. Since the yh09a04 cDNA clone selected from cosmid 37B2 is localised to the 16q24.3-qter region, it is a potential candidate for these genes. The further characterisation of this transcript is described in chapter 5.

4.5 CONCLUSION

The direct cDNA selection procedure is a powerful tool for the isolation of cDNAs encoded from large genomic regions. This method is especially useful for isolating candidate genes using the positional cloning strategy where the defined candidate interval has been delineated by linkage analysis, studying gene distribution and for the derivation of tissue specific transcription maps across large genomic regions. The technique is rapid and can be applied to many sources of cDNA and genomic clone pools in parallel. Furthermore, the method tends to normalise transcript levels, with a greater enrichment of rare messages than abundant ones. The cDNAs in the lowest abundance class which would otherwise be missed by conventional screening protocols have an increased likelihood of being identified using the direct cDNA selection technique. Although this protocol is not expected to identify all the genes in the source genome, it is an efficient protocol that complements other procedures and is likely to expedite the identification of the genes of an organism. This chapter describes a modification of the direct cDNA selection protocol that results in the isolation of longer cDNA fragments, compared to those isolated by exon trapping, which therefore increases the potential of this protocol. Techniques to localise genes including exon trapping, identifying evolutionarily conserved sequences and using CpG islands can be combined with direct cDNA selection to identify the majority of genes in a particular genomic region, since the efficiency of each technique is less than 100%.

CHAPTER 5

Characterisation of Two Novel Transcripts Localised to 16q24.3-qter

5.1 INTRODUCTION 179

5.2 MATERIALS AND METHODS 182

5.2.1 Isolation of DNA from cDNA Clones 182

5.2.2 Hybrid Cell Line Panel 182

5.2.3 Localisation of cDNAs to 16q24.3-qter 182

5.2.4 Extraction of RNA 183

5.2.4.1 Precautions Against Ribonucleases 183

5.2.4.2 From Tissue Samples 184

5.2.4.3 From Cells and Cell Lines 184

5.2.4.4 RNA Precipitation 185

5.2.5 Analysis of RNA 185

5.2.5.1 Northern Analysis 185

5.2.5.2 Reverse Transcriptase Polymerase Chain Reaction (RT-PCR) 186

5.2.6 Sequencing of cDNA Clones 189

5.3 RESULTS 191

5.3.1 Localisation of the cDNA Clones to Chromosome 16 191

5.3.2 Expression Studies 191

5.3.2.1 Northern Analysis 191

5.3.2.2 Reverse Transcriptase Polymerase Chain Reaction 197

5.3.3 Sequence Analysis 200

5.3.3.1 yh09a04 201

5.3.3.2 yc81e09 204

5.3.4 Isolation of cDNA Clones to Extend the Sequences of the Transcripts 204

5.3.5 Contributions to the Characterisation of Transcripts yh09a04 and yc81e09 From the Fanconi Anaemia/Breast Cancer Consortium 206

5.3.5.1 yh09a04 206

5.3.5.2 yc81e09 209

5.4 DISCUSSION 210

5.1 INTRODUCTION

Two disease genes have been localised to the 16q24.3-qter chromosomal region. Loss of heterozygosity (LOH) of this region has been demonstrated in sporadic breast tumours and indicates the presence of a tumour suppressor gene (TSG). Additionally, linkage analysis has localised the Fanconi anaemia group A (FAA) gene to the 16q24.3-qter region. At this time, this region did not contain mapped genes which could be considered candidate genes. Thus, the Fanconi Anaemia/Breast Cancer (FAB) consortium was established in 1995 to positionally clone the FAA and tumour suppressor genes localised to 16q24.3-qter. The identification of the breast cancer tumour suppressor gene will contribute to an understanding of the aetiology of breast cancer. The FAA gene will provide a basis for the understanding of the aetiology of the disease and a definitive diagnosis for this disease.

Initially, a cosmid based physical map of the 16q24 region was constructed using the Alu PCR strategy to isolate cosmids mapped to this region (see chapter 3). Alu PCR amplified products from the human/rodent somatic cell hybrids, CY2 and CY18A, which contain the distal part of chromosome 16q, were used as hybridisation probes for screening the chromosome 16 specific cosmid library (Stallings *et al.* 1992) to identify and localise cosmids to this chromosomal region. Since the tumour suppressor and FAA genes were localised to the 16q24.3-qter region, attention was subsequently concentrated on the construction of a cosmid based physical map across this critical region to expedite the positional cloning strategy for the identification of these genes (see chapter 3). This physical map was developed by screening the chromosome 16 cosmid library (Stallings *et al.* 1992) with sequence tagged sites (STSs) and expressed sequence tags (ESTs) already localised to this region. The end fragments of cosmids were used to probe the library to develop an integrated cosmid contig of about 650 kb in length. This contribution to the physical map represented the first step toward the cloning of the genes using the approach of positional cloning.

A transcript map of the 16q24 region was then constructed in order to identify all the genes and transcripts localised to the 16q24.3-qter region. These transcripts will then be screened for alterations in FA-A or breast cancer patient DNA, when compared to normal DNA. Initially, a selection of cosmid clones localised to 16q24 were utilised for the identification of transcribed sequences encoded by these cosmids, using the approach of direct cDNA selection (see chapter 4). Subsequently, the exon trapping procedure was used by members of the FAB consortium to identify and isolate cDNAs encoded by cosmids localised to the 16q24.3-qter genomic region. The investigation of cDNAs and genes localised to this region was performed by the FAB consortium, to determine whether differences could be detected in DNA of paired affected and normal samples from FA-A or breast cancer patients.

My contribution to the positional cloning of the tumour suppressor and FAA genes involved the detailed analysis of two novel cDNA clones localised to the most distal tip of chromosome 16q, between the CY18A(D2) somatic cell hybrid breakpoint and the telomere. The first clone, yh09a04, from a normalised oligo-dT primed infant brain cDNA library (Soares *et al.* 1994) was identified by direct cDNA selection and is encoded by cosmid 37B2 (chapter 4; table 4.2). The second clone, yc81e09, was identified by screening the arrayed normalised oligo-dT primed infant brain cDNA library (Soares *et al.* 1994) with the insert from cDNA clone ScDNA-A55. This was performed by Dr. Greg Lennon at the Lawrence Livermore National Laboratory, USA. Clone ScDNA-A55, localised between the CY18A(D2) somatic cell hybrid breakpoint and the telomere, was isolated from a hexamer primed heteronuclear cDNA library constructed from the CY18 mouse/human somatic cell hybrid (Whitmore *et al.* 1994).

Detailed analysis of cDNA clones yh09a04 and yc81e09 initially involved their precise localisation to the physical map of chromosome 16 using a panel of somatic cell hybrids containing only the distal tip of chromosome 16. Northern analysis was performed to determine the sizes of the two cDNA clones and whether they were true transcribed sequences. The tissue distribution of these clones was also determined by reverse transcriptase polymerase chain reaction (RT-PCR). RT-PCR is an alternative procedure to Northern analysis for determining tissue expression of transcribed sequences. It is a rapid PCR based technique in which a small amount of RNA is utilised, compared to the large quantities of RNA required for Northern analysis. Thus, it is possible to examine tissues from which low yields of RNA are isolated. RT-PCR also provides an increased sensitivity in examining sequences with low levels of expression.

The sequence of the clones was obtained and compared to sequences in accessible nucleotide databases for homologies using the BLAST-N program. Homologies to any known genes may provide clues for potential functions of the transcripts. Otherwise, homologies to cDNA clones may extend the sequence of the clones. Another approach to obtain full-length sequence of the transcripts was the isolation of overlapping cDNA clones by screening cDNA libraries. The determination of the full-length sequence is essential for prediction of the protein product encoded by each transcript, which may provide functional information on the genes.

This chapter describes the characterisation of the cDNA clones, yh09a04 and yc81e09, which map to 16q24.3-qter and are candidate genes for the Fanconi anaemia group A and the tumour suppressor gene involved in breast cancer.

5.2 METHODS

5.2.1 Isolation of DNA from cDNA Clones

cDNA clones were grown overnight in LB supplemented with ampicillin at 50 µg/ml. DNA was extracted using Qiagen tip-20 columns (2.3.1.1.2). The DNA was visualised on a 1% agarose gel (2.5.2) and the concentration (2.3.1.4.3) and purity of DNA samples (A260/A280 ratio) was determined. The A260/A280 ratio was consistently in the range 1.6-1.8.

5.2.2 Hybrid Cell Line Panel

The DNA from a number of somatic cell hybrids containing either a cytogenetically normal chromosome 16 or portion thereof, and human parental cell line DNA (ie. the cell line from which the somatic cell hybrid was derived) were used for Southern analysis. The portion of chromosome 16 present in each cell line used is described in table 3.1. The construction of somatic cell hybrids and details of the human parental cell lines have been previously described (Callen, 1986; Callen, 1990).

5.2.3 Localisation of cDNAs to 16q24.3-qter

The insert of each cDNA clone was PCR amplified (2.6) using vector primers, puc forward, 5'- TGTGAGCGGATAACAATTTCACACAGGA -3', and puc reverse, 5'- CACGACGTTGTAAAACGACGGCCAGT -3', in 50 µl reactions. Primers were synthesised and deprotected as described in sections 2.6.1 and 2.6.2. After an initial denaturation at 94°C for 4 minutes, thermal cycling was performed for 30 cycles under the following conditions: 94°C denaturation for 1 minute, 55°C annealing for 1 minute and 72°C extension for 3 minutes. A 10 minute extension was performed at the end of the last cycle.

Each PCR product was purified by electrophoresis through 1% low melting point agarose gels (2.5.2). The bands were excised from the gel and subsequently used as probes.

Human DNA, DNA from the human/mouse hybrid cell lines including CY18, CY18A, CY112, CY2, CY3, and the mouse cell line A9, was digested with HindIII (2.5.1). After electrophoresis of this DNA on an 0.8% agarose gel (2.5.2), a Southern blot filter was made by the methods described in 2.5.5.2 and 2.5.5.4. The filters were prehybridised then probed with labelled (2.5.4.1) cDNA insert in order to localise the cDNAs. Hybridisation was carried out at 42°C for 16 hours and filters were washed following the procedure outlined in 2.5.6.1.

5.2.4 Extraction of RNA

5.2.4.1 Precautions Against Ribonucleases

Standard precautions were taken to prevent the ribonuclease degradation of extracted RNA (Sambrook *et al.* 1989). All handling was performed using gloves and where possible, dedicated equipment was used. All RNA manipulations were performed using specifically prepared disposable centrifuge tubes and filtered Gilson pipette tips. All RNA work was conducted using fresh disposable gloves after every step of the protocol. Solutions contained diethyl pyrocarbonate (DEPC)-treated sterile water and stored at room temperature. All RNA samples were kept in screw top 1.5 ml microcentrifuge tubes in a separate, sealed box. They were handled on wet ice at all times and stored at -70°C when not in use.

5.2.4.2 From Tissue Samples

Total RNA was extracted from human tissue using a single-step method of RNA isolation modified from Chomczynski and Sacchi (1987). Tissue samples were dissected by Dr. Roger Byard (Department of Histopathology, Women's and Children's Hospital, South Australia) from an infant whose cause of death was Sudden Infant Death Syndrome. Tissues included liver, nerve and muscle, and the frontal lobe, occipital lobe, cerebellum and basal ganglia brain sections.

Tissue samples (0.5 g) sliced into small pieces were added to 10 mls of TRIzol (GIBCO BRL) in sterile 50 ml centrifuge tubes and placed on ice. Each sample was homogenised on ice at 30 second intervals with an Ultra-turrax probe which was rinsed with 0.1 M NaOH, DEPC-treated water and 70% ethanol between samples. One ml of chloroform was added, then each sample was vortexed and placed on ice for 15 minutes. The mixture was centrifuged at 12,000 rpm at 4°C for 15 minutes and the aqueous phase drawn off and placed in a sterile 50 ml centrifuge tube.

5.2.4.3 From Cells and Cell Lines

Total RNA from cell lines was extracted using TRIzol (GIBCO BRL) after trypsinisation of the cells from 75 cm^3 or 150 cm^3 flasks and three washes of the cells in PBS (see 2.2.4). RNA was also extracted from peripheral blood lymphocytes (PBL) after they were separated from whole blood using ficol-hypaque and washed three times in PBS.

Cells were resuspended in 1 ml of TRIzol /10^7 cells. The mixture was immediately vortexed and the lysed cells were split into sterile 1.5 ml microcentrifuge tubes. Chloroform (1/10th volume) was added to each tube. The suspension was then shaken vigorously for 15 seconds and allowed to stand on ice for 5 minutes. The mixture was centrifuged at 12,000 rpm at 4°C for 15 minutes and the aqueous phase was placed into a sterile 1.5 ml microcentrifuge tube.

5.2.4.4 RNA Precipitation

An equal volume of isopropanol was added to the aqueous phase. This was vortexed, placed on ice for 15 minutes and centrifuged at 12,000 rpm at 4°C for 15 minutes. The supernatant was removed, the precipitate washed with 8 mls of 75% ethanol and centrifuged for 15 minutes at 7,500 rpm at 4°C. The precipitate was then dried on ice and resuspended in 100 µl of 1 mM EDTA in DEPC-treated water. The RNA was then quantitated by spectroscopy (2.3.1.4.3). The A260/A280 ratios of the RNA samples were consistently in the range 1.6-2.0. A 1 µl aliquot of RNA was analysed on a 1% agarose gel (2.5.2) to assess the integrity of the 18S and 28S ribosomal RNA species prior to use.

5.2.5 Analysis of RNA

5.2.5.1 Northern Analysis

Northern hybridisation analysis of total RNA (20 µg) was carried out by denaturation on 1% agarose gels containing 1.1 M formaldehyde, 1 x MOPS buffer (see 2.2.4) and transfer onto Hybond^{TM} nylon membranes. An equal volume of RNA loading buffer (50% deionised formamide, 10% (v/v) glycerol, 1% MOPS buffer, 16.7% formaldehyde, 0.1% (w/v) bromophenol blue) was added to 20 µg of RNA sample then denatured at 65°C for 15 minutes and placed on ice. Samples were loaded on gels which were run in MOPS buffer at 20 mA for 5-6 hours.

Prior to transfer, gels were washed two times for 30 minutes in DEPC-treated water to remove formaldehyde. Transfer was carried out overnight in 10 x SSPE. Following transfer, the filters were baked for 45 seconds, high power, in a microwave oven. Filters were pre-hybridised for approximately 4 hours at 42°C in 50% formamide, 5 x SSC, 5 x Denhardt's solution and 200 µg/ml of sonicated salmon sperm DNA. Hybridisations were carried out for 18-24 hours under exactly the same conditions, except for the addition of radiolabelled cDNA insert (1-5 x 10^8 counts/µg) (see 2.5.4.2). Filters were washed in 2 x SSC, 0.1% SDS at room temperature for 5 and 15 minutes, followed by one wash in 2 x SSC, 0.1% SDS at 65°C for 15 minutes. After washing the membranes were exposed to X-ray film for varying times according to the probe used (16 hours to 7 days) at -70°C.

5.2.5.2 Reverse Transcriptase Polymerase Chain Reaction (RT-PCR)

The conditions for first strand cDNA synthesis were determined by examining a range of concentrations of starting RNA, from 1 µg to 15 µg. The PCR conditions for the synthesis of second strand cDNA were established according to the melting temperature (T_m) of the primers and the sizes of the PCR products (Newton and Graham, 1994).

First strand cDNA synthesis was performed with 5 µg of total RNA extracted from each tissue or cell line sample and 2 µl of oligo-dT primer at 500 ng/µl (GIBCO BRL) added to DEPC-treated water in a total volume of 40 µl. This was heated to 70°C for 10 minutes then placed on ice. Eight µl of 5 x first strand buffer (GIBCO BRL), 4 µl of 0.1M dithiothreitol and 2 µl of 10 mM dNTP mix (dATP, dTTP, dCTP, dGTP at neutral pH) were added, mixed and warmed to 37°C. Two µl (400U) of Superscript reverse transcriptase (GIBCO BRL) was added and the reaction was incubated at 37°C for 1 hour followed by inactivation of the enzyme at 95°C for 5 minutes. For each RNA/primer combination two control reactions were carried out: one without reverse transcriptase and one without RNA.

Second strand cDNA was synthesised by PCR (see 2.6) under the following conditions. Two and a half µl of first strand mixture and 150 ng of each primer specific to the sequence of the yc81e09 or yh09a04 cDNA clones were added in a total volume of 25 µl. The sequences of the primers used for second strand synthesis are listed in table 5.1. After an initial denaturation at 94°C for 5 minutes, 35 cycles were carried out with the following parameters: denaturation at 94°C for 1 minute, annealing at 60°C for 2 minutes and extension

Table 5.1

Primers used for second strand cDNA synthesis of cDNA clones yh09a04 and yc81e09. The locations of the primers in the sequence and the sizes of the PCR products are indicated.

cDNA clone	Primer Sequence (5' - 3')	Primer Position	Size of PCR Product (bp)	Size of cDNA clone (bp)
yh09a04	(F) TTC ACC TCC AGC TGG CAG GAG	bases 241-261	1047	1560
yh09a04	(R) TCA GAG GTG CGG AAC TAT AT	bases 1287-1268		
yc81e09	(F) CTG CCG GCT GGA TTA CCG CAG	bases 323-343	906	1537
yc81e09	(R) CTC CTT CCT CGG GCC TCT CCC C	bases 1228-1207		

at 72°C for 3 minutes. A 10 minute extension at 72°C was performed at the end of the last cycle. The sequences of the primers utilised in these reactions and the sizes of their corresponding PCR products are listed in table 5.1. The locations of the primers in the cDNA sequence were chosen such that the sequence encompassed by the primers in genomic DNA contained introns. The synthesis of a product from genomic DNA would be hampered by the physical distance between the primers, but a product could be synthesised from mRNA as only exonic sequences are present. PCR conditions for each primer pair were established by using DNA from each cDNA clone as template. DNA from each cDNA clone was subsequently included as a positive control to determine whether PCR product was produced in each experiment.

The housekeeping gene chosen as a control for the cDNA synthesis was Esterase D (ESD). The sequences of the ESD primers used in the PCR reaction are as follows: ESD-F, 5'-GGAGCTTCCCCAACTCATAAATGCC -3' and ESD-R, 5'-GCATGATGTCTGATGTGGTCAGTAA -3'. This control was initially analysed in establishing the technique. The annealing temperatures of the ESD, yh09a04 and yc81e09 specific primers were compatible, allowing the experiments to be performed simultaneously in the one thermal cycler. A negative control in which no template was added was also included.

Ten µl aliquots of the PCR products were electrophoresed on 1% agarose gels (2.5.2) and transferred to nylon membranes (2.5.5.3). PCR products from each cDNA, yh09a04 and yc81e09, were synthesised with the primers listed in table 5.1 under the conditions described and purified by electrophoresis through 1% low melting point agarose gels (2.5.2). The bands were excised from the gel and subsequently used as probes. Each cDNA PCR product was labelled by primer extension (2.5.4.1) and hybridised (2.5.6.1) to the appropriate membranes to visualise the product of cDNA synthesis.

5.2.6 Sequencing of cDNA Clones

The ends of the cDNA clones were sequenced first using dye primer kits (Perkin Elmer) (2.7.2). Primers specific to each cDNA clone were designed, synthesised and purified (2.6.1-2.6.2.1). They were used in dye terminator cycle sequencing reactions (Perkin Elmer) (2.7.2) to obtain the entire sequence of each cDNA clone. These primers are listed in table 5.2. The sequences were analysed using an automated DNA sequencer (ABI 373A) (2.7.4) and compared to nucleotide and protein databases using the BLAST-N and BLAST-X programs. Statistically significant sequence similarities were considered to be of P < 1 e -0.05.

Table 5.2

Primers used in dye terminator sequencing reactions to obtain the entire sequence of clones yh09a04 and yc81e09. The locations of the primers in the sequence are indicated.

cDNA clone	Primer Sequence (5' - 3')	Primer Position
yh09a04	TTC ACC TCC AGC TGG CAG GAG	bases 241-261
yh09a04	GAC CTG TCC CAG GAG CCT CAT C	bases 457-478
yh09a04	CAG GTG TGA GGT CTG TGG GTT C	bases 1116-1095
yh09a04	TCA GAG GTG CGG AAC TAT AT	bases 1287-1268
yc81e09	CTG CCG GCT GGA TTA CCG CAG	bases 323-343
yc81e09	TCC CTG AGT GTG AGC AGA GCT	bases 637-657
yc81e09	CTC CTT CCT CGG GCC TCT CCC C	bases 1228-1207

5.3 RESULTS

5.3.1 Localisation of the cDNA Clones to Chromosome 16

The locations of the two cDNA clones, yh09a04 and yc81e09, were analysed further by hybridisation to Southern blots consisting of a panel of somatic cell hybrids containing DNA of the most distal part of chromosome 16q. The autoradiographs of these Southern analyses are shown in figures 5.1 and 5.2. These hybridisations allowed the accurate map locations of the cDNA clones with respect to somatic cell hybrid breakpoints, shown in figure 5.3, to the most distal interval of the long arm of chromosome 16.

The localisation of cDNA clone yh09a04 to 16q24.3 was also confirmed by FISH analysis to metaphase chromosomes (data not shown). These results verify that these clones are of human chromosome 16 origin and confirm the localisations of ScDNA-A55, from which yc81e09 was derived, and cosmid 37B2, from which yh09a04 was derived. Both cDNA clones detected mouse bands in the mouse A9 cell line (lane 2, figure 5.1; lane 2, figure 5.2) suggesting a conservation of the expressed sequences across these two species.

5.3.2 Expression Studies

5.3.2.1 Northern Analysis

Total RNA from various human tissues and cell lines were used in Northern analysis of the two cDNA clones. Northern analysis was primarily conducted to determine the size of each transcript. Clone yc81e09 detected a transcript of approximately 3.7 kb in size in the lymphocyte and occipital lobe tissues and rhabdomyosarcoma and Hela cell lines (table 5.3). This Northern blot is shown in figure 5.4 A. The signal for this transcript was observed after a 16 hour exposure at -70^{o}C, indicating that it may have a relatively high level of expression. Clone yh09a04 detected a transcript of approximately 2.5 kb in frontal lobe and brain stem

Figure 5.1

Hybridisation of yh09a04 cDNA clone to HindIII digested DNA from human, mouse A9 cell line, chromosome 16 somatic cell hybrid CY18 which contains the whole chromosome 16, and somatic cell hybrids CY3, CY2, CY112 and CY18A containing the distal portion of chromosome 16.

A human band was demonstrated in somatic cell hybrids CY18, CY112 and CY2 but not in hybrids CY3 and CY18A. This localises the yh09a04 cDNA clone to the most distal interval of the chromosome 16 physical map.

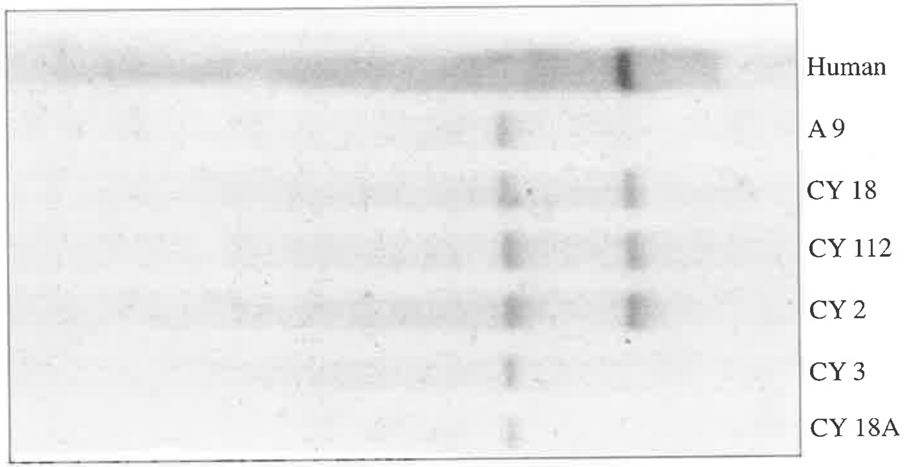

Human
A 9
CY 18
CY 112
CY 2
CY 3
CY 18A

Figure 5.2

Hybridisation of yc81e09 cDNA clone to HindIII digested DNA from human, mouse A9 cell line and chromosome 16 somatic cell hybrid CY18 which contains the whole chromosome 16, and somatic cell hybrids CY3, CY2, CY112 and CY18A containing the distal portion of chromosome 16.

A human band was demonstrated in somatic cell hybrids CY18, CY112 and CY2 but not in hybrids CY3 and CY18A. This localises the yc81e09 cDNA clone to the most distal interval of the chromosome 16 physical map.

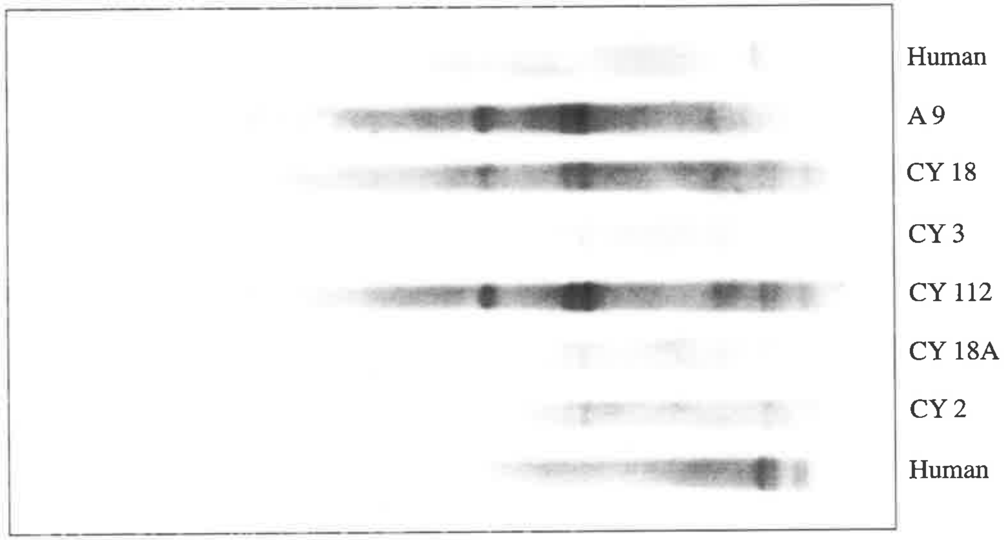

Human
A 9
CY 18
CY 3
CY 112
CY 18A
CY 2
Human

Figure 5.3

Ideogram of chromosome 16q24 showing the mapping positions of cDNA clones yh09a04 and yc81e09, localised by Southern analysis to the 16q24.3-qter region.

The locations of the FAA gene (Pronk *et al.* 1995) and putative tumour suppressor gene between APRT and D16S303, identified from LOH studies of 16q in breast cancer (Cleton-Jansen *et al.* 1994) are indicated.

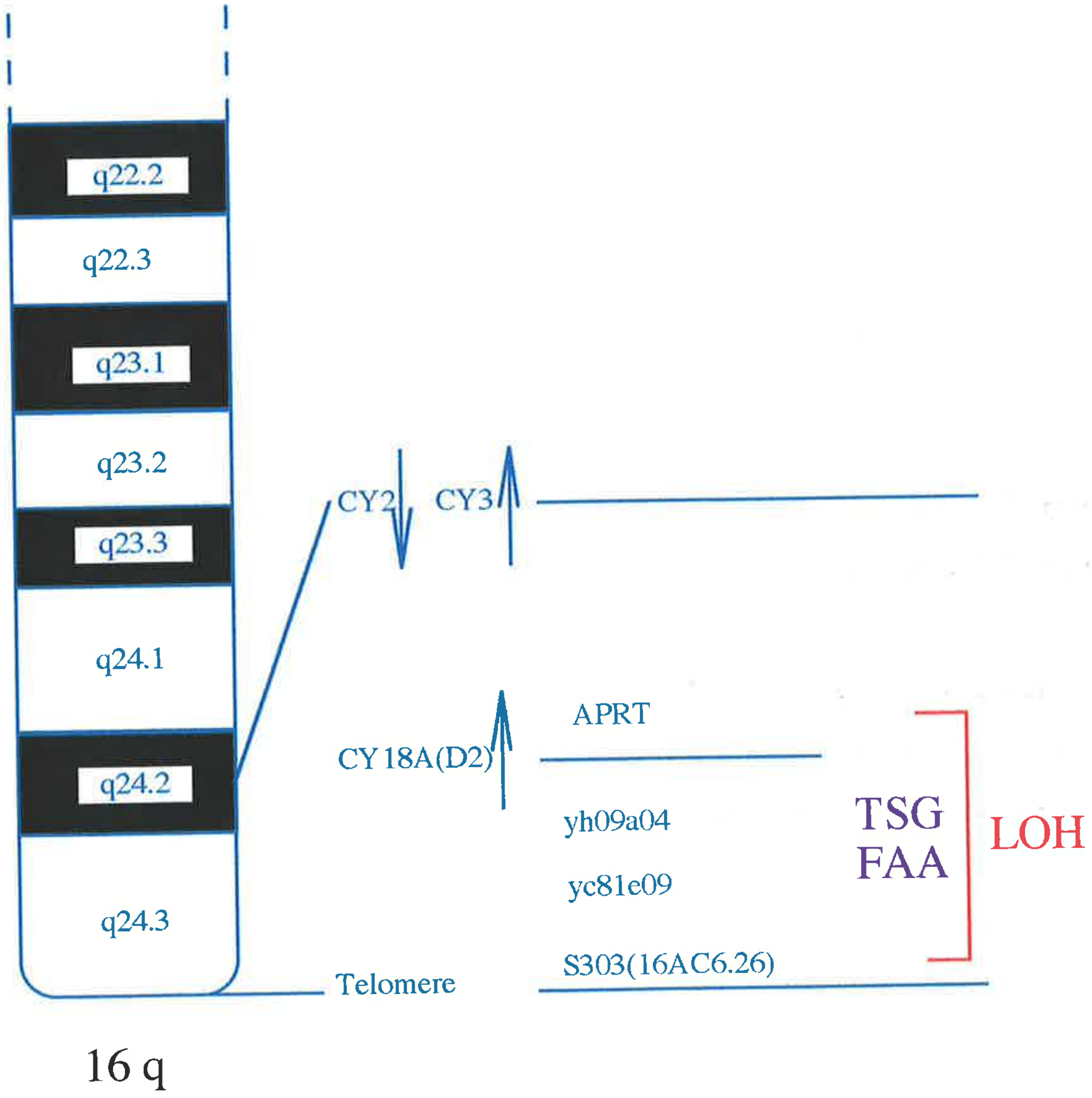

q22.2
q22.3
q23.1
q23.2
q23.3
q24.1
q24.2
q24.3
16 q
CY2
CY3
CY18A(D2)
APRT
yh09a04
yc81e09
S303(16AC6.26)
Telomere
TSG
FAA
LOH

Table 5.3

Results of Northern analysis with the indicated tissues and cell lines. A positive signal is denoted with a '+' and a negative signal is indicated by '-'. Sizes of the observed transcripts are given in kb.

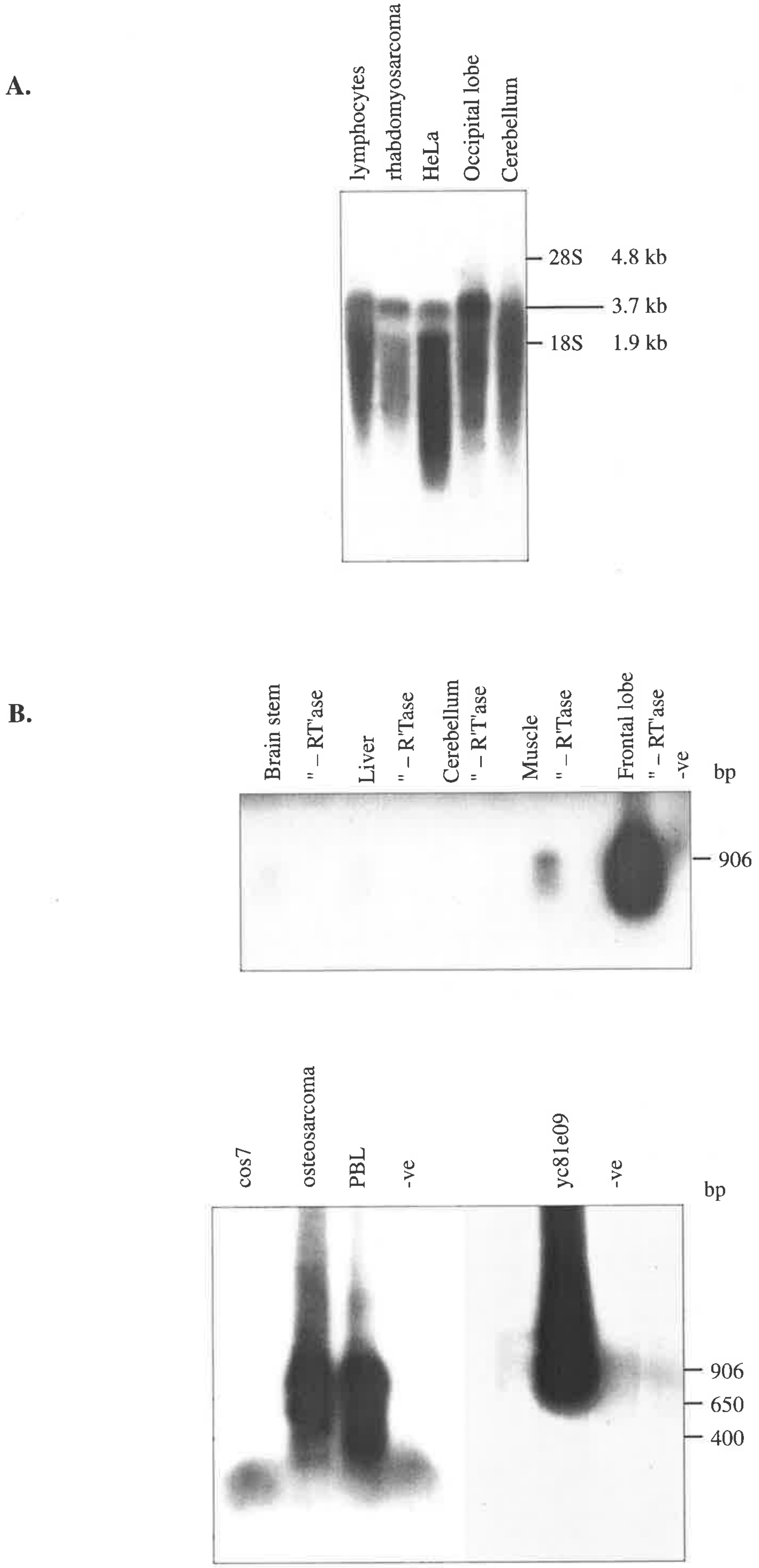
cDNA clone	frontal lobe	occipital lobe	brain stem	cerebellum	muscle	liver	PBL	Hela	rhabdomyosarcoma	Size (kb)
yh09a04	+		+	-	-	-				2.3
yc81e09		+					+	+	+	3.7

Figure 5.4

A. Northern blot hybridisation of yc81e09 cDNA probe localised to the 16q24.3-qter region. The Northern blot contains total human RNA from occipital lobe and cerebellum and the Hela and rhabdomyosarcoma cell lines. The cDNA identifies a 3.7 kb transcript. The sizes of the 28S and 18S ribosomal proteins, used as approximate size markers, are indicated.

B. An example of the autoradiographs of RT-PCR analyses of isolated cDNA sequences. These autoradiographs show amplified cDNA fragments generated from oligo-dT primed RNA from muscle, frontal lobe and peripheral blood lymphocytes (PBL) and the osteosarcoma cell line using primer sequences derived from the yc81e09 cDNA clone. The sizes of the PCR products are indicated in bp.

Lanes labelled as '- RT'ase' are reactions in which no Reverse Transcriptase was added. The lanes labelled '-ve' indicate the reactions in which no template was added to monitor any contamination which may occur during PCR. The lane labelled yc81e09 is the reaction in which yc81e09 was used as template.


A.
lymphocytes
rhabdomyosarcoma
HeLa
Occipital lobe
Cerebellum
28S 4.8 kb
3.7 kb
18S 1.9 kb
B.
Brain stem
" – RT'ase
Liver
" – R'Tase
Cerebellum
" – RT'ase
Muscle
" – R'Tase
Frontal lobe
" – RT'ase
-ve
bp
906
cos7
osteosarcoma
PBL
-ve
yc81e09
-ve
bp
906
650
400

tissues (see table 5.3). The level of expression of this transcript appeared to be low as a faint signal was observed after an exposure of seven nights at -70°C. This data is not shown as the signal on the autoradiograph was too faint to be scanned or photographed. The results from these Northern analyses also demonstrate that these cDNA clones are true transcribed sequences and not contaminating clones of the cDNA library from which they were isolated.

5.3.2.2 Reverse Transcriptase Polymerase Chain Reaction

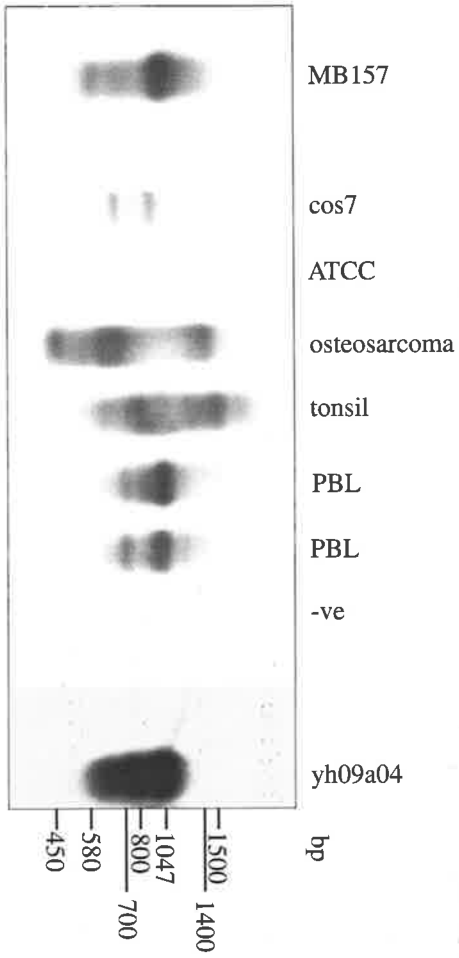
Northern analysis determined that clone yh09a04 was expressed at low levels, thus reverse transcriptase polymerase chain reaction (RT-PCR) was performed to acquire additional information on the expression patterns of this transcript. The RT-PCR technique provides an increased sensitivity, compared to Northern analysis, in examining sequences with low levels of expression. RT-PCR was also performed with the yc81e09 transcript to generate additional information on its expression pattern. PCR products of the double stranded cDNA syntheses generated with primers listed in table 5.1 were analysed on 1% agarose gels but a product was not visualised as the level of expression of both transcripts was low. Esterase D controls demonstrated a product of expected size (452 bp) for each RNA sample utilised, indicating that each RNA species was intact. Positive controls of cDNA as template indicated that the reaction conditions were optimal and demonstrated the correct sized products of 1047 bp for yh09a04 and 906 bp for yc81e09. No bands were demonstrated in reactions containing genomic DNA.

Following Southern blotting of the agarose gels and hybridisation with their respective radiolabelled PCR amplified cDNA sequence using the primers listed in table 5.1, signal was detected in various tissues. Examples of these autoradiographs are shown in figures 5.4 B and 5.5. RT-PCR using primers from clone yh09a04 (figure 5.5) demonstrated signal in brain stem, frontal lobe, muscle, tonsil and PBL. The osteosarcoma, monkey kidney cos 7 and breast carcinoma MB157 cell lines also demonstrated signal. These results are summarised in table 5.4. A band of expected size (1047 bp) was observed in muscle, frontal

Figure 5.5

An example of the autoradiographs of RT-PCR analyses of isolated cDNA sequences. These autoradiographs show amplified cDNA fragments generated from oligo-dT primed RNA from muscle, frontal lobe, brain stem, tonsil and peripheral blood lymphocytes (PBL) using primers derived from the yh09a04 cDNA clone. Products were also generated from the osteosarcoma, monkey kidney cos 7 and breast carcinoma MB157 cell lines. The sizes of the PCR products are indicated in bp.

Lanes labelled as '- RT'ase' are reactions in which no Reverse Transcriptase was added. The lanes labelled '-ve' indicate the reactions in which no template was added to monitor any contamination which may occur during PCR. The lane labelled yh09a04 is the reaction in which yh09a04 was used as template.


MB157
cos7
ATCC
osteosarcoma
tonsil
PBL
PBL
-ve
yh09a04
450
580
700
800
1047
1400
1500
bp

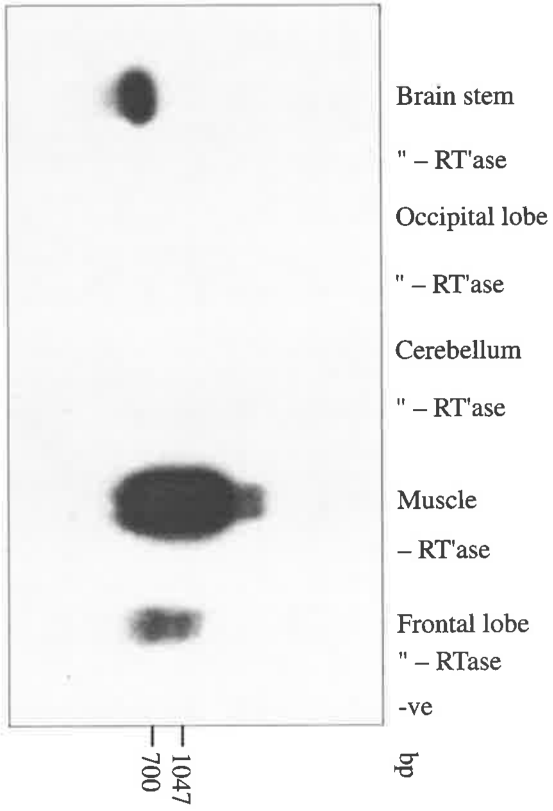

Brain stem
" – RT'ase
Occipital lobe
" – RT'ase
Cerebellum
" – RT'ase
Muscle
– RT'ase
Frontal lobe
" – RTase
-ve
700
1047
bp

Table 5.4

Results of RT-PCR with the indicated tissues and cell lines. The top table demonstrates the human tissues with which the transcripts were tested. The bottom table shows the cell lines used. A positive signal is denoted with a '+' and a negative signal is indicated by '-'.

cDNA clone	frontal lobe	occipital lobe	brain stem	cerebellum	muscle	lymphocytes	tonsil	nerve
yh09a04	+	-	+	-	+	+	+	-
yc81e09	+		-	-	+	+		-

cDNA clone	monkey kidney cos 7	osteosarcoma	breast carcinoma MB157	ATCC bladder carcinoma
yh09a04	+	+	+	-
yc81e09	-	+	+	+

lobe and PBL, and the MB157 and cos 7 cell lines. Different sized bands, which were most likely the products of alternative splicing, were observed in brain stem, muscle, frontal lobe and PBL and the cos 7, MB157 and osteosarcoma cell lines. The sizes of these bands were approximately 800 bp, 700 bp, 580 bp or 450 bp. Bands of 1.4 kb and 1.5 kb were observed in the tonsil and osteosarcoma cell line samples respectively and were probably due to non-specific signal.

RT-PCR using primers from clone yc81e09 demonstrated signal in muscle, frontal lobe and PBL. The osteosarcoma, ATCC bladder carcinoma and breast carcinoma MB157 cell lines also demonstrated a product. These results are summarised in table 5.4. A band of the expected size (906 bp) was observed in muscle, frontal lobe and the osteosarcoma cell line samples (figure 5.4 B). Bands of different size, probably due to alternative splicing, were observed in the MB157 and osteosarcoma cell lines and in PBL. The sizes of these bands were approximately 650 bp and 400 bp.

Signal was not observed in the reactions containing genomic DNA, or in the control reactions in which template was not added. Controls in which no reverse transcriptase was added to the reaction were also negative. This control is generally included in RT-PCR experiments as a monitor for genomic DNA contamination. However, the primers which were utilised are unlikely to amplify a product from genomic DNA as the size of this product would be too large to amplify, due to the presence of intronic sequences.

5.3.3 Sequence Analysis

Sequence of the two cDNA clones was first obtained from each end of the insert using primers homologous to the vector sequence adjacent to the cloning site. Primers were subsequently designed from the sequences generated to obtain sequence of the whole clone. Four primers were used to produce the entire sequence of the yh09a04 cDNA clone and three primers were used to generate the entire sequence of the yc81e09 cDNA clone. The primers that were utilised are listed in table 5.2. The total length of clone yh09a04 was 1659 nucleotides. The length of the yc81e09 clone was 1537 nucleotides. The non-redundant and dbEST nucleotide databases were screened using the BLAST-N algorithm to determine whether the clones demonstrated homology to sequences.

5.3.3.1 yh09a04

The sequences of the 3' and 5' ends of clone yh09a04 had previously been deposited in the GenBank database. The GenBank Accession numbers for the 5' and 3' ends of this clone were R61618 and R60950 respectively. The 5' end of clone yh09a04 showed identity to the sequence of the 3' end of clone yf14a03 which was deposited in the GenBank database (Accession No. R07162). An overlap of 580 nucleotides was observed. The yf14a03 cDNA clone, from an oligo-dT primed normalised infant brain cDNA library (Soares *et al.* 1994), was purchased from the IMAGE Consortium, USA, and the sequence of this clone was determined using vector primers on either side of the cloning site and the three primers listed in table 5.5. The total length of the clone was determined to be 1150 bp. The combined sequence of the overlapping clones yf14a03 and yh09a04, is 2140 bases. An exon, ET19, trapped from cosmid 352A12, performed by Mr. S. Whitmore, overlapped the 5' end of this sequence by 87 bases and extended the sequence by a further 171 bases. This total sequence of 2299 nucleotides, is depicted in figure 5.6. Exon ET19 spans bases 1-258. Clone yf14a03 (GenBank Accession Nos. R07162 and R07213) extends from bases 172-1323 and clone yh09a04 consists of bases 744-2299 and contains a poly(A) tail located at the 3' end. Examination of the nucleotide databases using the BLAST-N algorithm did not identify additional overlapping sequences which could extend this sequence further.

Table 5.5

Primers utilised for dye terminator sequencing reactions to determine the sequence of cDNA clone yf14a03. The locations of the primers in the sequence are indicated.

cDNA clone	**Primer Sequence (5' - 3')**	**Primer Position**
yf14a03	ATC ACT GCC CAC TTC TTC AGG	bases 236-256
yf14a03	AGG TCA CTC TCT GTC AAC TG	bases 647-628
yf14a03	ACA TGT CCA CAG CAA CAT GCA G	bases 767-746

Figure 5.6

Nucleotide sequence of exon ET19 and cDNA clones yh09a04 (GenBank Accession Nos. R61618 and R60950) and yf14a03 (GenBank Accession Nos. R07213 and R07162). The combined sequence of ET19, yf14a03 and yh09a04 totals 2299 nucleotides. ET19, which is underlined, spans bases 1-258 and overlaps the 5' end of yf14a03 by 87 bases. Clone yf14a03, shown in bold lettering, extends from bases 172-1323 and clone yh09a04, indicated with dashed underlining, consists of bases 744-2299. An overlap of 580 bases is observed between these cDNA clones.

```
CTCAAGGAGT TATGACCACT CAGAAAATTC TGATTTGGTC TTTGGTGGCC GCACAGGAAA   60
TGAGGATATT ATTTCCAGAT TGCAGGAGAT GGTAGCTGAC CTGGAGCTGC AGCAAGACCT  120
CATAGTGCCT CTCGGCCACA CCCCTTCCCA GGAGCACTTC CTCTTTGAGA T TTTCCGCAG  180
ACGGCTCCAG GCTCTGACAA GCGGGTGGAG CGTGGCTGCC AGCCTTCAGA GACAGAGGGA  240
GCTGCTAATG TACAAACGGA TCCTCCTCCG CCTGCCTTCG TCTGTCCTCT GCGGCAGCAG  300
CTTCCAGGCA GAACAGCCCA TCACTGCCAG ATGCGAGCAG TTCTTCCACT TGGTCAACTC  360
TGAGATGAGA AACTTCTGCT CCCACGGAGG TGCCCTGACA CAGGACATCA CTGCCCACTT  420
CTTCAGGGGC CTCCTGAACG CCTGTCTGCG GAGCAGAGAC CCCTCCCTGA TGGTCGACTT  480
CATACTGGCC AAGTGCCAGA CGAAATGCCC CTTAATTTTG ACCTCTGCTC TGGTGTGGTG  540
GCCGAGCCTG GAGCCTGTGC TGCTCTGCCG GTGGAGGAGA CACTGCCAGA GCCCGCTGCC  600
CCGGGAACTG CAGAAGCTAC AAGAAGGCCG GCAGTTTGCC AGCGAATTGG TTTTCCTTTT  660
CTTCTTCTCC TTGATGGGCC TGCTGTCGTC ACATCTGACC TCAAATAGCA CCACAGACCT  720
GCCAAAGGCT TTCCACGTTT GTGCAGCAAT CCTCGAGTGT TTAGAGAAGA GGAAGATATC  780
CTGGCTGGCA CTCTTTCAGT TGACAGAGAG TGACCTCAGG CTGGGGCGGC TCCTCCTCCG  840
TGTGGCCCCG GATCAGCACA CCAGGCTGCT GCCTTTCGCT TTTTACAGAC GCGGCCATCA  900
GGGAAGAGGC CTTCCTGCAT GTTGCTGTGG ACATGTACTT GAAGCTGGTC CAGCTCTTCG  960
TGGCTGGGGA TACAAGCACA GTTTCACCTC CAGCTGGCAG GAGCCTGGAG CTCAAGGGTC 1020
AGGCAGGGCA ACCCCGTGGA ACTGATAACA AAAGCTCGTC TTTTTCTGCT GCAGTTAATA 1080
CCTCGGTGCC CGAAAAAGAG CTTCTCACAC GTGGCAGAGC TGCTGGCTGA TCGTGGGGAC 1140
TGCGACCCAG AGGTGAGCGC CGCCCTCCAG AGCAGACAGC AGGCTGCCCC TGACGCTGAC 1200
CTGTCCCAGG AGCCTCATCT CTTCTGACGG GACCTGCCAC TGCACACCAG CCCAGCTCCC 1260
GTGTAAATAA TTTATTACAA GCATAACATG GAGCTCTTGT TGCACTAAAA AGTGGATTAC 1320
AAATCTCCTC GACTGCTTTA GTGGGGAAAG GAATCAATTA TTTATGAACT GTCCGGCCCC 1380
GAGTCACTCA GCGTTTGCGG GAAAATAAAC CACTGGTCCC AGAGCAGAGG AAGGCTACTT 1440
GAGCCGGACA CCAAGCCCGC CTCCAGCACC AAGGGCGGGC AGCACCCTCC GACCCTCCCA 1500
TGCGGGTGCA CACGAAGGGT GAGGCTGACA CAGCCACTGC GGAGTCCAGG CTGCTAGAGG 1560
TGCTCATCCT CACTGCCGTC CTCAGGTGGG TTCGGGCTTC ACCGCCTGGC CCTCTGTGGT 1620
CACAGAGGGG CTCGGTGGCC CAGGTGGTGG TTCCGCCTCC AGGGGCAGGG CCTTGTCCTG 1680
GGTCTGTGTC AGCGGGTGCA CCATGGACAT GTGTACATTG AGGTTGTGGG CCTTCTCAAA 1740
CCGCCGGCCA CACTGGTCAC AGGCAAAGTC CAGCTCAGTC TCAGCCTTGT GTTTGGTCAT 1800
GTGGTACTTG AGGGATGCCC GCTGCCTGCA CTGGAACCCA CAGACCTCAC ACCTGGGGGA 1860
CAGAGGCAGA TAAGAAGGTG CGAGGGCCAC AGCCCTGGGA GGGGGTCCTG ACTCACACTT 1920
ACTGCAAAGG CTTGGCTCCC GAATGTCGCA TTTGGTGGAC GAGAAGGTGC TTCCGCTGCT 1980
TGAAGGTTTG TCCACATTCG TCACAGATAT AGTTCCGCAC CTCTGAGAGG GGAGAGTCCA 2040
GTGAGTCCAG GCCCCTGATG CTCCAACCTC CCGGGGGGAC GACGATGACA ATGTGAAACC 2100
ATCACAGCTG GGAAGACATT TCTGCACATG GTTCACCATG CAGTGGGCCC AAGCAAGGGG 2160
CCTATGAGGG CCTCGTTTAT TAAGATCTTT AAACTGCTTT ATACACTGTC ACGTGGCTTC 2220
ATCAGCTGTG TGCATTTCAG GATGGTTTTT AAAGAAACCT CAGAAAGCTA TTTCCTTAAA 2280
AAAAAAAAAA AAAAAAAAA  2299
```

5.3.3.2 yc81e09

The sequences of the 3' and 5' ends of clone yc81e09 were previously deposited in the GenBank database. The GenBank Accession numbers for the 5' and 3' ends of this clone were T74058 and T87621 respectively. Clone yc81e09 detected overlaps with cDNA clones which include, yx82b02 (GenBank Accession No. N36323), zc48b07 (GenBank Accession No. W52106), zb82b10 (GenBank Accession Nos. W24223 and N95534), zb17g05 (GenBank Accession No. N78868), zb05a12 (GenBank Accession No. N78732) and yd48d10 (GenBank Accession Nos. T74058 and T90900) in the non-redundant and EST databases but these did not extend the sequence of clone yc81e09 at the 5' end. The sequence of this cDNA clone contains a poly(A) tail located at the 3' end and comprises a total of 1537 nucleotides. The sequence is shown in figure 5.7. Portions of the sequence of this clone were also confirmed by Mr. S. Whitmore and Ms. J. Crawford.

5.3.4 Isolation of cDNA Clones to Extend the Sequences of the Transcripts

Attempts were made to isolate additional cDNA clones to extend the existing sequence of the transcripts. One strategy involved a PCR-based method to obtain cDNA sequence covering additional portions of the two genes. Primers near the 5' end of the sequence of each cDNA clone were designed. Individual PCR reactions were performed with a specific primer corresponding to each cDNA combined with the λgt11-reverse vector primer. A 5'-stretch random and oligo-dT primed human foetal brain cDNA library cloned in the λgt11 vector was used as template (Clontech). Analysis of the PCR products on agarose gels revealed multiple bands of different sizes. These were probed with their corresponding cDNA clone then specific bands were excised from low melting point agarose gels and cloned into pGEM-T vector. Sequencing revealed that these clones overlapped with their corresponding cDNA clone, but unfortunately did not extend the sequence further.

Figure 5.7

Nucleotide sequence of cDNA clone yc81e09 (GenBank Accession Nos. T74058 and T87621). The sequence of this clone, which is shown in black and red lettering, begins at nucleotide 653 and extends to nucleotide 2189. This cDNA clone consists of a total of 1537 bp.

An additional 652 nucleotides, indicated in blue lettering, are shown at the 5' end of the yc81e09 cDNA clone. Exons ET17.14, bases 177-339, and ET17.12, bases 534-599, provided 475 nucleotides. These exons are indicated in bold blue lettering. An additional 177 nucleotides at the 5' end were identified from cDNA clone ze25b01 (GenBank Accession No. AA035673) following a search of non-redundant and EST nucleotide databases using BLAST-N. The total number of nucleotides in the sequence is 2189.

The longest continuous potential open reading frame (ORF) in the sequence begins at nucleotide 691 and is 759 bp long. This ORF is indicated in red lettering. The sequence surrounding the proposed translation initiation site, showing a degree of homology to the Kozak consensus sequence 'ACCATGG' or 'TCCATGTC' (Kozak, 1986) is underlined.

A possible polyadenylation signal may lie in the sequence TATAAAT located at bases 2141-2147, indicated with a dashed underline, which is 12 bp upstream of the poly(A) tract.

```
GCCCTCCGGA  GCTNGAGGGG  GGAACAGCGC  GGCCAGGAGC  CCCTCGCCCG   50
GCNCCTTGCA  TTTCGATCTC  CGTGATGACG  ATGACGCGGA  AGAAGAAGGG   100
CCCAAGCGGG  ANTTGTNGTC  CGGCGTCCCG  GGGGCGCANG  GAAGGAGGGC   150
GTCCGAGTCA  ACAACCGCTT  CGAGCTGATA  AACATTGACG  ATCTTGAGGA   200
TGACCCTGTG  GTGAACGGGG  AGAGGTCTGG  CTGTGCGCTC  ACAGACGCTG   250
TGGCACCAGG  GAACAAAGGA  AGGGGTCAGC  GTGGAAACAC  AGAGAGCAAG   300
ACGGATGGAG  ATGACACCGA  GACAGTGCCC  TCAGAGCAGT  CTCATGCAAG   350
TGGCAAACTC  CGGAAGAAGA  AAAAAAAACA  GAAAAACAAG  AAAAGCAGCA   400
CGGGAGAAGC  ATCGGAAAAC  GGACTAGAAG  ATATCGATCG  CATCCTAGAG   450
AGGATTGAGG  ACAGCACTGG  GTTGAACCGT  CCCGGCCCAG  CTCCCCTGAG   500
CTCCAGGAAG  CACGTTCTCT  ACGTGGAGCA  CAGACACTTG  AATCCAGACA   550
CAGAACTGAA  AAGGTATTTT  GGTGCCCGGG  CAATCCTGGG  GGAGCAAAGG   600
CCACGGCAGA  GACAACGTGT  GTACCCCAAG  TGCACATGGC  TGACCACCCC   650
TAAAGCACC   TGGCCCCGCT  ACAGCAAACC  AGGTCTGTCC  ATGCGGCTGC   700
TGGAATCAAA  AAAAGGCCTC  TCCTTCTTTG  CGTTTGAGCA  CAGTGAGGAG   750
TACCAGCAGG  CTCAGCACAA  GTTCCTGGTG  GCCGTGGAGT  CTATGGAGCC   800
GAACAACATC  GTGGTTCTGC  TCCAGACGAG  CCCTTACCAC  GTTGACTCAC   850
TCCTGCAGCT  CAGCGATGCC  TGCCGCTTTC  AAGAGGATCA  GGAGATGGCT   900
CGAGACCTCG  TAGAGAGAGC  GCTGTACAGC  ATGGAATGTG  CGTTCCCACC   950
CCCTGTTCAG  TCTCACCAGT  GGGGCTGCCG  GCTGGATTAC  CGCAGACCCG   1000
AGAACAGGAG  CTTCTACCTG  GCCCTCTACA  AGCAGATGAG  CTTCCTGGAG   1050
AAGCGAGGCT  GCCCGCGCAC  GGCGCTGGAG  TACTGCAAGC  TCATCCTGAG   1100
TCTCGAGCCG  GATGAGGACC  GCCTCTGCAT  GCTGCTGCTC  ATCGACCACC   1150
TGGCCTTGCG  GGCCCGGAAC  TACGAGTACC  TGATCCGCCT  CTTCCAGGAG   1200
TGGGAGGCTC  ATCGGAACCT  GTCCCAGCTC  CCTAATTTTG  CCTTCTCTGT   1250
TCCACTGGCG  TATTTCCTGC  TGAGCCAGCA  GACAGACCTC  CCTGAGTGTG   1300
AGCAGAGCTC  TGCCAGGCAG  AAGGCCTCTC  TCCTGATACA  GCAGGCGCTC   1350
ACCATGTTCC  CTGGAGTCCT  CCTTGCCCCT  GCTCGAGTCT  TGCAGTGTGC   1400
                                                      *
GGCCCGACGC  CAGCGTTTCC  AGTCACCGCT  TCTTTGGACC  CAATGCTGAA   1450
ATAAGCCAGC  CCCCTGCCCT  GAGCCAGCTG  GTGAACCTGT  ACCTTGGGAG   1500
GTCACACTTT  CTCTGGAAAG  AGCCCGCCAC  CATGAGCTGG  CTGGAGGAGA   1550
ACGTCCACGA  GGTTCTGCAA  GCAGTGGACG  CCGGGGACCC  AGCCGTGGAA   1600
GCCTGTGAGA  ACCGGCGGAA  GGTGCTCTAC  CAGCGTGCAC  CCAGGAATAT   1650
CCACCGCCAT  GTGATCCTCT  CTGAGATCAA  GGAAGCCGTC  GCTGCCCTGC   1700
CCCCGGACGT  GACCACGCAG  TCTGTGATGG  GGTTTGATCC  TCTGCCTCCT   1750
TCGGACACAA  TCTACTCCTA  CGTCAGGCCA  GAGAGGCTAA  GTCCTATCAG   1800
CCATGGAAAC  ACCATTGCTC  TCTTCTTCCG  GTCACTGTTT  CCAAACTAAA   1850
CCATGGAGGG  GGAGAGGCCC  GAGGAAGGAG  TGGCTGGGGG  TTCTGAACCG   1900
CAACCAGGGC  CTGAACAGGC  TGATGCTGGC  TGTGCGCGAC  ATGATGGCCA   1950
ACTTCCACTT  CAACGACCTG  GAGGCGCCGC  ACGAGGACGA  CGCTGAGGGG   2000
GAGGGGGAGT  GGGACTGAGC  GTCCGCAGAG  GTGACCGAAA  AGCCGTATGA   2050
TGATGTTCCC  GATTTCTCTG  TTGGTCGGAG  TCGGCCAGTT  GCCTGAAGTA   2100
GGGAAGCTGA  GTGTGTCGCT  CCCTGGTCCA  CTGTTTCTCC  TATAAATGTA   2150
AATGGGTCCA  AAAAAAAAAA  AAAAAAAAAA  AAAAAAAAA   2189
```

Another strategy to extend the sequence of the transcripts involved the use of the 5'-rapid amplification of cDNA ends (RACE) technique (GIBCO BRL) (Frohman *et al.* 1988). A primer specific for each cDNA, located near the 5' end of the existing sequence was used to synthesise first strand cDNA from total RNA isolated from PBL. This procedure proved to be unsuccessful as specific products were not generated.

5.3.5 Contributions to the Characterisation of Transcripts yh09a04 and yc81e09 From the Fanconi Anaemia/Breast Cancer Consortium

5.3.5.1 yh09a04

Exon trapping of cosmids c352A12 and c431F1, members of the yh09a04 contig (chapter 3), identified six potential exons, including ET19, distal to the direct selected cDNAs yh09a04/yf14a03. These exons were used to extend the sequence by RT-PCR and to screen cDNA libraries for larger clones. Sequencing of overlapping cDNA clones revealed an open reading frame of 4,365 nucleotides which encodes a protein of 1,455 amino acids. The sequence of the 5.5 kb transcript and the predicted amino acid sequence of the translation product are shown in figure 5.8. The compiled sequence of clones ET19, yf14a03 and yh09a04 begins at base 3026 of this sequence and ends at the poly (A) tail.

Comparison of the ET19/yf14a03/yh09a04 sequence, described in 5.3.3; figure 5.6, with the 5.5 kb transcript sequence revealed some differences. An insertion of 4 bases, GCAG, in the ET19/yf14a03/yh09a04 sequence was observed between position 4210 and 4211. This insertion was only observed in the sequence from the yh09a04 cDNA clone. Deletions in the ET19/yf14a03/yh09a04 sequence were observed from bases 3671-3811, a total of 139 bp, and 4055-4077, a total of 23 bp. The sequences that are deleted are highlighted in figure 5.8. The 139 bp deletion was only observed in the sequence of the yf14a03 cDNA clone, and the 23 bp deletion was only observed in the sequence of the yh09a04 cDNA clone.

Figure 5.8

Nucleotide (blue lettering) and deduced amino acid (red lettering) sequence of the 5.5 kb cDNA. The open reading frame which begins at base 45, consists of 4,365 nucleotides which encode a protein of 1,455 amino acids.

The combined sequence of clones ET19, yf14a03 and yh09a04 begins at base 3026 and ends at the poly(A) tail. The sequence of ET19 is highlighted in orange; sequence yf14a03 is underlined in black; sequence yh09a04 is highlighted in purple.

Sequence ET19/yf14a03/yh09a04 has an insertion of 4 bases, GCAG (black lettering), between bases 4210 and 4211. The two regions of sequence which were deleted in the ET19/yf14a03/yh09a04 sequence were from bases 3671-3811 and 4055-4077, indicated in bold blue lettering.

The 3' untranslated region consists of 973 nucleotides and contains a polyadenylation signal beginning at nucleotide 4589, which is underlined in red.

```
CTCGGGCGCAGGGAGCCGCCGCCGGGGCTGTAGGCGCCAAGGCCATGTCCGACTCGTGGGTCCCGAACTCCGCCTCGGGC  80
                                             Met Ser Asp Ser Trp Val Pro Asn Ser Ala Ser Gly

CAGGACCCAGGGGGCCGCCGGAGGGCCTGGGCCGAGCTGCTGGCGGGAAGGGTCAAGAGGGAAAAATATAATCCTGAAAG  160
Gln Asp Pro Gly Gly Arg Arg Arg Ala Trp Ala Glu Leu Leu Ala Gly Arg Val Lys Arg Glu Lys Tyr Asn Pro Glu Arg

GGCACAGAAATTAAAGGAATCAGCTGTGCGCCTCCTGCGAAGCCATCAGGACCTGAATGCCCTTTTGCTTGAGGTAGAAG  240
Ala Gln Lys Leu Lys Glu Ser Ala Val Arg Leu Leu Arg Ser His Gln Asp Leu Asn Ala Leu Leu Leu Glu Val Glu

GTCCACTGTGTAAAAAATTGTCTCTCAGCAAAGTGATTGACTGTGACAGTTCTGAGGCCTATGCTAATCATTCTAGTTCA  320
Gly Pro Leu Cys Lys Lys Leu Ser Leu Ser Lys Val Ile Asp Cys Asp Ser Ser Glu Ala Tyr Ala Asn His Ser Ser Ser

TTTATAGGCTCTGCTTTGCAGGATCAAGCCTCAAGGCTGGGGGTTCCCGTGGGTATTCTCTCAGCCGGGATGGTTGCCTC  400
Phe Ile Gly Ser Ala Leu Gln Asp Gln Ala Ser Arg Leu Gly Val Pro Val Gly Ile Leu Ser Ala Gly Met Val Ala Ser

TAGCGTGGGACAGATCTGCACGGCTCCAGCGGAGACCAGTCACCCTGTGCTGCTGACTGTGGAGCAGAGAAAGAAGCTGT  480
Ser Val Gly Gln Ile Cys Thr Ala Pro Ala Glu Thr Ser His Pro Val Leu Leu Thr Val Glu Gln Arg Lys Lys Leu

CTTCCCTGTTAGAGTTTGCTCAGTATTTATTGGCACACAGTATGTTCTCCCGTCTTTCCTTCTGTCAAGAATTATGGAAA  560
Ser Ser Leu Leu Glu Phe Ala Gln Tyr Leu Leu Ala His Ser Met Phe Ser Arg Leu Ser Phe Cys Gln Glu Leu Trp Lys

ATACAGAGTTCTTTGTTGCTTGAAGCGGTGTGGCATCTTCACGTACAAGGCATTGTGAGCCTGCAAGAGCTGCTGGAAAG  640
Ile Gln Ser Ser Leu Leu Leu Glu Ala Val Trp His Leu His Val Gln Gly Ile Val Ser Leu Gln Glu Leu Leu Glu Ser

CCATCCCGACATGCATGCTGTGGGATCGTGGCTCTTCAGGAATCTGTGCTGCCTTTGTGAACAGATGGAAGCATCCTGCC  720
His Pro Asp Met His Ala Val Gly Ser Trp Leu Phe Arg Asn Leu Cys Cys Leu Cys Glu Gln Met Glu Ala Ser Cys

AGCATGCTGACGTCGCCAGGGCCATGCTTTCTGATTTTGTTCAAATGTTTGTTTTGAGGGGATTTCAGAAAAACTCAGAT  800
Gln His Ala Asp Val Ala Arg Ala Met Leu Ser Asp Phe Val Gln Met Phe Val Leu Arg Gly Phe Gln Lys Asn Ser Asp

CTGAGAAGAACTGTGGAGCCTGAAAAAATGCCGCAGGTCACGGTTGATGTACTGCAGAGAATGCTGATTTTTGCACTTGA  880
Leu Arg Arg Thr Val Glu Pro Glu Lys Met Pro Gln Val Thr Val Asp Val Leu Gln Arg Met Leu Ile Phe Ala Leu Asp

CGCTTTGGCTGCTGGAGTACAGGAGGAGTCCTCCACTCACAAGATCGTGAGGTGCTGGTTCGGAGTGTTCAGTGGACACA  960
Ala Leu Ala Ala Gly Val Gln Glu Glu Ser Ser Thr His Lys Ile Val Arg Cys Trp Phe Gly Val Phe Ser Gly His

CGCTTGGCAGTGTAATTTCCACAGATCCTCTGAAGAGGTTCTTCAGTCATACCCTGACTCAGATACTCACTCACAGCCCT  1040
Thr Leu Gly Ser Val Ile Ser Thr Asp Pro Leu Lys Arg Phe Phe Ser His Thr Leu Thr Gln Ile Leu Thr His Ser Pro

GTGCTGAAAGCATCTGATGCTGTTCAGATGCAGAGAGAGTGGAGCTTTGCGCGGACACACCCTCTGCTCACCTCACTGTA  1120
Val Leu Lys Ala Ser Asp Ala Val Gln Met Gln Arg Glu Trp Ser Phe Ala Arg Thr His Pro Leu Leu Thr Ser Leu Tyr
```

```
CCGCAGGCTCTTTGTGATGCTGAGTGCAGAGGAGTTGGTTGGCCATTTGCAAGAAGTTCTGGAAACGCAGGAGGTTCACT 1200
 Arg Arg Leu Phe Val Met Leu Ser Ala Glu Glu Leu Val Gly His Leu Gln Glu Val Leu Glu Thr Gln Glu Val His

GGCAGAGAGTGCTCTCCTTTGTGTCTGCCCTGGTTGTCTGCTTTCCAGAAGCGCAGCAGCTGCTTGAAGACTGGGTGGCG 1280
Trp Gln Arg Val Leu Ser Phe Val Ser Ala Leu Val Val Cys Phe Pro Glu Ala Gln Gln Leu Leu Glu Asp Trp Val Ala

CGTTTGATGGCCCAGGCATTCGAGAGCTGCCAGCTGGACAGCATGGTCACTGCGTTCCTGGTTGTGCGCCAGGCAGCACT 1360
 Arg Leu Met Ala Gln Ala Phe Glu Ser Cys Gln Leu Asp Ser Met Val Thr Ala Phe Leu Val Val Arg Gln Ala Ala Leu

GGAGGGCCCCTCTGCGTTCCTGTCATATGCAGACTGGTTCAAGGCCTCCTTTGGGAGCACACGAGGCTACCATGGCTGCA 1440
  Glu Gly Pro Ser Ala Phe Leu Ser Tyr Ala Asp Trp Phe Lys Ala Ser Phe Gly Ser Thr Arg Gly Tyr His Gly Cys

GCAAGAAGGCCCTGGTCTTCCTGTTTACGTTCTTGTCAGAACTCGTGCCTTTTGAGTCTCCCCGGTACCTGCAGGTGCAC 1520
Ser Lys Lys Ala Leu Val Phe Leu Phe Thr Phe Leu Ser Glu Leu Val Pro Phe Glu Ser Pro Arg Tyr Leu Gln Val His

ATTCTCCACCCACCCCTGGTTCCCAGCAAGTACCGCTCCCTCCTCACAGACTACATCTCATTGGCCAAGACACGGCTGGC 1600
 Ile Leu His Pro Pro Leu Val Pro Ser Lys Tyr Arg Ser Leu Leu Thr Asp Tyr Ile Ser Leu Ala Lys Thr Arg Leu Ala

CGACCTCAAGGTTTCTATAGAAAACATGGGACTCTACGAGGATTTGTCATCAGCTGGGGACATTACTGAGCCCCACAGCC 1680
  Asp Leu Lys Val Ser Ile Glu Asn Met Gly Leu Tyr Glu Asp Leu Ser Ser Ala Gly Asp Ile Thr Glu Pro His Ser

AAGCTCTTCAGGATGTTGAAAAGGCCATCATGGTGTTTGAGCATACGGGGAACATCCCAGTCACCGTCATGGAGGCCAGC 1760
Gln Ala Leu Gln Asp Val Glu Lys Ala Ile Met Val Phe Glu His Thr Gly Asn Ile Pro Val Thr Val Met Glu Ala Ser

ATATTCAGGAGGCCTTACTACGTGTCCCACTTCCTCCCCGCCCTGCTCACACCTCGAGTGCTCCCCAAAGTCCCTGACTC 1840
 Ile Phe Arg Arg Pro Tyr Tyr Val Ser His Phe Leu Pro Ala Leu Leu Thr Pro Arg Val Leu Pro Lys Val Pro Asp Ser

CCGTGTGGCGTTTATAGAGTCTCTGAAGAGAGCAGATAAAATCCCCCCATCTCTGTACTCCACCTACTGCCAGGCCTGCT 1920
  Arg Val Ala Phe Ile Glu Ser Leu Lys Arg Ala Asp Lys Ile Pro Pro Ser Leu Tyr Ser Thr Tyr Cys Gln Ala Cys

CTGCTGCTGAAGAGAAGCCAGAAGATGCAGCCCTGGGAGTGAGGGCAGAACCCAACTCTGCTGAGGAGCCCCTGGGACAG 2000
Ser Ala Ala Glu Glu Lys Pro Glu Asp Ala Ala Leu Gly Val Arg Ala Glu Pro Asn Ser Ala Glu Glu Pro Leu Gly Gln

CTCACAGCTGCACTGGGAGAGCTGAGAGCCTCCATGACAGACCCCAGCCAGCGTGATGTTATATCGGCACAGGTGGCAGT 2080
 Leu Thr Ala Ala Leu Gly Glu Leu Arg Ala Ser Met Thr Asp Pro Ser Gln Arg Asp Val Ile Ser Ala Gln Val Ala Val

GATTTCTGAAAGACTGAGGGCTGTCCTGGGCCACAATGAGGATGACAGCAGCGTTGAGATATCAAAGATTCAGCTCAGCA 2160
  Ile Ser Glu Arg Leu Arg Ala Val Leu Gly His Asn Glu Asp Asp Ser Ser Val Glu Ile Ser Lys Ile Gln Leu Ser

TCAACACGCCGAGACTGGAGCCACGGGAACACATGGCTGTGGACCTCCTGCTGACGTCTTTCTGTCAGAACCTGATGGCT 2240
 Ile Asn Thr Pro Arg Leu Glu Pro Arg Glu His Met Ala Val Asp Leu Leu Leu Thr Ser Phe Cys Gln Asn Leu Met Ala

GCCTCCAGTGTCGCTCCCCCGGAGAGGCCGGGTCCCTGGGCTGCCCTCTTCGTGAGGACCATGTGTGGACGTGTGCTCCC 2320
 Ala Ser Ser Val Ala Pro Pro Glu Arg Pro Gly Pro Trp Ala Ala Leu Phe Val Arg Thr Met Cys Gly Arg Val Leu Pro
```

```
TGCAGTGCTCACCCGGCTCTGCCAGCTGCTCCGTCACCAGGGCCCGAGCCTGAGTGCCCCACATGTGCTGGGGTTGGCTG 2400
 Ala Val Leu Thr Arg Leu Cys Gln Leu Leu Arg His Gln Gly Pro Ser Leu Ser Ala Pro His Val Leu Gly Leu Ala
CCCTGGCCGTGCACCTGGGTGAGTCCAGGTCTGCGCTCCCAGAGGTGGATGTGGGTCCTCCTGCACCTGGTGCTGGCCTT 2480
Ala Leu Ala Val His Leu Gly Glu Ser Arg Ser Ala Leu Pro Glu Val Asp Val Gly Pro Pro Ala Pro Gly Ala Gly Leu
CCTGTCCCTGCGCTCTTTGACAGCCTCCTGACCTGTAGGACGAGGGATTCCTTGTTCTTCTGCCTGAAATTTTGTACAGC 2560
 Pro Val Pro Ala Leu Phe Asp Ser Leu Leu Thr Cys Arg Thr Arg Asp Ser Leu Phe Phe Cys Leu Lys Phe Cys Thr Ala
AGCAATTTCTTACTCTCTCTGCAAGTTTTCTTCCCAGTCACGAGATACTTTGTGCAGCTGCTTATCTCCAGGCCTTATTA 2640
  Ala Ile Ser Tyr Ser Leu Cys Lys Phe Ser Ser Gln Ser Arg Asp Thr Leu Cys Ser Cys Leu Ser Pro Gly Leu Ile
AAAAGTTTCAGTTCCTCATGTTCAGATTGTTCTCAGAGGCCCGACAGGCTCTTTCTGAGGAGGACGTAGCCAGCCTTTCC 2720
Lys Lys Phe Gln Phe Leu Met Phe Arg Leu Phe Ser Glu Ala Arg Gln Ala Leu Ser Glu Glu Asp Val Ala Ser Leu Ser
TGGAGACCCTTGCACCTTCCTTCTGCAGACTGGCAGAGAGCTGCCCTCTCTCTCTGGACACACAGAACCTTCCGAGAGGT 2800
 Trp Arg Pro Leu His Leu Pro Ser Ala Asp Trp Gln Arg Ala Ala Leu Ser Leu Trp Thr His Arg Thr Phe Arg Glu Val
GTTGAAAGAGGAAGATGTTCACTTAACTTACCAAGACTGGTTACACCTGGAGCTGGAAATTCAACCTGAAGCTGATGCTC 2880
  Leu Lys Glu Glu Asp Val His Leu Thr Tyr Gln Asp Trp Leu His Leu Glu Leu Glu Ile Gln Pro Glu Ala Asp Ala
TTTCAGATACTGAACGGCAGGACTTCCACCAGTGGGCGATCCATGAGCACTTTCTCCCTGAGTCCTCGGCTTCAGGGGGC 2960
Leu Ser Asp Thr Glu Arg Gln Asp Phe His Gln Trp Ala Ile His Glu His Phe Leu Pro Glu Ser Ser Ala Ser Gly Gly
TGTGACGGAGACCTGCAGGCTGCGTGTACCATTCTTGTCAACGCACTGATGGATTTCCACCAAAG CTCAAGGAGTTATGA 3040
 Cys Asp Gly Asp Leu Gln Ala Ala Cys Thr Ile Leu Val Asn Ala Leu Met Asp Phe His Gln Ser Ser Arg Ser Tyr Asp
CCACTCAGAAAATTCTGATTTGGTCTTTGGTGGCCGCACAGGAAATGAGGATATTATTTCCAGATTGCAGGAGATGGTAG 3120
  His Ser Glu Asn Ser Asp Leu Val Phe Gly Gly Arg Thr Gly Asn Glu Asp Ile Ile Ser Arg Leu Gln Glu Met Val
CTGACCTGGAGCTGCAGCAAGACCTCATAGTGCCTCTCGGCCACACCCCTTCCCAGGAGCACTTCCTCTTTGAGAT TTTC 3200
Ala Asp Leu Glu Leu Gln Gln Asp Leu Ile Val Pro Leu Gly His Thr Pro Ser Gln Glu His Phe Leu Phe Glu Ile Phe
CGCAGACGGCTCCAGGCTCTGACAAGCGGGTGGAGCGTGGCTGCCAGCCTTCAGAGACAGAGGGAGCTGCTAATGTACAA 3280
 Arg Arg Arg Leu Gln Ala Leu Thr Ser Gly Trp Ser Val Ala Ala Ser Leu Gln Arg Gln Arg Glu Leu Leu Met Tyr Lys
ACGGATCCTCCTCCGCCTGCCTTCGTCTGTCCTCTGCGGCAGCAGCTTCCAGGCAGAACAGCCCATCACTGCCAGATGCG 3360
  Arg Ile Leu Leu Arg Leu Pro Ser Ser Val Leu Cys Gly Ser Ser Phe Gln Ala Glu Gln Pro Ile Thr Ala Arg Cys
AGCAGTTCTTCCACTTGGTCAACTCTGAGATGAGAAACTTCTGCTCCCACGGAGGTGCCCTGACACAGGACATCACTGCC 3440
Glu Gln Phe Phe His Leu Val Asn Ser Glu Met Arg Asn Phe Cys Ser His Gly Gly Ala Leu Thr Gln Asp Ile Thr Ala
CACTTCTTCAGGGGCCTCCTGAACGCCTGTCTGCGGAGCAGAGACCCCTCCCTGATGGTCGACTTCATACTGGCCAAGTG 3520
  His Phe Phe Arg Gly Leu Leu Asn Ala Cys Leu Arg Ser Arg Asp Pro Ser Leu Met Val Asp Phe Ile Leu Ala Lys Cys
```

```
CCAGACGAAATGCCCCTTAATTTTGACCTCTGCTCTGGTGTGGTGGCCGAGCCTGGAGCCTGTGCTGCTCTGCCGGTGGA 3600
 Gln Thr Lys Cys Pro Leu Ile Leu Thr Ser Ala Leu Val Trp Trp Pro Ser Leu Glu Pro Val Leu Leu Cys Arg Trp

GGAGACACTGCCAGAGCCCGCTGCCCCGGGAACTGCAGAAGCTACAAGAAGGCCGGCAGTTTGCCAGCGA TTTCCTCTCC 3680
Arg Arg His Cys Gln Ser Pro Leu Pro Arg Glu Leu Gln Lys Leu Gln Glu Gly Arg Gln Phe Ala Ser Asp Phe Leu Ser

CCTGAGGCTGCCTCCCCAGCACCCAACCCGGACTGGCTCTCAGCTGCTGCACTGCACTTTGCGATTCAACAAGTCAGGGA 3760
 Pro Glu Ala Ala Ser Pro Ala Pro Asn Pro Asp Trp Leu Ser Ala Ala Ala Leu His Phe Ala Ile Gln Gln Val Arg Glu

AGAAAACATCAGGAAGCAGCTAAAGAAGCTGGACTGCGAGAGAGAGGAGCT ATTGGTTTTCCTTTTCTTCTTCTCCTTGA 3840
 Glu Asn Ile Arg Lys Gln Leu Lys Lys Leu Asp Cys Glu Arg Glu Glu Leu Leu Val Phe Leu Phe Phe Phe Ser Leu

TGGGCCTGCTGTCGTCACATCTGACCTCAAATAGCACCACAGACCTGCCAAAGGCTTTCCACGTTTGTG CAGCAATCCTC 3920
Met Gly Leu Leu Ser Ser His Leu Thr Ser Asn Ser Thr Thr Asp Leu Pro Lys Ala Phe His Val Cys Ala Ala Ile Leu

GAGTGTTTAGAGAAGAGGAAGATATCCTGGCTGGCACTCTTTCAGTTGACAGAGAGTGACCTCAGGCTGGGGCGGCTCCT 4000
 Glu Cys Leu Glu Lys Arg Lys Ile Ser Trp Leu Ala Leu Phe Gln Leu Thr Glu Ser Asp Leu Arg Leu Gly Arg Leu Leu

CCTCCGTGTGGCCCCGGATCAGCACACCAGGCTGCTGCCTTTCGCTTTTTACAG TCTTCTCTCCTACTTCCATGAAG ACG 4080
 Leu Arg Val Ala Pro Asp Gln His Thr Arg Leu Leu Pro Phe Ala Phe Tyr Ser Leu Leu Ser Tyr Phe His Glu Asp

CGGCCATCAGGGAAGAGGCCTTCCTGCATGTTGCTGTGGACATGTACTTGAAGCTGGTCCAGCTCTTCGTGGCTGGGGAT 4160
Ala Ala Ile Arg Glu Glu Ala Phe Leu His Val Ala Val Asp Met Tyr Leu Lys Leu Val Gln Leu Phe Val Ala Gly Asp

                                                                  GCAG
ACAAGCACAGTTTCACCTCCAGCTGGCAGGAGCCTGGAGCTCAAGGGTCAGGGCAACCCCGTGGAACTGATAACAAAAGC 4240
                                                                   ↑
 Thr Ser Thr Val Ser Pro Pro Ala Gly Arg Ser Leu Glu Leu Lys Gly Gln Gly Asn Pro Val Glu Leu Ile Thr Lys Ala

TCGTCTTTTTCTGCTGCAGTTAATACCTCGGTGCCCGAAAAAGAGCTTCTCACACGTGGCAGAGCTGCTGGCTGATCGTG 4320
  Arg Leu Phe Leu Leu Gln Leu Ile Pro Arg Cys Pro Lys Lys Ser Phe Ser His Val Ala Glu Leu Leu Ala Asp Arg

GGGACTGCGACCCAGAGGTGAGCGCCGCCCTCCAGAGCAGACAGCAGGCTGCCCCTGACGCTGACCTGTCCCAGGAGCCT 4400
Gly Asp Cys Asp Pro Glu Val Ser Ala Ala Leu Gln Ser Arg Gln Gln Ala Ala Pro Asp Ala Asp Leu Ser Gln Glu Pro

CATCTCTTCTGACGGGACCTGCCACT GCACACCAGCCCAGCTCCCGTGTAAATAATTTATTACAAGCATAACATGGAGCT 4480
His Leu Phe •

CTTGTTGCACTAAAAAGTGGATTACAAA TCTCCTCGACTGCTTTAGTGGGGAAAGGAATCAATTATTTATGAACTGTCCG 4560
GCCCCGAGTCACTCAGCGTTTGCGGGAAAATAAACCACTGGTCCCAGAGCAGAGGAAGGCTACTTGAGCCGGACACCAAG 4640
CCCGCCTCCAGCACCAAGGGCGGGCAGCACCCTCCGACCCTCCCATGCGGGTGCACACGAAGGGTGAGGCTGACACAGCC 4720
ACTGCGGAGTCCAGGCTGCTAGAGGTGCTCATCCTCACTGCCGTCCTCAGGTGGGTTCGGGCTTCACCGCCTGGCCCTCT 4800
GTGGTCACAGAGGGGCTCGGTGGCCCAGGTGGTGGTTCCGCCTCCAGGGGCAGGGCCTTGTCCTGGGTCTGTGTCAGCGG 4880
GTGCACCATGGACATGTGTACATTGAGGTTGTGGGCCTTCTCAAACCGCCGGCCACACTGGTCACAGGCAAAGTCCAGCT 4960
CAGTCTCAGCCTTGTGTTTGGTCATGTGGTACTTGAGGGATGCCCGCTGCCTGCACTGGAACCCACAGACCTCACACCTG 5040
GGGGACAGAGGCAGATAAGAAGGTGCGAGGGCCACAGCCCTGGGAGGGGGTCCTGACTCACACTTACTGCAAAGGCTTGG 5120
CTCCCGAATGTCGCATTTGGTGGACGAGAAGGTGCTTCCGCTGCTTGAAGGTTTGTCCACATTCGTCACAGATATAGTTC 5200
CGCACCTCTGAGAGGGGAGAGTCCAGTGAGTCCAGGCCCCTGATGCTCCAACCTCCCGGGGGGACGACGATGACAATGTG 5280
AAACCATCACAGCTGGGAAGACATTTCTGCACATGGTTCACCATGCAGTGGGCCCAAGCAAGGGGCCTATGAGGGCCTCG 5360
TTTATTAAGATCTTTAAACTGCTTTATACACTGTCACGTGGCTTCATCAGCTGTGTGCATTTCAGGATGGTTTTTAAAGA 5440
AACCTCAGAAAGCTATTTCCTTAAAAAAAAAAAAAAAAAAA 5481
```

Searches of the non-redundant and dbEST nucleotide and protein databases with the 5.5 kb transcript sequence using BLAST-N and BLAST-X failed to reveal strong homologies with genes of known function. Analysis of the amino acid sequence with PSORT, a program for the prediction of protein localisation sites, identified possible nuclear localisation signals that matched the nucleoplasmin consensus motif. The highest protein similarities detected were with the myosin 1B heavy chain and formin, the chicken limb deformity gene product [P(N) 0.20 and 0.24].

As stated, the hybridisation signal of the yh09a04 clone to the Northern blot was very faint (5.3.2.1). Thus, Northern analysis was repeated by members of the FAB consortium using a commercially available multiple tissue Northern blot of poly(A)+ RNA (Clontech) and a probe consisting of nucleotides 641-1403 of the 5.5 kb transcript. Northern blot analysis indicated that this gene is expressed in a variety of tissues, including pancreas, skeletal muscle, liver and placenta. Transcripts of several sizes were detected (7.5 kb, 5.5 kb, 4.7 kb, 3 kb and 2 kb) the most prominent of which was 4.7 kb in length. This result suggests a significant degree of alternative splicing or variations in the length of the 5' and 3' untranslated regions. The number of transcripts detected also depends on the region of the cDNA which is used as the probe for the Northern analysis. Generally, probes produced from the 5' end of the cDNA sequence are preferable. Either the 2 kb or 3 kb transcript observed in this Northern analysis may correspond to the Northern analysis result presented in 5.3.2.1. The size of 2.5 kb reported for this transcript was an estimate as accurate size markers were not utilised in the Northern blots described in 5.2.5.1.

5.3.5.2 yc81e09

Attempts were made to extend the sequence of clone yc81e09. The existing sequence comprised 1537 nucleotides, however, additional sequence was required as a transcript of approximately 3.7 kb was detected from Northern analysis. This Northern result was also confirmed by other members of the FAB consortium.

Exon trapping of cosmids c360D7 and c354E12, members of the yc81e09 contig (chapter 3), identified two potential exons (ET17.14 and ET17.12), mapping distal to the 5' end of the yc81e09 clone. RT-PCR was performed to join these exons. This procedure provided an additional 475 nucleotides at the 5' end of the yc81e09 sequence. A further 177 nucleotides from cDNA clone ze25b01 (GenBank Accession No. AA035673) were identified following searches of nucleotide databases using the BLAST-N algorithm. This clone provided a total of 398 nucleotides of which 221 overlapped the existing sequence. Thus a total of 2189 bases, presented in figure 5.7, have been obtained for this transcript. An additional 1.5 kb of sequence must be isolated to complete the sequence of this transcript. Searches of nucleotide and protein databases using BLAST-N and BLAST-X failed to reveal strong homologies with genes of known function, or to additional ESTs which could extend the sequence further.

5.4 DISCUSSION

Results from the sequencing, Northern analysis and RT-PCR experiments provide evidence that cDNA clones yh09a04 and yc81e09 are transcripts that represent two novel genes. These transcripts were demonstrated to map to the 16q24.3 region where the TSG and FAA gene have been localised, thus may be investigated as possible candidates for FAA and the TSG.

Expression studies demonstrated ubiquitous expression of the two transcripts. A high level of expression was observed for yc81e09, while a very low level of expression was demonstrated by yh09a04. Attempts made to isolate RNA from breast tissue were unsuccessful due to the high percentage of adipose tissue present in the sample. The study of expression in this tissue would have been advantageous to determine whether either of the transcripts were present in normal breast tissue.

Searches of the accessible non-redundant and EST nucleotide databases using BLAST-N provided some additional sequence information for both cDNA clones from sequences of cDNA clones deposited in these databases, but failed to reveal strong homologies with genes of known function. Analysis of the 5.5 kb transcript amino acid sequence identified protein similarities with the myosin 1B heavy chain and formin, the chicken limb deformity gene product but homology was not significant (The Fanconi Anaemia/Breast Cancer Consortium, 1996). The lack of homology of these transcripts to any other known gene is interesting and, as a consequence, the genes must encode novel proteins.

The open reading frame of the 5.5 kb transcript begins at position 45 and consists of 4,365 nucleotides which encode a protein of 1,455 amino acids. The 3' untranslated region of this transcript consists of 973 nucleotides and contains a polyadenylation signal beginning at nucleotide 4589. Comparison of the ET19/yf14a03/yh09a04 sequence with the 5.5 kb transcript sequence revealed deletions of the ET19/yf14a03/yh09a04 sequence in two separate sections. Clone yh09a04 contained a 23 bp deletion from position 4055-4077 in the

5.5 kb transcript (see figure 5.8). Clone yf14a03 contained a 139 bp deletion from position 3671-3811. A 4 bp insertion, GCAG, was observed in the yh09a04 clone. These deleted portions of DNA may correspond to exons that are spliced out of the mRNA sequence leading to the generation of alternatively spliced forms of the cDNA clones. The GCAG insertion in this sequence may be a result of incorrect splicing of RNA to mRNA at bases other than AG-GT.

Characterisation of the genomic organisation of this 5.5 kb transcript by the FAB consortium (Ianzano *et al.* in press) has determined that the 23 bp deletion corresponds to the 23 nucleotides at the 5' end of exon 41 and the GCAG insertion is located at the 3' portion of exon 41. Sequence analysis of exon 41 and its flanking regions identified cryptic donor and acceptor splice sites and this deletion is probably a result of the recognition of an alternative 5' acceptor sequence localised 23 bp from the start of exon 41. The GCAG insertion may be due to the use of the donor site AGgtgcaa as an alternative to the AGgcaggt site which results in the correct splicing of the intron. The presence of the 23 bp deletion and/or the 4 bp insertion create a frameshift and a downstream premature stop codon. The 139 bp deletion corresponds to an out of frame deletion of exon 37 (Ianzano *et al.* in press). It will be interesting to determine whether or not these alternatively spliced variants of the 5.5 kb transcript have a biological function, as this information may lead to insights into the role of the protein or proteins encoded by this gene.

The sequence of transcript yc81e09 is not complete, however, the longest continuous potential open reading frame (ORF) in the sequence, presented in figure 5.7, is 759 bp beginning at nucleotide 691. The sequences surrounding the proposed translation initiation site have a degree of homology to the Kozak consensus sequence (Kozak, 1986). The predicted 3' untranslated region comprises 740 bp which includes a possible polyadenylation signal preceding the poly(A) tail. This polyadenylation signal may lie in the sequence TATAAAT located at 2141-2147, 12 bp upstream of the poly(A) tract (see figure 5.7). It is possible that the sequence presented contains sequencing errors which may affect the size of the ORF. Additional sequence will provide information on the correct sized ORF.

The detailed characterisation of genes or transcribed sequences, mapped to regions where disease genes are localised by linkage analysis, is a necessary step for the positional cloning of disease genes of interest. Thus the detailed characterisation of cDNA clones yh09a04 and yc81e09 has provided useful information about these transcribed sequences which are potential candidates for the TSG or FAA gene. Although these sequences do not demonstrate significant homology with genes of known function, they should not be eliminated from further analysis. The TSG or FAA gene may be novel genes which encode proteins of unknown function. The next step in the characterisation of these transcripts and the positional cloning strategy is the mutation analysis of their sequence in FA-A and breast cancer patient samples, to determine if the sequence is altered when compared to normal DNA sequence.

CHAPTER 6

Mutation Analysis of Two Transcribed Sequences Localised to 16q24.3-qter

6.1 INTRODUCTION 215

6.1.1 Methods of Mutation Detection 215

6.1.1.1 Using Gel Electrophoresis 216

6.1.1.2 By Cleavage of DNA or RNA Molecules 220

6.1.2 Single Strand Conformation Polymorphism Analysis of Transcribed Sequences yh09a04/yf14a03 and yc81e09 223

6.1.3 Mutation Analysis of Genes and Transcribed Sequences Localised to 16q24.3-qter by the FAB Consortium 226

6.2 MATERIALS AND METHODS 228

6.2.1 Breast Tumour Samples 228

6.2.2 Isolation of RNA from Cells and Breast Tumour Cell Lines 228

6.2.3 Reverse Transcription Polymerase Chain Reaction (RT-PCR) 231

6.2.4 Amplification of cDNA Fragments for Use in Single Strand Conformation Polymorphism Analysis 235

6.2.5 Preparation of Samples for Single Strand Conformation Polymorphism Analysis 237

6.2.6 Single Strand Conformation Polymorphism Analysis and Electrophoresis 238

6.2.6.1 SSCP Analysis and Electrophoresis: Protocol I 238

6.2.6.2 SSCP Analysis and Electrophoresis: Protocol II 238

6.2.7 Preparation and Direct DNA Sequencing of PCR Products 239

6.2.8 Population Screening 239

6.3 RESULTS 241

6.3.1 Synthesis and Analysis of Large cDNA Fragments by RT-PCR 241

6.3.2 Single Strand Conformation Polymorphism Analysis of cDNA Fragments 243

6.3.3 Sequence Analysis of cDNA Fragments 245

6.3.4 Polymorphism Screening in DNA from Normal Individuals 251

6.4 DISCUSSION 256

6.4.1 Analysis of Other Candidate Tumour Suppressor Genes Localised to 16q24 262

6.1 INTRODUCTION

6.1.1 Methods of Mutation Detection

Disease causing mutations may be classified into two broad categories: those that cause a significant alteration in DNA structure (large deletions, insertions, inversions or duplications) and those that cause only a minimal alteration in DNA structure (small deletions, insertions, inversions, duplications and missense mutations). Methods for the identification of mutations that grossly alter DNA structure are well established. Cytogenetic analysis can easily identify mutations that grossly disturb chromosome structure. Techniques that assess genomic organisation, including PFGE, Southern analysis and restriction endonuclease mapping can be used to identify large deletions, insertions, inversions or duplications. Thus, it is advantageous to apply these techniques to gene segments prior to, or simultaneously with, the commencement of analyses for point mutations within a candidate gene.

Different strategies are required for the identification of mutations that cause only a minimal change in genomic structure. Identifying point mutations, small deletions, insertions, inversions and duplications in candidate genes is laborious. The sizes of candidate genes range from a few thousand to hundreds of thousands of base pairs in length. The coding sequence of these genes may be divided into many exons. Thus, the search for disease causing mutations in a large gene may require identification of a single nucleotide change among thousands of nucleotides.

A number of different strategies for screening candidate genes for mutations that alter only one or a few nucleotides have been developed. These methods for mutation detection can be divided into two groups. The first group of techniques can be used to efficiently identify known mutations. For example, population screening for carriers of the common cystic fibrosis mutations has been conducted by amplification of DNA encompassing a known mutation, followed by restriction enzyme digestion and analysis on polyacrylamide gels (Ng *et al.* 1991). The second group of methods involves scanning DNA sequences for unknown mutations, and will be described in this section. Although several useful strategies can be used to detect DNA sequence heterogeneity, no single strategy can be applied to all situations since each one does not detect 100% of the mutations. Choosing the most appropriate screening technique is influenced by the expected nature of the mutation, size and structure of the particular locus, availability of mRNA, degree of sensitivity required and available resources. Single base alterations are the most common type of mutation at most loci (Grompe, 1993). A number of techniques, which are described in this section, depend upon the amplification of candidate gene or cDNA segments by PCR and the subsequent characterisation of this amplified DNA to detect these subtle changes.

6.1.1.1 Using Gel Electrophoresis

Scanning methodologies to detect mutations by the conformational change they produce in DNA molecules include single strand conformation polymorphism (SSCP) analysis (Orita *et al.* 1989), heteroduplex analysis (HA) (White *et al.* 1992) and denaturing gradient gel electrophoresis (DGGE) (Myers *et al.* 1987). Each method relies on the detection of altered electrophoretic mobility of mutant DNA molecules compared to normal sequences. Some mutations alter the conformation of the DNA molecule and result in a dramatic mobility shift, but others produce little change. Change in conformation is determined in part by the nucleotide sequence flanking the mutation.

Due to its simplicity, capability of screening large numbers of samples, and relative sensitivity, SSCP is the most widely used of the scanning technologies. Regions of DNA are amplified by PCR, rendered single stranded by heating in a denaturing buffer, and fractionated on a non-denaturing polyacrylamide gel. The two single stranded DNA molecules from each denatured PCR product assumes a three-dimensional conformation, resulting from altered intrastrand base pairing. If a sequence difference exists between the normal and mutant DNA, this may alter the three-dimensional conformation and result in differential migration of the mutant strand. SSCP has been widely used to detect mutations in a variety of genes, for example p53 in astrocytoma (Fults *et al.* 1992) and CFTR in cystic fibrosis (Dean *et al.* 1990).

Most users of SSCP agree that 80-98% of mutations can be detected in fragments of the order of 200-300 bp if the analysis is performed using two different gel conditions (Forrest *et al.* 1995; Ravnik-Glavac *et al.* 1994; Sheffield *et al.* 1992; Grompe, 1993). Studies have shown that high percentage (10%) acrylamide gels improve resolution in SSCP analysis (Savov *et al.* 1992). The sensitivity of the method decreases with an increase in the size of the PCR product, and is less than 50% when fragments of > 400 bp are analysed. Analysis can become more efficient by the simultaneous investigation of several fragments in one lane (multiplexing) or by restriction digestion of large amplification products prior to electrophoresis (Iwahana *et al.* 1992). However, these techniques also have their limitations. The former technique can generate a complex patterns of bands that may be difficult to interpret. The partial digestion of DNA by restriction enzymes in the latter method may also cause problems with interpretation of results.

The technique of heteroduplex analysis (HA) (White *et al.* 1992) can be used to identify point mutations or single base polymorphisms in heterozygous individuals. This technique makes use of the fact that heteroduplex molecules that contain single base mismatches can be separated from nearly identical molecules that contain no mismatches under certain gel conditions. RNA or DNA from a potentially heterozygous individual is PCR amplified, denatured then allowed to renature to form heteroduplexes. These heteroduplexes are analysed on non-denaturing regular polyacrylamide gels (Nagamine *et al.* 1989) where hybrid molecules containing a mismatch migrate more slowly than their corresponding homoduplexes. New gel matrices, including mutation detection enhancement (MDE) (AT Biochem), considerably enhance the ability to detect mutation induced mobility shifts in heteroduplex molecules. This technique is simple and the detection of mutations is optimal, 80-90%, for PCR products of < 300 bp in size (White *et al.* 1992; Perry and Carrell, 1992).

Applications of the HA technique include the detection of mutations of the PAX-3 paired box gene in Waardenburg's syndrome (Tassebehji *et al.* 1992).

The separation principle of denaturing gradient gel electrophoresis (DGGE) (Myers *et al.* 1987) is based on the melting behaviour of the double helix of a given DNA fragment. This melting behaviour is sequence dependent and generally occurs in distinct domains, rather than at single bases. This is detected as a reduction in mobility of the DNA fragment as it moves through a parallel acrylamide gel containing a gradient of chemical denaturant (urea and formamide) with increasing concentration. DGGE can be used to detect minor sequence variations between two homoduplex DNA fragments. If homoduplexes cannot be distinguished on the gel, fragments from two different homoduplexes can be mixed to form heteroduplexes. Heteroduplex DNA is destabilised by a mismatch, and therefore melts at lower temperature than homoduplex molecules indicating that sequence variation exists between the two parent homoduplexes. The sensitivity of DGGE is greatly enhanced when heteroduplexes are analysed (Sheffield *et al.* 1989).

DGGE requires knowledge of nucleotide sequence or the melting profile and mutations can be found most reliably when the sequence heterogeneity lies within a domain of relatively low melting temperature. Theoretical melting profiles can be predicted with computer programs for virtually any sequence of interest (Lerman and Silverstein, 1987). The attachment of a GC-clamp to one primer pair has improved the number of mutations detected as this ensures the amplified sequence has a low dissociation temperature (Sheffield *et al.* 1989; Myers *et al.* 1985). The sizes of the PCR products analysed by DGGE should be up to 600 bp in length to detect 95% of single base differences in PCR products. DGGE has been used to detect mutations in a number of genes including β-thalassaemia (Losekoot *et al.* 1990) and factor VII deficiency (Higuchi *et al.* 1991).

Constant denaturant gel electrophoresis (CDGE) (Hovig *et al.* 1991) is a modification of DGGE which uses a single denaturant concentration determined to be optimal for the analysis of a particular fragment. It facilitates the analysis of specific melting domains and produces improved separation between homoduplexes. These gels can be used to screen a number of individuals for the presence of a mutation. Between 90-100% of all possible mutations in the area screened can be reliably detected using the DGGE and CDGE methods. However, the exact location of the mutation within the DNA fragment is uncertain because several domains are being screened at once.

SSCP, HA and DGGE are relatively simple techniques. The same primers can be used in heteroduplex and SSCP analyses but DGGE analysis requires specific primer construction. In general, these techniques are interchangeable and can be used to readily screen large numbers of samples. Each technique differs in its sensitivity for the detection of a mutation in a given DNA segment. The sensitivities vary with the length of the DNA segment being analysed and the gel conditions used to visualise the result. The most efficient mutation detection rate for the SSCP and HA techniques is observed in DNA fragments of 200-300 bp. The DGGE and CDGE methods can detect 90-95% of mutations in DNA fragments up to 600 bp in length. Thus, the DGGE technique is advantageous over the SSCP and HA techniques as it does not have their size limitations. Since larger DNA fragments can be used for mutation detection by DGGE, a smaller number of PCR reactions are required to screen a given DNA segment of interest.

HA is only appropriate for individuals heterozygous for a DNA polymorphism caused by a single base change and for identification of mutations in individuals with an autosomal dominant disease (Cotton, 1992). This method also requires that the mutation be previously identified and that sequence information is available. The SSCP method has a broader applicability in comparison to the HA method as previous mutation information is not required and homozygous and heterozygous individuals can be screened.

The DGGE technique requires knowledge of the nucleotide sequence together with the melting profile. Additionally, mutation detection is most efficient if the sequence heterogeneity lies within a domain of low melting temperature. If the nucleotide sequence is unknown, perpendicular DGGE can be performed using a broad range of denaturant (0-100%) to visualise the melting profile of a given DNA fragment. Suitable denaturation conditions for subsequent analyses based on the melting profile are then selected. The SSCP method does not require melting profile information or the synthesis of high melting temperature GC clamps, thus has advantages over the DGGE technique.

6.1.1.2 By Cleavage of DNA or RNA Molecules

Other scanning strategies rely on the cleavage of RNA or DNA molecules prior to analysis. In these techniques, heteroduplex molecules are specifically cleaved at sites of mismatched base pairs resulting from point mutations, using either ribonuclease A (RNase A) (Gibbs and Caskey, 1987) or chemical mismatch cleavage (CMC) (Cotton *et al.* 1988). RNase A is specific for single stranded RNA molecules and acts by cleaving at points of certain sequence mismatch in RNA-RNA or RNA-DNA hybrids (Myers *et al.* 1985). The RNA-DNA heteroduplexes are formed between a radioactive normal riboprobe and mutant DNA generated by PCR and mismatches are cut by RNase A. The products are analysed by electrophoresis and the presence and location of a mutation is indicated by a distinct cleavage band. RNase A has had a number of applications including the identification of mutations in the *apc* gene (Nishisho *et al.* 1991). However, this method has limitations as RNase A detects only 50-60% of mismatches (Myers *et al.* 1985). This is because RNase A is unable to cleave certain mismatches including G/A, G/T, G/G, A/A and A/C due to the effects of the adjacent sequences.

Mutation screening by CMC (Cotton *et al.* 1988) relies on the differential reactivity of mismatched cytosine (C) and thymidine (T) with the compounds hydroxylamine and osmium tetroxide, respectively, versus matched cytosine and thymidine bases. In this technique, unlabelled mutant DNA is hybridised with labelled normal DNA to form heteroduplexes. One aliquot is treated with hydroxylamine, and another aliquot is treated with osmium tetroxide, which react with mismatched C and T bases, respectively. Piperidine, which cleaves the strands at the sites where hydroxylamine and osmium tetroxide react, is then added. Cleaved fragments are analysed on denaturing polyacrylamide gels to identify the point of cleavage. Mismatched C and T bases are not directly detected but are transposed with a probe of opposite sense to the mismatched C and T bases respectively. This method can also detect mutations other than point mutations, as mutations can destabilise the double stranded structure of heteroduplexes not only at the site of mismatches, but also in the direct vicinity of the change. This method has been used for mutation detection in a number of genes including the BCL2 oncogene (Tanaka *et al.* 1992) and dystrophin (Roberts *et al.* 1992).

The techniques of RNase A cleavage and CMC are technically demanding and require more effort than methods based on gel mobility shifts. The use of the toxic chemical osmium tetroxide which requires use of a fume hood for part of the CMC procedure is also a disadvantage. The RNase A protocol is also limited as it cleaves only certain mismatches. Significant advantages of DNA cleavage methods are that large DNA fragments longer than 1 kb, up to 1.7 kb (Gibbs and Caskey, 1987; Zheng *et al.* 1991), can be scanned. The precise localisation is indicated by the size of the cleavage band and the nature of the change is also shown by the cleavage reagent, in the case of CMC. Chemical cleavage also possesses a very high mutation detection rate. It can detect > 95% of mismatches when only normal DNA is labelled and 100% of mutations when both normal and mutant DNAs are labelled (Cotton *et al.* 1988; Forrest *et al.* 1991).

Recently, enzymatic methods using either mismatch repair enzymes or resolvases have been developed and offer great promise for analysis of large fragments of DNA (Mashal *et al.* 1995; Youil *et al.* 1995). Bacteriophage resolvases recognise and cleave mismatched bases in radiolabelled heteroduplex double stranded DNA and digestion is monitored on a gel. Enzymes used in this technique are T4 endonuclease VII and T7 endonuclease I. The T4 endonuclease VII and T7 endonuclease I enzymes have been used to detect known heterozygous mutations in one of the following genes, familial adenomatous polyposis coli, p53 and the cystic fibrosis transmembrane receptor (Mashal *et al.* 1995). Youil *et al* (1995) studied the ability of the T4 enzyme to detect mismatched bases, rather than its ability to cleave specific mismatches. Both studies demonstrated a high sensitivity of detection of around 94%.

The enzyme mismatch cleavage (EMC) technique detects mismatches in individuals who are heterozygous at a given site and has the potential to be as easy and inexpensive as SSCP and as sensitive as DGGE. Added advantages are that it is applicable to fragments 1 kb or larger and the presence and estimated position of an alteration is revealed. The disadvantages of the technique include the tedious purification of DNA fragments before subjecting them to resolvase cleavage and the fact that a commercial source of the enzyme is not yet available. Homozygous mutant samples escape detection but this may be overcome by the addition of normal DNA. Another difficulty is that some mutations are poorly recognised by resolvases, resulting in digestion of only a small fraction of the DNA. Thus, several improvements need to be made, including a commercial source of resolvases, before EMC replaces the established methods.

All these methods, other than CMC, are capable of detecting mutations with varying efficiencies, but none precisely defines the nature of the change. Some sequence changes, for example those causing frame-shifts or chain terminations, are clearly functionally deleterious to a protein. Other alterations such as silent mutations leading to amino acid substitutions or base changes in introns and untranslated regions, can be either functionally relevant or represent sequences which are found in the normal population and can be considered human polymorphisms. DNA sequencing provides information on the biological significance of the alterations therefore, is a necessary final step of any mutation detection method. With the improvement of DNA sequencing methodologies it is becoming possible to rapidly screen directly for mutations using DNA sequencing approaches. Direct sequencing constitutes the analysis of PCR products of DNA segments containing potential mutations without prior subcloning. Additional information including family studies, population studies or functional assays after *in vitro* expression may be needed to determine the significance of any sequence alteration.

While all these methods have been used successfully for identification of disease causing alleles, SSCP (because of its simplicity and economy) and DDGE (because of its near 100% sensitivity) are the favoured techniques (Dean, 1995). As mentioned, the sensitivity of the different mutation detection procedures is dependent on the size of the fragment to be analysed. Even when scanning small DNA fragments, SSCP and HA are not as sensitive as DGGE, CMC and direct sequencing. Sequencing and CMC are the only methods that accurately localise the mutation within the examined fragment. In all procedures described, the detection methodology can be either radioactive or non-radioactive. It is possible to carry out SSCP, DGGE and HA non-radioactively with ethidium bromide staining of DNA. CMC and DNA sequencing require fluorescence detection. Of all the techniques clearly CMC has the most problems with the use of toxic chemicals.

Although many methods of mutation detection have been proposed and new novel procedures have been published, for example enzyme cleavage, these methods do not have the track record, technical simplicity and reliability of SSCP. Thus, SSCP remains a popular choice for screening DNA mutations in genes, but novel techniques, such as EMC, may be attempted particularly for variants that are not easily detected by SSCP.

6.1.2 Single Strand Conformation Polymorphism Analysis of Transcribed Sequences yh09a04/yf14a03 and yc81e09

The work presented in this chapter is the mutation analysis of the transcribed sequences yh09a04/yf14a03 and yc81e09 (see chapter 5) to determine whether either transcript represents the TSG involved in LOH at 16q24.3-qter in sporadic breast cancer. Single strand conformation polymorphism (SSCP) analysis was used exclusively in the analysis of these transcribed sequences. When SSCP analysis for sequence yh09a04/yf14a03 in breast tumour samples was initiated, the full 5.5 kb sequence of the transcript and its ORF had not been determined. The full sequence of the 5.5 kb transcript, which includes yh09a04/yf14a03, was obtained in June, 1996. As a consequence of obtaining the full-length coding sequence and determination of its open reading frame (ORF), SSCP mutation analysis was performed on the 4,365 bp ORF of the 5.5 kb transcript. A small segment of clone yc81e09 (see chapter 5) was also analysed by SSCP.

Since only the coding sequence of the two transcripts was known, mutation analysis was based wholly on the reverse transcription of RNA to cDNA. Mutation analysis was dependent on obtaining RNA from breast tumour samples displaying restricted LOH at 16q24.3-qter. LOH analysis of DNA from breast tumour and normal tissue from individual patients was performed by Dr. Anne-Marie Cleton-Jansen (Department of Pathology, University of Leiden, The Netherlands) to determine which tumour samples displayed LOH in the 16q24.3-qter region of interest. First strand cDNA synthesis by RT-PCR involved the synthesis of mRNA and random primed RNA from breast tumour RNA, in separate reactions. Primers specific for each cDNA clone were designed for amplification of the second cDNA strand from the breast tumour RNA.

The ORF of the 5.5 kb transcript was divided into two segments for SSCP analysis due to its large size. Mutation analysis of the 5' end of the ORF, comprising bases 1-1500, was performed by Dr. A-M. Cleton-Jansen. Ms. Joanna Crawford and myself performed mutation analysis of the 3' end of the ORF of the 5.5 kb transcript, comprising bases 1369-4447. The numbers correspond to those indicated in figure 5.8. Bases 975-1489 of transcript yc81e09 were only analysed, due to time constraints. The numbers correspond to those indicated in figure 5.7. Primers for second strand cDNA synthesis were designed to synthesise one large segment of the yc81e09 clone and three large overlapping segments of the 3' end of the 5.5 kb (yh09a04/yf14a03) cDNA.

SSCP analysis was confined to the coding sequence of the transcripts as mutations are more likely to be found in this region. Analysis of this sequence also avoids encountering the frequent polymorphisms found in introns (Orita *et al.* 1989; Sheffield *et al.* 1993). The sequence from the 3' untranslated region of a transcript frequently does not contain introns, but shows significantly higher polymorphisms between species and within different members of a family, therefore can create interpretational problems in mutation analysis if investigated.

Initially the sizes of the PCR products from a panel of seven LOH breast tumour samples that showed restricted LOH at 16q24.3-qter, were compared to PCR products from peripheral blood lymphocytes (PBLs) from a normal individual on agarose gels. This was performed to determine whether gross deletions were present in the tumour samples. Probing Southern blots of these RT-PCR products with the radiolabelled cDNA probe was also necessary to eliminate possible artefacts of the RT-PCR. The absence of a band, or detection of a band of reduced size in tumour samples, but its presence in normal PBL, indicates a likely candidate. Lack of, or a reduced gene transcript can be attributed to gene deletion. No detection of differences in band sizes leads to the search for point mutations or deletions in each transcript.

SSCP analysis performed for this study involved the use of the RT-PCR products from normal PBL and breast tumours as templates to amplify overlapping fragments of approximately 300 bp in size with the incorporation of radionucleotide label. Analysis of these cDNA PCR products was achieved with two gel systems, 10% acrylamide gels and MDE gels, followed by autoradiography. These gel systems have been widely used within the Department of Cytogenetics and Molecular Genetics and have been shown to be successful in mutation screening of various genes (Hayashi, 1991; Orita *et al.* 1989; H. Phillips and G. Hollway, personal communication). Any differences in migration of bands seen in SSCP were investigated further. Firstly, the procedure was repeated (artefactual bands are found) to see if the difference could be confirmed. The amplified PCR products from confirmed changes were sequenced to identify the mutational change.

SSCP analysis can be applied to both genomic DNA or cDNA samples. The advantages of using RT-PCR to obtain DNA sequences from tumour and normal samples for SSCP analysis are that long segments of peptide coding region, can be examined. Information on gene structure, intron/exon boundaries, is not required. Fewer PCR reactions are performed because more sequence can be analysed in one reaction. Analysis of genomic DNA sequence is achieved by amplification of individual exons. It is desirable to position primers close to the intron/exon boundary to minimise the inclusion of intronic sequences as they are more likely to contain non-coding polymorphisms.

The disadvantages of using RT-PCR are that the gene may not be expressed in the tissue which is accessible. Mutations in the promoter and intronic splice junctions, which are not part of the coding sequence, cannot be detected. Also, ribonuclease contamination may occur in RNA samples used for RT-PCR despite the precautions taken. This contamination may lead to spurious SSCP results. The disadvantages of analysing genomic DNA sequence are that knowledge of the genomic organisation of the gene is required. Also, if the gene contains many exons, this procedure can be tedious thus, the analysis of cDNA is advantageous as longer segments of DNA can be analysed rapidly.

6.1.3 Mutation Analysis of Genes and Transcribed Sequences Localised to 16q24.3-qter by the FAB Consortium

Additional genes and transcripts localised to 16q24.3-qter were analysed by other members of the FAB consortium to determine their involvement as the TSG in sporadic breast tumours with LOH at 16q24.3-qter. Further experiments conducted by members of the FAB consortium involved mutation detection experiments to determine whether transcripts yh09a04/yf14a03 and yc81e09, together with other genes and transcribed sequences localised to 16q24.3-qter, were mutated in Fanconi anaemia-A (FA-A) patients.

The 5.5 kb transcript, which includes yh09a04/yf14a03, was found to be partially deleted in RT-PCR products from an Italian patient with FA-A, by the FAB consortium. This cDNA was then investigated in more detail as a candidate for the FAA gene (The Fanconi Anaemia/Breast Cancer Consortium, 1996). Different mutations were demonstrated in this sequence in more FA-A patient samples and the 5.5 kb transcript was identified as the FAA gene. These results will be presented in more detail in section 6.4.

6.2 MATERIALS AND METHODS

6.2.1 Breast Tumour Samples

Total RNA from breast tissue displaying LOH at 16q24.3 was kindly provided by Dr. Anne-Marie Cleton-Jansen, Department of Pathology, University of Leiden, The Netherlands. The characterisation of these breast tumour samples for LOH with microsatellite markers was performed by Dr. Cleton-Jansen. The results are shown in table 6.1. The location of the markers used for analysis of the breast tumours are shown in figure 6.1. These tumour samples contained contaminating normal cells thus, Dr. Cleton-Jansen applied a flow sorting technique based on DNA ploidy differences and/or immunocytochemical cell lineage markers (keratin, vimentin) to enrich the tumour cell fractions.

6.2.2 Isolation of RNA from Cells and Breast Tumour Cell Lines

Cells were trypsinised from 75 cm^3 or 150 cm^3 flasks and washed three times in PBS (see 2.2.4). Total RNA from the breast tumour cell lines, MB157, MCF-7, ZR-75, T47D/110, MoA-231/51, was extracted using TRIzol (GIBCO BRL). RNA was also extracted from peripheral blood lymphocytes (PBL) from normal individuals in the laboratory after they were separated from whole blood using ficol-hypaque and washed three times in PBS. Cells were resuspended in 1 ml of TRIzol /10^7 cells. The procedure outlined in 5.2.4.3 was followed to isolate RNA. The RNA was precipitated and quantitated, as described in 5.2.4.4. A 1 µl aliquot of RNA was analysed on 1% agarose gel (2.5.2) to assess the integrity of the 18S and 28S ribosomal RNA species prior to use.

Table 6.1

Breast tumour samples used for SSCP mutation analysis. These breast tumours were characterised for loss of heterozygosity using microsatellite markers and STS markers localised to 16q24. This characterisation was performed by Dr. Anne-Marie Cleton-Jansen, Department of Pathology, University of Leiden, The Netherlands.

Tumour	S289	S449	S305	S1320	S413	gata 71F09	S7	S3023	APRT	S3026	S3121	CMAR	S303
BT 355	R	R	R	N	L	L	N	L	N			N	L
BT 413	R	R	R	R	L	N	L	L	N	L	L	N	L
BT 541		L	L	L	L	L	L	N	L	N	L	N	N
BT 555	R	R	N	L	L	N	L	N	N			L	L
BT 559	R	N	R	N	N	R	R	R	R	R	N	L	L
BT 819	N	R	R	R	R	N		N	N	N	L	N	L
BT 919	R	R	L	N	L	L			N				L

L = loss of heterozygosity

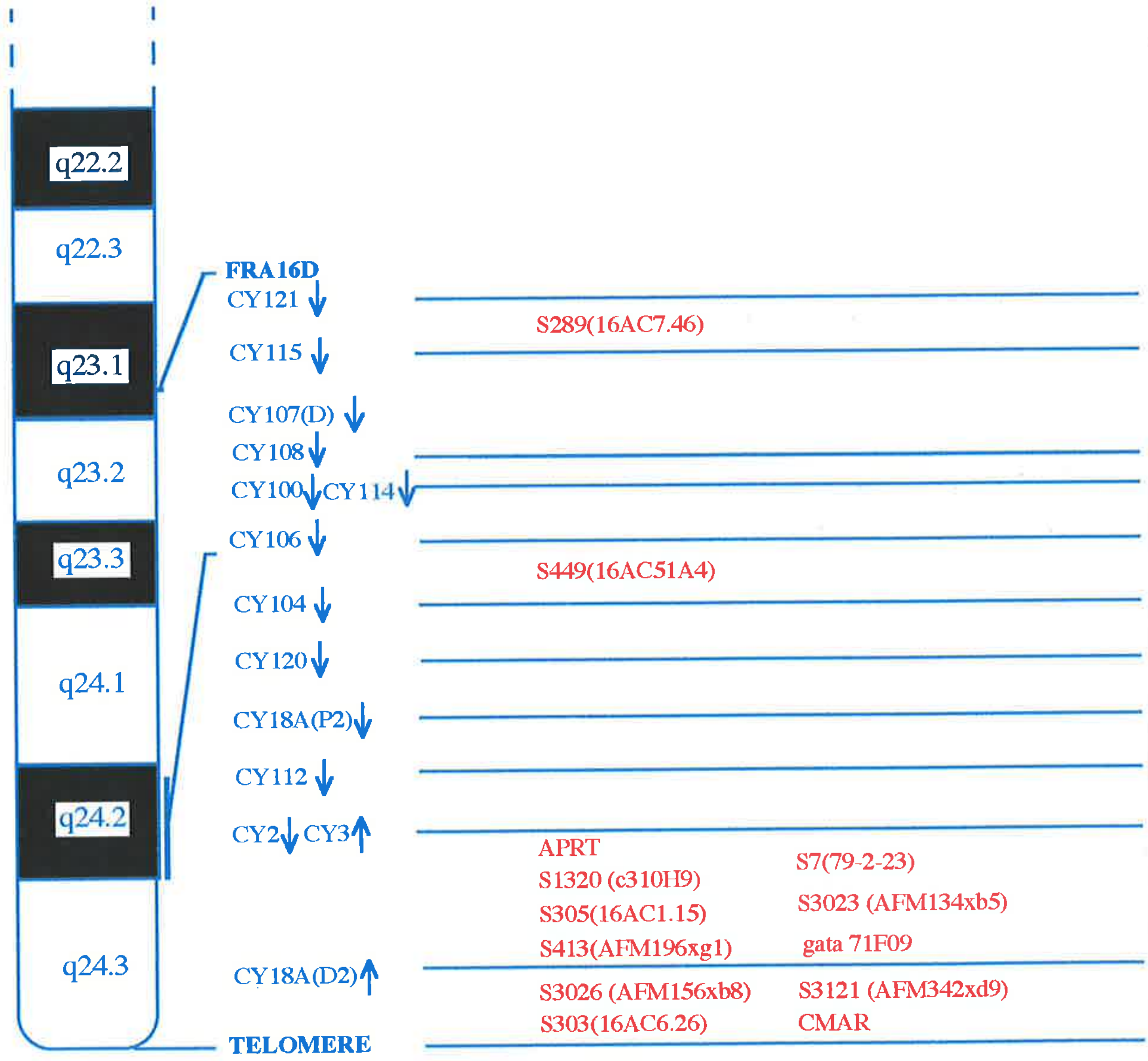
R = retention

N = non-informative

empty box = not tested or not interpretable

Figure 6.1

Ideogram of chromosome 16 displaying the localisation of markers and genes used to characterise the smallest region of LOH in breast tumours.


q22.2
q22.3
q23.1
q23.2
q23.3
q24.1
q24.2
q24.3
FRA16D
CY121
S289(16AC7.46)
CY115
CY107(D)
CY108
CY100
CY114
CY106
S449(16AC51A4)
CY104
CY120
CY18A(P2)
CY112
CY2
CY3
APRT
S1320 (c310H9)
S305(16AC1.15)
S413(AFM196xg1)
S7(79-2-23)
S3023 (AFM134xb5)
gata 71F09
CY18A(D2)
S3026 (AFM156xb8)
S303(16AC6.26)
S3121 (AFM342xd9)
CMAR
TELOMERE

6.2.3 Reverse Transcription Polymerase Chain Reaction (RT-PCR)

Primers located at the 3' end of the ORF of the 5.5 kb transcript, comprising bases 1369-4447, (the numbers correspond to those indicated in figure 5.8) were designed for second strand cDNA synthesis in RT-PCR. This segment of cDNA was divided into three large overlapping fragments. Primers FA5 + FA2 span bases 1369-2131 (fragment 1), FA3 + FA6 span bases 2061-2890 (fragment 2) and FA14 + FA24 span bases 2827-4447 (fragment 3). Only a small segment of clone yc81e09, bases 975-1489 amplified with primers yc81F1 + yc81R1 (fragment 4), was included for analysis, due to time constraints (the numbers correspond to those indicated in figure 5.7). The sequences and locations of these primers are listed in table 6.2. Figure 6.2 depicts the locations of the overlapping PCR products in the cDNA clone from which they are derived. All primers were synthesised and deprotected as described in sections 2.6.1 and 2.6.2.

The conditions for first strand cDNA synthesis using the breast tumour samples displaying LOH at 16q24.3 were determined by examining a range of concentrations of starting RNA, from 1 µg to 5 µg. The amount of this RNA was very limited, thus it was important to establish conditions using the least amount of material possible. First strand cDNA from normal PBL and breast tumour cell lines, and DNA from cDNA clones was used to determine the PCR conditions for the synthesis of second strand cDNA, as the breast tumour RNA was limited. These PCR conditions were established according to the melting temperature (T_m) of each primer pair and the sizes of the PCR products. These primers and their positions are listed in table 6.2.

First strand cDNA synthesis was achieved with 2.5 µg of total RNA extracted from normal PBL and each cell line sample. One to two µg of total RNA extracted from breast tumour samples was used. Two reactions for each sample were performed. One reaction consisted of RNA primed with oligo-dT primer, and the other with random hexamers. One µl of oligo-dT primer at 500 ng/µl (GIBCO BRL) or 1 µl of random hexamers (500 ng/µl - GIBCO BRL)

Table 6.2

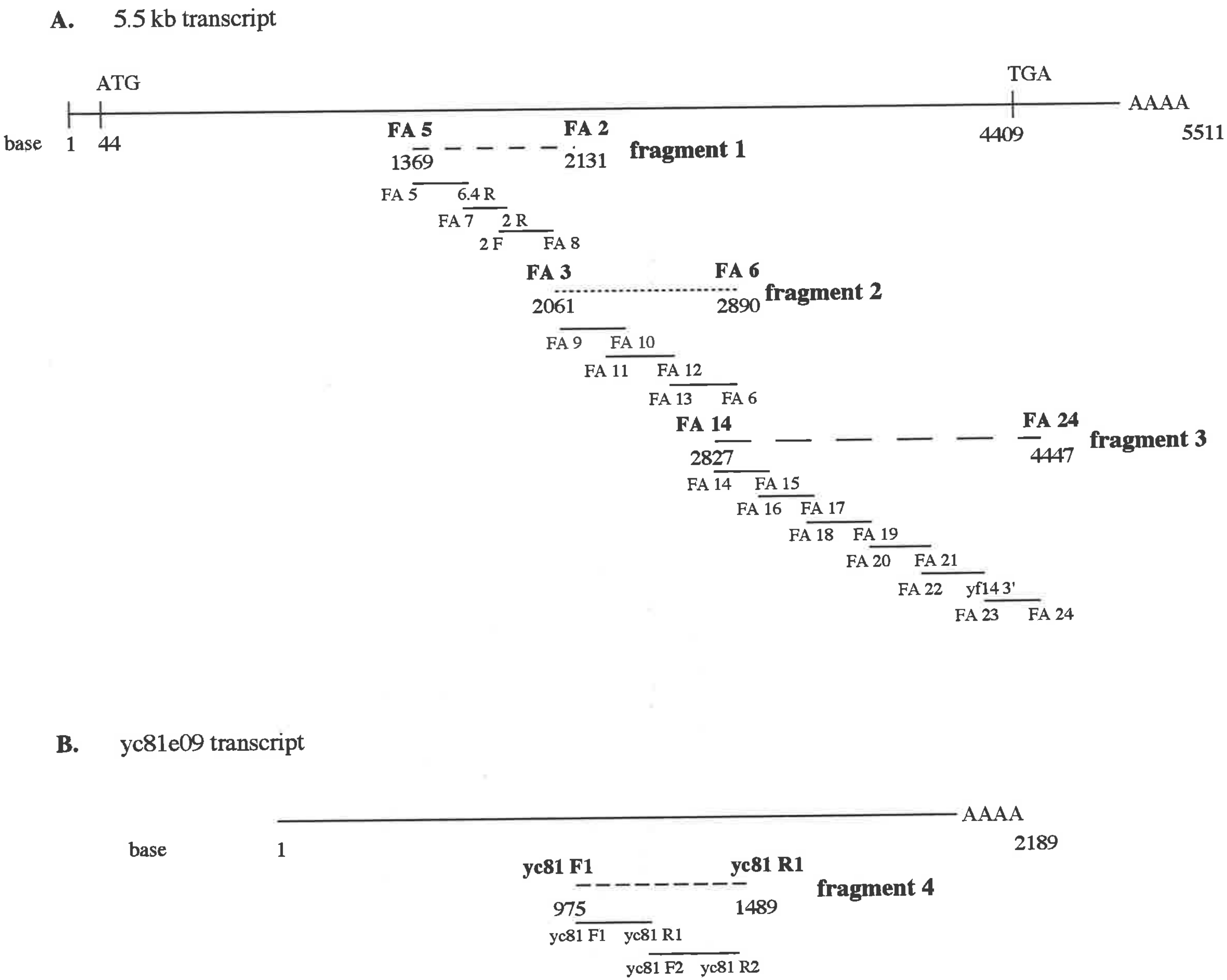
The sequences and locations of primers used for second strand cDNA synthesis in RT-PCR. The 5.5 kb transcript was divided into three segments, fragments 1 amplified with FA5 + FA2; fragment 2 amplified with FA3 + FA6; and fragment 3 amplified with FA14 + FA24 . One segment of clone yc81e09, fragment 4, was amplified with yc81F1 + yc81R1.

cDNA Clone	**Primer Name**	**Primer Sequence (5' - 3')**	**Primer Position**
5.5 kb transcript (yh09a04/yf14a03)	FA5	CCT CTG CGT TCC TGT CAT AT	1369 - 1388
5.5 kb transcript (yh09a04/yf14a03)	FA2	CTG CTG TCA TCC TCA TTG TGG C	2131 - 2110
5.5 kb transcript (yh09a04/yf14a03)	FA3	ATA TCG GCA CAG GTG GCA GTG A	2061 - 2082
5.5 kb transcript (yh09a04/yf14a03)	FA6	GTA TCT GAA AGA GCA TCA GCT	2890 - 2870
5.5 kb transcript (yh09a04/yf14a03)	FA14	CTT ACC AAG ACT GGT TAC ACC T	2827 - 2848
5.5 kb transcript (yh09a04/yf14a03)	FA24	ACG GGA GCT GGG CTG GTG TGC AGT	4447 - 4424
yc81e09	yc81F1	CTG CCG GCT GGA TTA CCG CAG	975 - 995
yc81e09	yc81R2	CAG GTT CAC CAG CTG GCT CAG G	1489 - 1468

Figure 6.2

A. Locations of the overlapping PCR products corresponding to fragments 1, 2 and 3 in the 5.5 kb transcript. The primers used to synthesise nested PCR products from each of the fragments are also indicated. The location of each primers is listed in table 6.3.

B. Location of the PCR product of fragment 4 with respect to the existing cDNA sequence of clone yc81e09. The primers used to synthesise nested PCR products from each of the fragments are also indicated. The location of each primers is listed in table 6.3.


A.
5.5 kb transcript
ATG
TGA
AAAA
base
1
44
4409
5511
FA 5
FA 2
1369
2131
fragment 1
FA 5
6.4 R
FA 7
2 R
2 F
FA 8
FA 3
FA 6
2061
2890
fragment 2
FA 9
FA 10
FA 11
FA 12
FA 13
FA 6
FA 14
FA 24
2827
4447
fragment 3
FA 14
FA 15
FA 16
FA 17
FA 18
FA 19
FA 20
FA 21
FA 22
yf14 3'
FA 23
FA 24
B.
yc81e09 transcript
AAAA
base
1
2189
yc81 F1
yc81 R1
975
1489
fragment 4
yc81 F1
yc81 R1
yc81 F2
yc81 R2

was added to DEPC-treated water in a total volume of 20 µl. This was heated to 70°C for 10 minutes then placed on ice. Four µl of 5 x first strand buffer (GIBCO BRL), 2 µl of 0.1M dithiothreitol and 1 µl of 10 mM dNTP mix (dATP, dTTP, dCTP, dGTP at neutral pH) were added, mixed and warmed to 37°C. One µl (400U) of Superscript reverse transcriptase (GIBCO BRL) was added and the reaction was incubated at 37°C for 1 hour followed by inactivation of the enzyme at 95°C for 5 minutes. Twenty µl of sterile water was added to each reaction.

Second strand cDNA was synthesised by PCR (see 2.6) under the following conditions. Three µl of first strand mixture and 150 ng of each primer specific to the sequence of the yc81e09 or 5.5 kb cDNA clones were added in a total volume of 30 µl. After an initial denaturation at 94°C for 5 minutes, 35 cycles were carried out with the following parameters: denaturation at 94°C for 1 minute, annealing at 58°C for 2 minutes and extension at 72°C for 3 minutes. A 10 minute extension at 72°C was performed at the end of the last cycle. The combinations of primers utilised in these reactions are as follows: FA5 + FA2, FA3 + FA6, FA14 + FA24 and yc81F1 + yc81R1. DNA from each cDNA clone was included as a positive control to determine whether PCR product was produced for each experiment and whether it was of the correct size. A control of no DNA was also included. The annealing temperatures of the primers specific to yc81e09 and the 5.5 kb transcript were compatible, allowing the experiments to be performed simultaneously in the one thermal cycler.

Ten µl aliquots of the PCR products were electrophoresed on 1% agarose gels (2.5.2) then Southern blotted (2.5.5.3). Each membrane was hybridised (2.5.6.1) with its corresponding labelled cDNA insert (2.5.4.1) to determine whether RT-PCR artefacts were present. Also, the sizes of the PCR products derived from the breast tumours and breast tumour cell lines were compared with the size of the product from normal PBL and their respective cDNA clone, for each primer pair.

6.2.4 Amplification of cDNA Fragments for Use in Single Strand Conformation Polymorphism Analysis

Primer pairs, spaced at approximately 200-400 bp intervals, with > 20 bp overlap at the 5' and 3' ends, were designed for the amplification of segments of cDNA from fragments 1, 2, 3 and 4 produced by RT-PCR (see 6.2.3). The primers were positioned with sufficient overlap to simplify the detection of mutations near the end of each segment, because point mutations within primer sequences are typically not detected after PCR. Figure 6.2 depicts the primers used to synthesise the overlapping nested PCR products from the cDNA fragment from which they are derived. The sequence and locations of these primers are listed in table 6.3. The primer pairs utilised in these reactions are as follows: Fragment 1 as template:- FA5 + 6.4R, FA7 + 2R, and 2F + FA8; Fragment 2 as template:- FA9 + FA10, FA11 + FA12, and FA13 + FA6; Fragment 3 as template:- FA14 + FA15, FA16 + FA17, FA18 + FA19, FA20 + FA21, FA22 + yf143', and FA23 + FA24; Fragment 4 as template:- yc81F1 + yc81R1, and yc81F2 + yc81R2.

To synthesise the various cDNA fragments, PCR conditions were established for each primer pair. For simplicity, the aim was to use the same conditions for all primer pairs. PCR products were visualised on 1% agarose gels to determine if the PCR conditions were optimal and whether the correct sized products were synthesised. The sizes of products from breast tumours were compared with those from breast cell lines, normal PBL and cDNA clones.

The autoradiographs of membranes containing the RT-PCR products (6.2.3) probed with cDNA were used to visually quantify the amount of cDNA present in each tube, according to the band intensity. In general, 1-2.5 µl of a 1/10 dilution of this PCR product was utilised in a PCR reaction to amplify products of approximately 200-400 bp. PCR (see 2.6) was performed under the following conditions. One to 2.5 µl of a 1/10 dilution of RT-PCR product was added to 75 ng of each primer in a total volume of 10 µl. For SSCP analysis 0.2

Table 6.3

Sequence and locations of primers used to amplify segments of cDNA from fragments 1, 2, 3 and 4 produced by RT-PCR. The products were analysed by SSCP to determine whether their sequences contained mutations.

Fragment No.	Primer Name	Primer Sequence (5' - 3')	Primer Position
1	FA5	CCT CTG CGT TCC TGT CAT AT	1369 - 1388
1	6.4R	CTG GCC TCC ATG ACG GTG AC	1759 - 1740
1	FA7	AAG GCC ATC ATG GTG TTT GAG C	1701 - 1722
1	2R	CTG GCA GTA GGT GGA GTA C	1913 - 1895
1	2F	ATC TCT GTA CTC CAC CTA CT	1889 - 1908
1	FA8	AGG ACA GCC CTC AGT CTT TCA GA	2107 - 2085
2	FA9	TGA AAG ACT GAG GGC TGT CCT	2087 - 2108
2	FA10	TGC ACG GCC AGG GCA GCC AA	2413 - 2394
2	FA11	ACC ATG TGT GGA CGT GTG CTC C	2298 - 2319
2	FA12	CTG AAC ATG AGG AAC TGA AAC	2665 - 2645

Fragment No.	Primer Name	Primer Sequence (5' - 3')	Primer Position
2	FA13	CAG TCA CGA GAT ACT TTG TGC A	2595 - 2616
2	FA6	GTA TCT GAA AGA GCA TCA GCT	2890 - 2870
3	FA14	CTT ACC AAG ACT GGT TAC ACC T	2827 - 2848
3	FA15	ACT ATG AGG TCT TGC TGC AGC T	3151 - 3130
3	FA16	TTG CAG GAG ATG GTA GCT GAC	3105 - 3125
3	FA17	ACC TCC GTG GGA GCA GAA GTT T	3395 - 3316
3	FA18	ATC ACT GCC AGA TGC GAG CAG T	3245 - 3366
3	FA19	AAA CTG CCG GCC TTC TTG TAG CT	3662 - 3640
3	FA20	TGC CGG TGG AGG AGA CAC TG	3591 - 3610
3	FA21	CTT CTC TAA ACA CTC GAG GAT T	3941 - 3914

Fragment No.	Primer Name	Primer Sequence (5' - 3')	Primer Position
3	FA22	TAG CAC CAC AGA CCT GCC AAA G	3871 - 3894
3	yf143'	ACA TGT CCA CAG CAA CAT GCA G	4125 - 4104
3	FA23	ATG AAG ACG CGG CCA TCA GGG AA	4072 - 4094
3	FA24	ACG GGA GCT GGG CTG GTG TGC AGT	4447 - 4424
4	yc81F1	CTG CCG GCT GGA TTA CCG CAG	975 - 995
4	yc81R1	ATT AGG GAG CTG GGA CAG GTT C	1236 - 1215
4	yc81F2	TAC GAG TAC CTG ATC CGC CTC T	1171 - 1192
4	yc81R2	CAG GTT CAC CAG CTG GCT CAG G	1489 - 1468

μl α^{32}P-dCTP (800 Ci/mmol) radionucleotide label was added to each reaction. Ten cycles were carried out with the following parameters: denaturation at 94°C for 1 minute, annealing at 60°C for 1.5 minutes and extension at 72°C for 1.5 minutes. A further 25 cycles were carried out with the following parameters: denaturation at 94°C for 1 minute, annealing at 55°C for 1.5 minutes and extension at 72°C for 1.5 minutes. A 10 minute extension at 72°C was performed at the end of the last cycle.

6.2.5 Preparation of Samples for Single Strand Conformation Polymorphism Analysis

Only breast tumour samples were used for SSCP analysis. With each primer pair utilised for the synthesis of a section of each cDNA, a PBL sample from a normal individual was routinely included as a control, and loaded into the first well of each gel. Samples were analysed on 10% acrylamide gels first. Any band shift observed was not considered indicative of a potential mutation or polymorphism unless it was reproducible on at least two separate occasions.

An equal volume of 10 x formamide loading buffer (96% deionised formamide, 20 mM EDTA, 0.1% xylene cyanol and 0.1% bromophenol blue) was added to each sample. The samples were centrifuged, then heat-denatured at 94-96°C for 5 minutes, placed immediately onto ice to prevent re-annealing of the single strands, and 3 μl loaded immediately into the wells. To avoid the re-annealing of single strands while other samples were being loaded into wells, the denaturing and loading of samples was performed in sets of eight.

6.2.6 Single Strand Conformation Polymorphism Analysis and Electrophoresis

The PCR fragments were analysed under two different gel systems to maximise the chance of detecting a band shift (Hayashi, 1991; Orita *et al.* 1989). Depending on the number of samples prepared, small or large gels were used. The dimension of the small gels was 50 cm x 21 cm x 4 mm and of the large gels, 50 cm x 37.5 cm x 4 mm. Contributions to the electrophoresis were made by Ms. J. Crawford.

6.2.6.1 SSCP Analysis and Electrophoresis: Protocol I

The samples were loaded onto a 10% non-denaturing gel containing: 24.5 ml of 40% acrylamide : 10 ml 2% bis-acrylamide (49:1 acrylamide : bis-acrylamide), 5% glycerol, 20 ml of 5 x TBE, 100 µl each of 25% APS and TEMED per 100 ml of gel solution. The gel required polymerisation for 2 hours before the samples were loaded. It was then run for a period of 24 hours that was determined by the size of the PCR fragment, at a constant voltage of 700 volts, at room temperature. The gel was dried on a vacuum slab dryer for 1-2 hours, before autoradiography for 30-60 minutes.

6.2.6.2 SSCP Analysis and Electrophoresis: Protocol II

A mutation detection enhancement (MDE) polyacrylamide gel was used for the SSCP analysis. MDE is a modified polyacrylamide based vinyl polymer. Its higher loading capacity allows for increased sensitivity without loss of resolution. The following reagents were used to make a 0.5 x MDE gel: 30 ml of MDE (stored at room temperature; AT Biochem) with 9.6 ml of 5 x TBE. When ready to pour, 80 µl of 25% APS and 80 µl of TEMED were added to a total volume of 80 ml. After polymerisation for 1 hour, the gel was run in 0.6 x TBE buffer without pre-electrophoresis. The samples were loaded, and the gel was run at 700 volts, constant voltage for 24 hours. The gel plates were separated, the gel dismantled, immediately dried then autoradiographed for 30-60 minutes.

6.2.7 Preparation and Direct DNA Sequencing of PCR Products

PCR products were purified using the reagents and protocols of the Qiaquick PCR Purification Kit Protocol (Qiagen) in a 100 µl reaction volume. The purity of the PCR products was determined by visualisation on 1% agarose gels (2.5.2). Their concentrations were determined as described in 2.3.1.4.3. PCR products were sequenced two times each from either end. Direct sequencing of PCR products was performed by dye primer cycle sequencing (2.7.3) with the forward and reverse primers to produce fluorescently labelled double stranded PCR products according to the conditions specified by the manufacturer. The purified labelled products were then separated and analysed by an Applied Biosystems Model 373A DNA Automated Sequencer (2.7.4).

6.2.8 Population Screening

Population screening was performed in all cases where a reproducible band shift (ie. the same differences in the pattern of migrating bands from tumour samples compared to normal samples observed on at least two separate occasions) was detected by SSCP analysis. In all such cases, genomic DNA from 50 normal individuals in the general population was examined for polymorphisms. When this part of the project was being performed, part of the genomic sequence of FAA had been elucidated. Thus the segments of DNA in which a band shift was demonstrated were amplified by PCR using primers flanking the exons.

Restriction endonuclease digestion (2.5.1) of PCR products was conducted at 37°C. Ten µl of PCR product were added to 2 µl of the corresponding 10 x enzyme buffer, 2 units of restriction endonuclease made up to 20 µl with sterile water. The reaction mix was incubated for 2-4 hours to ensure complete digestion. Reactions were terminated by the addition of 0.1 volume of 10 x agarose gel loading buffer.

Ten µl aliquots of the digested PCR products were analysed by electrophoresis on 3 % agarose gels as described in 2.5.2. Spp-1 bacteriophage restricted with EcoRI or puc19 DNA restricted with HpaII were used as markers on the gels to estimate the size of the product.

6.3 RESULTS

6.3.1 Synthesis and Analysis of Large cDNA Fragments by RT-PCR

RT-PCR was performed using RNAs extracted from breast tumour samples displaying LOH at 16q24.3-qter, PBLs from a normal individual, breast tumour cell lines and primers specific for each cDNA clone. Three large overlapping PCR products from the 3' end of the 5.5 kb transcript were generated. Random primed first strand cDNA was used as template with primers FA5 + FA2 (fragment 1) to synthesise the second cDNA strand. Attempted synthesis of this second cDNA strand using oligo-dT primed cDNA was not successful. This may have been due to the location of the FA5 primer, about 4 kb distal to the poly(A) tail, and the reverse transcriptase being unable to synthesise such long stretches of cDNA. If first strand cDNA was not produced to a length of 4 kb or more using the oligo-dT primer, the binding site for FA5 would not be present, leading to a lack of synthesised PCR products. Oligo-dT primed first strand cDNA was used as template with primers FA3 + FA6 (fragment 2) and FA14 + FA24 (fragment 3). A small segment of clone yc81e09 was amplified from oligo-dT primed first strand cDNA with primers yc81F1 + yc81R1 (fragment 4).

PCR products of the double stranded cDNA syntheses were analysed on 1% agarose gels. The gels were Southern blotted and hybridised with radiolabelled cDNA insert to determine the sizes of the specific products. Examples of these autoradiographs are shown in figure 6.3. The sizes of the PCR products from breast tumours and breast tumour cell lines were compared to those of the PBL sample from a normal individual. Comparison of PCR products from affected and normal tissue samples from an individual patient is the ideal procedure, but the normal tissue from each patient was not available in our laboratory. The PCR products from the sample of normal PBL and the breast tumour cell lines corresponded with the size of the PCR product synthesised from the respective cDNA clone. The sizes of the products from breast tumours also corresponded with the sizes of PCR product synthesised from each cDNA clone. Thus, bands which were totally deleted or of reduced

Figure 6.3

Representative autoradiographs of RT-PCR products synthesised with primers:

A. FA5 + FA2 - fragment 1 from the 5.5 kb transcript (763 bp)

B. FA3 + FA6 - fragment 2 from the 5.5 kb transcript (830 bp and 600 bp)

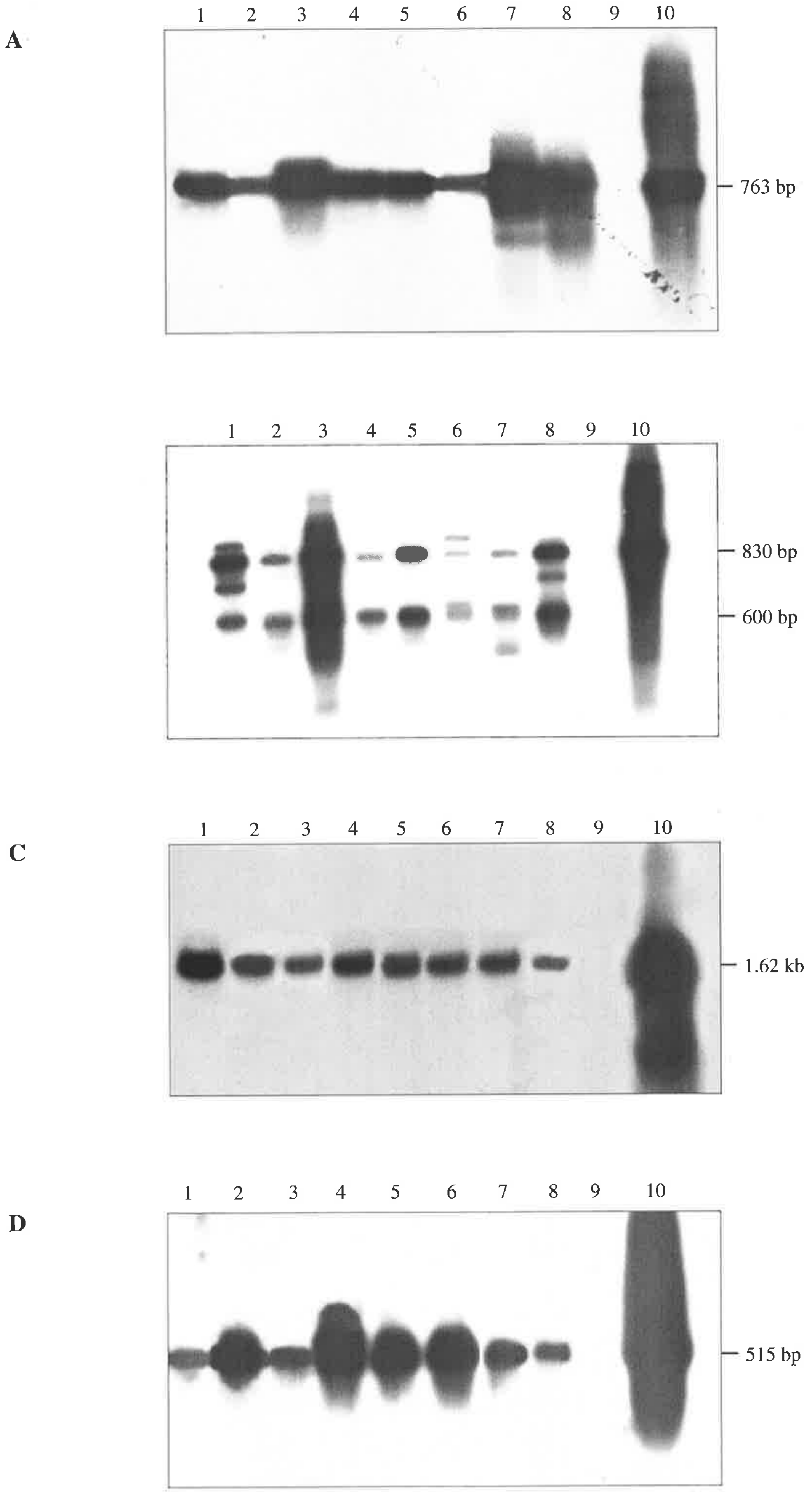
C. FA14 + FA24 - fragment 3 from the 5.5 kb transcript (1.62 kb)

D. yc81F1 + yc81R1 - fragment 4 from yc81e09 (515 bp)

The PCR products from double stranded cDNA syntheses were analysed on 1% agarose gels which were Southern blotted. Filters A, B and C were hybridised with radiolabelled yh09a04 cDNA insert and filter D was hybridised with radiolabelled yc81e09 insert.

The lanes represent:

Lane 1	Peripheral Blood Lymphocytes from a normal individual
Lane 2	Breast Tumour Sample (BT) 541
Lane 3	BT 355
Lane 4	BT 919
Lane 5	BT 819
Lane 6	BT 413
Lane 7	BT 555
Lane 8	BT 559
Lane 9	no DNA
Lane 10	cDNA clone


A
1 2 3 4 5 6 7 8 9 10
763 bp
B
1 2 3 4 5 6 7 8 9 10
830 bp
600 bp
C
1 2 3 4 5 6 7 8 9 10
1.62 kb
D
1 2 3 4 5 6 7 8 9 10
515 bp

size were not observed in the tumour samples. One band was observed for the PCR products of fragments 1 (763 bp), 3 (1.62 kb) and 4 (515 bp) in breast tumours, normal PBL, breast cell lines and the cDNA clone. Two main bands of 830 bp, which is the expected size, and about 600 bp were detected for PCR products of fragment 2 in breast tumours, normal PBL and breast cell lines. Additional bands were also detected in PBL from a normal individual and the breast tumour samples 555 and 559. Since all the breast tumour samples displayed a PCR product of the expected size (830 bp) and the sizes of the additional bands observed in the breast tumours also corresponded to the sizes of the bands in the normal PBL sample, they were considered to be normal variants of alternative splicing in the region encompassed by primers FA3 + FA6, not the products of deleted sequence resulting from a mutation.

Screening of the RT-PCR products with their respective radiolabelled cDNA probe detected bands homologous to each cDNA. Visualisation of the RT-PCR products on agarose gels demonstrated additional bands which were not detected by probing. These bands were attributed to artefacts related to degradation of the RNA. Signal was not observed in control reactions in which template was not added, indicating that contamination was not present in the reactions.

6.3.2 Single Strand Conformation Polymorphism Analysis of cDNA Fragments

Differences in band sizes of fragments 1-4, from breast tumours compared to PBL from a normal individual, were not noted, thus point mutations or deletions were searched for in each transcript by SSCP analysis. The RT-PCR products of fragments 1-4 were used as templates to amplify overlapping fragments of approximately 200-400 bp in size using primers listed in table 6.3. PCR products of the various cDNA fragments were visualised on 1% agarose gels to determine whether the PCR conditions were optimal. Examples of these PCR products are shown in figure 6.4. The sizes of the products corresponded to the size of the PCR product amplified from the cDNA clone indicating that the correct sized products

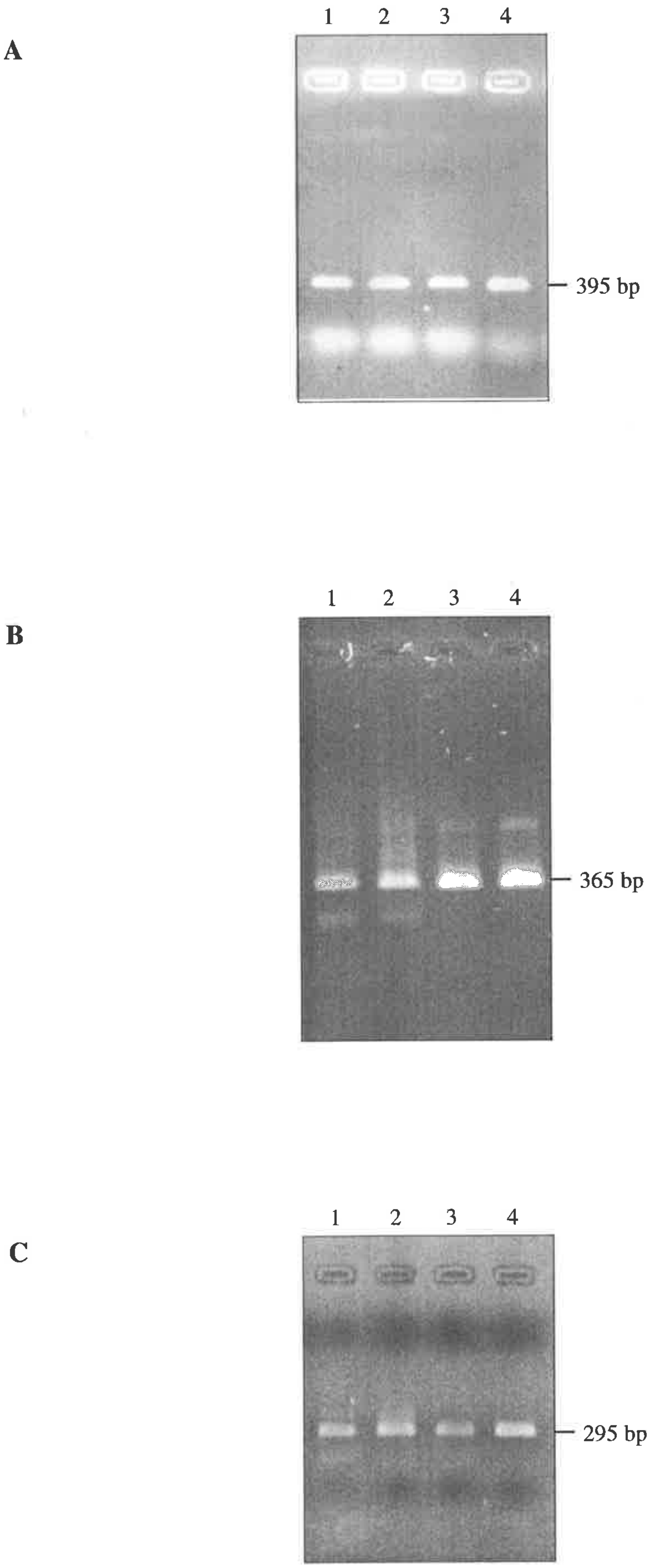
Figure 6.4

Examples of nested PCR products generated from RT-PCR products of fragments 1 and 2, visualised on 1% agarose gels.

Part A shows a 395 bp PCR product generated with primers FA5 + 6.4R from fragment 1.
Lanes 1-4 represent:
Peripheral blood lymphocytes from a normal individual (PBL), Breast tumour (BT) 541, BT 919, BT 555

Part B shows a 365 bp PCR product generated with primers FA11 + FA12 from fragment 2. Faint background bands were observed in these samples which may have been products of the additional bands synthesised using primers FA3 + FA6 to generate fragment 2 (see figure 6.3, part B).
Lanes 1-4 represent:
PBL, BT 559, BT 413, BT 919

Part C shows a 295 bp PCR product generated with primers FA13 + FA6 from fragment 2.
Lanes 1-4 represent:
PBL, BT 559, BT 555, BT 919


A
1 2 3 4
395 bp
B
1 2 3 4
365 bp
C
1 2 3 4
295 bp

were amplified. Since additional bands were present in the PCR products derived from primers FA3 + FA6 (fragment 2), extra bands were expected from the nested PCR but only faint background bands were observed (see figure 6.4 B). It is possible that any additional bands may have migrated off the gel as their size may have been small. An additional possibility is that, as a consequence of alternative splicing, the sequence homologous to the nested primers may not be present in the PCR product of fragment 2.

Following the examination of all primer conditions and their products, PCR reactions including radiolabel were prepared for SSCP analysis using 10% acrylamide and MDE gel electrophoresis. Band shifts were observed for cDNA (5.5 kb transcript) amplified with primers FA5 + 6.4R, FA9 + FA10, FA11 + FA12, FA13 + FA6, in various breast tumour samples when compared to PBL isolated from normal individuals. These results are summarised in table 6.4. Banding patterns of PCR products generated by FA5 + 6.4R were different in breast tumour samples 541, 355, 819 and 559 compared to normal PBL. Differences were also observed for breast tumour 919 from FA9 + FA10; breast tumours 541, 355, 919, 819, 555 and 559 from FA11 + FA12; and breast tumours 555 and 559 from FA13 + FA6. Examples of these autoradiographs are shown in figures 6.5, 6.6, 6.7 and 6.8. These differences were observed on both 10% acrylamide and MDE gels, but the banding patterns were more distinct on the 10% acrylamide gels. Band shifts were not observed for fragment 4, derived from clone yc81e09.

6.3.3 Sequence Analysis of cDNA Fragments

Reproducible band shifts observed on three or four separate occasions in DNA of breast tumour samples compared to DNA of PBL from normal individuals were further investigated by direct DNA sequencing of purified PCR products. The sequences derived from the PBL samples and their respective cDNA clone were compared with the sequences obtained from the breast tumour samples to determine whether they contained any differences. It was concluded that these sequences were accurate, as sequences from both directions of each

Table 6.4

Summary of band shifts observed for the 5.5 kb transcript. Fragments of cDNA amplified with primers FA5 + 6.4R, FA9 + FA10, FA11 + FA12, FA13 + FA6 demonstrated differences in breast tumour samples when compared to peripheral blood lymphocytes (PBL) from normal individuals. The '-' indicates the banding pattern seen in PBL and the '+' indicates a different banding pattern compared to PBL.

Tumour or PBL Sample	Primer Pairs			
	FA5 + 6.4R	FA9 + FA10	FA11 + FA12	FA13 + FA6
PBL	-	-	-	-
BT 541	+	-	+	-
BT 355	+		+	-
BT 919	-	+	+	-
BT 819	+	-	+	-
BT 413	-	-	-	-
BT 555	-	-	+	+
BT 559	+	-	+	+

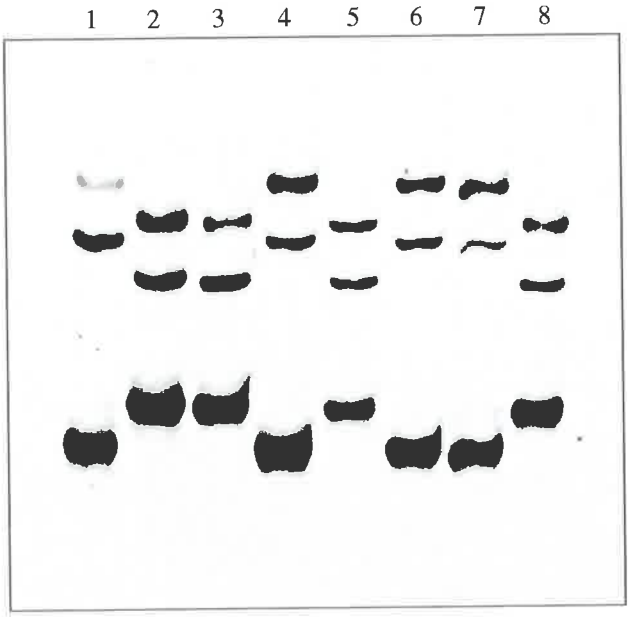
Figure 6.5

Autoradiograph of PCR products generated by FA5 + 6.4R analysed on a 10% acrylamide gel.

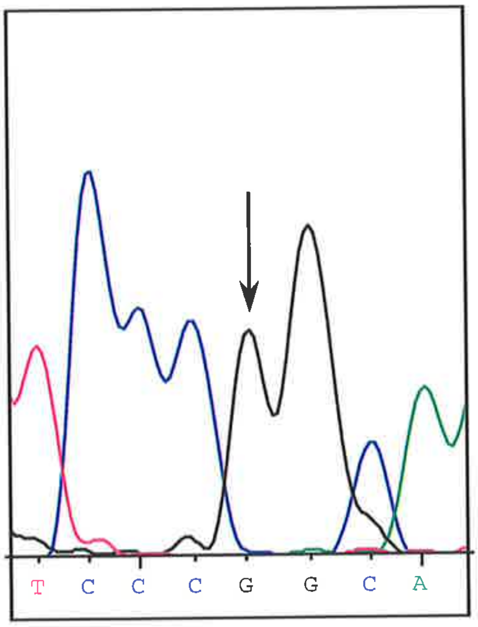
The lanes represent:

Lane 1	Peripheral Blood Lymphocytes from a normal individual
Lane 2	Breast Tumour Sample (BT) 541
Lane 3	BT 355
Lane 4	BT 919
Lane 5	BT 819
Lane 6	BT 413
Lane 7	BT 555
Lane 8	BT 559

The chromatograms show sequence differences demonstrated in PBL and BT 541.



BT 541



PBL

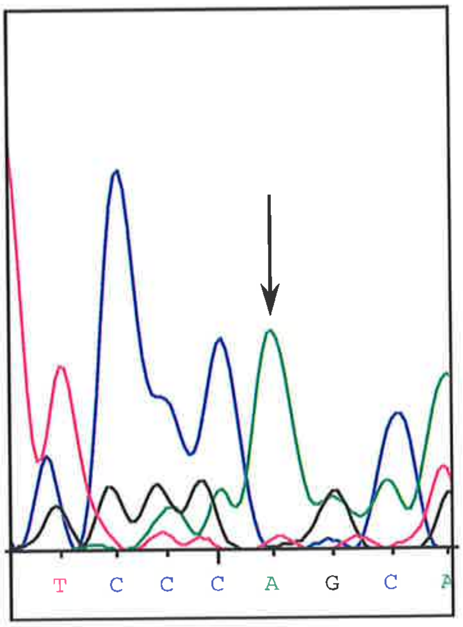


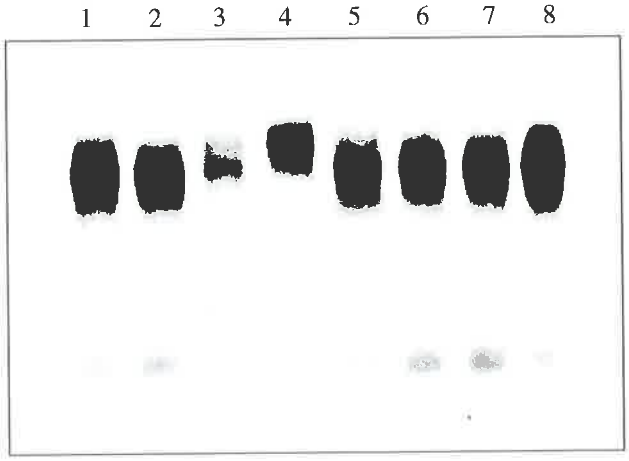
Figure 6.6

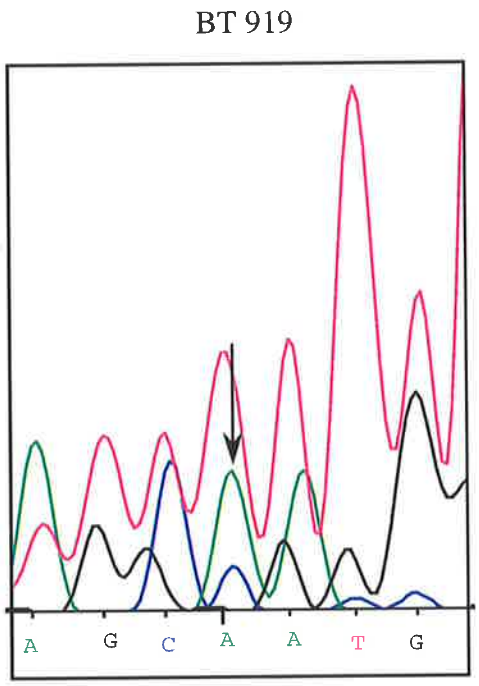
Autoradiograph of PCR products generated by FA9 + FA10 analysed on a 10% acrylamide gel.

The lanes represent:

Lane 1	Peripheral Blood Lymphocytes from a normal individual
Lane 2	Breast Tumour Sample (BT) 541
Lane 3	BT 355
Lane 4	BT 919
Lane 5	BT 819
Lane 6	BT 413
Lane 7	BT 555
Lane 8	BT 559

The chromatograms show the sequence differences demonstrated in PBL compared to BT 919. The reverse complement of the sequences to the 5.5 kb transcript sequence is shown.


1
2
3
4
5
6
7
8


BT 919
A
G
C
A
A
T
G

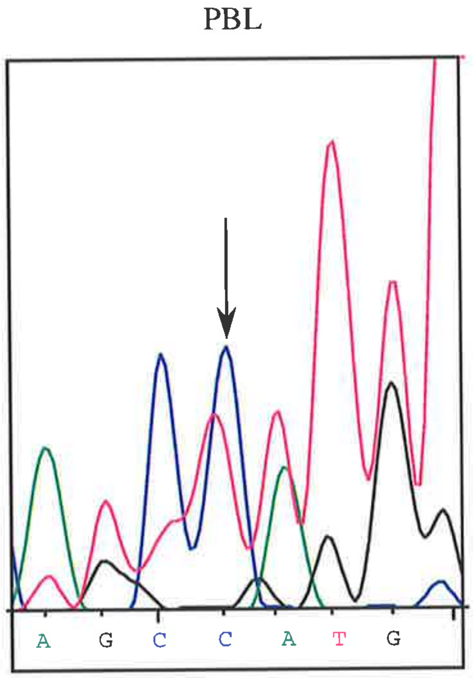

PBL
A
G
C
C
A
T
G

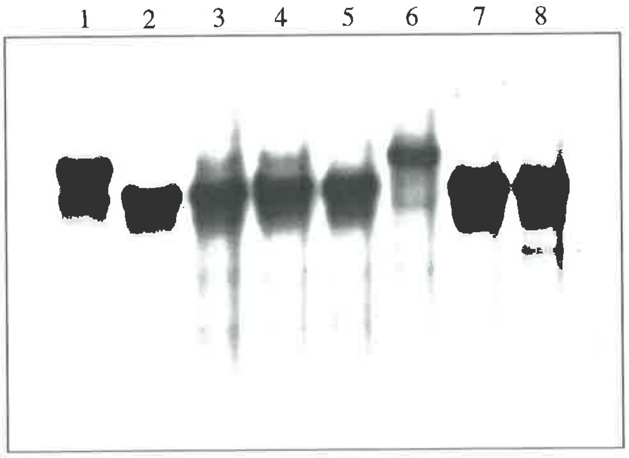
Figure 6.7

Autoradiograph of PCR products generated by FA11 + FA12 analysed on a 10% acrylamide gel.

The lanes represent:

Lane 1	Peripheral Blood Lymphocytes from a normal individual
Lane 2	Breast Tumour Sample (BT) 541
Lane 3	BT 355
Lane 4	BT 919
Lane 5	BT 819
Lane 6	BT 413
Lane 7	BT 555
Lane 8	BT 559

The chromatograms show differences demonstrated in PBL and BT 355.


1
2
3
4
5
6
7
8

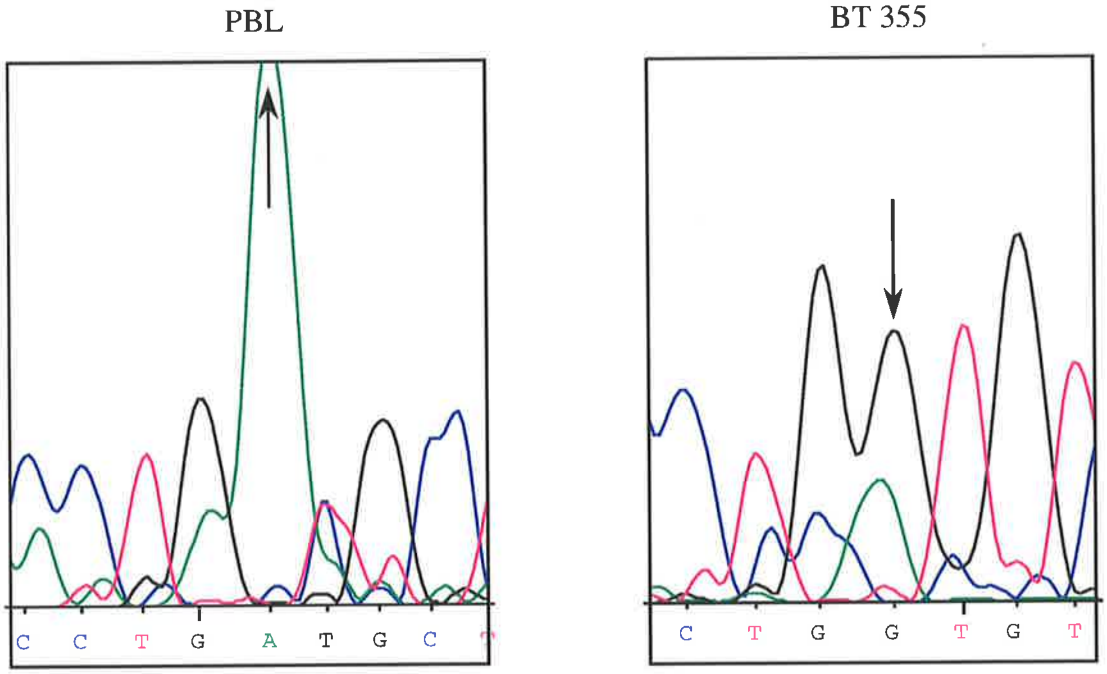
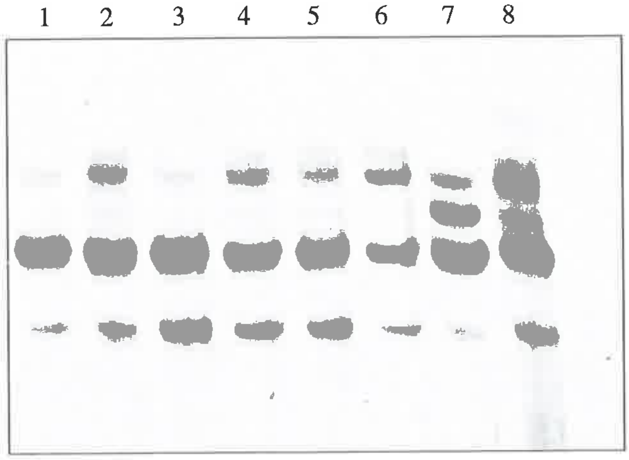

PBL
C C T G A T G C T
BT 355
C T G G T G T

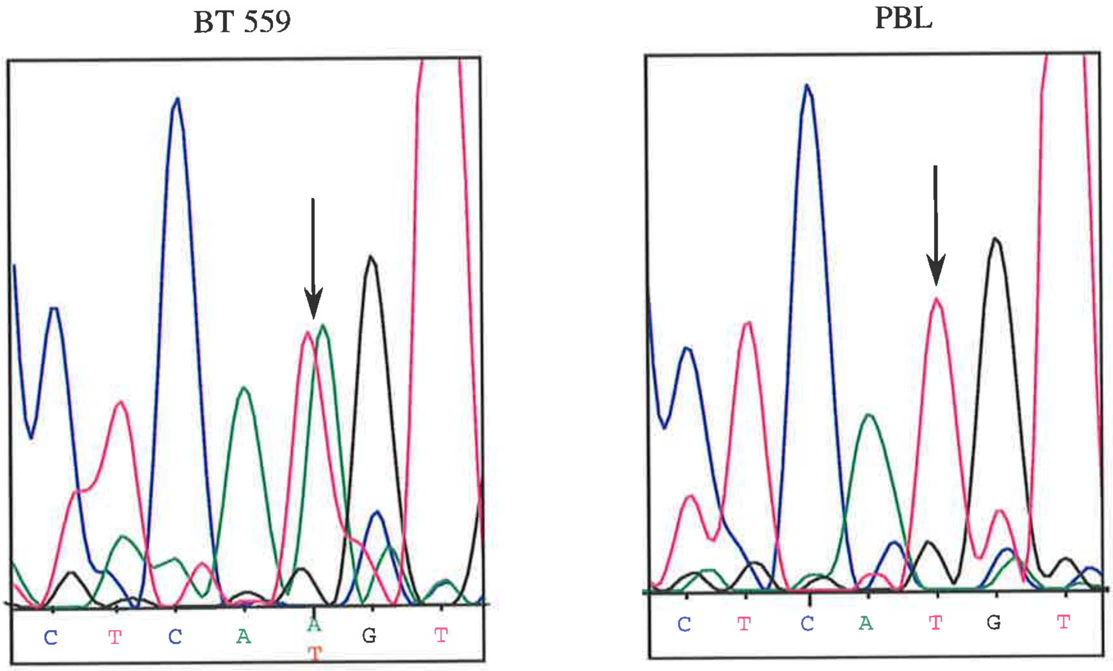
Figure 6.8

Autoradiograph of PCR products generated by FA13 + FA6 analysed on 10% acrylamide gel. The lanes represent:

Lane 1	Peripheral Blood Lymphocytes from a normal individual
Lane 2	Breast Tumour Sample (BT) 541
Lane 3	BT 355
Lane 4	BT 919
Lane 5	BT 819
Lane 6	BT 413
Lane 7	BT 555
Lane 8	BT 559

The chromatograms show differences demonstrated in PBL and BT 559.


1
2
3
4
5
6
7
8


BT 559
C T C A A G T
T
PBL
C T C A T G T

PCR product were observed to be complementary to each other. Sequence changes detected in the PCR products from breast tumour and PBL from normal individuals are listed in table 6.5. The sequence differences in the 5.5 kb transcript included nucleotides 'A' or 'G' at position 1545, 'G' or 'T' at position 2195, and 'G' or 'A' at 2470. The difference in the sequence of each PCR product analysed was in agreement with the difference in the banding pattern of the respective PCR product, demonstrated by gel electrophoresis.

One result which remains unclear is that of the similar band shift seen in PCR products from breast tumour samples 555 and 559 (figure 6.8, lanes 7 and 8 respectively), generated with primers FA13 + FA6. Nucleotides 'A' and 'T' were displayed at position 2659 on the chromatogram for sample 559, indicating variance in the sequence (figure 6.8). Only the 'T' nucleotide was observed at position 2659 in sample 555. Additionally, the 'T' nucleotide was present at position 2659 in sequences from the 5.5 kb transcript and PBL from normal individuals. The SSCP gel banding pattern of PBL from a normal individual was clearly different to the banding patterns of samples 555 and 559. Thus, sequences from breast tumour samples 555 and 559 were expected to be different to the sequence from PBL. The sequence obtained for sample 555 does not correlate with the banding pattern observed on the SSCP gel. These results were verified on three occasions using newly synthesised PCR product samples for SSCP analysis and sequencing.

The nucleotide alterations observed in the DNA sequence obtained from the breast tumours resulted in substitutions of amino acids at the corresponding positions in the protein sequence. These are listed in table 6.6.

6.3.4 Polymorphism Screening in DNA from Normal Individuals

A panel of fifty normal genomic DNA samples was screened to identify polymorphisms in the general population. While these experiments were conducted, the genomic structure of the 5.5 kb transcript, FAA gene, was being elucidated. Based on this sequence, it was now

Table 6.5

Sequence changes detected in PCR products generated from breast tumours which were observed to have a different banding pattern of the 5.5 kb transcript compared to peripheral blood lymphocytes from normal individuals.

	Primer Pairs			
Tumour / PBL Sample	FA5 + 6.4R	FA9 + FA10	FA11 + FA12	FA13 + FA6
Position	1545	2195	2470	2659
5.5 kb transcript	CCCAGC	CATTGC	CTGGTG	TCATGT
PBL	CCCAGC	CATGGC	CTGATG	TCATGT
BT 541	CCCGGC	CATGGC		
BT 355	CCCGGC	CATGGC	CTGGTG	
BT 919	CCCAGC	CATTGC	CTGGTG	TCATGT
BT 819	CCCGGC	CATGGC	CTGGTG	
BT 413		CATGGC	CTGATG	
BT 555	CCCAGC	CATGGC		TCATGT
BT 559	CCCGGC	CATGGC	CTGGTG	TCATGT A

Table 6.6

Restriction endonuclease sites which were created or destroyed by the nucleotide changes observed in the cDNA sequence derived from breast tumours. Amino acid substitutions resulting from base alterations are also indicated.

DNA	Restriction Enzyme		Protein	Exon	Frequency	
	Created	Destroyed			Allele (+)	Allele (-)
A 1545 G	HpaII / MspI		Ser 501 Gly	16	0.60	0.40
G 2195 T	NlaIII		Met 717 Ile	23		
G 2470 A		MvaI / BstNI	Gly 809 Asn	26	0.70	0.30
T 2659 A		NlaIII	Met 872 Lys			

possible to PCR amplify exons containing the divergent sequence of the FAA gene, described in section 6.3.3, using primers homologous to the known exon sequence. Polymorphisms were detected with the use of restriction enzymes, since the nucleotide changes in the divergent sequences either created or destroyed restriction endonuclease sites. Table 6.6 lists the restriction enzyme sites altered by the observed nucleotide changes.

PCR amplified exons were restricted with the appropriate enzyme and visualised on 3% agarose gels, together with unrestricted samples, to determine if the divergent sequence comprised a mutation of the FAA gene or a polymorphism in the general population. Figure 6.9 shows an example of PCR products amplified from the panel of normal DNA samples which were restricted with the appropriate enzyme. Due to time restrictions, only the sequences at positions 1545 and 2195 in the general population were characterised. These two nucleotide differences detected in the breast tumour samples were subsequently identified to be frequent polymorphisms in the general population.

PCR products amplified from breast tumour samples restricted with the appropriate enzyme were also included in this experiment. An example is shown in figure 6.9. These reactions were performed to determine the restriction pattern of each tumour sample and to compare this with their SSCP banding pattern and in all cases an altered SSCP banding pattern corresponded with an altered restriction pattern. The SSCP, restriction endonuclease and sequence data were all in agreement when compared to each other.

Further investigation of the sequence variation at position 2659, observed in breast tumour samples 555 and 559 (see 6.3.3) and PBL from normal individuals, by restriction enzyme analysis may have been beneficial in providing additional information for the elucidation of the differences between these samples. Unfortunately, due to time constraints, this analysis was not performed.

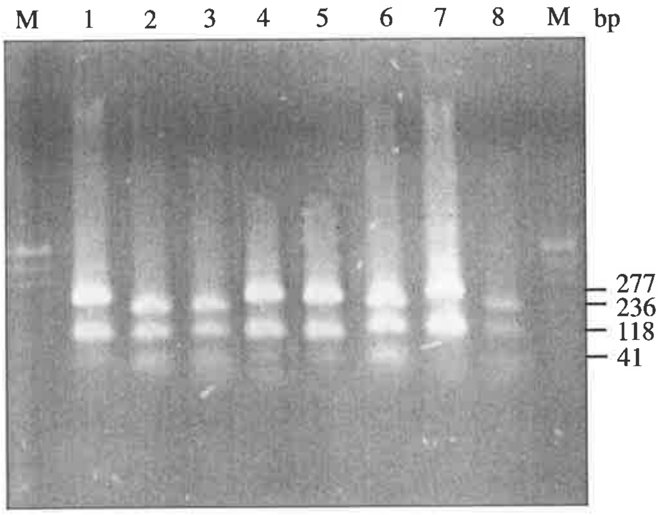
Figure 6.9

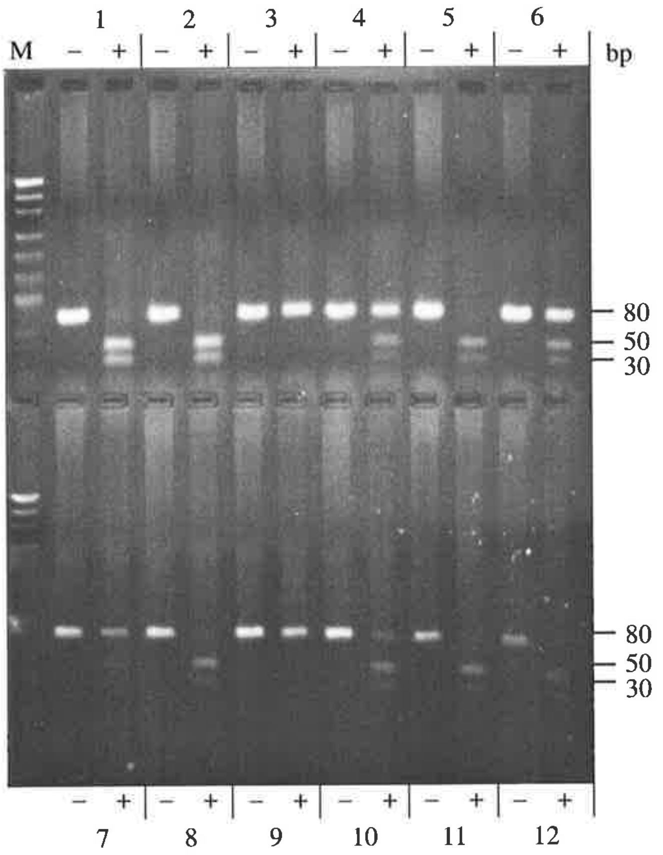
A. (Top) PCR products amplified from breast tumour samples using FA5 + 6.4R primers. The products were restricted with HpaII to determine the restriction pattern of each tumour sample and to compare this with the SSCP banding pattern.

The lanes represent:

Lane M	HpaII restricted puc19 size markers
Lane 1	Peripheral blood lymphocytes from a normal individual
Lane 2	Breast tumour (BT) 541
Lane 3	BT 355
Lane 4	BT 919
Lane 5	BT 413
Lane 6	BT 819
Lane 7	BT 555
Lane 8	BT 559

B. (Bottom) Example of PCR products amplified from a panel of normal DNA samples using exon 15 primers, 15F 5'- TCTCTCCACACAGGACACTG -3' and 15R 5'- TTGGGGAGGCCAAGGCAGTC -3', to characterise the base alteration at position 1545. The 80 bp PCR products were restricted with HpaII enzyme to determine whether DNA samples from the general population possessed the restriction enzyme site. The '+' and '-' indicate the presence or absence of restriction enzyme respectively, in the DNA sample. Restriction of DNA at a HpaII site generated fragments of 50 bp and 30 bp. The presence of 3 bands (80 bp, 50 bp and 30 bp), observed in samples 4, 6 and 10, indicated that the individual examined was heterozygous and the presence of 2 bands (50 bp and 30 bp), observed in samples 1, 8 and 5. indicated a homozygous individual.


M
1
2
3
4
5
6
7
8
M
bp
277
236
118
41


1
2
3
4
5
6
M
− + − + − + − + − + − +
bp
80
50
30
80
50
30
− + − + − + − + − + − +
7
8
9
10
11
12

6.4 DISCUSSION

Single strand conformation polymorphism analysis was conducted on the 5.5 kb transcript to determine if mutations could be detected in a panel of seven breast tumour samples which had restricted LOH of the 16q24.3-qter region, and whether the mutations were restricted to breast tumours. The analysis of a number of short 200-400 bp overlapping segments of cDNA by SSCP using 10% acrylamide and MDE gels was performed in order to achieve a high detection rate of band shifts. The combination of these conditions is reported to detect 80-98% of mutations (Forrest *et al.* 1995; Ravnik-Glavac *et al.* 1994; Sheffield *et al.* 1992; Grompe, 1993). SSCP mutation analysis was also performed on a short segment of the cDNA clone yc81e09.

Four base alterations, nucleotides 'A' or 'G' at position 1545, 'G' or 'T' at position 2195, 'G' or 'A' at 2470 and 'A' and 'T' at position 2659, were found in the 3' end of the ORF of the 5.5 kb sequence between position 1369 and 4447. All these alterations resulted in the substitution of an amino acid. Base changes at positions 1545 and 2470 were demonstrated to be polymorphisms which occurred in the general population. Subsequent data from the FAB consortium supported this result and provided evidence that the alteration observed at position 2195 can also be considered to be a polymorphism. These polymorphisms will be useful in refining the region of LOH at 16q24.3 (see 1.6.1) by screening breast tumour samples since polymorphisms lose or create restriction enzyme sites. The other base change at position 2659 was not screened for polymorphisms in the general population due to time constraints, thus it is not possible to conclude if this is a mutation or a polymorphism. This alteration was not detected in the work conducted by the FAB consortium. The paired normal sample of breast tumour 559 is currently being analysed to determine whether its sequence differs from the sequence observed in this breast tumour sample.

Conjointly with this work, the FAB consortium initiated a search for mutations in a panel of FA-A patients by RT-PCR. One cDNA that contained the sequence from clones yf14a03 and yh09a04 was found to contain six exons trapped from cosmids c352A12 and c431F1 (see 5.3.5.1), and was partially deleted in RT-PCR products from an Italian patient (The Fanconi Anaemia/Breast Cancer Consortium, 1996). This 5.5 kb cDNA clone, which contained the compiled sequence of clones ET19, yf14a03 and yh09a04 at its 3' end (nucleotides 3026-5481), was then investigated in more detail as a candidate for the FAA gene. Also, through personal communication with Dr. Hans Joenje (Department of Human Genetics, Free University, The Netherlands), it was discovered that this sequence is virtually identical to a cDNA isolated from an expression library by functional complementation in FA-A cells (Lo Ten Foe *et al.* 1996) but contains an additional 13 bp of 5' untranslated sequence.

SSCP analysis and direct sequencing of RT-PCR products from FA-A patients were then conducted to provide further evidence that this cDNA was derived from the FAA gene (The Fanconi Anaemia/Breast Cancer Consortium, 1996). Four different mutations were observed, all of which were expected to disrupt the function of the encoded protein. The first mutation, detected in a patient of Italian origin, was a deletion of 274 bp (nucleotides 1671-1944), resulting in a shift of the reading frame and the creation of a premature termination codon 6 amino acids downstream. Preliminary characterisation of the structure of the gene indicates that this deletion removes 3 exons. A second mutation in a British FA-A family was a deletion of one of two direct repeats of the sequence TTGG at nucleotides 1155-1162. This would result in a shift of the reading frame and the production of a termination codon 42 amino acids downstream. An African-American FA-A patient was found to have a deletion of 156 bp involving nucleotides 1515-1670. This deletion, which spans 2 exons in genomic DNA, does not cause a frameshift, but would remove 52 amino acids from the encoded protein. It was also observed in two other FA families with an unknown complementation group. A 113 bp deletion (nucleotides 938-1050), was detected in an FA patient of North European origin and undefined complementation group who is a compound heterozygote for this deletion and the 156 bp deletion. The 113 bp deletion removes a single exon from the coding sequence, and creates a premature termination codon 2 amino acids downstream. This mutation is likely to be a genomic deletion rather than a splice site mutation, since sequencing of genomic DNA from this exon and its flanking intronic sequence in the patient showed only the normal sequence. These results are summarised in table 6.7.

Thus, single strand conformation polymorphism analysis of the 5.5 kb transcript sequence in FA-A cases resulted in the identification of four different intragenic mutations, three of which produce truncated proteins. These mutations demonstrated that the 5.5 kb sequence which was isolated was the FAA gene (The Fanconi Anaemia/Breast Cancer Consortium, 1996). The fact that three of the four mutations detected in this study are intragenic deletions removing one or more exons is surprising, since most mutations in autosomal recessive disorders are either point mutations or deletions and insertions of a few base pairs. These mutations may be due to the introns of the FAA gene containing repetitive, unstable sequences which are susceptible to deletions. Detailed analysis of the deletion breakpoints in these and other FA-A cases is required to resolve this question.

Polymorphisms detected by the SSCP analysis and direct sequencing of RT-PCR products from FA-A patients included changes of nucleotide 'G' to 'A' at position 840, and 'T' to 'A' at position 1133. Three of the nucleotide changes which were detected by myself, at positions 1545, 2195 and 2470 using breast tumour samples, were also detected by the FAB consortium in FA-A samples. An additional alteration, 'G' to 'T' at position 4119, was detected in FA-A samples but not in breast tumour samples. Subsequent screening of all FAA exons using primers located at the intron/exon boundaries has shown that this gene is not involved in breast cancer (The Fanconi Anaemia/Breast Cancer Consortium, personal communication).

To determine the possible role of another transcript in breast cancer, single strand conformation polymorphism analysis was performed on the 515 bp segment of the yc81e09 cDNA clone. This did not display any band shifts in the breast tumour samples compared to

Table 6.7

Summary of results obtained from analysis of the 5.5 kb transcript in FA-A patients. The location and type of mutations detected are listed.

Nucleotides Affected	DNA Mutation Detected	Protein Alteration
1671-1944	274 bp deletion (3 exons)	premature termination of codon 6 amino acids downstream
1155-1162	deletion of TTGG (reading frame shift)	premature termination codon 42 amino acids downstream
1515-1670	deletion of 156 bp (spans 2 exons in genomic DNA)	removal of 52 amino acids from encoded protein
938-1050	113 bp deletion (1 exon)	premature termination codon 2 amino acids downstream

DNA from a normal individual. Once the complete 3.7 kb sequence of this transcript is isolated and its ORF determined, SSCP mutation analysis on the remaining portion of the ORF can be performed to determine its possible involvement as the TSG in breast cancer.

Mutation analysis by SSCP is a simple and effective technique for the detection of single base substitutions (Sheffield *et al.* 1993). The sensitivity of SSCP is dependent on several factors and varies dramatically with the size of the DNA fragment analysed, in addition to the electrophoretic gel composition. Short overlapping 200-400 bp segments of cDNA were used in the analysis of the transcripts in this study. High percentage acrylamide gels have been reported to improve resolution in SSCP analysis (Savov *et al.* 1992) and this finding is supported by the results of the research conducted in the Department of Cytogenetics and Molecular Genetics where the detection of band shifts on 10% acrylamide gels was superior to that of both 6% and 4.5% acrylamide gels. The alternate MDE gel system has several advantages over the 10% acrylamide gels including greater ability than conventional polyacrylamide gels to resolve and detect point mutations. The MDE gel system is however, more expensive than the polyacrylamide/glycerol gels. Through my experience with the two gel systems, band shifts which were detected with MDE gels were also seen with the 10% gels. The banding pattern was much clearer with 10% gels than the distorted banding patterns observed with MDE.

Close to 100% of mutations can be detected in fragments of the order of 250-300 bp if the analysis is performed using two different gel conditions (Forrest *et al.* 1995; Sheffield *et al.* 1992; Grompe, 1993). Even though these guidelines were followed, the polymorphism detected at position 4119 in FA-A samples by the FAB consortium, was not seen in the SSCP studies conducted in breast tumour samples. Two fragments encompassing this position, created by FA22 + yf143' (254 bp) and FA23 + FA24 (375 bp), were analysed using the two gel systems, but band shifts were not detected. This may be due to the SSCP technique not being 100% efficient or this polymorphism may not be present in the breast tumour samples used in this analysis.

A disadvantage of using RT-PCR in SSCP analysis is that ribonuclease contamination may occur in RNA samples which may lead to confusion in the interpretation of the results. This contamination may be observed as additional or missing bands and may lead to spurious SSCP results. Fortunately, the artefactual bands observed in RT-PCR products from breast tumours did not appear to affect the SSCP analyses which were performed.

The presence of RNA from contaminating normal tissue in breast tumour samples is an additional problem which can lead to false negative results from SSCP mutation analysis based on RT-PCR. Microdissection of the tumour from surrounding normal tissue is an option which can reduce this problem. The tumour samples used in these experiments were flow sorted, based on DNA ploidy differences and/or immunocytochemical cell lineage markers, by Dr. Cleton-Jansen to enrich the tumour cell fractions to assist the interpretation of the mutation analysis results. All the tumour samples are likely to contain some normal tissue which have the constitutive genotype, and such contamination would only have been evident as a faint band in the SSCP results.

Alternative splicing variants may also affect the nested PCR step and subsequent SSCP mutation analysis if the nested primers span a spliced exon. Additional bands were observed for the portion of the 5.5 kb transcript encompassed by the primers FA3 + FA6 (figure 6.3 B). However, the correct sized PCR products were observed in nested PCR. The sequence analysis of these PCR products confirmed that these corresponded to the correct segments of the transcript.

The fidelity of *Taq* DNA polymerase in PCR potentially can affect the recognition of true mutations. Direct sequencing of PCR amplified DNA generates a consensus sequence independent of errors that may occur during amplification, allowing the detection of mutations (McMahon *et al.* 1987; Engelke *et al.* 1988). The bulk of PCR product sequences will be correct at each base position and the erroneous copies are likely to be at levels too low to interfere with screening for mutations. The exception to this situation occurs when products are cloned for sequencing which may yield unacceptable alterations within the isolated clone. Considering these factors, sequence of greatest fidelity is more likely to be generated by directly sequencing double stranded PCR products rather than the cloned PCR products. Thus, to avoid encountering any *Taq* DNA polymerase errors, direct sequencing of PCR amplified DNA was performed in these experiments. The sequence data obtained from PCR products was verified by investigation of restriction enzyme recognition sites present in the sequence of interest. Base pair changes which could be attributed to *Taq* DNA polymerase errors were not observed in the sequences obtained.

In conclusion, the sequence at the 3' end of the 5.5 kb transcript (ET19, yf14a03 and yh09a04), corresponding to the FAA gene, and a section of the transcript represented by clone yc81e09 were analysed for mutations in a set of breast tumours with restricted LOH at 16q24.3. Mutations were not observed for the cDNA sequences examined, thus it is likely that these transcripts are not candidates for the TSG localised to 16q24.3. However, mutation analysis of the 5.5 kb transcript in FA-A patients by the FAB consortium demonstrated mutations in this sequence, suggesting that this cDNA is the FAA gene (The Fanconi Anaemia/Breast Cancer Consortium, 1996).

6.4.1 Analysis of Other Candidate Tumour Suppressor Genes Localised to 16q24

Members of the FAB consortium have also analysed the genes BBC1 and CMAR (see 1.6.4) for mutations in breast tumour samples displaying restricted LOH at 16q24.3-qter, to determine whether they represent the TSG. This was achieved by direct sequencing of PCR products derived from the tumours and comparison of the sequences with the normal gene sequences. All sequences obtained from the breast tumours were identical to the sequence from the normal transcript. Thus, the BBC1 and CMAR genes were eliminated as candidates for the breast cancer TSG.

The H-cadherin gene was previously localised to 16q24 by FISH analysis (Lee, 1996) and was considered to be a favourable candidate for the TSG. Refined mapping of this gene to 16q24 using the chromosome 16 somatic cell hybrid panel demonstrated the localisation of H-cadherin to be at 16q24.2, proximal to the CY3 / CY2 breakpoint and therefore proximal to the region of LOH. Thus, this gene is not a candidate for the TSG localised to 16q24.3-qter.

The FAB consortium has now constructed an expressed sequence and gene map of the region of loss of heterozygosity at 16q24.3-qter. SSCP mutation analysis with breast tumour samples is currently being conducted with the individual genes and transcripts to determine whether any of them is the TSG.

CHAPTER 7

Isolation and Identification of Transcripts in the Familial Mediterranean Fever Candidate Region

7.1 INTRODUCTION 265

7.1.1 Refined Localisation of the MEF Gene 265

7.1.2 Aim of the Project 267

7.2 MATERIALS AND METHODS 271

7.2.1 Preparation of cDNA and Cosmid DNA 271

7.2.2 Hybridisation and Isolation of Specific cDNAs 272

7.2.3 Analysis of cDNA Clones 273

7.2.4 Sequencing of cDNA Clones 274

7.3 RESULTS 275

7.3.1 Analysis of cDNA Enrichment 275

7.3.2 Isolation and Identification of Transcribed Sequences 277

7.3.3 Localisation of cDNAs to Cosmid DNA 279

7.3.4 Sequence Analysis of Positive cDNA Clones 282

7.4 DISCUSSION 283

7.4.1 Progress in the Positional Cloning of MEF 286

7.1 INTRODUCTION

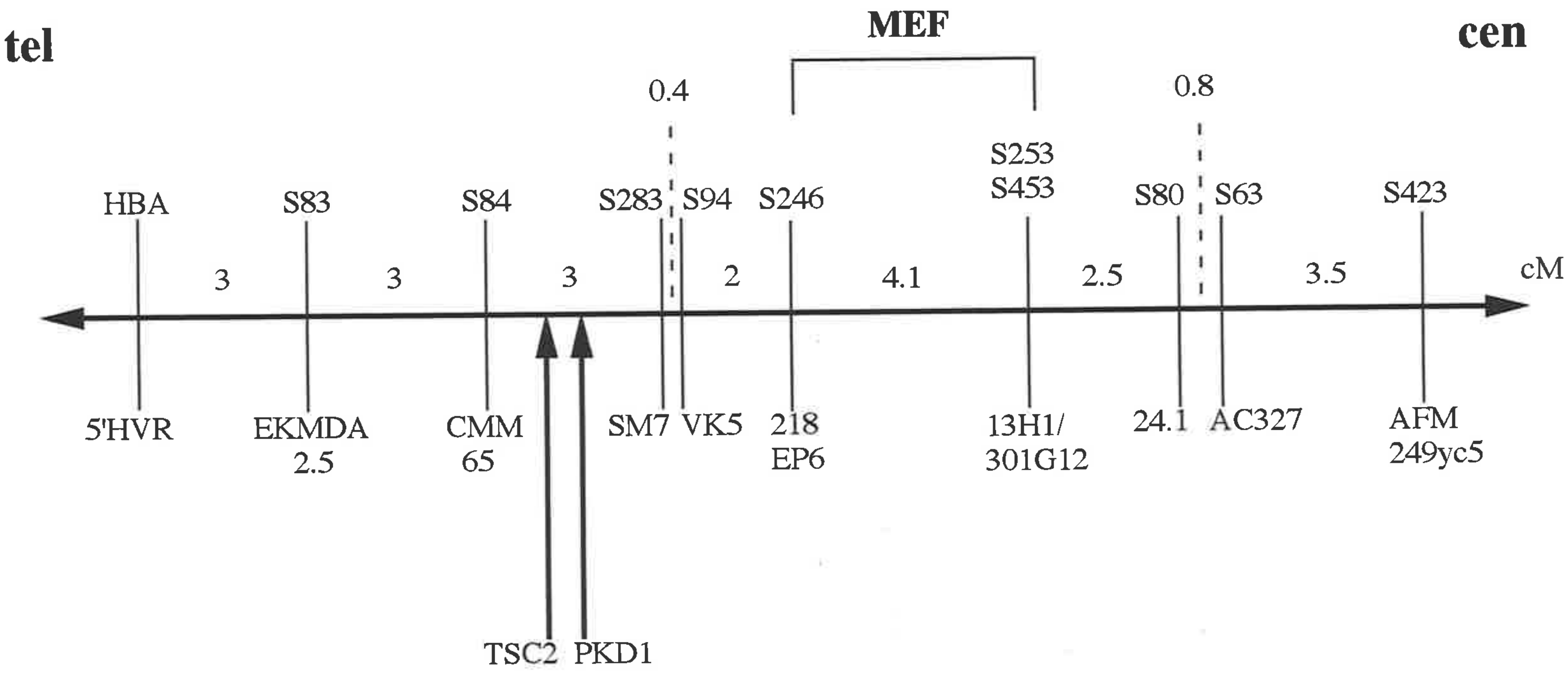
7.1.1 Refined Localisation of the MEF Gene

A consortium of laboratories (Levy *et al.* 1996) has been involved in refining the genetic localisation of the familial Mediterranean fever (FMF) susceptibility gene, MEF (see 1.7.2). A panel of 65 families of Jewish, Armenian and Arab ancestry were genotyped for 8 markers on chromosome 16p, including 5 not examined previously (figure 7.1). All markers showed linkage to MEF with maximal lod scores over all the families ranging from 14.6 to 40.7. The HOMOZ program (Kruglyak *et al.* 1995) uses a newly described algorithm for the analysis of recessive diseases in nuclear families. Multipoint lod scores were calculated for the 65 families between MEF and the 8 polymorphic markers. The maximal lod score of 49.2 was obtained at a location 1.6 cM centromeric to D16S246. Following the analysis of recombinant families, the most likely position of MEF is within the 4.1 cM (sex-averaged) interval between D16S246 and D16S523.

Linkage disequilibrium and haplotype analysis were utilised to determine the allelic distribution of markers in the population to extract additional genetic information from the FMF families. The panel of families used by Levy *et al* (1996) included two Jewish sub populations as well as Armenians and Arabs, making it possible to test whether common alleles or haplotypes were associated with the disease in these different ethnic groups. This is expected to be the case for markers close to the disease gene, if the haplotype is from a single ancestral chromosome. A strong haplotype association was previously seen among 76% of Moroccan Jews (Aksentijevich *et al* 1993a). This haplotype was also overrepresented, but to a lesser extent, 32%, in non-Moroccan Jews but was not associated with FMF in the Armenian or Arab populations. This haplotype, D16S283-D16S94-D16S246 was confirmed by Levy *et al* (1996). The closest flanking locus in this haplotype, D16S246, showed a highly significant association of a 2.5 kb allele among both Moroccan Jews and non-

Figure 7.1

Map of polymorphic markers on chromosome 16p. Sex-averaged map distances are based on CEPH data (Kozman *et al.* 1995) or are estimated from the study by Levy *et al* (1996).


tel
MEF
cen
0.4
0.8
S253
S453
HBA
S83
S84
S283
S94
S246
S80
S63
S423
3
3
3
2
4.1
2.5
3.5
cM
5'HVR
EKMDA
2.5
CMM
65
SM7
VK5
218
EP6
13H1/
301G12
24.1
AC327
AFM
249yc5
TSC2
PKD1

Moroccan Jews. This suggests that a large percentage of these Moroccan and non-Moroccan Jewish FMF patients may have the same ancestral mutation.

Luria-Delbruck formulas are used to estimate genetic distances from a disease locus to closely linked markers on the basis of linkage disequilibrium in recently founded populations in which there is one major disease causing haplotype. Luria-Delbruck analysis was successfully used in the positional cloning of the gene causing diastrophic dysplasia (Hastbacka *et al.* 1994) and was employed in the refined mapping of several recessively inherited disorders including autoimmune polyglandular disease type I (Aaltonen *et al.* 1994) and congenital nephrotic syndrome (Kestila *et al.* 1994). Thus, the strategy of refining the genetic localisation of a gene using Luria-Delbruck formulas was applied to the Moroccan linkage disequilibrium data (Levy *et al.* 1996) since there was good evidence that most Moroccan Jewish FMF carrier chromosomes were descended from a common ancestor. The method placed the MEF gene within 0.305 cM of D16S246 (2-lod-unit range 0.02-0.64 cM). Based on the genome wide average of 1 cM = 1 Mb, this suggests that MEF is within 640 kb of D16S246.

7.1.3 Aim of the Project

An international consortium of laboratories (The FMF consortium), including the Department of Cytogenetics and Molecular Genetics, Women's and Children's Hospital, South Australia, has been involved in the localisation of the MEF gene with Dr. D.L. Kastner's group in National Institute of Arthritis and Musculoskeletal and Skin Diseases (Bethesda, USA). The collaboration began in early 1992, when Kastner's group found the linkage between the markers D16S84 and D16S85, and the MEF gene on chromosome 16pl3.3 (Pras *et al.* 1992). More recently this consortium has refined the localisation of MEF (Levy *et al.* 1996) to within 0.305 cM of D16S246 through genetic analysis of the Moroccan Jewish population using additional microsatellite markers. The MEF gene is thought to be within 640 kb of D16S246.

The FMF consortium has continued to further refine the localisation of MEF. In May 1995, recombination data and new markers placed MEF between the microsatellite D16S468 and STS marker RT70 with no recombinants observed for STS marker 57E1. A specific haplotype shared between non-Moroccan and Moroccan families placed the gene between D16S468 and RT70 (see figure 7.2). Linkage disequilibrium data on Armenian, Arab and Turkish families suggested that the gene may be between 57E1 and RT70. Luria-Delbruck calculations and Poisson branching placed MEF closer to 57E1 than any other marker yet tested.

Members of the FMF consortium have constructed a cosmid contig extending from D16S246 to RT70 in order to identify the MEF gene. This contig is shown in figure 7.2. Initially, this region was covered by three contigs. Contig 1 extended approximately 500 kb centromeric from VK5 to 414G4, Contig 2 extended approximately 200 kb from RT70 telomeric to 367E12 and Contig 3 extended approximately 100 kb from RT194 to 441H9. Contigs 1, 2 and 3 were subsequently linked with cosmids identified by screening the ten times coverage chromosome 16 cosmid library (LANL) with inter-Alu products from YACs localised to this region. The total length of the FMF contig exceeds 1 Mb, with the FMF critical region being adjacent to 57E1.

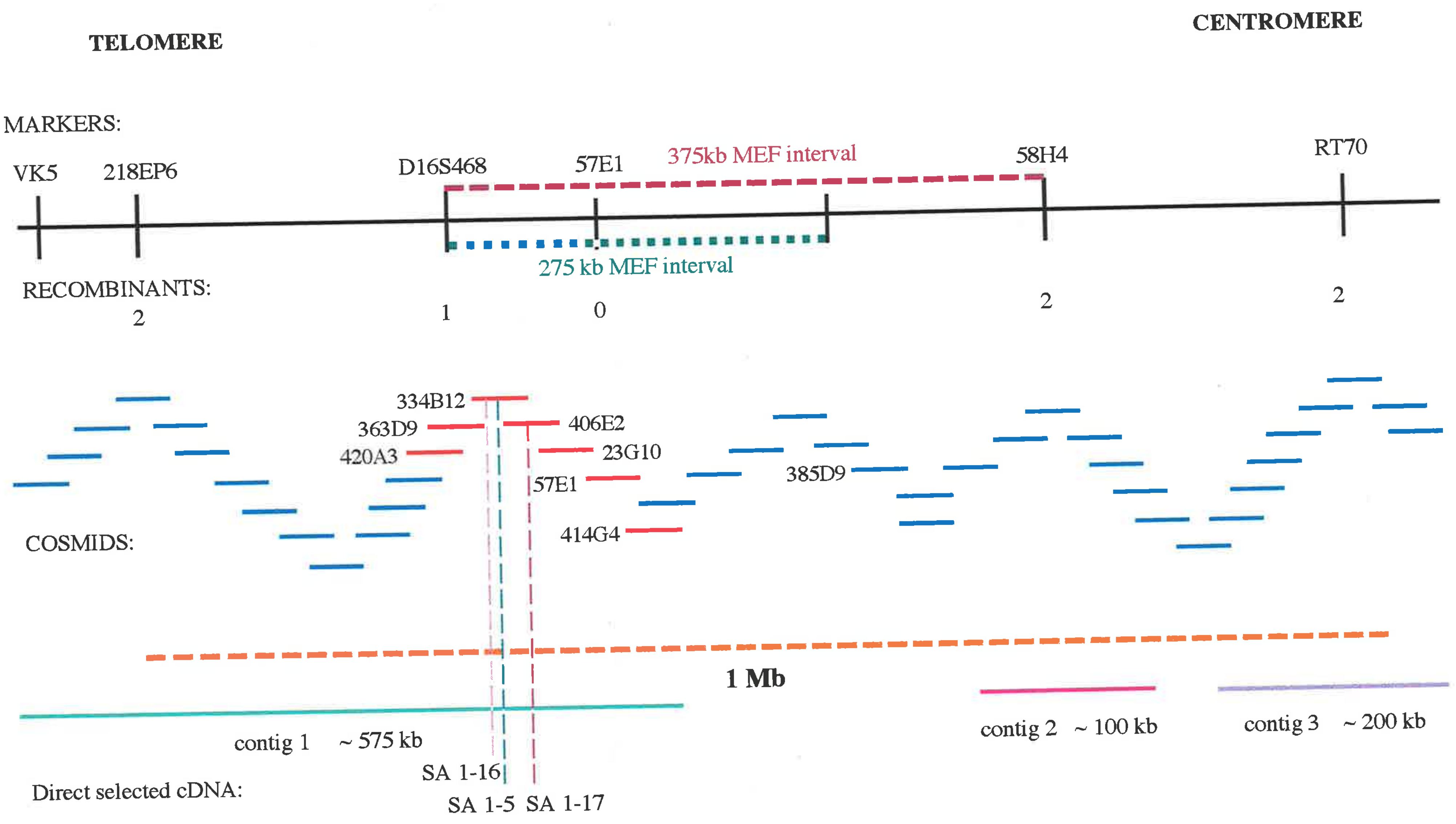
At the stage of my involvement in this project, the aim was to construct a transcript map in the FMF candidate region. My contribution to this project was to identify and isolate transcribed sequences encoded by cosmids localised to this region using the method of direct cDNA selection. Other groups in the consortium were using this technique with additional cosmids in the contig in addition to the complementary technique of exon trapping. The construction of a transcript map and the identification of genes in the region of interest is beneficial for the eventual identification of the MEF gene.

Figure 7.2

Cosmid contig extending over the 1 Mb FMF candidate region. The three cosmid contigs initially constructed are shown. Contig 1 extends approximately 500 kb centromeric from VK5 to 414G4. Contig 2, extends approximately 200 kb from RT70 telomeric to 367E12. Contig 3 extends approximately 100 kb from RT194 to 441H9.

The MEF interval has been refined to 275 kb, indicated in green, between D16S468 and 385D9 using the (AC)n marker from cosmid 385D9.

Cosmids in contig 1 which were utilised in the direct cDNA selection procedure are indicated in red. The positions of the direct selected cDNA clones in the contig are also indicated.


TELOMERE
CENTROMERE
MARKERS:
VK5
218EP6
D16S468
57E1
375kb MEF interval
58H4
RT70
275 kb MEF interval
RECOMBINANTS:
2
1
0
2
2
334B12
363D9
406E2
420A3
23G10
57E1
385D9
414G4
COSMIDS:
1 Mb
contig 2 ~ 100 kb
contig 3 ~ 200 kb
contig 1 ~ 575 kb
SA 1-16
Direct selected cDNA:
SA 1-5
SA 1-17

During the course of this work, the FMF consortium reported that the physical map of the FMF candidate region was essentially finished since 80% of the FMF candidate region was estimated to be represented by transcripts. Thus, characterisation of all the direct selected cDNAs isolated in this study and my contribution to the project concluded at this stage.

7.2 MATERIALS AND METHODS

7.2.1 Preparation of cDNA and Cosmid DNA

A modification of the method of direct cDNA selection described in chapter 4, section 4.2 was used. Inserts from approximately $2x10^6$ clones from a variety of tissue sources were amplified by PCR (2.6) using λgt10 or λgt11 vector primers, depending on the vector of the individual cDNA library, in a final volume of 100 µl. These libraries were 5' stretch foetal brain in λgt11 vector, 5' stretch endothelial vein in λgt11, 5' stretch kidney in λgt10, foetal kidney in λgt10 and lung fibroblast in λgt11. All libraries are oligo-dT and random primed and were purchased from Clontech. The primer sequences were as follows:

λgt10-F	5'- AGCAAGTTCAGCCTGGTTAAG -3',
λgt10-R	5'- CTTATGAGTATTTCTTCCAGGGTA -3'
λgt11-F	5'- GGTGGCGACGACTCCTGGAGCCCG -3'
λgt11-R	5'- TTGACACCAGACCAACTGGTAATG -3'

After an initial denaturation at 94°C for 2 minutes, thermal cycling was performed for 30 cycles under the following conditions: 94°C denaturation for 30 seconds, 55°C annealing for λgt10 primers, or 60°C annealing for λgt11 primers, for 30 seconds and 72°C extension for 1 minute and 30 seconds plus an additional 5 seconds extension per cycle using the automatic segment extension option on the 9600 Perkin Elmer DNA thermal cycler. A 10 minute extension was performed at the end of the last cycle.

A 10 µl aliquot of PCR amplified inserts from each library was analysed by electrophoresis on a 1.2% agarose gel (2.5.2) to determine the extent of the amplification. The PCR products were cleaned of excess primers, dNTPs and *Taq* enzyme using Wizard PCR prep minicolumns (Promega) according to the manufacturer's instructions, except that the columns were eluted with 50 µl TE at 60°C. The DNA was ethanol precipitated, lyophilised

(2.3.1.4.2) and resuspended in TE. The DNA concentration was determined (2.3.1.4.3) then adjusted to 1 µg/µl.

Cosmids were grown overnight in LB supplemented with kanamycin at 50 µg/ml. Cosmid DNA was extracted using Qiagen columns described in 2.3.1.1.2. Gel electrophoretic patterns of EcoRI digested cosmid DNAs were analysed after ethidium bromide staining to determine the integrity of the DNA.

One µg aliquots of DNA from each cosmid was digested (2.5.1) with each of the enzymes EcoRI and HindIII, in separate reactions for 2 hours at 37°C. The EcoRI and HindIII digests of each cosmid were pooled and 500 ng of this DNA was biotinylated by nick translation (BRL Bionick labelling system). The final concentration of the cosmid DNA was 20 ng/µl.

7.2.2 Hybridisation and Isolation of Specific cDNAs

Biotinylated cosmid DNAs were grouped together to form three independent pools. Each cDNA group contained one set of cDNA inserts derived from λgt11 vector and another set from λgt10 vector. Group 1 consisted of foetal kidney (λgt10) and 5' stretch foetal brain (λgt11); Group 2 consisted of 5' stretch kidney (λgt10) and lung fibroblast (λgt11) and Group 3 consisted only of 5' stretch endothelial vein (λgt11).

Human repeat and vector sequences of the cDNA library inserts and the genomic DNAs were each blocked by pre-reassociation with an excess of unlabelled human placental DNA added to a cot value of 20 and sCos1 vector DNA. Two µg of sheared human placental DNA and 500 ng of sCos1 DNA were added to 2 µg of each cDNA group. Identical concentrations of human placental DNA and sCos1 DNA were also added to 200 ng of biotinylated cosmid DNA. Hybridisation solution consisting of 0.12 M sodium phosphate pH7.0 was added to a total volume of 15 µl. The two DNA samples were denatured for 5 minutes at 95°C then hybridised for 4 hours at 60°C. The pre-reassociated cDNA (15 µl) and the pre-reassociated biotinylated cosmid DNA (15 μl) were hybridised in a total volume of 30 μl at 60°C for 24 hours.

Specific cDNAs bound to biotinylated cosmid DNAs were captured using streptavidin-coated magnetic beads (Dynal). The washing and elution procedures outlined in section 4.2.2, chapter 4 were followed. The cosmids and cDNAs were subjected to only one round of enrichment to generate libraries of selected cDNA fragments.

The selected cDNA fragments were analysed to determine whether enrichment had occurred. Ten μl of the selected cDNA fragments were PCR amplified (2.6) in 100 μl reactions under the conditions described in 7.2.1. The cDNA fragments from each tissue source, present in cDNA groups 1, 2 and 3, were PCR amplified using primers homologous to the vector sequence of each cDNA library (λgt10 or λgt11). These PCR products, along with those from each starting cDNA library, were electrophoresed on a 1% agarose gel (2.5.2) to compare the sequence complexities of the two cDNA samples corresponding to each library. This gel was subsequently Southern blotted (2.5.5.3) and hybridised with radiolabelled pre-reassociated 420A3 cosmid DNA (2.5.4.1, 2.5.4.2 and 2.5.6.2) to investigate the presence of sequences homologous to this cosmid in the enriched cDNA libraries.

7.2.3 Analysis of cDNA Clones

PCR products from one round of selection were cloned into pGEM-T vector (2.6.3). One hundred recombinant colonies were arrayed on ampicillin plates and allowed to grow overnight at 37°C. Colony lifts were performed (2.5.5.1) for use in screening the cDNA clones with radiolabelled total human DNA (2.5.4.1 and 2.5.6.1) to determine which clones contained high copy repetitive elements. Positive clones were eliminated from further analysis.

Inserts of the remaining cDNA clones were PCR amplified directly from colonies. Cells from each colony were transferred to the PCR reaction mixture using a sterile pipette tip. Inserts were amplified by PCR (2.6) using puc forward, 5'-TGTGAGCGGATAACAATTTCACACAGGA -3', and puc reverse, 5'-CACGACGTTGTAAAACGACGGCCAGT -3', vector primers in 50 µl reactions. After an initial denaturation at 94°C for 4 minutes, thermal cycling was performed for 30 cycles under the following conditions: 94°C denaturation for 1 minute, 55°C annealing for 1 minute and 72°C extension for 3 minutes. A 10 minute extension was performed at the end of the last cycle. Fifteen µl of each PCR product were electrophoresed on 1% agarose gels (2.5.2) to assess the sizes of the inserts.

To determine which cosmid fragments were specifically homologous to the cDNA aliquots of PCR amplified inserts of cDNA clones were purified by electrophoresis through 1% low melting point agarose gels (2.5.2). The bands were excised from the gel and subsequently used as probes.

Filters containing 500 ng of cosmid DNA restricted with either EcoRI, HindIII, HpaII or HaeIII were hybridised with radiolabelled total human DNA (2.5.4.1 and 2.5.6.1) to determine which cosmid fragments contained high copy repeats and with sCos1 to identify which bands contained vector sequence. The filters were subsequently probed with individual cDNA inserts labelled by primer extension (2.5.4.1 and 2.5.6.2).

7.2.4 Sequencing of cDNA Clones

DNA of cDNA clones was prepared using Qiagen columns (2.3.1.2.2). cDNA clones were sequenced using dye primer kits (Applied Biosystems Inc.) (2.7.2). The sequences were analysed using an automated DNA sequencer (ABI 373A) (2.7.4) and compared to nucleotide databases using the BLAST-N program. Statistically significant sequence similarities were considered to be of P < 1 e -0.05.

7.3 RESULTS

7.3.1 Analysis of cDNA Enrichment

Seven cosmids from contig 1, localised to the telomeric end of the FMF candidate region, including the critical region adjacent to 57E1 were utilised in the direct cDNA selection procedure. These cosmids are highlighted in red in figure 7.2. These overlapping cosmids spanned 181 kb of genomic DNA. Libraries of selected cDNA fragments were generated using two independent cosmid pools. Cosmid pool 1 consisted of 363D9, 334B12, 420A3; and cosmid pool 2 consisted of 23G10, 406E2, 57E1, 414G4. Visualisation of the PCR products from one round of enrichment on an agarose gel demonstrated a smear of DNA fragments ranging in size between 0.3-1.1 kb. The sizes of the enriched cDNAs from the various libraries were smaller than the cDNA inserts of the starting libraries, whose sizes were up to 2 kb. This was an indication that the sequence complexity in the selected libraries was reduced when compared to the starting libraries. PCR products were not observed in control reactions in which template was not added (data not shown).

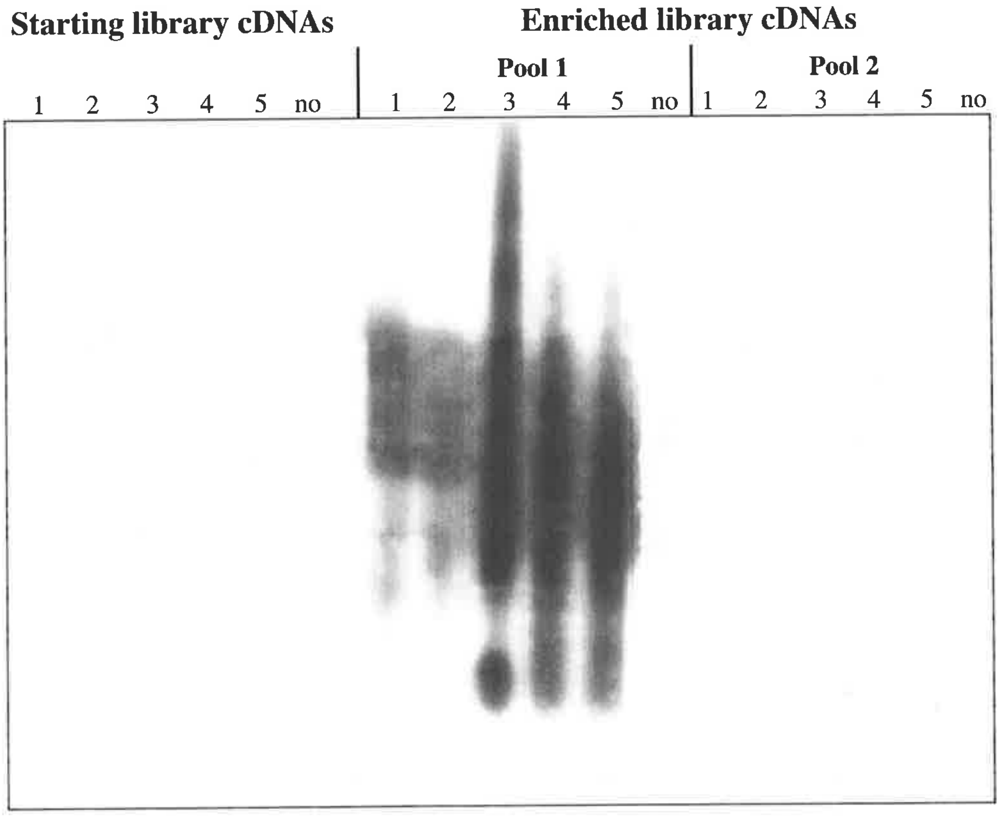
To determine whether the enriched cDNA sequences were homologous to the genomic DNA from which they were derived, a Southern blot consisting of the PCR products from each enriched library together with PCR products from each starting cDNA library was hybridised with pre-reassociated 420A3 cosmid, which was a member of cosmid pool 1. The results are shown in figure 7.3. A positive signal was demonstrated in the enriched cDNA products from cosmids in pool 1, but not pool 2. The tissue sources in which a positive signal was most prominent were lung fibroblast, foetal brain and endothelial vein. This indicated that the signal was due to the specific homology of the amplified cDNAs to the 420A3 cosmid DNA, and not to repeat or non-specific DNA sequences. The signal was also stronger for the enriched PCR products compared to the starting cDNAs which suggested that the selection procedure was successful in the enrichment of specific cDNA products.

Figure 7.3

PCR products from each starting cDNA library and enriched cDNA library hybridised with pre-reassociated 420A3 cosmid, a member of cosmid pool 1. The first group of six lanes represent PCR amplified inserts from starting cDNA libraries. The second group of six lanes represent PCR amplified inserts from the enriched cDNA libraries derived from cosmids in pool 1. The third group of six lanes represent PCR amplified inserts from cDNA libraries derived from cosmids in pool 2.

Lanes 1-6 represent PCR amplified inserts from:
foetal kidney cDNA library, 5' stretch kidney cDNA library, lung fibroblast cDNA library, 5' stretch foetal brain cDNA library, 5' stretch endothelial vein cDNA library, no DNA.

A positive signal is demonstrated only in the enriched cDNA products and not in the starting libraries. The tissue sources in which a positive signal was most prominent were lung fibroblast, foetal brain and endothelial vein. Signal is not observed in the no DNA control.


Starting library cDNAs
Enriched library cDNAs
Pool 1
Pool 2
1 2 3 4 5 no
1 2 3 4 5 no
1 2 3 4 5 no

cDNA enrichment was also investigated by including positive control cosmids, 340E8, 380C10, and 315F12, containing the N-acetyl galactosamine-6-sulfatase (GALNS) gene in the experiment (as in 4.3.2, chapter 4). After one round of selection, these PCR products were cloned in pGEM-T vector. Colony lifts of 100 arrayed recombinant colonies were screened for homology to total human DNA. Thirteen clones (13%) containing high copy repetitive elements were identified and eliminated from subsequent analysis. Probing the remainder of the clones with a full length single copy GALNS cDNA clone identified 43 positive clones, or 50% of the non-repetitive clones, which demonstrated that enrichment of the GALNS cDNA from a total cDNA library was successful. These clones were not analysed further.

7.3.2 Isolation and Identification of Transcribed Sequences

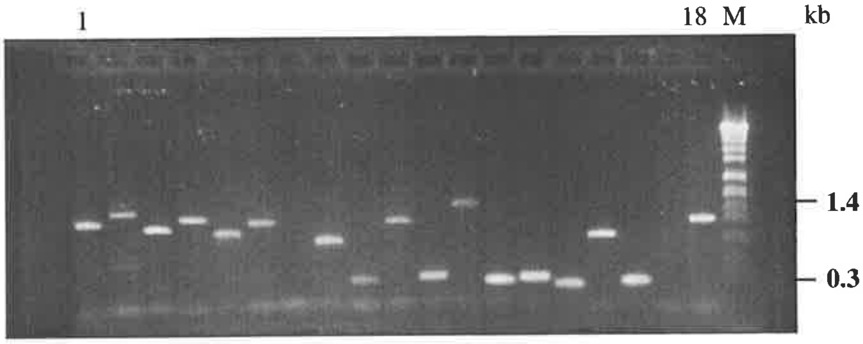
The initial study involved the characterisation of the enriched products derived from the direct selection of cDNAs from the 5' stretch foetal brain library. Instead of screening a normalised library with the enriched PCR products to obtain longer cDNA clones, as in 4.2.2 (chapter 4), the products were cloned into the pGEM-T vector. Colony lifts of 100 arrayed recombinant colonies from each cosmid pool were screened for homology to total human DNA to identify clones containing high copy repetitive elements. A total of 18 clones (18%) derived from cosmids in pool 1, and 17 clones, (17%), from pool 2, were identified and eliminated from subsequent analysis.

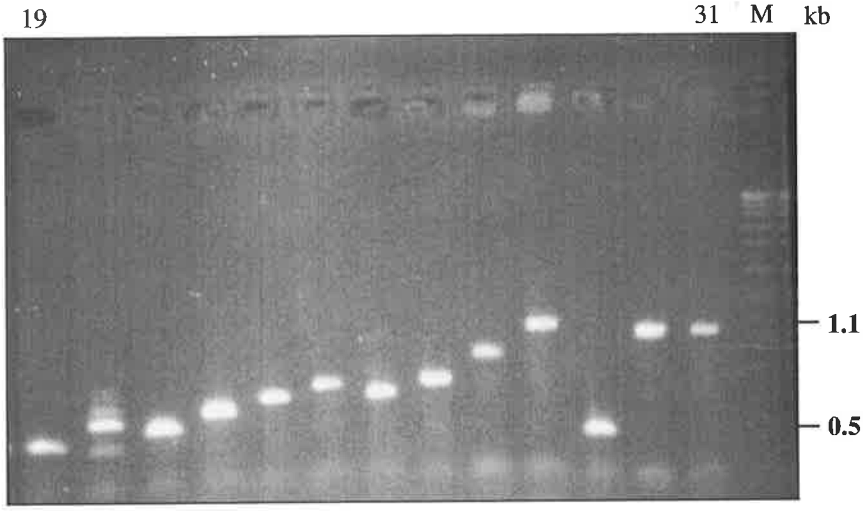
Inserts of the remaining clones were sized on 1% agarose gels. The sizes of the inserts of the cDNA clones, including pGEM-T vector sequence surrounding the cloning site, ranged from about 0.3 kb to 1.2 kb. An example of the gels with PCR amplified cDNA inserts are shown in figure 7.4.

Figure 7.4

A selection of 31 PCR amplified inserts from the cloned enriched cDNA fragments from the foetal brain cDNA analysed on 1% agarose gels. The sizes of the inserts of the cDNA clones, including pGEM-T vector sequence surrounding the cloning site, range from about 0.3 kb to 1.2 kb.

The M lane represents Spp-1 markers.


1
18
M
kb
1.4
0.3


19
31
M
kb
1.1
0.5

7.3.3 Localisation of cDNAs to Cosmid DNA

To determine whether the cDNAs were homologous to the cosmid DNA from which they were derived, each non-repetitive purified cDNA insert was labelled and used as a probe on Southern blots containing EcoRI, HindIII, HaeIII or HpaII digests of the cosmids. It was assumed that inserts of similar size represented the same cDNA fragment, thus these cDNAs were grouped into 33 groups and one clone from each group was analysed. This step was undertaken to avoid analysing the same cDNA fragment repeatedly. Of these cDNAs, three had a hybridisation pattern consistent with the recognition of specific cosmid fragments. Clone SA 1-5 is homologous to cosmids 334B12 and 406E2. Clone SA 1-16 is homologous to cosmid 334B12. Clone SA 1-17 is homologous to cosmid 406E2. Examples of these hybridisations are shown in figure 7.5. The sizes of clones SA 1-5, SA 1-16 and SA 1-17 are 74 bp, 144 bp and 150 bp respectively. A list of these cDNA clones and the cosmid fragments to which they are homologous is presented in table 7.1. The positions of these clones in the cosmid contig spanning the FMF candidate region are shown in figure 7.2.

Eight of the cDNA clones analysed were demonstrated to have hybridisation patterns suggesting that they recognised human low copy repetitive elements. Five clones demonstrated homology to vector sequence only. Inserts of clones SA 1-5, SA 1-16 and SA 1-17 were hybridised individually to the membranes containing the arrayed recombinant clones from the enriched libraries to identify homologous clones. A significant number of the clones were positive for these probes (6 clones for SA 1-5, 6 clones for SA 1-16, 5 clones for SA 1-17), indicating the presence of multiple copies of these selected cDNAs.

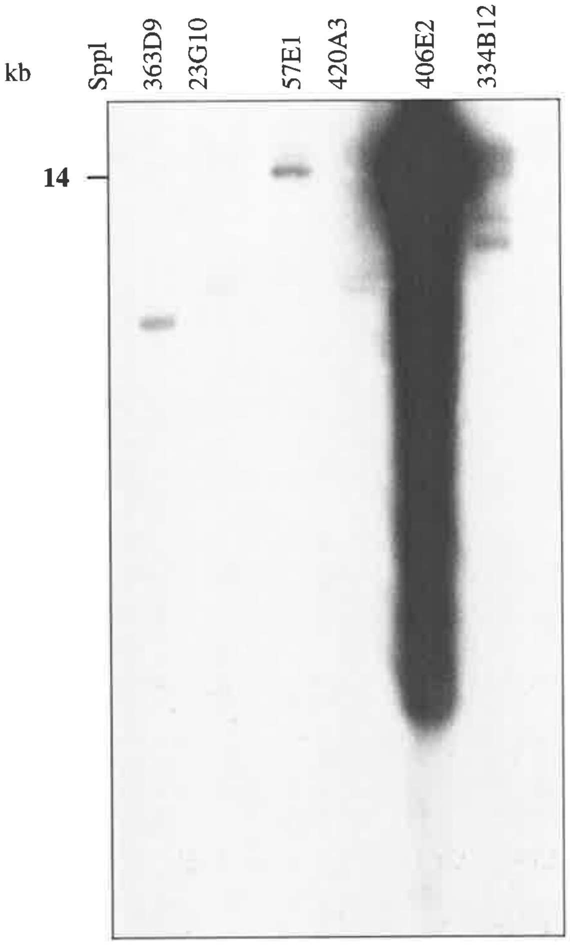
Figure 7.5

Hybridisation of selected cDNA clones to cosmid DNAs. The autoradiographs represent hybridisations of cDNA clones to restricted cosmid DNA. These are as follows:

1. cDNA clone SA 1-17 hybridised to EcoRI digested cosmid DNA.
   This clone demonstrates homology to cosmid 406E2.

2. cDNA clone SA 1-5 hybridised to EcoRI digested cosmid DNA.
   This clone is homologous to cosmids 334B12 and 406E2.

SA1-17


kb
Sppl
363D9
23G10
57E1
420A3
406E2
334B12
14

SA1-5

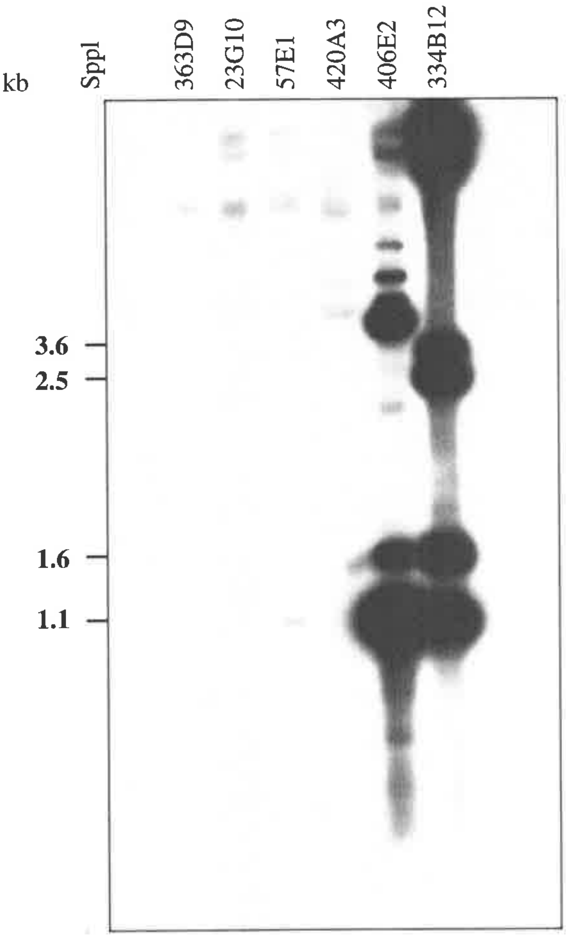

kb
Sppl
363D9
23G10
57E1
420A3
406E2
334B12
3.6
2.5
1.6
1.1

Table 7.1

A list of cDNA clones identified by direct cDNA selection using cosmids spanning 181 kb in the FMF candidate region. The sizes and sequences of each cDNA clone are indicated. Cosmids to which these cDNA clones are homologous are also shown.

cDNA Clone	**Size (bp)**	**Homologous Cosmids**	
SA 1-5	74	334B12 406E2	CTGTGGTCAC CGTCTCTCCT GGACCTGCCT AGATCCCAAA GCCAGCCCTG GAAGGAACAC CTCTCATTCT CAAG
SA 1-16	144	334B12	TTGCAGAGCC TAGGGGCCAG GAGGAACAGC TGGGAGGGCA CTTCTCACCA TTTCTGAGGC CACGCTCAAC CGGCCCGGCA GAGGCTCCCG TCTTCCTCCA CTCCGTACCT CCACTGACTC CCACAGACCA CCCTCCCCAT TCAG
SA 1-17	150	406E2	TTCCAGGTCG TGGACATCCT CATCAGGAAA TGCCTTCTAA GCTGGGGGAG GCGGTACCTT CAGGGGACGC TCAGGAGTCA CTGCACATTA AGATGGAGCC CGAAGAGCCA CACTCCGAGG GGGCATCGCA GGAGGATGGG GCTCCAAGGT

7.3.4 Sequence Analysis of Positive cDNA Clones

DNA sequence from both ends of each of the cDNA clones, SA 1-5, SA 1-16 and SA 1-17, was obtained and compared to the accessible nucleotide databases using the BLAST-N program. The sequences are shown in table 7.1. The sequences identified by this analysis did not show significant homologies to any known sequence deposited in the databases.

7.4 DISCUSSION

The direct cDNA selection procedure was utilised for the identification of transcribed sequences in the FMF candidate region. These isolated cDNA clones may serve as candidate genes for MEF. Two pools of cosmids covering 181 kb of genomic DNA were used to identify three cDNA clones ranging in size from 74 bp to 150 bp, which hybridised specifically to individual cosmids. It is unknown whether these cDNA fragments represent pseudogenes or transcribed sequences as the methodology of direct cDNA selection is based on hybridisation by sequence homology. Further mapping, expression studies and sequencing can determine whether the cDNAs are actively transcribed genes or pseudogenes.

The results of the hybridisation of cosmid 420A3, a member of pool 1, to cDNAs from the starting libraries and the enriched libraries derived from cosmid pools 1 and 2 demonstrated specific hybridisation to the cDNAs enriched from cosmids in pool 1. Signal was not observed in the starting libraries or in enriched cDNAs derived from pool 2 cosmids. This suggested that conditions utilised in the direct cDNA selection procedure were successful in the amplification and isolation of specific cDNA fragments. Additionally, these results indicated that one round of enrichment, instead of two rounds which were performed for the study described in chapter 4, was sufficient to isolate specific cDNA fragments.

The selection conditions utilised were also demonstrated to be successful in enriching GALNS cDNA clones from the starting cDNA library. The frequency of clones containing high copy repeats was 13% which is similar to frequencies found in other studies. These frequencies vary from 1% to 20-50% (Lovett *et al.* 1991; Fan *et al.* 1993; Baens *et al.* 1995; Cheng *et al.* 1994). In the experiment described in chapter 4, only the repetitive sequences of the cDNA inserts were suppressed, which led to the identification of a high frequency, 30-50%, of the selected cDNA clones containing high copy repeats. In this study, the repetitive sequences of both the cDNA inserts and the genomic DNAs were suppressed in order to reduce the number of cDNA clones containing high copy repeats. The frequency of selected cDNA clones containing high copy repeats from cosmid pools 1 and 2 was demonstrated to be 18%. This frequency may be attributed to blocking the repeats in both DNA sources. However, this frequency could be a consequence of the one round of selection that was performed in this study.

The approach of screening cDNA clones of an arrayed normalised cDNA library with pools of selected cDNA PCR products to identify long cDNA clones, described in chapter 4, was not used in this study. Instead, the PCR products were cloned and subsequently analysed. The purpose of this was to avoid analysis of cDNA clones which may contain repetitive elements in the 3' untranslated region (UTR). The disadvantage of this approach is that screening cDNA libraries is required for the identification of longer cDNA clones. However, the random primed foetal brain cDNA library used in this study should possess overlap between cDNA clones, therefore is the cDNA source of choice for building a collection of overlapping molecules to obtain long cDNA clones.

An estimation of the number of clones expected to be isolated from the cosmids in this study can be calculated according to results obtained from previous studies. The overall density of selected cDNA fragments, with an average size of 500 bp, has been demonstrated to be one in every 20 kb in cosmids (Petersen *et al.* 1994; Cheng *et al.* 1994). Thus, approximately nine cDNA clones may be expected to be isolated from these cosmids which cover 181 kb of genomic DNA. However, it must be emphasised that gene density can vary considerably in different regions of the genome and this number is only an estimate.

Three cDNA clones which hybridised to individual cosmids were identified in this study. This small number of identified clones may be related to the fact that one round of enrichment may not have been sufficient to amplify the necessary transcripts. It is thought that enrichments between 1,000 fold and 10,000 fold can be achieved in the first cycle of selection (Lovett *et al.* 1991; Parimoo *et al.* 1991). In this study, two rounds of selection may have been more appropriate. Also, the clones of identical size which were considered to contain identical sequence and were eliminated from analysis, may have contained different sequences. Thus, further analysis of additional clones may lead to the identification of more transcripts. Another explanation for the small number of identified clones may be because the amplified commercial foetal brain cDNA library which was utilised in this study was of limited complexity. The ratio of target molecules to molecules of selectable cDNA depends on the abundance of the cDNA in the amplified insert library. This abundance varies depending on a number of factors including the expression of the cDNA in a particular tissue source, the length of the exons and the hybridisation conditions (Korn *et al.* 1992).

In this study, enriched products from a commercial foetal brain cDNA library were analysed. In order to select all encoded cDNAs from a given region, the ideal would be to use a cDNA library representing all possible transcribed sequences. In this regard pooled libraries from multiple tissue sources and from various developmental stages, tagged with different linkers, can be used to approach a complete representation cDNA library. Thus, the products selected from the other cDNA libraries used in this study should be analysed to determine whether additional cDNA fragments can be identified.

The three transcripts identified by direct cDNA selection were found to be identical to exons trapped by other members of the FMF consortium. SA 1-5 is identical to PL81, SA 1-16 is identical to JMA and SA 1-17 is identical to S468CE2. This observation was promising as it indicated that the two approaches, exon trapping and direct cDNA selection, used in different laboratories were producing the same results. Comparison of the two techniques demonstrated that all clones identified by exon trapping are not identified by direct cDNA selection and vice versa (Yaspo *et al.* 1995). Thus, these two techniques are complementary to each other and it is beneficial to use them both for the isolation of transcripts in regions of interest.

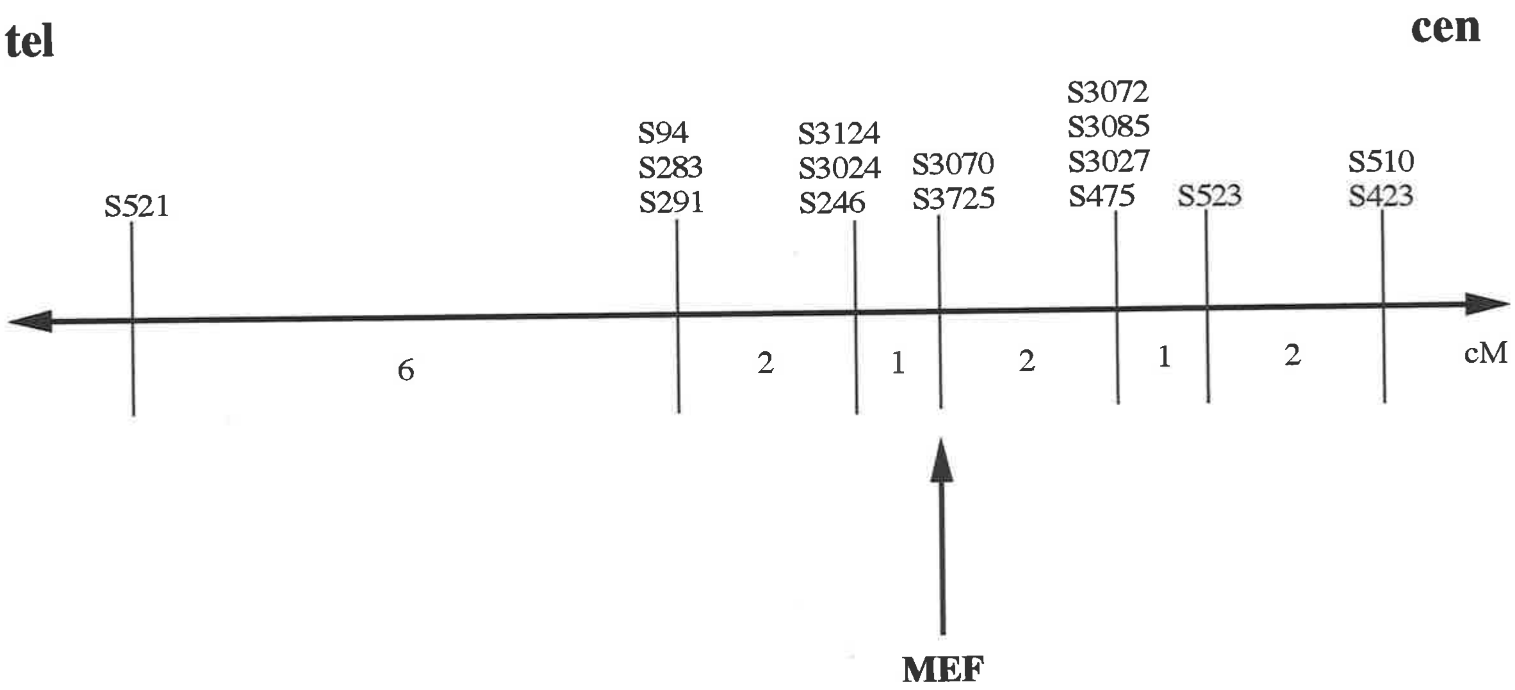
7.4.1 Progress in the Positional Cloning of MEF

In February 1996, the FMF consortium isolated six new microsatellite markers from the cosmids in the contig. The (AC)n marker from cosmid 385D9 refined the MEF interval from 375 kb to 275 kb, between D16S468 and 385D9 (figure 7.2). Additional genetic markers include 41E08 which maps to the same EcoRI restriction fragment as 57E1 and 49B4 which is approximately 15 kb centromeric of D16S468. The order of these markers is tel-D16S468-49B4 (57E1)-385D9-cen. Three recombinants have been identified at 385D9 and 49B4 in the North African Jewish population comprising Moroccans, Libyan, Tunisians and Egyptians.

In September 1996, the French FMF Consortium published an extensive study on 26 non-Ashkenazi Jewish families to also refine the location of MEF by using new markers mapping to 16p13.3. Haplotype analysis relative to critical recombination events placed the MEF locus between D16S3124 (telomeric) and D16S475 (centromeric) with the integration of 8 polymorphic markers in this map, 5 of which were incorporated between D16S523 and MEF (figure 7.6). Colocation of D16S246, the known distal limit of the MEF region (Levy *et al.* 1996), and D16S3124 in the same YAC was consistent with genetic linkage analysis placing both D16S246 and D16S3124 1-2 cM telomeric to MEF. Two markers, D16S3070 and D16S3275, a microsatellite marker isolated from a YAC that also contains D16S3070, showed no recombination with the disease. Linkage disequilibrium and haplotype analysis demonstrated a founder haplotype in this population. Alleles bearing D16S3275 were found to be derived from the founder haplotype. The ancestral alleles were preserved at the core loci D16S3070 and D16S3275 in 71% of the chromosomes studied. Furthermore, identification of ancestral recombination events in these pedigrees indicated that MEF is located between these two loci. These markers are found in a 250 kb genomic fragment. These data support Kastner's FMF consortium results.

Figure 7.6

Polymorphic markers of the FMF region. Sex-averaged intermarker distances are estimated from CEPH data (Dib *et al.* 1996 and The French FMF consortium study (1996), except for the D16S94 marker, which was placed from meiotic analysis of MEF families (Aksentijevich *et al.* 1993b; Levy *et al.* 1996). D16S246, a microsatellite marker, and D16S3275 were physically mapped.


tel
cen
S521
S94
S283
S291
S3124
S3024
S246
S3070
S3725
S3072
S3085
S3027
S475
S523
S510
S423
6
2
1
2
1
2
cM
MEF

Recent data from the FMF consortium shows that the 1 Mb FMF candidate region now consists of exons and cDNA fragments which have been grouped into twelve distinct transcripts. It has been estimated that 80% of the FMF candidate region is represented by transcripts. The FMF consortium found that both the exon trapping and direct cDNA selection procedures have their limitations. The FMF consortium identified 20% of the 1213 clones from direct cDNA selection to contain either repeat or ribosomal protein sequences. 20-30 genes have been identified with a high rate of redundancy. The isolation of long transcripts using these techniques has also been difficult. The FMF consortium is now concentrating on obtaining full-length transcripts using the 5' RACE procedure. The clones are also being screened for expression by RT-PCR and Northern analysis to determine whether they are true transcribed sequences.

Mutation analysis to determine whether any of the isolated transcripts represents the MEF gene is also being performed by the FMF consortium, but thus far mutations have not been identified in any of the transcripts analysed. The isolation of MEF and identification of mutations in this gene will hopefully shed light both on the clinical variability in FMF and the pathophysiology of the periodic initiation and termination of the inflammatory pathways, which are characteristic hallmarks of the disease.

CHAPTER 8

Concluding Remarks

CONCLUDING REMARKS

The development of resources and technologies in the Human Genome Project has led to the construction of high resolution genetic, physical and transcript maps that have facilitated the identification of disease genes through the approach of positional cloning. Positional cloning begins with the localisation of a disease gene to a particular region in the genome by genetic linkage analysis, and is followed by molecular analysis of this region. This includes finer genetic mapping, physical mapping, DNA isolation, transcript identification, cDNA cloning and mutation analysis of candidate genes.

The construction of a detailed physical map of the 16q24 chromosomal region was presented in this thesis. Such physical maps form the basis for attempts to positionally clone candidate disease genes to specific chromosomal regions. The focus to map the 16q24 chromosomal region was in view of its high gene density and the localisation of two disease genes to the 16q24.3-qter region. The Fanconi anaemia group A gene (FAA) was mapped to 16q24.3-qter by linkage analysis. Fanconi anaemia (FA) is a rare autosomal recessive disorder associated with progressive bone marrow failure and an increased risk of developing acute myeloid leukaemia or a variety of solid tumours, and FA-A is the most prevalent complementation group observed in FA patients (Buchwald, 1995). Also, a tumour suppressor gene (TSG) was localised to 16q24.3-qter by loss of heterozygosity (LOH) studies in sporadic breast tumours. Hence, work towards the positional cloning of FAA and the TSG was presented in this thesis.

The first step toward constructing the physical map of 16q24 and the positional cloning of FAA and the TSG was achieved with an Alu PCR strategy to identify and localise genomic cosmid clones to this chromosomal region. This involved PCR amplification of human sequences from two human/rodent somatic cell hybrids that contained the distal part of 16q as the only human chromosome 16 material. These Alu PCR amplified products were hybridised to a chromosome 16 specific cosmid library to identify cloned DNAs from this predefined chromosomal region. The amount of 16q24 represented by these cosmids, together with their contigs, was estimated to be 2 megabases. Further FAA mapping and LOH studies using sporadic breast tumour samples indicated that FAA and the TSG were localised in the 16q24.3-qter region of chromosome 16, therefore, attention was focussed on this region. Additional cosmids mapping to the 16q24.3-qter region were identified by using expressed sequences, microsatellite repeats, and cosmid ends of singleton cosmids and cosmid contigs already mapped to this region to probe the chromosome 16 cosmid library. This ultimately resulted in the construction of a cosmid contig that extended over 650 kb in the 16q24.3-qter region.

Further resources for the physical map and positional cloning of FAA and the TSG were obtained through the approach of direct cDNA selection to identify transcribed sequences encoded by cosmids localised to 16q24. These transcribed sequences were used to construct a transcript map that greatly benefited the efforts to positionally clone the TSG and FAA assigned to this region of interest. Five cDNA clones were identified to be encoded by cosmids localised to 16q24.

Two novel cDNA clones, yc81e09 and yh09a04, localised to 16q24.3-qter were further characterised as they were possible candidates for FAA and the TSG. Clone yc81e09 detected a transcript of approximately 3.7 kb in size and clone yh09a04 detected transcripts of several sizes including one of 5.5 kb by Northern analysis. Full length sequence of the 5.5 kb transcript corresponding to clone yh09a04, which was isolated by using the direct cDNA selection technique, was obtained. The sequence of the 5' end of clone yh09a04 demonstrated identity to the sequence of the 3' end of a cDNA clone, yf14a03, that was deposited in the GenBank database. The combined sequence of the overlapping yf14a03 and yh09a04 clones was 2140 bases and corresponded to the 3' end of the 5.5 kb transcript. Partial sequence of the 3.7 kb transcript, corresponding to clone yc81e09, was also obtained. Comparison of these sequences to sequences in accessible nucleotide and protein databases, corresponding to yh09a04 and no differences were detected for yc81e09 in the segments that were investigated. Thus, it is likely that these transcripts are not candidates for the TSG localised to 16q24.3 and further work is required to successfully complete the positional cloning of this gene. This may include the isolation or use of new polymorphic markers that may further restrict the region of LOH at 16q24.3-qter in sporadic breast tumour samples. Moreover, SSCP mutation analysis of additional transcripts localised to this region, using sporadic breast tumour samples with restricted LOH at 16q24.3-qter will also be conducted. These transcripts will include the complete 3.7 kb transcript corresponding to clone yc81e09. The identification of the TSG will have an impact on the diagnosis of individuals and the development of new therapeutic strategies. It will also provide more knowledge of the functional roles of both TSGs and oncogenes in the genesis of cancer and is critical for the understanding of cancer progression.

This thesis also presented contributions to the positional cloning of the MEF gene responsible for familial Mediterranean fever (FMF) localised to 16p13.3 by genetic linkage mapping. The international FMF consortium constructed a YAC/cosmid contig encompassing the FMF candidate region. Direct cDNA selection was also applied to the positional cloning of MEF to identify and isolate transcripts encoded by cosmids localised to the FMF candidate region and to aid the construction of a transcript map. Three transcripts demonstrated homology to cosmids from which they were derived. The FMF consortium is continuing with positional cloning of the MEF gene by screening transcripts mapped in the FMF candidate region for disease specific mutations. The isolation of MEF and identification of mutations in this gene may provide insights to the clinical variability in FMF and the pathophysiology of the inflammatory pathways which are characteristic of the disease.

In conclusion, the approach of screening the chromosome 16 specific cosmid library with Alu PCR products from chromosome 16 somatic cell hybrids was successful in the isolation of cosmids in a specific region of chromosome 16. These cosmids formed the basis for generating a cosmid contig of approximately 650 kb in the 16q24.3-qter region. The direct cDNA selection technique was modified by utilising a pool of selected cDNAs to screen a normalised infant brain cDNA library. The yh09a04 cDNA clone, isolated by using this approach, was identified as the 3' end of the FAA gene. Two transcripts, yh09a04 and yc81e09, were largely eliminated as candidates for the TSG involved in sporadic breast cancer by SSCP analysis of breast tumour samples displaying restricted LOH at 16q24.3-qter.

Barakat, M.H., Karnik, A.M., Majeed, H.W.A., el-Sobki, N.I., Fenech, F.F. (1986).
Familial Mediterranean fever (recurrent hereditary polyserositis) in Arabs: a study of 175 patients and review of the literature.
Q. J. Med. 60: 837-847.

Barber, J.C.K., Mahl, H., Portch, J., Crawford, M. (1991).
Interstitial deletions without phenotypic effect: prenatal diagnosis of a new family and brief review.
Prenat. Diagn. 11: 411-416.

Bates, G., Lehrach, H. (1994).
Trinucleotide repeat expansions and human genetic disease.
Bioessays 16: 277-284.

Beckmann, J.S., Tomfohrde, J., Barnes, R.I., Williams, M., Broux, O., Richard, I., Weissenbach, J., Bowcock, A.M. (1993).
A linkage map of human chromosome 15 with an average resolution of 2 cM and containing 55 polymorphic microsatellites.
Hum. Mol. Genet. 2: 2019-2030.

Behrens, J. (1993).
The role of adhesion molecules in cancer invasion and metastasis.
Breast Cancer Res. Treat. 24: 175-184.

Benton, W.D., Davis, R.W. (1977).
Screening lambda gt recombinant clones by hybridisation to single plaques *in situ*.
Science 196: 180-182.

Bergerheim, U.S.R., Kunimi, K., Collins, V.P., Ekman, P. (1991).
Deletion mapping of chromosome 8, 10, and 16 in human prostatic carcinoma.
Genes Chrom. Cancer 3: 215-220.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., Rodier, F. (1985).
The mosaic genome of warm blooded vertebrates.
Science 228; 953-968.

Bernardi, G. (1989).
The isochore organization of the human genome.

Annu. Rev. Genet. 23: 637-661.

Berry, R., Stevens, T.J., Walter, N.A., Wilcox, A.S., Rubano, T., Hopkins, J.A., Weber, J., Goold, R., Soares, M.B., Sikela, J.M. (1995).
Gene-based sequence-tagged-sites (STSs) as the basis for a human gene map.
Nature Genet. 10: 415-423.

Berx, G., Cleton-Jansen, A.M., Nollet, F., de Leeuw, W.J., van de Vijver, M., Cornelisse, C., van Roy, F. (1995).
E-cadherin is a tumour/invasion suppressor gene mutated in human lobular breast cancers.
EMBO J. 14: 6107-6115.

Berx, G., Cleton-Jansen, A.M., Strumane, K., de Leeuw, W.J., Nollet, F., van Roy, F., Cornelisse, C. (1996).
E-cadherin is inactivated in a majority of invasive human lobular breast cancers by truncation mutations throughout its extracellular domain.
Oncogene 13: 1919-1925.

Bilofsky, H.S., Burks, C. (1988).
The GenBank genetic sequence data bank.
Nucl. Acids Res. 16: 1861-1863.

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

Birnboim, H.C., Doly, J. (1979).
A rapid alkaline extraction procedure for screening recombinant plasmid DNA.
Nucl. Acid Res. 7: 1513-1522.

Bodmer, W. (1981).
Gene clusters, genome organization and complex phenotypes. When the sequence is known what will it mean?
Am. J. Hum. Genet. 33: 664-682.

Boguski, M.S., Lowe, M.J., Tolstoshev, C.M. (1993).
dbEST database for 'expressed sequence tags'.
Nat. Genet. 4: 332-333.

Botstein, D., White, D.L., Skolnick, M., Davis, R.W. (1980).
Construction of a genetic linkage map in man using restriction fragment length polymorphisms.
Am. J. Hum. Genet. 32: 314-331.

Brennan, M.B., Hochgeschwender, U. (1995).
So many needles, so much hay.
Hum. Mol. Genet. 4: 153-156.

Britten, R.J., Baron, W.F., Stout, D.B. Davidson, E.H. (1988).
Source and evaluation of human Alu repeated sequences.
Proc. Natl. Acad. Sci. USA 85: 4770-4774.

Britten, R.J., Davidson, E.H. (1985).
Nucleic acid hybridization - A practical approach (Eds. Hames, B.D. and Higgins, S.J.)
IRL, Oxford. pp 3-15.

Bronson, S.K., Pei, J., Taillon-Miller, P., Chroney, M., Geraghty, D.E., Chaplin, D. (1991).
Isolation and characterization of yeast artificial chromosome clones linking the HLA-B and HLA-C loci.
Proc. Natl. Acad. Sci. USA 88: 1676-1680.

Brown, W.R.A., Bird, A.P. (1986).
Long-range restriction site mapping of mammalian genomic DNA.
Nature 322: 477-481.

Buchwald, M. (1995).
Complementation groups: one or more per gene?
Nature Genet. 11: 228-230.

Buckler, A.J., Chang, D.D., Graw, S.L., Brook, J.D., Haber, D.A., Sharp, P.A., Housman, D.E. (1991).
Exon amplification: a strategy to isolate mammalian genes based on RNA splicing.
Proc. Natl. Acad. Sci. USA 88: 4005-4009.
Buetow, K.H., Weber, J.L., Ludwigsen, S., Scherpbier-Heddema, T., Duyk, G.M., Sheffield, V.C., Wang, Z., Murray, J.C. (1994).
Integrated human genome-wide maps constructed using the CEPH reference panel.
Nat. Genet. 6: 391-393.

Bullrich, F., MacLachlan, T.K., Sang, N., Druck, T., Veronese, M-L., Allen, S.L., Chiorazzi, N., Koff, A., Huebner, K., Croce, C.M., Giordano, A. (1995).
Chromosomal mapping of members of the cdc2 family of protein kinases, cdk3, cdk6, PISSLRE, and PITALRE, and a cdk inhibitor, p27^{Kip1}, to regions involved in human cancer.
Cancer Res. 55: 1199-1205.

Burke, D.T., Carle, G.F., Olson, M.V. (1987).
Cloning of large exogenous DNA into yeast by means of artificial chromosome vectors.
Science 236: 806-812.

Burks, C., Fickett, J.W., Goad, W.B., Kanehisa, M., Lewitter, F.I., Rindone, W.P., Swindell, C.D., Tung, C.S., Bilofsky, H.S. (1985).
The GenBank nucleic acid sequence database.
Comput. Appl. Biosci. 1: 225-233.

Buxton, J., Shelbourne, P., Davies, J., Jones, C., Van Tongeren, T., Aslanidis, C., de Jong, P., Jansen, G., Anvret, M., Riley, B., et al. (1992).
Detection of an unstable fragment of DNA specific to individuals with myotonic dystrophy.
Nature 355: 547-548.

Callen, D.F. (1986).
A mouse - human hybrid cell panel for mapping human chromosome 16.
Ann. Genet. 29: 235-239.

Callen, D.F., Hyland, V.J., Baker, E.G., Fratini, A., Simmers, R.N., Mulley, J.C., Sutherland, G.R. (1988).
Fine mapping of gene probes and anonymous DNA fragments to the long arm of chromosome 16.
Genomics 2: 144-153.

Callen, D.F., Hyland, V.J., Baker, E.G., Fratini, A., Gedeon, A.K., Mulley, J.C., Fernandez, K.E.W., Breuning, M.H., Sutherland, G.R. (1989).
Mapping the short arm of human chromosome 16.
Genomics 4: 348-354.

Callen, D.F., Baker, E., Eyre, H.J., Lane, S.A. (1990).
An expanded mouse-human hybrid cell panel for mapping human chromosome 16.
Ann. Genet. 33: 190-195.

Callen, D.F., Doggett, N.A., Stallings, R.L., Chen, L.Z., Whitmore, S.A., Lane, S.A., Nancarrow, J.K., Apostolou, S., Thompson, A.D., Lapsys, N.M. Baker, E.G., Shen, Y., Holman, K., Phillips, H., Richards, R.I., Sutherland, G.R. (1992).
High-resolution cytogenetic-based physical map of human chromosome 16.
Genomics 13: 1178-1185.

Callen, D.F., Lane, S.A., Kozman, H., Kremmidiotis, G., Whitmore, S.A., Lowenstein, M., Doggett, N.A., Kenmochi, N., Page, D.C., Maglott, D.R., Nierman, W.C., Murakawa, K., Berry, R., Sikela, J.M., Houlgatte, R., Auffray, C., Sutherland, G.R. (1995).
Integration of transcript and genetic maps of chromosome 16 at near-1-Mb resolution: Demonstration of a "hot spot" for recombination at 16p12.
Genomics 29: 503-511.

Casanova, J-L., Pannetier, C., Kourilsky, P. (1990).
Optimal conditions for directly sequencing double-stranded PCR products with Sequenase.
Nucleic Acids Res. 18: 4028

Castilla, L.H., Couch, F.J., Erdos, M.R., Hoskins, K.F., Calzone, K., Garber, J.E., Boyd, J., Lubin, M.B., Deshano, M.L., Brody, L.C., Collins, F.S., Weber, B.L. (1994).
Mutations in the BRCA1 gene in families with early-onset breast and ovarian cancer.
Nature Genet. 8: 387-391.

Cavenee, W.K., Dryja, T.P., Philips, R.A., Benedict, W.F., Godbout, R., Gallie, B.L., Murphree, A.L., Strong, L.C., White, R. (1983).
Expression of recessive alleles by chromosomal mechanisms in retinoblastoma.
Nature 305: 779-784.

Cavenee, W.K., Hansen, M.F., Nordenskjold, M., Kock, E., Maumenee, I., Squire, J., Phillips, R.A., Gallie, B.L. (1985).
Genetic origin of mutations predisposing to retinoblastoma.
Science 228: 501-503.

Ceccherini, I., Romeo, G., Lawrence, S., Breuning, M.H., Harris, P.C., Himmelbauer, H., Frischauf, A.M., Sutherland, G.R., Germino, G.G., Reeders, S.T., Morton, N.E. (1992).
Construction of a map of chromosome 16 by using radiation hybrids.
Proc. Natl. Acad. Sci. USA 89: 104-108.

Chen, L-C., Dollbaum, C., Smith, H.S. (1989).
Loss of heterozygosity on chromosome 1q in human breast cancer.
Proc. Natl. Acad. Sci. USA 86: 7204-7207.

Chen, L.Z., Harris, P.C., Apostolou, S., Baker, E., Holman, K., Lane, S.A., Nancarrow, J.K., Whitmore, S.A., Stallings, R.L., Hildebrand, C.E., Richards, R.I., Sutherland, G.R., Callen, D.F. (1991).
A refined physical map of the long arm of human chromosome 16.
Genomics 10: 308-312.

Chia, W., Scott, M.R.D., Rigby, P.W.J. (1982).
The construction of cosmid libraries of eukaryotic DNA using the Homer series of vectors.
Nucl. Acid Res. 10: 2503-2520.

Chien, A., Edgar, D.B., Trela, J.M. (1976).
Deoxyribonucleic acid polymerase from the extreme thermophile *Thermus aquaticus*.
J. Bacteriology 127: 1550-1556.

Chomczynski, P., Sacchi, N. (1987).
Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction.
Anal. Biochem. 162: 156-159.

Chumakov, I., Le Gall, I., Billault, A., Ougen, P., Soularue, P., Rigault, P., Bui, H., De Tand, M-F., Barillot, E., Abderrahim, H., Cherif, D., Berger, R., Le Paslier, D., Cohen, D. (1992a).
Isolation of chromosome 21-specific yeast artificial chromosomes from a total human genomic library.
Nat. Genet. 1: 222-225.

Chumakov, I., Rigault, P., Guillou, S., Ougen, P., Billaut, A., Guasconi, G., Gervy, P., Legall, I., Soularue, P., Grinas, L., Bougueleret, L., Bellanne-Chantelot, C., Lacroix, B., Barillot, E., Gesnouin, P., Pook, S., Vaysseix, G., Frelat, G., Schmitz, A., Sambucy, J.L., Bosch, A., Estivill, X., Weissenbach, J., Vignal, A., Riethman, H., Cox, D., Patterson, D., Gardiner, K., Hattori, M., Sakaki, Y., Ichikawa, H., Ohki, M., Le Paslier, D., Heilig, R., Antonarakis, S., Cohen, D. (1992b).
Continuum of overlapping clones spanning the entire human chromosome 21q.
Nature 359: 380-387.

Chumakov, I., Rigault, P., Le Gall, I., Bellanne-Chantelot, C., Billaut, A., Guillou, S., Soularue, P., Guasconi, G., Poullier, E., Gros, I., et al. (1995).
A YAC contig map of the human genome.
Nature 377: 175-297.

Chaudhari, N., Hahn, W.E. (1983).
Genetic expression in the developing brain.
Science 220: 924-928.

Cheng, J-F., Boyartchuk, V., Zhu, Y. (1994).
Isolation and mapping of human chromosome 21 cDNA: Progress in constructing a chromosome 21 expression map.
Genomics 23: 75-84.

Church, D.M., Stotler, C.J., Rutter, J.L., Murrell, J.R., Trofatter, J.A., Buckler, A.J. (1994).
Isolation of genes from complex sources of mammalian DNA using exon amplification.
Nature Genet. 6: 98-105.

Cinkosky M.J., Fickett J.W., Gilna, P., Burks, C. (1991).
Electronic data publishing and GenBank.
Science 252: 1273-7

Cleton-Jansen, A-M., Moerland, E., Kulpers-Dijkshoorn, N., Callen, D.F., Sutherland, G.R., Hansen, B., Devilee, P., Cornelisse, C. (1994).
At least two different regions are involved in allelic imbalance on chromosome arm 16 in breast cancer.
Genes Chrom. Cancer 9: 101-107.

Cleton-Jansen, A-M., Moerland, E., Callen, D.F., Doggett, N.A., Devilee, P., Cornelisse, C.J. (1995).
Mapping of the basic breast conserved gene (D16S444E) to human chromosome band 16q24.3.
Cytogenet. Cell Genet. 68: 49-51.

Cohen, D., Chumakov, I., Weissenbach, J. (1993).
A first-generation physical map of the human genome.
Science 258: 52-59.

Coles, C., Condie, A., Chetty, U., Steel, C.M., Evans, H.J., Prosser, J. (1992).
p53 mutations in breast cancer.
Cancer Res. 52: 5291-5298.

Collins, F.S. (1992).
Positional cloning: let's not call it reverse anymore.
Nat. Genet. 1: 3-6.

Collins, F.S. (1995).
Positional cloning moves from perditional to traditional.
Nat. Genet. 9: 347-350.

Murray, J.C., Buetow, K.H., Weber, J.L., Ludwigsen, S., Scherpbier-Heddema, F., Manion, J., Quillen, V.C., Sheffield, S., Sunden, G.M., Duyk, G.M. et al. (1994).
A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC).
Science 265: 2049-2054.

Cornelisse, C.J., Kuipers-Dijkshoorn, N., Van Vliet, M., Hermans, J., Devilee, P. (1992).
Fractional allelic imbalance in human breast cancer increases with tetraploidization and chromosome loss.
Int. J. Cancer 50: 544-548.

Cotton, R.G. (1992).
Detection of mutations in DNA.
Curr. Opin. Biotechnol. 3: 24-30.

Cotton, R.G., Rodrigues, N.R., Campbell, R.D. (1988).
Reactivity of cytosine and thymine in single-base-pair mismatches with hydroxylamine and osmium tetroxide and its application to the study of mutations.
Proc. Natl. Acad. Sci. USA 85: 4397-4401.

Couturier, J., Morichon-Delvallez, N., Dutrillaux, B. (1985).
Deletion of band 13q21 is compatible with normal phenotype.
Hum. Genet. 70: 87-91.

Cox, D.R., Burmeister, M., Price, E.R., Kim, S., Myers, R.M. (1990).
Radiation hybrid mapping: a somatic cell genetic method for constructing high resolution maps of mammalian chromosomes.

Science 250: 245-250.

Cox, L.A., Chen, G., Lee, EY-HP. (1994).
Tumor suppressor genes and their roles in breast cancer.
Breast Cancer Res. Treat. 32: 19-38.

Crampton, J.M., Davies, K.E., Knapp, T.F. (1981).
The occurrence of families of repetitive sequences in a library of cloned cDNA from human lymphocytes.
Nucl. Acids Res. 9: 3821-3833.

Dawson, D.S., Murray, A.W., Szostak, J.W. (1986).
An alternative pathway for meiotic chromosome segregation in yeast.
Science 234: 713-720.

Dean M. (1995).
Resolving DNA mutations.
Nature Genet. 9: 103-104.

Dean M., White, M.B., Amos, J., Gerrard, B., Stewart, C., Khaw, K-T., Leppert, M. (1990).
Multiple mutations in highly conserved residues are found in mildly affected cystic fibrosis patients.
Cell 61: 863-870.

Denizot, F., Mattei, M.G., Vernet, C., Pontarotti, P., Chimini, G. (1992).
YAC-assisted cloning of a putative G-protein mapping to the MHC class I region.
Genomics 14: 857-862.

Devilee, P., Van den Broek, M., Kuipers-Dijkshoorn, N., Kolluri, R., Meera Khan, P.M., Pearson, P.L., Cornelisse, C.J. (1989).
At least four different chromosomal regions are involved in loss of heterozygosity in human breast cancer.
Genomics 5: 554-560.

Devilee, P., Van Vliet, M., Van Sloun, P., Kuipers-Dijkshoorn, N., Hermans, J., Pearson, P.L., Cornelisse, C.J. (1991).
Allelotype of human breast carcinoma: a second major site for loss of heterozygosity is on chromosome 6q.

Oncogene 6: 1705-1711.

Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., Lathrop, M., Gyapay, G., Morissette, J., Weissenbach, J. (1996).
A comprehensive genetic map of the human genome based on 5,264 microsatellites [see comments]
Nature 380: 152-154.

Dietz, H.C., Cutting, G.R., Pyeritz, R.E., Maslen, C.L., Sakai, L.Y., Corson, G.M., Puffenberger, E.G., Hamosh, A., Nanthakumar, E.J., Curnstin, S.M., Stetten, G., Meyers, D.A., Francomano, C.A. (1991).
Marfan syndrome caused by a recurrent de novo missense mutation in the fibrillin gene.
Nature 352: 337-339.

Digweed, M. (1993).
Human genetic instability syndromes: Single gene defects with increased risk of cancer.
Toxicol. Letters 67: 259-281.

Doggett, N.A., Callen, D.F. (1995).
Report of the third international workshop on human chromosome 16 mapping 1994.
Cytogenet. Cell Genet. 68: 166-177.

Doggett, N.A., Goodwin, L.A., Tesmer, J.G., Meincke, L.J., Bruce, D.C., Clark, L.M., Altherr, M.R., Ford, A.A., Chi, H-C., Marrone, B.L., Longmire, J.L., Lane, S.A., Whitmore, S.A., Lowenstein, M.G., Sutherland, R.D., Mundt, M.O., Knill, E.H., Bruno, W.J., Macken, C.A., Torney, D.C., Wu, J-R., Griffith, J., Sutherland, G.R., Deaven, L.L., Callen, D.F., Moyzis, R.K. (1995).
An integrated physical map of human chromosome 16.
Nature 377 (Suppl): 335-339.

Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T.P., Bowden, D.W., Smith, D.R., Lander, E.S., Botstein, D., Akots, G., Rediker, K.S., Gravius, T., Brown, V.A., Rising, M.B., Parker, C., Powers, J.A., Watt, D.E., Kauffman, E.R., Bricker, A., Phipps, P., Muller-Kahle, H., Fulton, T.R., Ng, S., Schumm, J.W., Braman, J.C., Knowlton, R.G., Barker, D.F., Crooks, S.M., Lincoln, S.E., Daly, M.J., Abrahamson, J. (1987).
A genetic linkage map of the human genome.
Cell 51: 319-337.

Dorion-Bonnet, F., Mautalen, S., Hostein, I., Longy, M. (1995).
Allelic imbalance study of 16q in human primary breast carcinomas using microsatellite markers.
Genes Chrom. Cancer 14: 171-181.

Dunn, J.M., Phillips, R.A., Zhu, X., Becker, A., Gallie, B.L. (1989).
Mutations in the RB1 gene and their effects on transcription.
Mol. Cell. Biol. 9: 4596-4604.

Dutrillaux, B. (1973).
Nouveau systeme de marquage chromosomique: les bandes T.
Chromosoma 41: 395-402.

Dutrillaux, B. Gerbault-Seureau, M., Zafrani, B. (1990).
Characterization of chromosomal anomalies in human breast cancer.
Cancer Genet. Cytogenet. 49: 203-217.

Duyk, G.M., Kim, S., Myers, R.M., Cox, D.R. (1990).
Exon trapping: a genetic screen to identify candidate transcribed sequences in cloned mammalian genomic DNA.
Proc. Natl. Acad. Sci. USA 87: 8995-8999.

El-Ashry, D., Lippman, M.E. (1994).
Molecular biology of breast carcinoma.
World J. Surg. 18: 12-20.

Eliakim, M., Levy, M., Ehrenfeld, M. (1981).
Recurrent polyserositis. Elsevier/North-Holland Biomedical Press, Amsterdam, The Netherlands.

Engelke, D.R., Hoener, P.A., Collins, F.S. (1988).
Direct sequencing of enzymatically amplified human genomic DNA.
Proc. Natl. Acad. Sci. USA 85: 544-548.

Erisman, M.D., Astrin, S.M. (1988).
The myc oncogene. In The Oncogene Handbook, Roddy, E.P., Shalka, A.M., Curran, T. Eds. Amsterdam, Elsevier. pp 341-346.

Escot, C., Theillet, C., Lidereau, R. (1986).
Genetic alteration of the c-myc protooncogene (MYC) in human primary breast carcinomas.
Proc. Natl. Acad. Sci. USA 83: 4834

Evans, G.A., Lewis, K., Rothenberg, B.E. (1989).
High efficiency vectors for cosmid microcloning and genomic analysis.
Gene 79: 9-20.

Fan, W-F., Wei, X., Shukla, H., Parimoo, S., Xu., Sankhavaram, P., Li, Z., Weissman, S.M. (1993).
Application of cDNA selection techniques to regions of the human MHC.
Genomics 17: 575-581.

Fearon, E.R., Vogelstein, B., Feinberg, A.P. (1984).
Somatic deletion and duplication of genes on chromosome 11 in Wilms' tumor.
Nature 309: 176-178.

Fearon, E.R., Vogelstein, B. (1990).
A genetic model for colorectal tumorigenesis.
Cell 61: 759

Feinberg, A.P., Vogelstein, B. (1983).
A technique for radiolabelling DNA restriction fragments to high specific activity.
Analyt. Biochem. 132: 6-13.

Fidler, I.J., Hart, I.R. (1982).
Biological diversity in metastatic neoplasms: origins and implications.
Science 217: 998-1003.

Fischel-Ghodsian, N., Bu, X., Prezant, T.R., Oeztas, S., Huang, Z-S., Bohlman, M.C., Rotter, J.I., Shohat, M (1993).
Regional mapping of the gene for familial Mediterranean fever on human chromosome 16p13.3.
Am. J. Med. Genet. 46: 689-693.

Fishel, R., Lescoe, M.K., Rao, M.R., Copeland, N.G., Jenkins, N.A., Garber, J., Kane, M., Kolodner, R. (1993).
The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer.

Cell 75: 1027-1038.

Fong, K.M., Kida, Y., Zimmerman, P.V., Ikenga, M., Smith, P.J. (1995).
Loss of heterozygosity frequently affects chromosome 17q in non-small cell lung cancer.
Cancer Res. 55: 4268-4272.

Foote, S., Vollrath, D., Hilton, A., Page, D.C. (1992).
The human Y chromosome: overlapping DNA clones spanning the euchromatic region.
Science 258: 60-66.

Forrest, S., Dahl, H.H., Howells, D.W., Dianzani, I, Cotton, R.G. (1991).
Mutation detection in phenylketonuria by using chemical cleavage of mismatch: importance of using probes from both normal and patient samples.
Am. J. Hum. Genet. 49: 175-183.

Forrest, S., Cotton, R., Landegren, U., Southern, E. (1995).
How to find all those mutations.
Nature Genet. 10: 375-376.

Forster, A., Rabbitts, T.H. (1993).
A method for identifying genes within yeast artificial chromosomes: application to isolation of MLL fusion cDNAs from acute leukaemia translocations.
Oncogene 8: 3157-3160.

Francke, U. (1978).
Retinoblastoma and chromosome 13.
Birth Defects 12: 131-137.

Friedman, L.S., Ostermeyer, E.A., Szabo, C.I., Dowd, P., Lynch, E.D., Rowell, S.E., King, M-C. (1994).
Confirmation of BRCA1 by analysis of germline mutations linked to breast and ovarian cancer in ten families.
Nature Genet. 8: 399-404.

Frohman, M.A., Dush, M.K., Martin, G.R. (1988).
Rapid production of full-length cDNAs from rare transcripts: Amplification using a single gene-specific oligonucleotide primer.
Proc. Natl. Acad. Sci. USA 85: 8998-9002.

Fults, D., Brockmeyer, D., Tullous, M.W., Pedone, C.A., Cawthorn, R.M. (1992).
p53 mutation and loss of heterozygosity on chromosome 17 and 10 during human astrocytoma progression.
Cancer Res. 52: 674-679.

Futreal, P.A., Liu. Q., Shattuck-Eidens, D., Cochran, C., Harshman, K., Tavtigian, S., Bennett, L.M., Haugen-Strano, A., Swensen, J., Miki, Y., Eddington, K., McClure, M., Frye, C., Weaver-Feldhaus, J., Ding, W., Gholami, Z., Soderkvist, P., Terry, L. (1994).
BRCA1 mutations in primary breast and ovarian carcinomas.
Science 266: 120-122.

Gecz, J., Villard, L., Lossi, A.M., Millaseau, P., Djabali, M., Fontes, M. (1993).
Physical and transcriptional mapping of DXS56-PGK1 1 Mb region: identification of three new transcripts.
Hum. Mol. Genet. 2: 1389-1396.

Gibbs, R.A., Caskey, C.T. (1987).
Identification and localisation of mutations at the Lesch-Nyhan locus by ribonuclease A cleavage.
Science 230: 1242-1246.

Gibson, R.A., Buchwald, M., Roberts, R.G., Mathew, C.G. (1993a).
Characterisation of the exon structure of the Fanconi anaemia group -C gene by vectorette PCR.
Hum. Mol. Genet. 2: 35-38.

Gibson, R.A., Hanjianpour, A., Murer-Orlando, M., Buchwald, M., Mathew, C.G. (1993b).
A nonsense mutation and exon skipping in the Fanconi anaemia group-C gene.
Hum. Mol. Genet. 2: 797-799.

Grana, X., Claudio, P.P., De Luca, A., Sang, N., Giordano, A. (1994).
PISSLRE, a novel member of the cdk family of protein/serine threonine kinases.
Oncogene 9: 2097-2103.

Green, E.D., Olson, M.V. (1990).
Chromosomal region of the cystic fibrosis gene in yeast artificial chromosomes: a model for human genome mapping.
Science 250: 94-98.

Grompe, M. (1993).
The rapid detection of unknown mutations in nucleic acids.
Nature Genet. 5: 111- 117.

Grunstein, M. Hogness, D.S. (1975).
Colony hybridisation: a method for the isolation of cloned DNAs that contain a specific gene.
Proc. Natl. Acad. Sci. USA 72: 3961-3965.

Gusella, J.F., Keys, C., Varsanyi-Breiner, A., Kao, F-T., Jones, C., Puck, T.T., Housman, D. (1980).
Isolation and localization of DNA segments from specific human chromosomes.
Proc. Natl. Acad. Sci. USA 77: 2829-2833.

Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R., Sakaguchi, A.Y., Young, A.B., Shoulson, I., Bonnilla, E., Martin, J.B. (1983).
A polymorphic DNA marker genetically linked to Huntington's disease.
Nature 306: 234-238.

Guyer, M.S., Collins, F.S. (1995).
How is the Human Genome Project doing, and what have we learned so far?
Proc. Natl. Acad. Sci. USA 92: 10841-10848.

Gyapay, G., Morisette, J., Vignal, A., Dib, C., Fizames, C., Millaseau, P., Marc, S., Bernardi, G., Lathrop, M., Weissenbach, J. (1994).
The 1993-94 Genethon human genetic linkage map.
Nat. Genet. 7: 246-339.

Gyapay, G., Schmitt, K., Fizames, C., Jones, H., Vega-Czarny, N., Spillett, D., Muselet, D., Prud'Homme, J.F., Dib, C., Auffray, C., Morisette, J., Weissenbach, J., Goodfellow, P.N. (1996).
A radiation hybrid map of the human genome.
Hum. Mol. Genet. 5: 339-346.

Hainsworth, P.J., Raphael, K.L., Stillwell, R.G., Bennett, R.C., Garson, O.M. (1991).
Cytogenetic features of twenty-six primary breast cancers.
Cancer Genet. Cytogenet. 52: 205-218.

Hall, J.M., Lee, M.K., Newman, B., Morrow, J.E., Anderson, L.A., Huey, B., King, M-C. (1990).
Linkage of early-onset familial breast cancer to chromosome 17q21.
Science 250: 1684-1689.

Harada, Y., Katagiri, T., Ito, I., Akiyama, F., Sakamoto, G., Kasumi, F., Nakamura, Y., Emi M. (1994).
Genetic studies of 457 breast cancers.
Cancer 74: 2281-2286.

Harris, H., Hopkinson, D.A., Edwards, Y.H. (1977).
Polymorphism and the subunit structure of enzymes: A contribution to the neutralist-selectionist controversy.
Proc. Natl. Acad. Sci. USA 74: 698-701.
Harris, H. (1988).
The analysis of malignancy by cell fusion: the position in 1988.
Cancer Res. 48: 3302-3306.

Harwood, J., Tachibana, T., Davis, R., Bhattacharyya, N.P., Meuth, M. (1993).
High rate of multilocus deletion in a human tumor cell line.
Hum. Mol. Genet. 2: 165-171.

Hastbacka, J., de la Chapelle, A., Mahtani, M.M., Clines, G., Reeve-Daly, M.P., Daly, M., Hamilton, B.A., Kusumi, K., Trivedi, B., Weaver, A., et al.(1994).
The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping.
Cell 78: 1073-1087.

Hayashi, K. (1992).
PCR-SSCP: A method for detection of mutations.
GATA (Genetic Analysis: Techniques and Applications). 9: 73-79.

Higuchi, M., Kazazian, H.H., Kasper, J.A., Phillips, A., Antonarakis, S.E. (1991).
Towards complete characterization of factor VII mutations: Analysis of 19 of 26 exons by DGGE.
Ped. Res. 27: Abs. 777.

Hirayama, T. (1989).
Genetic epidemiology of cancer. In Genetic epidemiology of cancer. Lynch, H.T., Hirayama, T. Eds. CRC Press, Inc. Boca Raton, Fl. pp 69-101.

Hogervorst, F.B.L., Cornelis, R.S., Bout, M., van Vliet, M., Oosterwijk, J.C., Olmer, R., Bakker, B. et al. (1995).
Rapid detection of BRCA1 mutations by the protein truncation test.
Nat. Genet. 10: 208-212.

Hohn, B., Collins, J. (1980).
A small cosmid for efficient cloning of large DNA fragments.
Genet. 11: 291-298.

Horowitz, J.M., Yandell, D.W., Park, S.H., Canning, S., Whyte, P., Buchkovich, K., Harlow, E., Weinberg, R.A., Dryja, T.P. (1989).
Point mutational inactivation of the retinoblastoma antioncogene.
Science 243: 937-940.

Hovig, E., Smith-Sorensen, B., Brogger, A., Borrensen, A.L. (1991).
Constant denaturant gel electrophoresis, a modification of denaturing gradient gel electrophoresis, in mutation detection.
Mutat. Res. 262: 63-71.

Hsu, L-C., Kennan, W.S., Shepel, L.A., Jacob, H.J., Szpirer, C., Szpirer, J., Lander, E.S., Gould, M.N. (1994).
Genetic identification of Msc-1, a rat mammary carcinoma suppressor gene.
Cancer Res. 54: 2765-2770.

Hudson, T.J., Stein, L.D., Gerety, S.S., Ma, J., Castle, A.B., Silva, J., Slonim, D.K., Baptista, R., Kruglyak, L., Xu, SH., et al. (1995).
An STS-based map of the human genome. [see comments].
Science 270: 1945-1954.

Hunter, T., Pines, J. (1994).
Cyclins and cancer. II. Cyclin D and cdk inhibitors come of age.
Cell 79: 573-582.

Huntington's Disease Research Collaborative (1993).
A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes.
Cell 72: 971-983.

Huttner, W.B., Gerdes, H-H., Rosa, P. (1991).
The granin (chromogranin/secretogranin) family.
Trends Biochem. Sci. 16: 27-30.

Ianzano, L., d'Apolito, M., Centra, M., Savino, M., Levran, O., Auerbach, A.D., Cleton-Jansen, A-M., Doggett, N.A., Pronk, J.C., Tipping, A.J., Gibson, R.A., Mathew, C.G., Whitmore, S.A., Apostolou, S., Callen, D.F., Zelante, L., Savoia, A. (1997).
The genomic organization of the Fanconi anaemia group A (FAA) gene.
Genomics (in press).

Ioannou, P.A., Amemiya, C.T., Garnes, J., Kroisel, P.M., Shizuya, H., Chen, C., Batzer, M.A., de Jong, P.J. (1994).
A new bacteriophage P1-derived vector for the propagation of large human DNA fragments.
Nat. Genet. 6: 84-89.

Iwahana, H., Yoshimoto, K., Itakara, M. (1992).
Detection of point mutations by SSCP of PCR-amplified DNA after endonuclease digestion.
Biotechniques 12: 64.

Iwaya, K., Tsuda, H., Hiraide, H., Tamaki, K., Tamakuma, S., Fukutomi, T., Mukai, K., Hirohashi, S. (1991).
Nuclear p53 immunoreaction associated with poor prognosis of breast cancer.
Jpn. J. Cancer Res. 82: 835-840.

Jeffreys, J.J., Wilson, V., Thein, S.L. (1985).
Hypervariable 'minisatellite' regions in human DNA.
Nature 314: 67-73.

Jelenik, W.R., Schmid, C.W. (1982).
Repetitive sequences in eukaryotic DNA and their expression.
Annu. Rev. Biochem. 51: 813-844.

Jensen, R.A., Thompson, M.E., Jetton, T.L., Szabo, C.I., van der Meer, R., Helou, B., Tronick, S.R., Page, D.L., King, M-C., Holt, J.T. (1996).

BRCA1 is secreted and exhibits properties of a granin.
Nat. Genet. 12: 303-308.

Joenje, H., Lo ten Foe, J.R., Oostra, A.B., van Berkel, G.G., Rooimans, M.A., Schroeder-Kurth, T., Wegner, R.D., Gille, J.J., Buchwald, M., Arwert, F. (1995).
Classification of Fanconi anemia patients by complementation analysis: evidence for a fifth genetic subtype.
Blood 86: 2156-2160.

Kainulainen, K., Pulkkinen, L., Savolainen, A., Kaitila, I., Peltonen, L. (1990).
Location on chromosome 15 of the gene defect causing Marfan syndrome [see comments].
New Engl. J. Med. 323: 935-939.

Kanai, Y., Oda, T., Tsuda, H., Ochiai, A., Hirohashi, S. (1994).
Point mutation of the E-cadherin gene in invasive lobular carcinoma of the breast.
Jpn. J. Cancer Res. 85: 1035-1039.

Kao, F.T., Yu, J.W. (1991).
Chromosome microdissection and cloning in human genome and genetic disease analysis.
Proc. Natl. Acad. Sci. USA 88: 1844-1848.

Kariya, Y., Kato, K., Hayashizaki, Y., Himeno, S., Tarui, S., Matsubara, K. (1987).
Revision of consensus sequence of human Alu repeats--a review.
Gene 53: 1-10.

Kestila, M., Mannikko, M., Holmberg, C., Gyapay, G., Weissenbach, J., Savolainen, E-R., Peltonen, L., Tryggvason, K (1994).
Congenital nephrotic syndrome of the Finnish type maps to the long arm of chromosome 19.
Am. J. Hum. Genet. 54: 757-764.

Knudson, A.G. (1971).
Mutation and cancer: statistical study of retinoblastoma.
Proc. Natl. Acad. Sci. USA 68: 820-823.

Knudson, A.G. (1989).
Hereditary cancers: clue to mechanisms of carcinogenesis.
British J. Cancer 59: 661-666.

Korenberg, J.R., Rykowski, M.C. (1988).
Human genome organization: Alu, LINES, and the molecular structure of metaphase chromosome bands.
Cell 53: 391-400.

Korn, B., Sedlacek, Z., Manca, A., Kioschis, P., Konecki, D., Lehrach., Poustka, A. (1992).
A strategy for the selection of transcribed sequences in the Xq28 region.
Hum. Mol. Genet. 1: 235-242.

Koyama, K., Mitsuru, E., Nakamura, Y. (1993).
The cell adhesion regulator (CAR) gene, Taq I and insertion/deletion polymorphisms, and regional assignment to the peritelomeric region of 16q by linkage analysis.
Genomics 16: 264-265.

Kozak, M. (1986).
Point mutations define a sequence flanking the AUG initiator codon that modulates translation of eukaryotic ribosomes.
Cell 44: 283-292.

Kozman, H.M., Phillips, H.A., Callen, D.F., Sutherland, G.R., Mulley, J.C. (1993).
Integration of the cytogenetic and genetic linkage maps of human chromosome 16 using 50 physical intervals and 50 polymorphic loci.
Cytogenet. Cell Genet. 62: 194-198,

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

Krizman, D.B., Berget, S.M. (1993).
Efficient selection of 3'-terminal exons from vertebrate DNA.
Nuc. Acids Res. 21: 5198-5202.

Kruglyak, L., Daly, M.J., Lander, E.S. (1995).
Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping.
Am. J. Hum. Genet. 56: 519-527.

Kunimi, K., Bergerheim, U.S.R., Larsson, I-L., Ekman, P., Collins, V.P. (1991).
Allelotyping of human prostatic adenocarcinoma.
Genomics 11: 530-536.

Kwok, K., Ledley, F.D., Dilella, A.G., Robson, K.J.H., Woo, S.L.C. (1985).
Nucleotide sequence of a full-length complementary DNA clone and amino acid sequence of human phenylalanine hydroxylase.
Biochemistry 24: 556-561.

Landegent, J.E., Jansen in de Wal, N., Dirks, R.W., Baas, F., van der Ploeg, M. (1987).
Use of whole cosmid cloned genomic sequences for chromosomal localization by non-radioactive *in situ* hybridization.
Hum. Genet. 77: 366-370.

Lane, D.P. (1992).
p53, guardian of the genome.
Nature 358: 15

Larsen, F., Gundersen, G., Lopez, R., Prydz, H. (1992).
CpG islands as gene markers in the human genome.
Genomics 13: 1095-1107.

Larsen, F., Gundersen, G., Prydz, H. (1992b).
Choice of enzymes for mapping based on CpG islands in the human genome.
GATA 9: 80-85.

Larsson, C., Brystrom, C., Skoog, L., Rotstein, S., Nordenskjold, M. (1990).
Genomic alterations in human breast carcinomas.
Genes Chrom. Cancer 2: 191-197.

Lawrence, J.B., Singer, R.H., McNeil, J.A. (1990).
Interphase and metaphase resolution of different distances within the human dystrophin gene.
Science 249: 928-32.

Lawrence, S., Morton, N.E., Cox, D.R. (1991).
Radiation hybrid mapping.
Proc. Natl. Acad. Sci. USA 88: 7477-7480.

Lee, E.Y.P., To, H., Shew, J.Y., Brookstein, R., Scully, P., Lee, W.H. (1988).
Inactivation of Rb susceptibility gene in human breast cancer.
Science 241: 218-221.

Lee, S.W. (1996).
H-cadherin, a novel cadherin with growth inhibitory functions and diminished expression in human breast cancer.
Nature Med. 2: 776-782.

Lefebvre, S., Burglen, L., Reboullet, S., Clermont, O., Burlet, P., Viollet, L., Benichou, B., Cruaud, C., Millaseau, P., Zeviani, M. et al. (1995).
Identification and characterization of a spinal muscular atrophy-determining gene.
Cell 80: 155-165.

Lerman, L.S., Silverstein, K. (1987).
Computational simulation of DNA melting and its application to denaturing gel electrophoresis.
Meth. Enzymol. 155: 482-501.

Levine, A.J., Momand, J., Finlay, C.A. (1991).
The p53 tumor suppressor gene.
Nature 351: 453

Levy, E.N., Shen, Y., Kupelian, A., Kruglyak, L., Aksentijevich, I., Pras, E., Balow, J.E., Linzer, B., Chen, X., Shelton, D.A., Gumucio, D., Pras, M., Shohat, M., Rotter, J.I., Fischel-Ghodsian, N., Richards, R.I., Kastner, D.L. (1996).
Linkage disequilibrium mapping places the gene causing familial Mediterranean fever close to D16S246.
Am. J. Hum. Genet. 58: 523-534.

Lichter, P., Ledbetter, S.A., Ledbetter, D.H., Ward, D.C. (1990).
Fluorescent *in situ* hybridization using Alu-PCR and L1-PCR probes for rapid characterization of human chromosomes in hybrid cell lines.
Proc. Natl. Acad. Sci. USA 87: 6634-6638.

Lindblom, A., Rotstein, S., Skoog, L., Nordenskjold, M., Larsson, C. (1993).
Deletions on chromosome 16 in primary familial breast carcinomas are associated with development of distant metastases.
Cancer Res. 5: 3707-3711.

Liu, P., Legerski, R., Siciliano, J. (1989).
Isolation of human transcribed sequences from human-rodent somatic cell hybrids.
Science 246: 813-815.

Liu, J.M., Buchwald, M., Walsh, C.E., Young, N.S. (1994).
Fanconi Anemia and novel strategies for therapy.
Blood 84: 3995-4007.

Lo Ten Foe, J.R., Rooimans, M.A., Bosnoyan-Collins, L., Alon, N., Wijker, M., Parker, L., Lightfoot, J., Carreau, M., Callen, D.F., Savoia, A., Cheng, N.C., van Berkel, C.G.M., Strunk, M.H.P., Gille, J.J.P., Pals, G., Kruyt, F.A.E., Pronk, J.C., Arwert, F., Buchwald, M., Joenje, H. (1996).
Expression cloning of a cDNA for the major Fanconi anaemia gene, FAA.
Nat. Genet. 14: 320-323.

Longmire, J.L., Brown, N.C., Meincke, L.J., Campbell, M.L., Albright, K.L., Fawcett, J.J., Campbell, E.W., Moyzis, R.K., Hildebrand, C.E., Evans, G.A., Deaven, L.L. (1991).
Construction and characterization of partial digest DNA libraries made from flow-sorted human chromosome 16.
GATA 10: 69-76.

Losekoot, M., Fodde, R., Harteveld, C.I., Van Heeren, H., Giordano, P.C., Bernini, L.F. (1990).
Denaturing gradient gel electrophoresis and direct sequencing of PCR amplified genomic DNA: A rapid and reliable diagnostic approach to beta thalassaemia.
Br. J., Haematol. 76: 269-274.

Lovett, M., Kere, J., Hinton, L.M. (1991).
Direct selection: a method for the isolation of cDNAs encoded by large genomic regions.
Proc. Natl. Acad. Sci. USA 88: 9623-9627.

Ludecke, H.J., Senger, G., Claussen, U., Horsthemke, B. (1989).
Cloning defined regions of the human genome by microdissection of banded chromosomes and enzymatic amplification.
Nature 338: 348-350.

Lundberg, C., Skoog, L., Cavenee, W.K., Nordenskjold, M. (1987).
Loss of heterozygosity in human ductal breast tumors indicates a recessive mutation on chromosome 13.
Proc. Natl. Acad. Sci. USA 84: 2372-2376.

Mackay, J., Elder, P.A., Porteous, D.J., Steel, C.M., Hawkins, R.A., Going, J.J., Chetty, U. (1988).
Partial deletion of chromosome 11p in breast cancer correlates with size of primary tumor and estrogen receptor level.
Br. J. Cancer 58: 710-714.

Magenis, R.E., Maslen, C.L., Smith, L., Allen, L., Sakai, L.Y. (1991).
Localization of the fibrillin (FBN) gene to chromosome 15, band q21.1.
Genomics 11: 346-351.

Malkin, D., Li, F.P., Strong, L.C., Fraumeni, J.F., Nelson, C.E., Kim, D.H., Kassel, J., Gryka, M.A., Bischoff, F.Z., Tainsky, M.A., Friend, S.H. (1990).
Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms.
Science 250: 1233-1238.

Mansouri, M., Spurr, N., Goodfellow, P.N., Kemler, R. (1988).
Characterization and chromosomal localization of the gene encoding the human cell adhesion molecule uvomorulin.
Differentiation 38: 67-71.

Marquis, S.T., Rajan, J.V., Wynshaw-Boris, A., Xu, J., Yin, G-Y., Abel, K.J., Weber, B.L., Chodosh, L.A. (1995).
The developmental pattern of BRCA1 expression implies a role in differentiation of the breast and other tissues.
Nat. Genet. 11: 17-26.

Mashal, R.D., Koontz, J., Sklar, J. (1995).
Detection of mutations by cleavage of DNA heteroduplexes with bacteriophage resolvases.
Nat. Genet. 9: 177-183.

Maw, M., Grundy, P.E., Millow, L.J., Eccles, M.R., Dunn, R.S., Smith, P.J., Feinberg, A.P., Law, D.J., Paterson, M., Telzerow, P.E., Callen, D.F., Thompson, A.D., Richards, R.I., Reeve, A.E. (1992).

A third Wilms' tumor locus on chromosome 16q.
Cancer Res. 52: 3094-3098.

McCormick, M.K., Campbell, E., Deaven, L., Moyzis, R. (1993).
Low-frequency chimeric yeast artificial chromosome libraries from flow-sorted human chromosomes 16 and 21.
Proc. Natl. Acad. Sci. USA 90: 1063-1067.

McInnis, M.G., Chakravarti, A., Blaschak, J., Petersen, M.B., Sharma, V., Avramopoulos, D., Blouin, J.L., Konig, U., Brahe, C., Matise, T.C., Warren, A.C., Talbot, C.C., Van Broeckhoven, C., Litt, M., Antonarakis, S.E. (1993).
A linkage map of human chromosome 21; 43 PCR markers at average intervals of 2.5 cM.
Genomics 16: 562-571.

McKusick, V.A. (1992).
Autosomal recessive phenotypes. In Mendelian Inheritance in Man. Tenth Edition, The Johns Hopkins University Press, Baltimore and London. pp. 1518-1520.

McMahon, G., Davis, E., Wogan, G.N. (1987).
Characterisation of c-Ki-ras oncogene alleles by direct sequencing of enzymatically amplified DNA from carcinogen-induced tumors.
Proc. Natl. Acad. Sci. USA 84: 4974-4978.

Merajver, S.D., Pham, T.M., Caduff, R.F., Chen, M., Poy, E.L., Cooney, K.A., Weber, B.L., Collins, F.S., Johnston, C., Frank, T.S. (1995).
Somatic mutations in the BRCA1 gene in sporadic ovarian tumours.
Nat. Genet. 9: 439-443.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L.M., Ding, W., Bell, R., Rosenthal, J., Hussey, C., Tran, T., McClure, M., Frye, C., Hattier, T., Phelps, R., Haugen-Strano, A., Katcher, H., Yakumo, K., Gholami, Z., Shaffer, D., Stone, S., Bayer, S., Wray, C., Bogden, R., Dayananth, P., Ward, J., Tonin, P., Narod, S., Bristow, P.K., Norris, F.H., Helvering, L., Morrison, P., Rosteck, P., Lai, M., Barret, C., Lewis, C., Neuhausen, S., Cannon-Albright, L., Goldgar, D., Wiseman, R., Kamb, A., Skolnick, M.H. (1994).
A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1.
Science 266: 66-71.

Monaco, A.P., Neve, R.L., Colletti-Feener, C., Bertelson, C.J., Kurnit, D.M., Kunkel, L.M. (1986).
Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene.
Nature 323: 646-650.

Moore, M., Query, C.C., Sharp, P.A. (1992).
The RNA world. Cold Spring Harbor Press, Cold Spring Harbor, N.Y. pp 1-30.

Morgan, J.G., Dolganov, G.M., Robbins, S.E., Hinton, L.M., Lovett, M. (1992).
The selective isolation of novel cDNAs encoded by the regions surrounding the human interleukin 4 and 5 genes.
Nuc. Acids Res. 20: 5173-5179.

Morris, C.P., Guo, X-H., Apostolou, S., Hopwood, J.J., Scott, H.S. (1994).
Morquio A syndrome: Cloning, sequence and structure of the human N-acetylgalactosamine 6-sulfatase (GALNS) gene,
Genomics 22: 652-654.

Morton, N.E. (1991).
Parameters of the human genome.
Genomics 88: 7474-7476.

Mouchiroud, D., D'Onofrio, G., Aissani, B., Macaya, G., Gautier, C., Bernardi, G. (1991).
The distribution of genes in the human genome.
Gene 100: 181-187.

Munroe, D.J., Haas, M., Bric, E., Whitton, T., Aburatani, H., Hunter, K., Ward, D., Housman, D.E. (1994).
IRE-Bubble PCR: A rapid method for efficient and representative amplification of human genomic DNA sequences from complex sources.
Genomics 19: 506-514.

Myers R.M., Larin, Z., Maniatis T. (1985).
Detection of single base substitutions by ribonuclease cleavage at mismatches in RNA : DNA duplexes.
Science 230: 1242-1246.

Myers R.M., Maniatis T., Lerman L.S. (1987).
Detection and localisation of single base changes by denaturing gradient gel electrophoresis.
Meth Enzymol. 155: 501-527.

Nagai, M.A., Yamamoto, L., Salaorni, S., Pacheco, M.M., Brentani, M.M., Barbosa, E.M., Brentani, R.R., Mazoyer, S., Smith, S.A., Ponder, B.A. et al (1994).
Detailed deletion mapping of chromosome segment 17q12-21 in sporadic breast tumours.
Genes Chrom. Cancer 11: 58-62.

Nagai, M.A., Medeiros, A.C., Brentani, M.M., Marques, L.A., Mazoyer, S., Mulligan, L.M. (1995).
Five distinct deleted regions on chromosome 17 defining different subsets of human primary breast tumors.
Oncology 52: 448-453.

Nagamine, C.M., Chan, K., Lau, Y.F.C. (1989).
A PCR artifact: Generation of heteroduplexes.
Am. J. Hum. Genet. 45: 337-339.

Nahmias, J., Hornigold, N., Fitzgibbon, J., Woodward, K., Pilz, A., Griffin, D., Henske, E.P., Nakamura, Y., Graw, S., Florian, F. et al. (1995).
Cosmid contigs spanning 9q34 including the candidate region for TSC1.
Eur. J. Hum. Genet. 3: 65-67.

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E. et al. (1987).
Variable number of tandem repeat (VNTR) markers for human gene mapping.
Science 253; 1616-1622.

Nakamura, Y., Carlson, M., Krapcho, K., Kanamori, M., White, R. (1988).
New approach for isolation of VNTR markers.
Am. J. Hum. Genet. 43: 854-859.

Narod, S.A., Ford, D., Devilee, P., Barkardottir, R.B., Lynch, H.T., Smith, S.A., Ponder, B.A., Weber, B.L., Garber, J.E., Birch, J.M. et al. (1995).
An evaluation of genetic heterogeneity in 145 breast-ovarian cancer families. Breast Cancer Linkage Consortium.
Am. J. Hum. Genet. 56: 254-264.

Naylor, S.L., Johnson, B.E., Minna, J.D., Sakaguchi, A.Y. (1987).
Loss of heterozygosity of chromosome 3p markers in small cell lung cancer.
Nature 329: 451-454.

Nelson, D.L., Ledbetter, S.A., Corbo, L., Victoria, M.F., Ramirez-Solis, R., Webster, T.D., Ledbetter, D.H., Caskey, C.T. (1980).
Alu polymerase chain reaction: A method for rapid isolation of human-specific sequences from complex DNA sources.
Proc. Natl. Acad. Sci. USA 86: 6686-6690.

Newman, B., Austin, M.A., Lee, M., King, M-C. (1988).
Inheritance of human breast cancer: Evidence for autosomal dominant transmission in high-risk families.
Proc. Natl. Acad. Sci. USA 85: 3044-3048.

Newton, C.R., Graham, A. (1994).
PCR - Polymerase Chain Reaction. First Edition. βIOS Scientific Publishers, Oxford, UK.

Ng, I.S., Pace, R., Richard, M.V., Kobayashi, K., Kerem, B., Tsui, L.C., Beaudet, A.L. (1991).
Methods for analysis of multiple cystic fibrosis mutations.
Hum. Genet. 87: 613-617.

Nigro, J.M., Baker, S.J., Preisinger, A.C., Jessup, J.M., Hostetter, R., Cleary, K., Bigner, S.H., Davidson, N., Baylin, S., Devilee, P., Glover, T., Collins, F.S., Weston, A., Modali, R., Harris, C.C., Vogelstein, B. (1989).
Mutations in the p53 gene occur in diverse human tumour types.
Nature 342: 705-707.

Nishisho, I., Nakamura, Y., Myoshi, Y., Miki, H., Ando, A.K., Horii, J., Utsunomiya, S., Baba, P., Hedge, P., Markham, A. (1991).
mutation of chromosome 5q21 genes in FAP and colorectal cancer patients.
Science 253: 665-666.

Niwa, M., Berget, S.M. (1991).
Mutation of the AAUAAA polyadenylation signal depresses *in vitro* splicing of proximal but not distal introns.
Genes and Dev. 5: 2086-2095.

Nowak, R. (1994).
Mining treasures from 'junk DNA' [news].
Science 263: 608-610.

Oka, H., Shioaki, H., Kobayashi, K., Inoue, M., Tahara, H., Kobayashi, T., Takatsuka, Y., Matsuyoshi, N., Hirano, S., Takeichi, M., et al. (1993).
Expression of E-cadherin cell adhesion molecules in human breast cancer tissues and its relationship to metastasis.
Cancer Res. 53: 1696-1701.

Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y., Matsubara, K. (1993).
Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression.
Nat. Genet. 2: 173-179.

Olson, M., Hood, L., Cantor, C., Botstein, D. (1989).
A common language for physical mapping of the human genome.
Science 245: 1434-1435.

Olvera, J., Wool, I.G. (1994).
The primary structure of rat ribosomal protein L13.
Biochem. Biophys. Res. Comm. 201: 102-107.

Orita, M., Suzuki, Y., Sekiya, T., Hayashi, K. (1989).
Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction.
Genomics 5: 874-879.

Ozdemir, A.I., Sokmen, C. (1969).
Familial Mediterranean fever among the Turkish people.
Am. J. Gastroenterol. 51: 311-316.

Parimoo, S., Patanjali, S.R., Shukla, H., Chaplin, D.D. (1991).
cDNA selection: Efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments.
Proc. Natl. Acad. Sci. USA 88: 9623-9627.

Park, S.Y., Kang, Y.S., Kim, B.G., Lee, S.H., Lee, E.D., Lee, K.H., Park, K.B., Lee, J.H. (1995).
Loss of heterozygosity on the short arm of chromosome 17 in uterine cervical carcinomas.
Cancer Genet. Cytogenet. 79: 74-78.

Perry, D.J., Carrell, R.W. (1992).
Hydrolink gels: A rapid and simple approach to the detection of DNA mutations in thromboembolic disease.
J. Clin. Pathol. 45: 158-160.

Peterson, A., Patli, N., Robbins, C., Wang, L., Cox, D.R., Myers, R.M. (1994).
A transcript map of the Down Syndrome critical region on chromosome 21.
Hum. Mol. Genet. 3: 1735-1742.

Pierce, J.C., Sternberg, N.L. (1992).
Using bacteriophage P1 system to clone high molecular weight genomic DNA.
Methods Enzymol. 216: 549-574.

Polymeropoulos, M.H., Xiao, H., Glodek, A., Gorski, M., Adams, M.D., Moreno, R.F., Fitzgerald, M.G., Venter, J.C., Merril, C.R. (1992).
Chromosomal assignment of 46 brain cDNAs.
Genomics 12: 492-496.

Pras, E., Aksentijevich, I., Gruberg, L., Balow, J.E., Prosen, L., Dean, M., Steinberg, A.D. et al. (1992).
Mapping of a gene causing familial Mediterranean fever to the short arm of chromosome 16.
N. Engl. J. Med. 326: 1509-1513.

Pronk, J.C., Gibson, R.A., Savoia, A., Wijker, M., Morgan, N.V., Melchionda, S., Ford, D., Temtamy, S., Ortega, J.J., Jansen, S., Havnga, C., Cohn, R.J., de Ravel, T.J., Roberts, I., Westerveld, A., Easton, D.F., Joenje, H., Mathew, C.G., Arwert, F. (1995).
Localisation of the Fanconi anaemia complementation group A gene to chromosome 16q24.3.
Nature Genet. 11: 338-340.

Pullman, W.E., Bodmer, W.F. (1992).
Cloning and characterization of a gene that regulates cell adhesion.
Nature 356: 529-532.

Race, R.R., Sanger, R. (1968).
Blood groups in man. Blackwell Scientific publications, Oxford.

Ravnik-Glavak, M., Glavac, D., Dean, M. (1994).
Sensitivity of single-strand conformation polymorphism analysis and heteroduplex method for mutation detection in the cystic fibrosis gene.
Hum. Mol. Genet. 3: 801-807.

Reed, K.C. and Mann, D.A. (1985).
Rapid transfer of DNA from agarose gels to nylon membranes.
Nucl. Acids Res. 13: 7207-7221.

Reeders, S.T., Breuning, M.H., Davies, K.E., Nicholls, R.D., Jarman, A.P., Higgs, D.R., Pearson, P.L., Weatherall, D.J. (1985).
A highly polymorphic DNA marker linked to adult polycystic kidney disease on chromosome 16.
Nature 317: 542-544.

Rich, A., Nordheim, A., Wang, A. (1984).
The chemistry and biology of left-handed Z-DNA.
Ann. Rev. Biochem. 53: 791-846.

Riordan, J.R., Rommens, J.M., Kerem, B.S., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., Chou, J.L., Drumm, M.L., Iannuzzi, M.L., Collins, F.S., Tsui, L.C. (1989).
Identification of the cystic fibrosis gene: cloning and characterization of complementary cDNA.
Science 245: 1066-1073.

Roberts, R.G., Bobrow, M., Bentley, D.R. (1992).
Point mutations in the dystrophin gene.
Proc. Natl. Acad. Sci. USA 89: 2331-2335.

Rodgers, C.S., Hill, S.M., Hulten, M.A. (1984).
Cytogenetic analysis in human breast carcinoma. I. Nine cases in the diploid range investigated using direct preparations.
Cancer Genet. Cytogenet. 13: 95-119.100

Rogers, D.B., Shohat, M., Petersen, G.M., Bickal, J., Congleton, J., Schwade, A.D., Rotter, J.I. (1989).
Familial Mediterranean fever in Armenians: autosomal recessive inheritance with high gene frequency.
Am. J. Med. Genet. 34: 168-172.

Rohme, D., Siden, T., van der Maarel, S.M., Cremers, F.P., Tantravahi, U., Marinoni, J.C., Ropers, H.H., Schwartz, C.E. (1994).
Radiation hybrids for the proximal long arm of the X chromosome and their use in the derivation of an ordered set of cosmid markers from a defined subregion in proximal Xq13.1.
Somat. Cell Mol. Genet. 20: 1-10.

Rommens, J.M., Lin, B., Hutchinson, G.B., Andrew, S.E., Goldberg, Y.P. Glaves, M.L., Graham, R., Lai, V., McArthur, J., Nasir, J., Theilmann, J., McDonald, H., Kalchman, M., Clarke, L.A., Schappert, K., Hayden, M.R. (1993).
A transcription map of the region containing the Huntington disease gene.
Hum. Mol. Genet. 2: 901-907.

Rose, E.A (1990)
Complete physical map of the WAGR region of 11p13 localizes a candidate Wilm's tumor gene.
Cell 60: 495-508.

Rosendorff, J., Bernstein, R., Macdougall, L., Jenkins, T. (1987).
Fanconi anaemia: another disease of high prevalence in the Afrikaans population of South Africa.
Am. J. Med. Genet. 27: 793-797.

Rossiter, B.J.F., Caskey, T.C. (1995).
Impact of the human genome project on medical practice.
Annals Surg. Oncol. 2: 14-25.

Royer-Pokora, B., Kunkel, L.M., Monaco, A.P., Goff, S.C., Neuburger, P.E., Baehner, R.L., Coles, F.S., Curnutte, J.T., Orkin, S.H. (1986).
Cloning the gene for an inherited human disorder- chronic granulomatous disease- on the basis of its chromosomal location.
Nature 322; 32-38.

Royle, N.J., Clarkson, R.E., Wong, Z., Jeffreys, A.J. (1988).
Clustering of hypervariable minisatellites in the proterminal regions of human autosomes.
Genomics 3: 352-360.

Ruddle, F.H. (1981).
A new era in mammalian gene mapping: somatic cell genetics and recombinant DNA methodologies.
Nature 294: 115-120.

Rynditch, A., Kadi, F., Geryk, J., Zoubak, S., Svoboda, J., Bernardi, G. (1991).
The isopycnic, compartmentalized integration of Rous sarcoma virus sequences.
Gene 106: 165-172.

Saccone, S., De Sario, A., Della Valle, G., Bernardi, G. (1992).
The highest gene concentrations in the human genome are in T-bands of metaphase chromosomes.
Proc. Natl. Acad. Sci. USA 89: 4913-4917.

Sack, G.H. (1988).
Serum amyloid A (SAA) gene variations in familial Mediterranean fever.
Mol. Biol. Med. 5: 61-67.

Sack, G.H., Talvot, Jr. C.C., McCarthy, B.G., Harris, E.L., Kastner, D., Gruberg, L. Pras, M. (1991).
Exclusion of linkage between familial Mediterranean fever and the human serum amyloid A (SAA) gene cluster.
Hum. Genet. 87: 506-508.

Sambrook, J., Fritsch, E.F., Maniatis, T. (1989).
Molecular cloning: A laboratory manual. Second Edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Sato, T., Tanigami, A., Yamakawa, K., Akiyama, F., Kasumi, F. (1990).
Allelotype of breast cancer: cumulative allele losses promote tumor progression in primary breast cancer.
Cancer Res. 50: 7184-7189.

Sato, T., Akiyama, F., Sakamoto, G., Kasumi, F., Nakamura, Y. (1991).
Accumulation of genetic alterations and progression of primary breast cancer.

Cancer Res. 51: 5794-5799.

Savov, A., Angelicheva, A., Jordanova, A., Eigel, A., Kalaydjieva, L. (1992).
High percentage acrylamide gels improve resolution in SSCP analysis.
Nucleic Acids Res. 20: 6741-6742.

Sawadogo, M., Van Dyke, M.W. (1991).
A rapid method for the purification of deprotected oligodeoxynucleotides.
Nucl. Acids. Res. 19: 674.

Schlessinger, D. (1990).
Yeast artificial chromosomes: tools for mapping and analysis of complex genomes.
Trends in Genetics 6: 248-258.

Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., Bentolila, S., Birren, B.B., Butler, A. et al. (1996).
A gene map of the human genome.
Science 274: 540-546.

Schwabe, A.D., Peters, R.S. (1974).
Familial Mediterranean fever in Armenians: analysis of 100 cases.
Medicine 53: 453-462.

Scott, A.F., Schmeckpeper, B.J., Abdelrazik, M., Comey, T.C., O'Hara, B., Rossiter, J., Couley, T., Heath, P., Smith, K.D., Margolet, L. (1987).
Origin of the human L1 elements: Proposed progenitor genes deduced from a consensus DNA sequence.
Genomics 1: 113-125.

Scott, I.C., Halila, R., Jenkins, J.M., Mehan, S., Apostolou, S., Winqvist, R., Callen, D.F., Prockop, D.J., Peltonen, L., Kadler, K.E. (1996).
Molecular cloning, expression and chromosomal localization of a human gene encoding a 33 kDa putative metallopeptidase (PRSM1).
Gene 174: 135-43.

Sealy, P.G., Whittaker, P.A., Southern, E.M. (1985).
Removal of repeated sequences from hybridisation probes.
Nucl. Acids Res. 13: 1905-1922.

Shattuck-Eidens, D., McClure, M., Simard, J., Labrie, F., Narod, S., Couch, F., Weber, B., Castilla, L., Brody, L., Friedman, L., Ostermeyer, E. et al. (1995).
A collaborative survey of 80 mutations in the BRCA1 breast and ovarian cancer susceptibility gene: Implications for presymptomatic testing and screening.
J. Am. Med. Assoc. 273: 535-541.

Sheffield, V.C., Cox, D.R., Lerman, L.S., Myers, R.M. (1989).
Attachment of a 40 base pair G + C-rich sequence (GC clamp) to genomic DNA fragments by the polymerase chain reaction results in improved detection of single base changes.
Proc. Natl. Acad. Sci. USA 86: 232-236.

Sheffield, V.C., Beck, J.S., Kwitek, A.E. (1992).
Analysis of the efficiency of single base substitution detection by SSCP. Cold Spring Harbor Laboratory. pp. 1-149.

Sheffield, V.C., Beck, J.S., Kwitek, A.E., Sandstrom, D.W., Stone, E.M. (1993).
The sensitivity of single-strand conformation polymorphism analysis for the detection of single base substitutions.
Genomics 16: 325-332.

Shen, Y., Kozman, H.M., Thompson, A., Phillips, H.A., Holman, K., Nancarrow, J., Lane, S., Chen, L.Z., Apostolou, S., Doggett, N., Callen, D.F., Mulley, J.C., Sutherland, G.R., Richards, R.I. (1994).
A PCR-based genetic linkage map of human chromosome 16.
Genomics 22: 68-76.

Sherrington, R., Rogaev, E.I., Liang, Y., Rogaeva, E.A., Levesque, G., Ikeda, H., Chi, H., Lin, C., Li, G. et al. (1995).
Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease.
Nature 375: 754-760.

Shizuya, H., Birren, B., Kim, U.J., Mancino, V., Slepak, T., Tachiiri, Y., Simon, M. (1992).
Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector.
Proc. Natl. Acad. Sci. USA 89: 8794-8797.

Shohat, M., Shohat, T., Rotter, J.I., Schlesinger, M., Petersen, G.M., Pribyl, T., Sack, G., Schwabe, A.D., Korenberg, J.R. (1990a).

Serum amyloid A and P protein genes in familial Mediterranean fever.
Genomics 8: 83-89.

Shohat, M., Bu, X., Shohat, T., Fischel-Ghodsian, N., Magal, N., Nakamura, Y., Schwabe, A.D. Schlesinger, M., Danon, Y., Rotter, J.I. (1992).
The gene for familial Mediterranean fever in both Armenians and non-Ashkenazi Jews is linked to the a-globin complex on 16p: evidence for locus homogeneity.
Am. J. Hum. Genet. 51: 1349-1354.

Shohat, T., Shohat, M., Petersen, G.M., Sparkes, R.S., Langfield, D., Bickal, J., Korenberg, J.R., Schwabe, A.D., Rotter, J.I. (1990b).
Genetic marker family studies in familial Mediterranean fever (FMF) in Armenians.
Clin. Genet. 38: 332-339.

Silverman, G.A., Jockel, J.I., Domer, P.H., Mohr, R.M., Taillon-Miller, P., Korsmeyer, S.J. (1991).
Yeast artificial chromosome cloning of a two-megabase-size contig within chromosomal band 18q21 establishes physical linkage between BCL2 and plasminogen activator inhibitor type-2.
Genomics 9: 219-228.

Skirnisdottir, S., Eiriksdottir, G., Baldursson, T., Barkardottir, R.B., Egilsson, V., Ingvarrson, S. (1995).
High frequency of allelic imbalance at chromosome region 16q22-23 in human breast cancer: correlation with high PgR and low S phase.
Int. J. Cancer 64: 112-116.

Smith, C.L., Klco, S.R., Cantor, C.R. (1988).
Pulsed-field gel electrophoresis and the technology of large DNA molecules. In: Genome analysis: A practical approach. Ed. K.E. Davies. Oxford, IRL Press. pp. 41-72.

Smith, S.A., Easton, D.F., Evans, D.G.R., Ponder, B.A.J. (1992).
Allele losses in the region 17q12-21 in familial breast and ovarian cancer involve the wild-type chromosome.
Nat. Genet. 2: 128-131.

Soares, M.B., Bonaldo, M.F., Jelene P; Su, L., Lawton, L., Efstratiadis, A. (1994).
Construction and characterization of a normalized cDNA library.
Proc. Natl. Acad. Sci. USA 91: 9228-32.

Sohar, E., Gafni, J., Pras, M., Heller, H. (1967).
Familial Mediterranean Fever: a survey of 470 cases and review of the literature.
Am. J. Med. 43: 227-253.

Sommers, C.L., Thompson, E.W., Torri, J.A., Kemler, R., Gelmann, E.P., Byers, S.W. (1991).
Cell adhesion molecule uvomorulin expression in human breast cancer cell lines: relationship to morphology and invasive capacities.
Cell Growth Differ. 2: 365-372.

Stack, M., Jones, D., White, G., Liscia, D.S., Venesio, T., Casey, G., Crichton, D., Varley, J., Mitchell, E., Heighway, J., et al. (1995).
Detailed mapping and loss of heterozygosity analysis suggests a suppressor locus involved in sporadic breast cancer within a distal region of chromosome band 17p13.3.
Hum. Mol. Genet. 4: 2047-2055.

Stallings, R.L., Torney, D.C., Hildebrand, C.E., Longmire, J.L., Deaven, L.L., Jett, J.H., Doggett, N.A., Moyzis R. K. (1990).
Physical mapping of human chromosomes by repetitive sequence fingerprinting.
Proc. Natl. Acad. Sci. USA 87: 6218-6222.

Stallings, R.L., Ford, A.F., Nelson, D., Torney, D.C., Hildebrand, C.E., Moyzis, R.K. (1991).
Evolution and distribution of (GT)n repetitive sequences in mammalian genomes.
Genomics 10: 807-815.

Stallings, R.L., Doggett, N.A., Callen, D., Apostolou, S., Chen, L.Z., Nancarrow, J.K., Whitmore, S.A., Harris, P., Michison, H., Breuning, M., Saris, J.J., Fickett, J., Cinkosky, M., Torney, D.C., Hildebrand, C.E. and Moyzis, R.K. (1992a).
Evaluation of a cosmid contig physical map of human chromosome 16.
Genomics 13: 1031-1039.

Stallings, R.L., Doggett, N.A., Okumura, K., Ward, D.C. (1992b).
Chromosome 16 - specific repetitive DNA sequences that map to chromosomal regions known to undergo breakage/rearrangement in Leukemia cells.
Genomics 13: 332-338.

Stallings, R.L., Whitmore, S.A., Doggett, N.A., Callen, D.F. (1993).
Refined physical mapping of chromosome 16-specific low-abundance repetitive DNA sequences.
Cytogenet. Cell Genet. 63: 97-101.

Stanbridge, E.J. (1992).
Functional evidence for human tumour suppressor genes: chromosome and molecular genetic studies.
Cancer Surv. 12: 5-24.

Stewart, T.A., Pattengale, P.K., Leder, P. (1984).
Spontaneous mammary adenocarcinoma in transgenic mice that carry and express MTV-myc fusion genes.
Cell 38: 627-637

Strathdee, C.A., Buchwald, M. (1992).
Molecular and cellular biology of Fanconi Anemia.
Am. J. Pediatr. Hematol. Oncol. 14: 177-185.

Strathdee, C.A., Duncan, A.M.V., Buchwald, M. (1992b).
Evidence for at least four Fanconi Anemia genes including FACC on chromosome 9.
Nature Genet. 1; 196-198.

Strathdee, C.A., Gavish, H., Shannon, W.R., Buchwald, M. (1992c).
Cloning of cDNAs for Fanconi's anaemia by functional complementation.
Nature 356: 763-767.

Swanson, G.M., Ragheb, N.E., Chen-Sheng, L., Hankey, B., Miller, B., Horn-Ross, P., White, E., Liff, J.M., Harlan, L.C., McWhorter, W.P., Mullan, P., Key, C.R. (1993).
Cancer 72: 788-798.

Szabo, C.I., King, M-C. (1995).
Inherited breast and ovarian cancer.
Hum. Mol. Genet. 4: 1811-1817.

Tagle, D.A., Swaroop, M., Lovett, M., Collins, F.S. (1993).
Magnetic bead capture of expressed sequences encoded within large genomic segments.
Nature 361: 751-753.

Tagle, D.A., Swaroop, M., Elmer, L., Valdes, J., Blanchard-McQuate, K., Bates, G., Baxendale, S., Snell, R., MacDonald, M., Gusella, J., Lehrach, H., Collins, F.S. (1994).
Magnetic bead capture of cDNAs: A strategy for isolating expressed sequences encoded within large genomic segments.
In: Advances in biomagnetic separation. (Eds. Uhlen, M., Hornes, E., Olsvik, O.) Eaton Publishing Co., Natick, MA, USA. pp. 107-111.

Takeichi, M. (1991).
Cadherin cell adhesion receptors as a morphogenetic regulator.
Science 251: 1451-1455.

Takeichi, M. (1993).
Cadherins in cancer: Implications for invasion and metastasis.
Curr. Opin. Cell. Biol. 5: 806-811.

Tanaka, S., Louie, D.C., Kant, J.A., Reed, J.C. (1992).
Frequent incidence of somatic mutations in translocated BCL2 oncogenes of non-Hodgkin's lymphomas.
Blood 79: 229-237.

Tangir, J., Muto, M.G., Berkowitz, R.S., Welch, W.R., Bell, D.A., Mok, S.C. (1996).
A 400 kb novel deletion unit centromeric to the BRCA1 gene in sporadic epithelial ovarian cancer.
Oncogene 12: 735-740.

Tassebehji, M., Read, A.P., Newton, V.E., Harris, R., Balling, R., Gruss, P., Strachan, T. (1992).
Waardenburg's syndrome patients have mutations in the human homologue of the Pax-3 paired box gene.
Nature 355: 635.

Tassone, F., Xu, H., Burkin, H., Weissman, S., Gardiner, K. (1995).
cDNA selection from 10 Mb of chromosome 21 DNA: efficiency in transcriptional mapping and reflections of genome organization.
Hum. Mol. Genet. 4: 1509-1518.

Taub, R., Kirsch, I., Morton, C., Lenoir, G., Swan, D., Tronick, S., Aaronson, S., Leder, P. (1982).

Translocation of the c-myc gene into the immunoglobulin heavy chain locus in human Burkitt lymphoma and murine plasmacytoma cells.
Proc. Natl. Acad. Sci. USA 79: 7837-7841.

The Fanconi Anaemia/Breast Cancer Consortium (1996).
Group 1: Apostolou, S., Whitmore, S.A., Crawford, J., Lennon, G., Sutherland, G.R., Callen, D.F.
Group 2: Ianzano, L., Savino, M., D'Apolito, M., Notarangelo, A., Meneo, E., Piemontese, M.R., Zelante, L., Savoia, A.
Group 3: Gibson, R.A., Tipping A.J., Morgan, N.V., Hassock, S., Jansen, S., de Ravel, T.J., Van Berkel, C., Pronk, J., Easton, D.F., Mathew, C.G.
Group 4: Levran, O., Verlander, P.C., Dev Batish, S., Erlich, T., Auerbach, A.D.
Group 5: Cleton-Jansen, A., Moerland, E.W., Cornelisse, C.J.
Group 6: Doggett, N.A., Deaven, L.L., Moyzis, R.K.
Positional cloning of the Fanconi anaemia group A gene.
Nat. Genet. 14: 324-328.

Thomas, G.A., Raffel, C. (1991).
Loss of heterozygosity on 6q, 16q, and 17p in human central nervous system primitive neuroectodermal tumors.
Cancer Res. 51: 639-643.

Thompson, M.E., Zimmer, W.E., Haynes, A.L., Valentine, D.L., Forss-Petter, S., Scammell, J.G. (1992).
Prolactin granulogenesis is associated with increased secretogranin expression and aggregation in the golgi apparatus of GH4C1 cells.
Endocrinol. 131: 318-326.

Trask, B.J. (1991).
Fluorescence *in situ* hybridization: Applications in cytogenetics and gene mapping.
Trends Genet. 7: 149-154.

Trask, B., Christensen, M., Fertitta, A., Bergmann, A., Ashworth, L., Branscomb, E., Carrano, A., Van Den Engh, G. (1992).
Fluorescence *in situ* hybridization mapping of human chromosome 19: mapping and verification of cosmid contigs formed by random restriction enzyme fingerprinting.
Genomics 14: 162-167.

Tsuda, H., Zhang, W.D., Shimosato, Y., Yokota, J., Terada, M., Sugimura, T., Miyamura, T., Hirohashi, S. (1990).
Allele loss on chromosome 16 associated with progression of human hepatocellular carcinoma.
Proc. Natl. Acad. Sci. USA 87: 6791-6794.

Tsuda, H., Hirohashi, S., Shimosato, Y., Hirota, T., Tsugane, S., Watanabe, S., Terada, M., Yamamoto, H. (1990b).
Correlation between histological grade of malignancy and copy number of c-erbB-2 gene in breast carcinoma.
Cancer 65: 1794-1800.

Tsuda, H., Callen, D.F., Fukutomi, T., Nakamura, Y., Hirohashi, S. (1994).
Allele loss on chromosome 16q24-qter occurs frequently in breast cancers irrespectively of differences in the phenotype and extent of spread.
Cancer Research 54: 513-517.

Tsurugi, K., Mitsui, K. (1991).
Bilateral hydrophobic zipper as a hypothetical structure which binds acidic ribosomal protein family together on ribosomes in yeast Saccharomyces cerevisiae.
Biochem. Biophys. Res. Comm. 174: 1318-1323.

Uberbacher, E.C., Mural, R.J. (1991).
Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach.
Proc. Natl. Acad. Sci. USA 88: 11261-11265.

Van der Vijver, M.J. (1993).
Molecular genetic changes in human breast cancer.
Adv. Cancer Res. 61: 25-56.

Van Ness, J., Hahn, W.E. (1980).
Sequence complexity of cDNA transcribed from a diverse mRNA population.
Nuc. Acids Res. 8: 4259-4270.

Verlander, P.C., Lin, J.D., Udono, M.U., Zhang, Q., Gibson, R.A., Mathew, C.G., Auerbach, A.D. (1994).
Mutation analysis of the Fanconi anemia gene FACC.
Am. J. Hum. Genet. 54: 595-601.

Vogelstein, B., Fearon, E.R., Hamilton, S.R. Ken, S.E., Preisinger, A.C., Leppert, M., Nakamura, Y., White, R., Smits, A.M., Bos, J.L. (1988).
Genetic alterations during colorectal-tumor development.
N. Engl. J. Med. 319: 525-532.

Vogelstein, B., Kinzler, K.W. (1992).
p53 function and dysfunction.
Cell 70: 523

Vulpe, C., Levinson, B., Whitney, S., Packman, S., Gitschier, J. (1993).
Isolation of a candidate gene for Menkes disease and evidence that it encodes a copper transporting ATPase.
Nat. Genet. 3: 7-13.

Walker, A.P., Muscatelli, F., Monaco, A.P. (1993).
Isolation of the human Xp21 glycerol kinase gene by positional cloning.
Hum. Mol. Genet. 2: 107-114.

Warrington, J.A., Hall. L., Hinton, L.M., Miller, J.N., Wasmuth, J.J., Lovett, M. (1991).
Radiation hybrid map of 13 loci on the long arm of chromosome 5.
Genomics 11: 701-708.

Weber, J.L., May, P.E. (1989).
Abundant class of human DNA polymorphisms which can be typed using Polymerase Chain Reaction.
Am. J. Hum. Gen. 44: 388-396.

Weinberg, R. (1991).
Tumor suppressor genes.
Science 254: 1138-1146.

Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G., Lathrop, M. (1992).
A second-generation linkage map of the human genome.
Nature 359: 794-801.

Whitmore, S.A., Apostolou, S., Lane, S., Nancarrow, J.K., Phillips, H.A., Richards, R.I., Sutherland, G.R., Callen, D.F. (1994).

Isolation and characterization of transcribed sequences from a Chromosome 16 hn-cDNA library and the physical mapping of genes and transcribed sequences using a high resolution somatic cell hybrid panel of human chromosome 16.
Genomics 20: 169-175.

Wilcox, A.S., Khan, A.S., Hopkins, J.A., Sikela, J.M. (1991).
Use of 3' untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: implications for an expression map of the genome.
Nucl. Acids Res. 19: 1837-1843.

Williams, B.G., Blattner, F.R. (1979).
Construction and characterisation of the hybrid bacteriophage lambda Charon vectors for DNA cloning.
J. Virology 29: 555-575.

White, M.B., Carvalho, M., Derse, D., O'Brien, S.J., Dean, M. (1992).
Detecting single base substitutions as heteroduplex polymorphisms.
Genomics 12: 301-306.

Witt, D.R., Lew, S.P., Mann, J. (1988).
Heritable deletion of band 16q21 with normal phenotype: Relationship to late replicating DNA.
Am. J. Hum. Genet. 43: A127.

Wooster, R., Mangion, J., Eeles, R., Smith, S., Dowsett, M., Averill, D., Barrett-Lee, P., Easton, D.F., Ponder, B.A., Stratton, M.R. (1992).
A germline mutation in the androgen receptor gene in two brothers with breast cancer and Reifenstein syndrome.
Nature Genet. 2: 132-134.

Wooster, R., Neuhausen, S.L., Mangion, J., Quirk, Y., Ford, D., Collins, N., Nguyen, K., Seal, S., Tran, T., Averill, D., Fields, P., Marshall, G., Narod, S., Lenoir, G.M., Lynch, H., Feunteun, J., Devilee, P., Cornelisse, C.J., Menko, F.H., Daly, P.A., Orminston, W., McManus, R., Pye, C., Lewis, C.M., Cannon-Albright, L.A., Peto, B., Ponder, B., Skolnick, M.H., Easton, D.F., Golgar, D., Stratton, M. (1994).
Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13.
Science 265: 2088-2090.

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C., Micklem, G., Barfoot, R., Hamoudi, R., Patel, S., Rice, C., Biggs, P., Hashim, Y., Smith, A., Connor, F., Arason, A., Gudmundsson, J., Ficenec, D., Kelsell, D., Ford, D., Tonin, P., Bishop, D.T., Spurr, N.K., Ponder, B.A.J., Eeles, R., Peto, J., Devilee, P., Cornelisse, C., Lynch, H., Narod, S., Lenoir, G., Egilsson, V., Barkadottir, R.B., Easton, D.F., Bentley, D.R., Futreal, P.A., Ashworth, A., Stratton, M. (1995).
Identification of the breast cancer susceptibility gene BRCA2.
Nature 378: 789-792.

Wyman, A.R., White, R. (1980).
A highly polymorphic locus in human DNA.
Proc. Natl. Acad. Sci. USA 77: 6754-6758.

Yanisch-Perron, C., Vieira, J., Messing, J. (1985).
Improved M13 phage cloning vectors and host strains: nucleotide sequence of the M13mp18 and puc19 vectors.
Genet. 33: 103-119.

Yaspo, M-L., Gellen, L., Mott, R., Korn, B., Nizetic, D., Poustka, A., Lehrach, H. (1995).
Model for a transcript map of human chromosome 21: isolation of new coding sequences from exon and enriched cDNA libraries.
Hum. Mol. Genet. 4: 1291-1304.

Youil, R., Kemper, B.W., Cotton, R.G.H. (1995).
Screening for mutations by enzyme mismatch cleavage with T4 endonuclease VII.
Proc. Natl. Acad. Sci. USA 92: 87-91.

Yulig, I.G., Yulig, A., Fisher, E.M.C. (1995).
The frequency and position of Alu repeats in cDNAs, as determined by database searching.
Genomics 27: 544-548.

Zbar, B., Brauch, H., Talmadge, C., Linehan, M. (1987).
Loss of alleles of loci on the short arm of chromosome 3 in renal cell carcinoma.
Nature 327: 721-724.

Zheng, H., Hasty, P., Brenneman, M.A., Grompe, M., Gibbs, R.A., Wilson, J.H., Bradley, A. (1991).

Fidelity of targeted recombination in human fibroblasts and murine embryonic stem cells.
Proc. Natl. Acad. Sci. USA 88: 8067-8071.

Zucman, J., Delattre, O., Desmaze, C., Azambuja, C., Rouleau, G., De Jong, P., Aurias, A., Thomas, G. (1992).
Rapid isolation of cosmids from defined subregions by differential Alu-PCR hybridization on chromosome 22-specific library.
Genomics 13: 395-401.