Positional Cloning of Genes Associated with Human Disease

Scott Anthony Whitmore B.Sc.

Thesis submitted for the Degree of Doctor of Philosophy

Department of Cytogenetics and Molecular Genetics,

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

Faculty of Medicine, Department of Paediatrics,

University of Adelaide, South Australia.

January, 1999.

Table of Contents

	Page
Summary	I
Declaration	IV
List of Publications	V
Abbreviations	VI
Acknowledgements	VIII
Chapter 1: Literature Review	1
Chapter 2: Materials and Methods	58
Chapter 3: Physical Mapping of Chromosome 16q24.3	92
Chapter 4: Identification of Transcribed Sequences at 16q24.3	120
Chapter 5: Characterisation of the *GAS11* and *C16orf3* Genes	156
Chapter 6: Characterisation of Transcription Unit 6 (T6)	196
Chapter 7: Cloning of the Gene for Cystinosis	227
Chapter 8: General Discussion and Future Directions	248
References	255
Appendix: Publications	

Summary

Genetic linkage analysis has been successful in mapping the gene responsible for Fanconi anaemia type A (*FAA*) to chromosome 16q24.3. In addition, loss of heterozygosity (LOH) studies have localised a tumour suppressor gene involved in the development of sporadic breast cancer to the same chromosomal region. The main aim of the thesis was to isolate the gene(s) responsible for these disorders using a positional cloning strategy. At the start of the project there was a lack of cloned DNA and candidate genes located at 16q24.3, therefore a detailed physical map of this region was constructed based on overlapping cosmid, BAC, and PAC clones. The resulting map extends approximately 1.1 Mb from the telomere of chromosome 16q and consists of a minimum overlapping set of 35 cosmids, 2 PACs, and 1 BAC clone. This physical map encompasses the genetic markers that define the region most likely to contain the *FAA* gene and a breast cancer tumour suppressor gene.

Refined LOH and linkage analysis by collaborating laboratories was successful in narrowing down the candidate region of both diseases to approximately 750 kb between D16S303 and D16S3026. Selected cosmids spanning the region between these markers were used as templates for exon trapping experiments to enable identification of candidate genes. Additional transcript data was obtained from the analysis of ESTs mapped to 16q24.3 as part of the Human Gene Map construction at NCBI, and dbEST database screening of partial cosmid sequence (generated from collaborating laboratories). A total of 71 unique exons were trapped from 26 cosmid templates. Each one was successfully mapped back to its cosmid of origin allowing an integration of the physical and transcript map. Twenty eight exons showed significant homology to ESTs, while 7 corresponded to genes previously mapped to chromosome 16q24.3. A group of five exons could be linked based on their EST homology

and physical map proximity. The corresponding gene (*FAA*) was later shown by collaborators to be deleted for 2 of these 5 trapped exons in an Italian patient with Fanconi anaemia. Further screening of individuals affected with the disorder have shown a multitude of mutations associated with this gene. However, comparison of normal and breast tumour DNA failed to identify mutations in *FAA* restricted to tumour DNA. Therefore it was concluded that *FAA* was unlikely to have a role in breast tumourigenesis.

The region between D16S303 and D16S3026, containing at least 20 transcripts, was then screened for the presence of a breast cancer tumour suppressor gene. Two candidate transcripts were examined in detail. This involved the isolation of the complete sequence of the corresponding gene, followed by the identification of its genomic structure. Single stranded conformation polymorphism (SSCP) analysis was used to screen for mutations restricted to breast tumours. One of these transcripts, subsequently referred to as the growth arrest-specific 11 (*GAS11*) gene, was a likely candidate since the mouse homologue was expressed specifically during cell growth arrest. *GAS11* was found to consist of 11 exons, one of which was identified by exon trapping, which span approximately 25 kb of genomic DNA. SSCP analysis of breast tumour DNA failed to identify nucleotide sequence alterations when compared to corresponding normal DNA from the same individuals. In addition, Southern analysis of breast cancer cell line DNA failed to identify homozygous deletions encompassing *GAS11* and RT-PCR eliminated exon skipping as a disease causing mechanism in these cell lines as well as in affected individuals. These results suggest this gene is not involved in breast carcinogenesis. Another gene, *C16orf3* (chromosome 16 open reading frame 3), was found to lie within intron 2 of *GAS11*. This gene is expressed at a low level as a 1.2 kb mRNA, is intronless, and codes for a 125 amino acid protein which exists in two isoforms based on the presence or absence of two copies of an imperfect tetrapeptide repeat. SSCP mutation analysis

also indicated this gene was not mutated in breast tumours.

Analysis of a second transcript, T6, showed that this gene consisted of 18 exons (5 were identified from exon trapping) which code for a protein of 669 amino acids that does not show homology to any previously characterised proteins. SSCP mutation analysis of T6 again failed to identify SSCP changes specific for breast tumour DNA alone, suggesting this gene is also not involved in breast carcinogenesis.

The approach of exon trapping was also applied to a separate collaborative project, the positional cloning of a gene responsible for nephropathic cystinosis. The collaborators successfully constructed a cosmid contig across the ~500 kb candidate region on chromosome 17p13 and selected clones were subsequently used for exon trapping to identify candidate genes. A total of eight overlapping cosmids were used to identify 29 exons, 10 of which corresponded to characterised genes previously mapped to this chromosomal region. Of the 15 exons that did not display homology to database sequences, 6 were subsequently shown by collaborators to belong to a gene (*CTNS*) which was located within a region of homozygous deletion seen in a cystinosis family. SSCP analysis of individuals affected with cystinosis has subsequently identified 11 different disease-causing mutations within this gene.

The technique of exon trapping has therefore enabled the cloning of two disease genes, *FAA*, and *CTNS*, from a positional cloning approach. It has also allowed the identification of many other novel transcripts that are candidates for a tumour suppressor gene mapping to 16q24.3 and which have not yet been further characterised. The eventual identification of all genes (including tumour suppressor genes associated with LOH) involved in the oncogenic pathway in breast cancer will improve our understanding of tumour progression in this disease.

Declaration

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.

Signed: Date: 28/1/99

List of Publications

Most of the work presented in this thesis can be found in the following publications. A copy of each manuscript is given in the appendix.

The FAB Consortium. (1996). Positional cloning of the Fanconi anaemia group A gene. *Nature Genet.* **14**: 324-328.

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., and Savoia, A. (1997). The genomic organisation of the Fanconi anemia group A (*FAA*) gene. *Genomics* **41**: 309-314.

Whitmore, S.A., Crawford, J., Apostolou, S., Eyre, H., Baker, E., Lower, K.M., Settasatian, C., Goldup, S., Seshadri, R., Gibson, R.A., Mathew, C.G., Cleton-Jansen, A-M., Savoia, A., Pronk, J.C., Auerbach, A.D., Doggett, N.A., Sutherland, G.R., and Callen, D.F. (1998). Construction of a high-resolution physical and transcription map of chromosome 16q24.3: a region of frequent loss of heterozygosity in sporadic breast cancer. *Genomics* **50**: 1-8.

Whitmore, S.A., Settasatian, C., Crawford, J., Lower, K.M., McCallum, B., Seshadri, R., Cornelisse, C.J., Moerland, E.W., Cleton-Jansen, A-M., Tipping, A.J., Mathew, C.G., Savino, M., Savoia, A., Verlander, P., Auerbach, A.D., Van Berkel, C., Pronk, J.C., Doggett, N.A., and Callen, D.F. (1998). Characterisation and screening for mutations in breast cancer of the growth arrest-specific 11 (*GAS11*) and *C16orf3* genes at 16q24.3. *Genomics* **52**: 325-331.

Town, M., Jean, G., Cherqui, S., Attard, M., Forestier, L., Whitmore, S.A., Callen, D.F., Gribouval, O., Broyer, M., Bates, G.P., van't Hoff, W., and Antignac, C. (1998). A novel gene encoding an integral membrane protein is mutated in nephropathic cystinosis. *Nature Genet.* **18**: 319-324.

Abbreviations

BAC:	bacterial artificial chromosome.
bp:	base pairs.
BLAST:	basic local alignment search tool.
C16orf:	chromosome 16 open reading frame.
cDNA:	complementary deoxyribonucleic acid.
cM:	centimorgan.
cR:	centiray.
CGH:	comparative genomic hybridisation.
dbEST:	database of expressed sequence tags.
dNTP:	deoxynucleotide triphosphate.
DNA:	deoxyribonucleic acid.
DCIS:	ductal carcinoma *in situ*.
EST:	expressed sequence tag.
ET:	trapped exon.
FA:	Fanconi anaemia.
FAA:	Fanconi anaemia group A gene.
FAB:	Fanconi anaemia/Breast cancer.
FISH:	fluoresence *in situ* hybridisation.
GAS:	growth arrest-specific
hnRNA:	heteronuclear ribonucleic acid.
kb:	kilobase pairs.
LCIS:	lobular carcinoma *in situ*.
LOH:	loss of heterozygosity.
Mb:	megabase pairs.

µg:	microgram.
µl:	microlitre.
mg:	milligram.
ml:	millilitre.
mRNA:	messenger ribonucleic acid.
NCBI:	National Centre for Biotechnology Information.
ng:	nanograms.
ORF:	open reading frame.
PAC:	P1 artificial chromosome.
PCR:	polymerase chain reaction.
PFGE:	pulsed field gel electrophoresis.
RACE:	rapid amplification of cDNA ends.
RFLP:	restriction fragment length polymorphism.
RH:	radiation hybrid.
RNA:	ribonucleic acid.
RT:	reverse transcription.
SSCP:	single stranded conformation polymorphism.
STRP:	short tandem repeat polymorphism.
STS:	sequenced tagged site.
THC:	tentative human consensus sequence.
TIGR:	The Institute of Genomic Research.
UTR:	untranslated region.
VNTR:	variable number of tandem repeats.
WCH:	Women's and Children's Hospital (Adelaide).
YAC:	yeast artificial chromosome.

Acknowledgments

This study was performed in the Department of Cytogenetics and Molecular Genetics at the Women's and Children's Hospital in Adelaide. I am therefore extremely grateful to the Department for providing me with the opportunity and resources to conduct this research project. In particular, I would like to express my sincere thanks to my principal supervisors, Professor Grant Sutherland and Associate Professor David Callen, who have provided me with tremendous encouragement and support not only throughout this study, but since the first day I began in the Department. I also wish to thank them for advice and critical review of this thesis. Thankyou also to the Department of Paediatrics at the University of Adelaide who coordinated all aspects of the PhD program.

I would also like to thank all members of the Department who helped create a wonderful working environment. In particular, Kavita Bhalla, Rebecca Bilton, Joanna Crawford, Marina Kochetkova, Gabriel Kremmidiotis, Ingrid Lensink, Karen Lower, Jason Powell, and Chatri Settasatian, who made the lab an enjoyable place to be in. I would also like to thank Dr. Jozef Gecz for sharing his much appreciated technical knowledge and expert advice throughout the course of this study, and Liz Baker, Erica Woollatt, and Helen Eyre for all FISH work.

Thanks to all consortium members, in particular Dr. Anne-Marie Cleton-Jansen in Leiden for all LOH data and especially for providing the precious breast tumour DNA samples; Joanna Crawford and Dr. Sinoula Apostolou for their contributions to this thesis, as indicated in the text; Dr. Ram Seshadri, Sandra Goldup, and Brett McCallum from the Flinders Medical Centre in Adelaide for additional LOH data; Dr. Norman Doggett and Dr. Judy Tesmer for the relentless requests for cosmid clones; and all other international consortium members and

collaborators too numerous to mention.

Finally I would like to thank my family who have shown genuine interest and provided enormous support to me during this project. In particular, thankyou to my wife Helen, who has had to endure this thesis as much as I have. Your patience and love has helped keep me going which I appreciate so much.

Chapter 1

Literature

Review

Table of Contents

Page

1.1 Introduction 1

1.2 Mapping Human Chromosomes 5

1.2.1 Genetic Linkage Mapping 6

1.2.1.1 Protein Polymorphisms 6

1.2.1.2 Restriction Fragment Length Polymorphisms (RFLPs) 6

1.2.1.3 Microsatellite Markers 8

1.2.1.4 Return of the Single Nucleotide Polymorphism (SNP) 10

1.2.2 Physical Mapping 10

1.2.2.1 Cytogenetic Maps 11

1.2.2.2 Somatic Cell Hybrid Maps 12

1.2.2.3 Radiation Hybrid (RH) Maps 13

1.2.2.4 Whole-Genome Based STS Maps 14

1.2.2.5 Clone-Based Maps 14

1.2.2.5.1 Restriction Enzyme Based Maps 15

1.2.2.5.2 Repetitive Sequence Fingerprinted Maps 16

1.2.2.5.3 YAC-Based STS Content Maps 16

1.2.2.5.4 Bacterial Clone Maps 18

1.2.3 Transcription Maps 18

1.2.3.1 Classical Approaches 19

1.2.3.2 Chromosome Specific cDNA Libraries 20

1.2.3.3 Hybridisation-Based Approaches 21

1.2.3.4 Direct Selection/cDNA Selection 22

1.2.3.5 Exon Trapping/Exon Amplification 24

1.2.3.6 Complementary DNA Sequencing and Mapping 26

1.2.3.7 Identification of Genes by Sequence Homology 29

1.3 Mapping of Genes Associated With Disease 30

1.3.1 Positional Cloning of Disease Genes 30

1.3.2 Positional Candidate Cloning 30

1.4 Complex Disorders: Identification of Genes Involved in Cancer 31

1.4.1 Recessive Mutations in Cancer: The Retinoblastoma Paradigm 33

1.4.2 Loss of Heterozygosity and Other Tumour Types 34

1.5 Breast Cancer 35

1.5.1 Cytogenetic Studies of Breast Cancer 36

1.5.2 Familial Breast Cancer 37

1.5.3 Other Genes Involved in Breast Cancer: Prognostic Implications 40

1.5.4 Sporadic Breast Cancer and Loss of Heterozygosity 43

1.5.5 Loss of Heterozygosity and Chromosome 16q 45

1.5.6 Candidate Genes for 16q Loss of Heterozygosity 52

1.6 Project Aims 56

1.1 Introduction

In the past, the molecular analysis of human inherited disorders has occurred primarily through the identification and characterisation of specific proteins and their corresponding genes. For the majority of diseases where such information exists, the gene responsible has now been successfully cloned. However, for many inherited diseases with interesting phenotypes the biochemical basis for the disorder is unknown. Positional cloning provides an approach to characterise such disease genes. This approach does not rely on information regarding a chemical effect, but rather is initiated by mapping the disorder to a particular chromosomal region and subsequently cloning the gene responsible based on this localisation (Ruddle, 1984; Orkin, 1986; Collins, 1992). The increasing availability and resolution of integrated cytogenetic, genetic, physical and transcript maps provides the necessary reagents and information needed for disease gene localisation, and the subsequent cloning of the gene involved.

The Human Genome Project has in essence provided these resources for the positional cloning of genes associated with inherited disease. One of the major justifications and the premise behind the Human Genome Project is that the knowledge of our complete DNA structure will further our basic understandings of the role that various genes play in health and disease, both directly and through interactions with the environment. At the onset of this project, it was proposed that the ultimate goal would be to obtain the complete DNA sequence of each human chromosome, an aim that was at that time theoretically possible, but unacceptable in terms of cost in both time and money. This led to the establishment of a series of short-term goals, which apart from being fundamental for the concerted sequencing efforts that are beginning to dominate the later years of the project, have provided valuable tools towards the identification

and mapping of all human genes, particularly those associated with disease. Some of the specific objectives related to the human genome for the 5-year plan released in 1993 are listed below (Collins and Gallas, 1993; Dizikes, 1995). Each goal has been successfully completed or in fact exceeded at the end of this 5-year period (Collins *et al.*, 1998).

- **Genetic mapping:** Complete a human genetic map with markers spaced on average, 2 to 5 cM apart, with each marker being identified by a sequence-tagged site (STS).
- **Physical mapping:** Create detailed STS maps of all human chromosomes with markers spaced at 100 kb intervals. Generate overlapping sets of cloned DNA with continuity over 2 Mb for large parts of the human genome.
- **DNA sequencing:** Develop more efficient approaches to allow sequencing of large stretches of DNA and increase the sequencing capacity by increasing the number of groups actively involved in large-scale sequence production.
- **Gene identification:** Develop efficient methods for the identification of genes and for the placement of known genes on physical maps.
- **Informatics:** Develop effective software and database designs to support large-scale mapping and sequencing projects. Create database tools that provide easy access to up-to-date physical, genetic and chromosome mapping and sequencing information.

Positional cloning is a multi-step process beginning with the localisation of the disease gene using linkage analysis on pedigrees in which the responsible gene is segregating. In some cases, the disease gene may also be localised by the identification of specific associated chromosome abnormalities such as fragile sites, deletions, duplications or translocations. For cancer related

genes more complex methods such as comparative genomic hybridisation or loss of heterozygosity (LOH) studies can be used. Once the disease gene is localised to a candidate region, physically mapped DNA clones located within these sites then serve as substrates to identify genes lying within them using a variety of methods. A major impact in positional cloning has been the advent of single-pass end sequencing of random cDNA clones (Adams *et al.*, 1991,1995; Hillier *et al.*, 1996; Touchman *et al.*, 1997). From this sequence expressed sequence tags (ESTs) were developed and mapped to chromosomal regions (Polymeropolous *et al.*, 1992,1993; Berry *et al.*, 1995; Schuler *et al.*, 1996). This rapidly increasing resource has led to the positional candidate approach to disease gene isolation (Ballabio, 1993). In this approach, the information obtained from a gene sequence regarding its possible function, in combination with its localisation, is used to select candidate genes for a disease that has been mapped to the same chromosomal location. The steps involved in the positional cloning of disease genes is summarised in Figure 1.1.

Genetic linkage studies, which utilise polymorphic genetic markers, have been successful in mapping the gene responsible for Fanconi anaemia Type A (*FAA*) to chromosome 16q24.3. In addition, a tumour suppressor gene involved in the development of sporadic breast cancer has been mapped to the same region based on the results of detailed LOH studies. However, detailed physical and expressed sequence maps covering the critical region within this cytogenetic interval were not sufficiently developed to allow immediate identification of candidate genes. Therefore, the general aim of this study was to identify the gene(s) responsible for these disorders by a positional cloning approach. This involves the development of a high-density physical map of the 16q24.3 region, followed by the use of exon trapping to identify new transcripts which are potential candidates for both these disorders. In addition, the

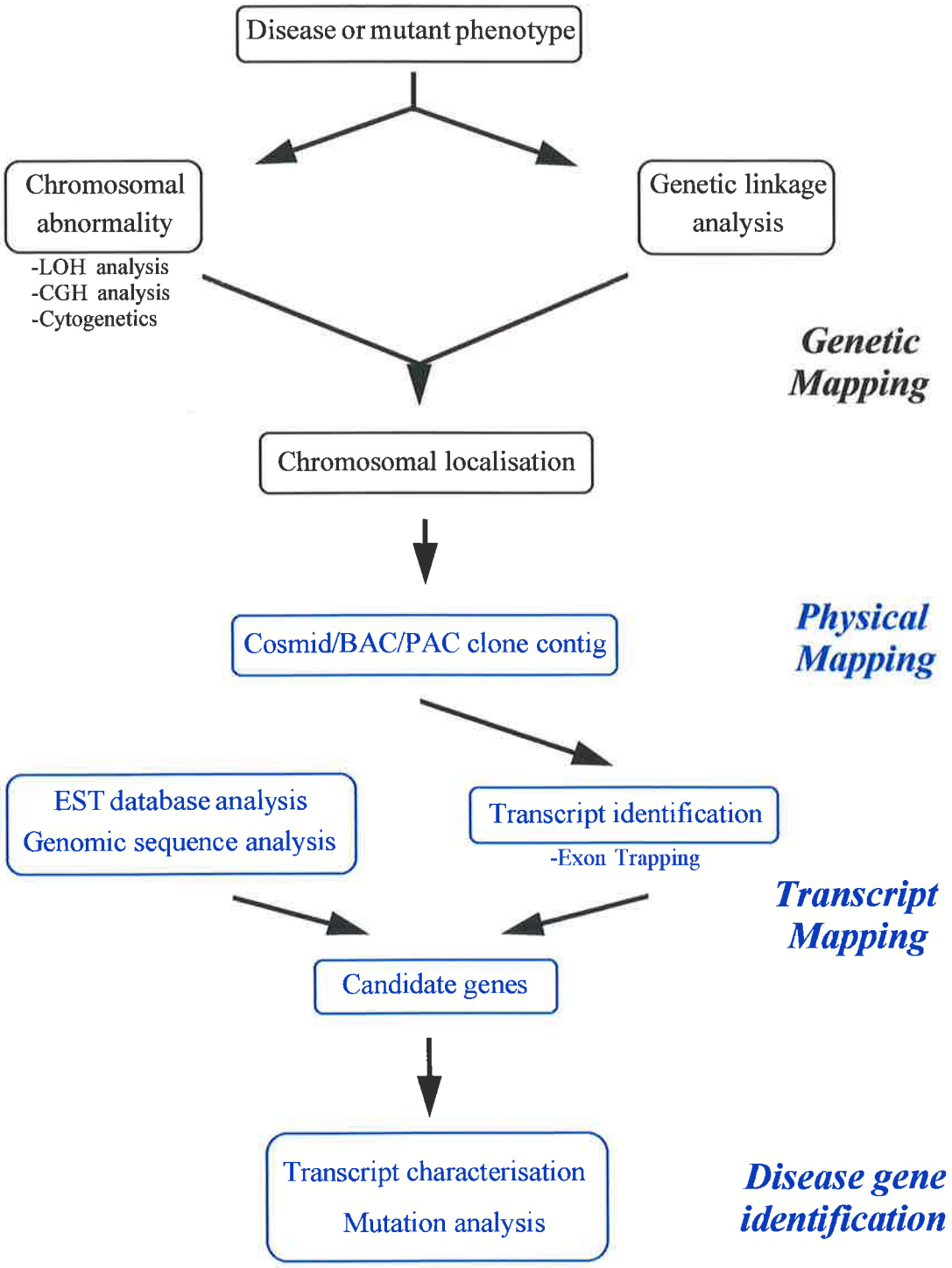


Figure 1.1: Schematic outline of the steps involved in the positional cloning of genes associated with disease. Blue text indicates steps in which the candidate has contributed and is described in this thesis.

gene responsible for the disease cystinosis has been mapped to chromosome 17p13 by genetic linkage analysis. The exon amplification technique will therefore be used towards the cloning of this gene also. Figure 1.1 summarises the contribution of the candidate in the overall positional cloning approach.

This literature review will initially discuss developments in the mapping of human chromosomes that has allowed the cloning of genes associated with disease. The latter part of this review will discuss the identification of genes involved in tumourigenesis, in particular those involved in the development of sporadic breast cancer. Finally, the specific aims of the thesis will be presented.

1.2 Mapping Human Chromosomes

In order to understand the biological basis of human disease and the role genes play in the determination of biological structure and function, knowledge is required of how the nucleotide sequence of all human chromosomes is organised. The construction and integration of maps at all levels of resolution is a vital first step in the determination of this DNA sequence. However, it is anticipated that the sequence of the entire genome will not become available until the year 2003 (Collins *et al.*, 1998). Therefore methods of gene identification which make use of the presently available integrated high-resolution maps are paramount for isolation of disease genes using positional cloning. The genome can be mapped by either genetic or physical methods to both order and determine the relative distance between markers. However, until the modern era of molecular biology, the only physical markers available were whole chromosomes and chromosomal bands. As a consequence particular phenotypic traits were mapped relative to one another by genetic linkage analysis.

1.2.1 Genetic Linkage Mapping

By determining the frequency of meiotic recombination between two genes, their proximity can be estimated. In principal, the further two genes are physically apart, the higher the frequency of genetic recombination between them. On the linkage map, distance between markers is estimated as a recombination frequency (θ), which can be transformed into a genetic distance measured in centimorgans (cM). In general, one cM on a genome average is approximately equivalent to one megabase of DNA sequence. However, a human genetic map was never going to be based solely on genes, but rather would have to be based on polymorphic markers that were not necessarily related to genes.

1.2.1.1 Protein Polymorphisms

The first polymorphic markers to be used for genetic mapping in humans were protein polymorphisms from blood group antigens and serum proteins (Renwick and Schulze, 1965). However, relatively few of these were available, their informativeness was low, and recombination occurred between only a few pairs of markers which were often not assigned to a chromosome. It was not until individual-to-individual variations in DNA sequence were identified, that it became practical to construct complete genetic linkage maps of humans (Botstein *et al.*, 1980).

1.2.1.2 Restriction Fragment Length Polymorphisms (RFLPs)

Single nucleotide polymorphisms occurring in the genome have the potential to create or destroy restriction enzyme sites. The variation in DNA sequence between individuals can therefore be observed as an alteration in the mobility of restriction fragments on agarose gels as detected by Southern analysis with single-copy DNA probes contained within these fragments. RFLPs were first used for genetic analysis in the study of temperature-sensitive

mutations of adenovirus in 1974 by Grodzicker *et al.* In humans, early studies identified RFLPs that were recognised by their relationship to the DNA sequence of particular genes, notably the γ-*globin* (Maniatis *et al.,* 1978) and β-*globin* (Kan and Dozy, 1978) genes.

In 1980, Botstein *et al.* suggested that for the more general purpose of mapping genetic loci, RFLPs did not have to encode the gene of interest, but only be located sufficiently nearby to display genetic linkage. They proposed that if sufficient numbers of polymorphic regions could be identified, then all genes could be linked to a region containing an RFLP and therefore be mapped. Subsequently, through the study of inheritance of randomly selected RFLPs in human families, linkages to a number of diseases were detected, including Duchenne muscular dystrophy (Davies *et al.*, 1983), Huntington's disease (Gusella *et al.*, 1983), retinoblastoma (Sparkes *et al.*, 1983), and adult polycystic kidney disease (Reeders *et al.*, 1985).

The realisation that the availability of a linkage map of the human genome would greatly assist the power of this approach to disease gene mapping, led to the construction of the first whole-genome genetic linkage map (Donis-Keller *et al.*, 1987). This study involved following the pattern of inheritance of 403 polymorphic loci, including 393 RFLPs, in a panel of DNAs from 21 three-generation families. Their studies established genetic maps for each of the 22 autosomes and the X chromosome, with an average spacing of 10 to 20 cM between markers. The limitation of this initial map was the relatively large distance between some adjacent markers, and the fact that RFLPs are not very informative and are difficult to type.

While most RFLP variants are the consequence of base-pair changes that create or destroy a restriction site, some are due to changes internal to the restriction fragment due to the presence of a variable number of short, tandemly repeated, DNA sequences (VNTRs). These

hypervariable minisatellite markers were first reported in 1980 by Wyman and White and in 1985, Jeffreys *et al.* developed probes from these markers that were able to detect many unique hypervariable loci. Although they were highly polymorphic they tended to cluster near the ends of chromosomes (Royle *et al.*, 1988) and were therefore not ideal markers for the generation of comprehensive genetic linkage maps.

1.2.1.3 Microsatellite Markers

By 1991, around 3,000 human polymorphic markers had been identified but only 10% of these had a heterozygosity of greater than 50% (Weissenbach, 1993). This, coupled with the fact that these markers were relatively unevenly spaced throughout the genome, meant that the current human linkage map was far from ideal. The identification of short tandem repeat polymorphisms (STRPs), or microsatellites, in 1989, and the ability for these markers to be assayed by the polymerase chain reaction (Litt and Luty, 1989; Weber and May, 1989), was a significant step towards the production of the comprehensive linkage maps that would supersede these first generation RFLP maps. The polymorphic nature of these STRPs, or sequence tagged site (STS) markers, is the result of a variation in the number of tandemly repeated units from one allele to another. Different classes of STRP loci have been reported, including di-, tri-, and tetra-nucleotides (Litt and Luty, 1989; Weber and May, 1989; Edwards *et al.*, 1991). With the exception of A_n:T_n multimers, $(CA)_n$:$(TG)_n$ repeats are the most frequent STRPs found in the human genome and have been the most common microsatellite used for the construction of genetic linkage maps. Both tri- and tetra-nucleotide repeats are easier to type than the CA di-nucleotide repeats, however they occur less frequently in the human genome and therefore have not been as widely used. For genetic mapping, these microsatellite markers, compared with RFLP and VNTR markers, were advantageous since

they were often highly polymorphic, were ubiquitous and abundant in the genome, and could be typed using significantly less DNA than with the hybridisation-based RFLP and VNTR probes. A further advantage was that information regarding each microsatellite marker could be distributed and implemented by knowledge of their DNA sequence and did not require the distribution of DNA clones as was necessary with RFLP based markers. Consequently, increasingly dense genome-wide human linkage maps were developed (NIH/CEPH collaborative mapping group, 1992; Weissenbach *et al.*, 1992; Buetow *et al.*, 1994; Gyapay *et al*, 1994; Matise *et al.*, 1994).

Due to the fact that different research groups often used different genetic markers to construct genetic maps of the same chromosome, it was vital that these studies utilised the same set of reference families. This enabled linkage data from individual laboratories to be pooled to generate maps of increasingly higher resolution. The most widely used set of reference families are those distributed through the Centre d'Etude du Polymorphisme Humain (CEPH), which now consist of sixty, three generation families. Consequently, in 1994, Murray *et al.* were able to compile an integrated human linkage map based on the efforts of three large groups and 110 CEPH collaborators. This map consisted of 5,840 loci, of which, 970 were uniquely ordered and spaced at an average density of 0.7 cM. Subsequent to this report, Dib *et al.* (1996) produced a comprehensive genetic map based on 5,264 CA di-nucleotide polymorphisms, of which 2,032 could be ordered with an odds ratio of 1,000:1 against alternative orders. The average interval size produced was 1.6 cM with only 1% of the genome containing intervals greater than 10 cM. These maps now provide significant resources and a framework to map simple Mendelian diseases to within small chromosomal intervals and also allow a first search for genetic causes of the more complex polygenic disorders.

1.2.1.4 Return of the Single Nucleotide Polymorphism (SNP)

The search for genetic linkage to a disease gene requires the typing of a large number of individuals with a large number of markers, a task that is time consuming, costly and not possible for many laboratories. With this in mind, recent attention has focussed on returning to SNPs as a source of genetic markers. Although, as with RFLPs, they have low heterozygosities, technology today has allowed identification of SNPs in sufficient numbers such that advantages over microsatellites now exist. They are highly abundant and have been predicted to occur on average every one kb (Cooper *et al.*, 1985). The fact that these markers can be genotyped by a simple plus or minus assay, suggests this process could be done in an automated fashion and some non-gel-based assays have already been proposed (Nickerson *et al.*, 1990; Livak *et al.*, 1995). It has been suggested that a map of 700-900 moderately polymorphic biallelic markers is equivalent to the current 300-400 microsatellite marker sets (Reed *et al.*, 1994; Dubovsky *et al.*, 1995) with 1,500-3,000 markers being superior (Kruglyak, 1997). Analysis of 2.3 Mb of human genomic sequence by both gel-based systems and DNA arrays has identified a total of 3,241 candidate SNPs with 2,227 being used to construct a genetic map (Wang *et al.*, 1998). This study also simultaneously typed 500 SNPs using a genotyping "chip" indicating the feasibility of future genetic linkage studies being automated and done *en masse*. The availability of dense SNP maps and the stabilisation of SNP genotyping technologies will ultimately assist in mapping genes responsible for genetically complex phenotypes through disease association studies (Kruglyak, 1997).

1.2.2 Physical Mapping

While the development of high-resolution genetic maps has assisted in the identification of disease loci by the use of linkage analysis, physical maps based on contiguous overlapping sets of DNA clones provide the basis for the subsequent cloning of the disease gene(s) involved.

High-density physical maps based on cloned DNA afford immediate access to any DNA segment of the genome that can be defined genetically. Therefore, integration of these maps is vital for the investigation of genetic disease and studying the large scale organisation of the genome.

1.2.2.1 Cytogenetic Maps

In essence, the first physical map of the human genome came from the identification of the correct diploid number of chromosomes (Tijo and Levan, 1956). Early cytogenetic maps grouped chromosomes according to their size and relative position of their centromere. However, not until the development of chromosomal banding techniques was it possible to uniquely distinguish all the chromosome pairs and define regions within chromosome arms (Caspersson *et al.*, 1970). Methods to hybridise cloned DNA fragments to metaphase chromosomes *in situ* were soon developed which allowed the mapping of unique probes to particular chromosomes. Initially, *in situ* hybridisation utilised tritium labelled DNA which was later followed by the development of fluorescently labelled probes and the fluorescence *in situ* hybridisation (FISH) technique (Landegent *et al.*, 1985). Development of prometaphase chromosome preparations allowed high-level banding resolution, and therefore localisation by FISH in the order of 2 to 3 Mb. More recent techniques using FISH on interphase or extended pronuclear chromatin have increased the mapping resolution to ~50 kb (Engh *et al.*, 1992; Heng *et al.*, 1992) while the use of FISH on extended DNA fibres (Fibre-FISH) has increased this resolution even further (Florijn *et al.*, 1995). Together, these techniques have enabled the mapping of individual DNA fragments to chromosomal regions, the ordering of DNA clones with respect to each other, and the ability to recognise chromsomal abnormalities associated with disease, such as fragile sites, chromosomal rearrangements, deletions, and amplifications.

1.2.2.2 Somatic Cell Hybrid Maps

Physical maps based on somatic cell hybrid panels derived from naturally occurring translocation or interstitial deletion chromosomes also enabled the physical mapping of cloned DNA fragments. Hybrid panels for a number of chromosomes were constructed by the fusion of human cells containing the chromosomal abnormality with either mouse or hamster cells deficient for a selectable gene marker such as *APRT* or *HPRT*. Somatic hybrid cell clones were then identified with tissue culture selection for this marker (Callen, 1986; Callen *et al.*, 1990*b*; Wagner *et al.*, 1991; Washington *et al.*, 1993). This allowed cloned DNA probes to be physically located to a defined chromosomal interval by Southern blot hybridisations to DNA isolated from the somatic cell hybrids and produced an ordering of the breakpoints and the probes (Callen *et al.*, 1988,1989). This task was a laborious approach but the advent of PCR and the development of STSs as genomic landmarks (Olson *et al.*, 1989), allowed this mapping to proceed rapidly and with the utilisation of fewer resources (Richards *et al.*, 1991; Callen *et al.*, 1995). The resolution of mapping by this approach was dependant on the size of the panel however an average mapping resolution of 1 Mb was achieved for the entire chromosome 16 (Callen *et al.*, 1995), while localised panels in some regions, for example Xq28, provided even higher resolution (Suthers *et al.*, 1990). Unfortunately, further high-resolution mapping was restricted by the identification and availability of additional translocation and deletion events, and a limitation in the selection systems that could be used.

The National Institute of General Medical Science (NIGMS) produced, and made available, a somatic cell hybrid mapping panel consisting of human and either mouse or hamster hybrids in which each hybrid had only one or two human chromosomes. As a general resource for the mapping of unique probes to individual chromosomes this panel has been quite useful either by Southern hybridisations (Fukushima *et al.*, 1994) or simple PCR based assays (Kremmidiotis *et*

al., 1998).

1.2.2.3 Radiation Hybrid (RH) Maps

An alternative approach for physical mapping was the construction and use of radiation hybrids (RH) which were originally developed in 1975 by Goss and Harris. These hybrids were generated by breaking the DNA of donor human cells into fragments whose size depended on the dose of irradiation applied. The radiated cell was then rescued by fusion to rodent recipient cells. The original approach was later modified by using, as the donor cell, a somatic cell hybrid containing a single human chromosome instead of a whole diploid human cell (Cox *et al.*, 1990). Following irradiation, the human and hamster chromosomal fragments rapidly rejoin, resulting in the formation of complex chromosomal rearrangements. The observation that these random fragments of human material could be stably retained in RHs provided the basis for RH mapping. In principal, the closer two markers were to each other on a chromosome, the greater the chance they would be present in the same RH. By screening a panel of such RHs, the distance between markers could be estimated in centiRays (cR). An advantage these maps had over somatic cell hybrids was that manipulating the dose of irradiation administered could create panels of varying resolution. To date a number of STS-based RH maps of individual chromosomes have been constructed (Cox *et al.*, 1990; James *et al.*, 1994; Shaw *et al.*, 1995; Dear *et al.*, 1998).

The development and availability of "whole-genome" panels represented by ~100 RHs has greatly facilitated the construction of a high-resolution framework with which to locate and order additional fragments of DNA (Walter *et al.*, 1994). At present, three different hamster-human RH panels of the entire genome have been constructed using different doses of irradiation resulting in a variation in the resolution of each map. These include the Genebridge

4 panel (Gyapy *et al.*, 1996) as well as the G3 and TNG panels, which were developed at the Stanford Human Genome Centre.

1.2.2.4 Whole-Genome Based STS Maps

One advantage of somatic cell hybrid and RH mapping compared with genetic linkage mapping is that the STS markers used were not required to be polymorphic. This enables mapping of both genetic polymorphic markers and non-polymorphic STSs to specific chromosomal intervals allowing the integration of genetic and physical maps (Kozman *et al* 1993; Mulley and Sutherland, 1993; Ferrero *et al.*, 1995). In 1995, Hudson *et al.* reported on the construction of a physical map of the entire genome based on 15,086 STSs (landmarks used to anchor the map) to give an average spacing between markers of 199 kb. This study involved the generation of a RH map of STSs generated from sequencing of random human genomic clones, the frequently used genetic linkage markers of Weissenbach and others, and STSs developed from ESTs (see 1.2.3.6). Another study developed an STS-based RH map of the human genome based on 10,478 markers (Stewart *et al.*, 1997). Of these, 5,994 were mapped such that odds of an alternate order were 1,000:1 giving rise to a framework map consisting of 1,776 "bins" with an average spacing of 500 kb between the 5,994 marker set. Both of these studies contribute greatly to the goal of the Human Genome Project of 30,000 STS spaced at an average of 100 kb apart (see 1.1).

1.2.2.5 Clone-Based Maps

The construction of the ultimate physical map, the complete nucleotide sequence of the genome, requires a series of cloned DNA fragments to be assembled which collectively provide full representation of the DNA to be sequenced. A number of approaches have been taken to construct genomic physical maps based on cloned DNA fragments with all of these techniques

defining overlaps between clones allowing the reconstitution of the original genomic order.

1.2.2.5.1 Restriction Enzyme Based Maps

The identification of site specific restriction endonucleases allowed the construction of the first genetic linkage maps in man, and in addition they have been instrumental in the development of the first clone-based physical maps involving *E. coli* (Kohara *et al.*, 1987), yeast (Olson *et al.*, 1986) and *C. elegans* (Coulson *et al.*, 1986). These maps were based on overlapping cosmid or lambda phage clones aligned by analysis of their fragment sizes produced when cleaved using a number of frequently cutting restriction enzymes. This technique was ideal for genomes of small to medium complexity and had also been applied to small regions of the human genome (Steinmetz *et al.*, 1986; Rommens *et al.*, 1989; Fearon *et al.*, 1990). With the use of pulsed field gradient gel electrophoresis (PFGE) and restriction enzymes that cut less frequently in DNA, the analysis of much larger stretches of DNA was made possible (Schwartz and Cantor, 1984). This technique, when used in combination with chromosome jumping or linking libraries (Collins and Weissman, 1984; Poustka *et al.*, 1987), allowed the construction of restriction maps based on such rare cutting enzymes as *Not*I (Ichikawa *et al.*, 1993; Allikmets *et al.*, 1994).

A *Not*I linking clone is defined as a short genomic clone, containing a single *Not*I restriction site, that identifies two adjacent genomic *Not*I fragments when used as a hybridisation probe. Therefore, two clones that hybridise to the same *Not*I fragment are derived from neighbouring *Not*I sites, the size of which can be determined by PFGE. This approach allowed the analysis of much larger regions of the genome, and provided an advantage in that actual physical distances between restriction sites could be determined.

1.2.2.5.2 Repetitive Sequence Fingerprinted Maps

The human genome is interspersed with numerous repetitive DNA elements such as *Alu*, LINE-1 and $(CA)_n$. It was suggested these could aid in the assembly of overlapping DNA fragments into contigs by providing "signatures" for restriction fragments and so allowing recognition of shared fragments between clones (Moyzis *et al.*, 1989). Although an attempt to utilise this approach to construct contigs of cosmids originating from an entire chromosome 16 proved to be successful (Stallings *et al.*, 1990), most contigs produced were small and had large overlaps. Therefore this technique was not suitable for producing physical maps of the whole genome and was finally abandoned when mapping technologies evolved further.

1.2.2.5.3 YAC-Based STS Content Maps

Unfortunately, a limitation of physical maps based on lambda or cosmid clones, was the restricted size of the inserts, which were usually a maximum of 40 kb. Entire human genome physical maps were made practical with the construction of a new type of cloning vector, the yeast artificial chromosome (YAC), which was capable of holding up to 1 to 2 Mb of DNA (Burke *et al.*, 1987), a significant improvement on previous vectors. The detailed STS maps that were being produced from the integration of genetic linkage and physical map construction therefore provided an excellent framework with which to isolate YAC clones and construct comprehensive clone contigs (Smith and Cantor, 1989).

Single copy landmark screening was first used to construct a physical map of the human Y chromosome and the long arm of chromosome 21 (Chumakov *et al.*, 1992; Foote *et al.*, 1992). In both these studies the physical map produced was based entirely on YAC clones identified by STS markers previously mapped to these two chromosomes. The resulting order of STSs was consistent with the previously established cytogenetic and mapping data and therefore it

was suggested that such an approach could be applied to the whole human genome. Subsequently in 1992, Bellanne-Chantelot *et al.* published a report toward a whole-genome YAC map and in 1993 a first-generation physical map of the entire human genome based on YAC clones was established (Cohen *et al.*, 1993). In this study, a 33,000 clone (10 genome equivalents) CEPH YAC library was screened with 2,100 polymorphic STS markers comprising the most recent genetic linkage map (Weissenbach *et al.*, 1992; Weissenbach, 1993). As well as using STSs to identify YAC clones, they isolated additional clones by generating specific sequences from the YACs already identified by PCR amplification between the ubiquitous *Alu* repeats contained within these YACs (Nelson *et al.*, 1989; Chumakov *et al.*, 1992; Doggett *et al.*, 1995). The integrity of the overlaps was confirmed with repetitive sequence fingerprinting, and FISH on metaphase chromosomes was used to locate individual contigs to specific chromosomal regions. In 1995, an update of this massive study indicated that they had successfully developed a reliable YAC contig map covering about 75% of the human genome consisting of 225 contigs having an average size of about 10 Mb (Chumakov *et al.*, 1995). To date, comprehensive YAC based physical maps covering a large number of individual human chromosomes have been subsequently established (Fan *et al.*, 1994; Collins *et al.*, 1995*b*; Doggett *et al.*, 1995; Gemmill *et al.*, 1995; Krauter *et al.*, 1995; Quackenbush *et al.*, 1995; Bouffard *et al.*, 1997).

One disadvantage of YAC cloning is a large percentage (up to 60%) of clones are chimaeric and therefore contain non-contiguous segments of the human genome. When chimaeric clones linked segments from different chromosomes they could be readily identified by FISH analysis, however syntenic chimaeric clones were more difficult to detect. A further problem was the instability of YAC clones containing certain areas of the genome, a reflection of the base composition at these regions (Foote *et al.*, 1992). This was particularly evident at the G-C rich

(gene rich) regions of the genome where few YAC clones were identified (Doggett *et al.*, 1995; De Sario *et al.*, 1996). This may be attributed to the yeast genome having a G-C base composition of only ~38%, suggesting the yeast replication machinery may have difficulty in replicating higher G-C content regions. Indeed, analysis of the whole genome YAC contig map produced by Chumakov *et al.* (1995), demonstrated that the large majority of gene dense bands were poorly represented (Saccone *et al.*, 1996). This problem has now been overcome by the emerging use of large-insert bacterial clones.

1.2.2.5.4 Bacterial Clone Maps

Although YAC maps have greatly contributed to the progress of detailed physical maps of the human genome, YACs were not a reliable source of representative genomic DNA and did not provide manageable substrates for sequencing purposes. As an alternative, large-insert cloning approaches using bacteriophage P1-based vectors (PACs) and bacterial artificial chromosomes (BACs) were established (Ioannou *et al.*, 1994; Shizuya *et al.*, 1992). The advantages of using libraries of the human genome generated with these vectors was the ease of DNA isolation and the ability to directly sequence from them. In addition, the clones were found to be highly stable, chimaerism was minimal, and inserts up to 300 kb could be achieved. As a result, YAC based maps of human chromosomes are being rapidly transformed into maps based on BACs (Kim *et al.*, 1996) and maps of genomic regions based solely on BACs are being constructed (Schmitt *et al.*, 1996; Marra *et al.*, 1997) in preparation for the large scale sequencing projects that are underway.

1.2.3 Transcription Maps

Methods which allow the rapid identification of the protein coding sequences contained within genomic DNA are essential for the construction of transcript maps covering large genomic

regions. Since the availability of the sequence of the whole genome is not estimated to be available until the year 2003 (Collins *et al.*, 1998), the development of transcript maps produced in the absence of this information is vital for the identification of disease causing genes.

1.2.3.1 Classical Approaches

Initially, one of the main approaches towards the isolation of protein coding regions in cloned genomic DNA was the isolation of single-copy DNA fragments within the region of interest. These fragments could first be used as hybridisation probes on Southern blot filters containing DNA isolated from a number of different species to test for evolutionary conservation of a particular DNA segment. These fragments could also be tested to determine if they originated from transcripts by hybridisation to Northern blots of RNA isolated from human tissues. If the DNA segment was part of a gene then it could be used as a probe to identify corresponding clones from a cDNA library generated from the relevant tissue(s). The use of cross-species hybridisation or "zoo blots" was important for the cloning of many disease genes, including identification of candidate cDNA clones for Duchenne muscular dystrophy (Monaco *et al.*, 1986), cloning of the *DCC* gene altered in colorectal cancers (Fearon *et al.*, 1990), and the cystic fibrosis gene (Rommens *et al.*, 1989).

Short regions of genomic DNA containing many sites for the restriction enzyme *Hpa*II, first referred to as *Hpa*II Tiny Fragment (HTF) islands, were shown to be extremely G-C rich (60 to 70%) compared with the average for the human genome (40%). These CpG islands remain unmethylated, are found in short regions of 1 to 2 kb, and account for ~2% of the genome. CpG islands have been shown to co-localise with the 5' end of genes (Bird, 1987; Antequera and Bird, 1993), and in most cases contain the promoter and one or more exons of the gene. In

humans, CpG islands have been found to be associated with all house keeping genes and about 40% of tissue-restricted genes (Larsen *et al.*, 1992; Antequera and Bird, 1993). The isolation of CpG island DNA was therefore utilised as a way in which genes could be identified. As the majority of CpG residues in genomic DNA are methylated, the use of restriction enzymes that target G-C rich DNA will therefore predominantly digest CpG island regions which remain unmethylated. The construction of linking libraries from a number of regions of the genome (see 1.2.2.5.1) has allowed identification and cloning of CpG islands, and has aided in the identification of many new genes (Rommens *et al.*, 1989; Tribioli *et al.*, 1994).

1.2.3.2 Chromosome Specific cDNA Libraries

Another method for the identification of transcribed sequences from a defined region of the genome is the construction of chromosome specific or region specific cDNA libraries. The availability of somatic cell hybrids that contain the human genomic region of interest as its only human DNA content have typically been used. The cDNA libraries are made from unprocessed (heteronuclear) messenger RNA (hnRNA) by using primers that bind to 5' intron consensus splice site sequences to initiate cDNA synthesis (Liu *et al.*, 1989; Whitmore *et al.*, 1994). The principal behind this procedure was that hnRNA still contains introns and exons, enabling the initiation of cDNA synthesis from intron/exon boundaries. The presence of human specific *Alu* repeats contained within the introns, and therefore within some cDNA clones, provided a means to screen for human specific cDNAs within the background of mouse clones. In a similar approach, cDNA synthesis was primed from hnRNA by oligonucleotides derived from conserved regions of human *Alu* repeats (Corbo *et al.*, 1990). However using this method, a high proportion of rodent clones were found in the cDNA libraries produced due to non-specific priming events. The specificity of *Alu* oligonucleotide binding was subsequently improved by the double-purification of hnRNA through oligo-dT cellulose columns (Lagoda *et*

al., 1994). Unfortunately, these approaches require extensive screening and subsequent verification of the human specific cDNA clones. In addition, the libraries produced only represent the human genes expressed in the originating somatic cell hybrid and therefore have limited use in the construction of whole-genome or even whole-chromosome transcription maps.

1.2.3.3 Hybridisation-Based Approaches

The human genes expressed in somatic cell hybrids have also been specifically cloned by subtractive hybridisation. The approach is based on the fact that cDNA fragments produced from non-coding segments of mature human RNA will not form stable heteroduplexes with their rodent homologues under high-stringency hybridisation conditions due to the low sequence homologies between the species in the 3' untranslated regions of their transcripts. Jones *et al.* (1992) used a somatic cell hybrid retaining a small region of human chromosome 17 to produce oligo-dT primed cDNA fragments. These were subsequently hybridised in solution with an excess of RNA from a similar somatic cell hybrid derived from this chromosome, such that the only difference between the two hybrids was the ~4 Mb human region under analysis. Under high stringency hybridisation conditions, the cDNA mixture, now enriched for human sequences expressed in one hybrid but not the other, could be used as a probe to screen conventional cDNA libraries and human cDNAs encoded within the non-overlap region could be obtained. This procedure was successful in the identification of nine expressed genes from within the non-overlap region and was suggested could be an approach applied to larger regions of non-overlap. However, as a general procedure for the isolation of transcribed sequences, its greatest utilisation would be for smaller defined regions of the genome.

A number of groups have used whole YAC inserts to screen cDNA libraries directly (Elvin *et al.*, 1990; Geraghty *et al.*, 1993). This involved labelling the YAC insert DNA to a high specific activity, blocking out high level repeat sequences, and then screening cDNA libraries to identify regions within the YAC that contain homology to expressed sequences. This approach, while successful in the identification of some new genes (Geraghty *et al.*, 1993; Kahloun *et al.*, 1993), had disadvantages because of the lack of suppression of low level repetitive sequences within the YAC clones. The hybridisation technique is close to the limits of sensitivity, resulting in small insert cDNA clones unlikely to be detected and YACs with large inserts being less efficient in detecting cDNAs. In one study (Elvin *et al.*, 1990), a 180 kb YAC containing the human aldolase reductase gene was used to directly screen a cDNA library in which the target cDNA was moderately abundant (1 in 10,000 clones). However, the screen succeeded in detecting only ~10% of cDNAs that were previously identified using an aldolase reductase cDNA probe.

Despite these problems, a modification of this approach was developed such that cDNA clones within libraries were first enriched before being screened by hybridisation with large genomic regions such as YACs, an approach termed direct selection (Lovett *et al.*, 1991) or cDNA selection (Parimoo *et al.*, 1991).

1.2.3.4 Direct Selection/cDNA Selection

The first development of direct selection was based on the initial amplification of inserts from a cDNA library of interest using vector specific primers. Products from this amplification are then labelled with ^{32}P, quenched for repetitive elements, and used as a probe to YAC or cosmid inserts that have been immobilised on nylon membranes. The cDNA inserts that specifically hybridise to the cloned genomic DNA are then eluted from the membranes, re-

amplified with vector specific primers, and the process repeated two or three times. The resulting cDNA sub-libraries, enriched for expressed sequences from the genomic region, are then cloned and analysed further. Using this strategy, a rare cDNA species encoded by a gene on a YAC from the *HLA* region, was enriched up to 7,000 fold with two rounds of selection (Parimoo *et al.*, 1991), and 1,000 to 2,000 fold enrichment for the *EPO* and *GNB2* cDNA clones respectively, were obtained from a human chromosome 7 YAC (Lovett *et al.*, 1991). Although the successful enrichment of certain cDNA clones cannot be disputed, these clones only contribute to a fraction of the resulting selected cDNA library due to simultaneous selection of cDNAs containing repeat sequences and other non-specific hybridisation events.

An alternative approach was to select the cDNAs in solution rather than by using filter-bound templates (Korn *et al.*, 1992). The method was based on the hybridisation in solution of amplified cDNA inserts to biotinylated cosmid DNA. Selected cDNA clones were then captured by binding to streptavidin coated magnetic beads, which could be isolated by passing over a magnet, while non-selected cDNA clones were washed from the hybridisation. After subsequent rounds of capture, the eluted cDNAs can be re-amplified by PCR, cloned, and analysed further. This method has been successful in the identification of many new genes (Rommens *et al.*, 1993; Peterson *et al.*, 1994; Baens *et al.*, 1995) including those responsible for a number of diseases (Gecz *et al.*, 1993; Onyango *et al.*, 1998).

This technique had several disadvantages. Abundant cDNA clones (due to highly expressed genes) present in a given library may be preferentially amplified, resulting in a further under-representation of rarer transcripts. Also, the presence of repetitive sequences in cDNA clones, particularly those of low abundance that are not sufficiently suppressed with human cot-1 pre-annealing, can give rise to false positive clones. A further disadvantage is that the origin of

cDNA clones is limited to those expressed in the tissue from which the library was constructed. Genes expressed at limited developmental stages or in restricted tissues therefore will not be identified by this technique.

1.2.3.5 Exon Trapping/Exon Amplification

A method of identifying potential protein coding regions of the genome that does not have these limitations is exon trapping or exon amplification. This method is based on the selection of functional spilce sites flanking exon coding sequences present within genomic DNA. Cloned genomic DNA is ligated into a vector that contains splicing sequences that can act in conjunction with splice site sequences that may be present in the cloned DNA fragment. Expression in mammalian cells in culture results in splicing of intronic DNA within the genomic fragment into a mature mRNA which can be subsequently isolated. An early version of this system (Duyk *et al.*, 1990) used a retroviral shuttle vector that contained a splice donor site, while the genomic fragment had to provide the splice acceptor site. This method relied on the presence of just one splice site within the genomic DNA fragment, resulting in the potential for the trapping of false positive exons due to cryptic splice acceptor sites present within the insert.

Another approach again made use of the presence of splice acceptor sites alone, within the cloned genomic fragment, to isolate 3' terminal exons from DNA cloned into a mammalian expression vector (Krizman and Berget, 1993). The advantage of this procedure was that terminal exons contain a polyadenylation signal in addition to a splice acceptor site and both were needed to allow correct splicing of the mRNA. The production of clones as a result of cryptic splicing events would therefore be eliminated, as they would not be likely to have polyadenylation signals located downstream of them.

An alternative exon trapping vector constructed by Buckler *et al.* (1991) required the presence of both a splice donor and splice acceptor site within the cloned genomic fragment for exons to be incorporated into the resulting mRNA. This vector, called pSPL1, contained a cloning site within an intron derived from the HIV-1 *tat* gene, whose flanking exons and splice sites were substituted for an exon of the rabbit β-globin gene (see Figure 4.1). Transcription was initiated from an SV40 early promoter with the polyadenylation signal also provided by SV40. Upon transfection of this vector (containing fragments of genomic DNA) into mammalian cells (COS-7), RNA transcripts were generated and the *tat* intron sequences spliced to produce a polyadenylated cytoplasmic RNA. When a fragment containing an intact exon is cloned into pSPL1 in the sense orientation, the exon should be retained (or trapped) in the mature poly(A)$^+$ RNA produced. The trapped products are then reverse transcribed and amplified using PCR (exon amplification). This vector has been used very successfully in the identification of a number of disease related genes including the gene for Huntington's disease (The Huntington's Disease Collaborative Research Group, 1993), the gene for Menkes disease (Vulpe *et al.*, 1993), the neurofibromatosis 2 tumour suppressor gene (Trofatter *et al.*, 1993), and the Batten disease gene (International Batten Disease Consortium, 1995).

Although this vector was quite effective, improvements were needed for its application for analyses of larger genomic regions. The first problem was the abundance of trapped products containing only pSPL1 sequences due to genomic fragments lacking intact exons, resulting in competition among PCR templates for this small and most abundant template. Another problem was the generation of false positives due to the presence of a cryptic splice site in the HIV-*tat* intron. These limitations were addressed leading to the construction of improved vectors (Church *et al.*, 1994; Burn *et al.*, 1995) that had the cryptic splice site removed, had additional restriction enzyme sites in the cloning region, and contained *Bst*XI half sites adjacent

to the splice sites of the vector. Clones derived from genomic DNA lacking exons would result in the formation of an intact *Bst*XI site that could be eliminated from further analysis by digestion with this enzyme. Additional modifications have allowed a more efficient approach to the isolation of exons from BAC DNA clones (Burn *et al.*, 1995; Hu *et al.*, 1997).

As well as being a useful technique for the identification of transcribed sequences in small genomic regions where disease genes have been mapped, exon trapping has also been applied to the identification of genes from whole chromosomal arms (Chen *et al.*, 1996*a*). In this study, 559 individual potential exons were trapped from 1,194 cosmid DNA clones derived from chromosome 21. They were able to trap exons from 13 of the 30 mapped chromosome 21 genes and estimated that they had identified portions of up to ~40% of all genes on this chromosome.

A distinct advantage of this procedure over cDNA selection is the avoidance of tissue or developmental specificity of gene expression. However, the isolation of exons depends on the distribution of restriction enzyme sites within particular genes, and competition between exons both *in vivo* and during PCR amplification may reduce the number of different exons trapped. In addition, the products generated are often too small in size for Northern analysis and therefore have to be confirmed as being parts of genes by linking with adjacent exons using RT-PCR or through database homologies to sequenced cDNA clones.

1.2.3.6 Complementary DNA Sequencing and Mapping

The large-scale sequencing of random cDNA clones to generate expressed sequence tags (ESTs) was seen as an important step towards the identification of a catalogue of human genes (Brenner, 1990). Just as STSs serve as physical mapping landmarks, PCR assays developed for

short stretches of cDNA sequence, provided an additional feature of developing an STS which points directly to an expressed gene (Adams *et al.*, 1991; Wilcox *et al.*, 1991). The sequencing of the 3' untranslated region (UTR) of mRNAs was chosen as they are generally free of introns (Hawkins, 1988) allowing PCR amplification from genomic DNA. Since the 3' UTR shows significant sequence divergence between human and rodent homologues, such ESTs avoid cross-species homology and so allow physical mapping using radiation hybrid and somatic cell hybrid panels. High-throughput cDNA sequencing began in 1991 (Adams *et al.*, 1991), with the mapping of a limited number of ESTs soon following (Khan *et al.*, 1992; Polymeropoulos *et al.*, 1992), and in 1992, a database called dbEST was established at the National Centre for Biotechnology Information (NCBI), to serve as a collection point for the large amount of cDNA sequence data being generated (Boguski *et al.*, 1993).

In 1995, Venter and colleagues at the Institute for Genomic Research (TIGR), generated 174,472 partial cDNA sequences from both the 3' and 5' ends of cDNA clones, which totalled more than 52 million nucleotides of sequence (Adams *et al.*, 1995). These data were generated from the use of 300 human cDNA libraries that collectively represented all major organs, several developmental stages and disease states, and many cell types. A project of similar scale funded by the Merck Company and carried out at the Washington University Genome Sequencing Centre, generated 319,311 sequence reads from the 3' and 5' ends of 194,031 human cDNA clones obtained from 17 different tissues representing 3 developmental states (Hillier *et al.*, 1996). Other large projects have also contributed to the human cDNA sequence data available in dbEST (Houlgatte *et al.*, 1995; Touchman *et al.*, 1997) and as of August 28, 1998, 1,086,919 human EST entries had been deposited in dbEST, with the majority of these being generated from individual cDNA clones. Many of these cDNA clones have been made available by the Integrated Molecular Analysis of Genomes and Their Expression (IMAGE)

consortium, who have been arraying and distributing the sequenced cDNA clones (Lennon *et al.*, 1996).

The integrity of the EST data generated depends to a large extent on the cDNA libraries that are used. In general, the frequency of occurrence of a cDNA clone in a library is equivalent to that of its corresponding mRNA, which results in a bias against genes of low expression. To overcome this problem, normalised cDNA libraries were constructed and used for sequencing purposes such that the frequency of all clones were within a narrow range (Soares *et al.*, 1994; Bonaldo *et al.*, 1996). Unfortunately, redundant identification of genes that are expressed in multiple tissues cannot be avoided from normalisation alone. As a result, subtractive cDNA libraries enriched for genes expressed at low levels are being developed (Bonaldo *et al.*, 1996). Initial results have shown a significant reduction in the representation of ~5,000 IMAGE consortium clones from a fetal liver/spleen subtracted cDNA library. With further work, the generation of complete subtracted and normalised cDNA libraries enriched for novel sequences should become possible, facilitating the isolation of many more human cDNAs not yet identified.

In order to map the large number of sequenced cDNA clones, individual ESTs were assembled into overlapping contigs based on sequence, with each contig likely to represent a unique gene. One approach has been the construction of the Genexpress Index (Houlgatte *et al.*, 1995), where 18,698 cDNA sequences from human skeletal muscle and brain libraries were clustered into 5,750 distinct gene transcripts according to their homology. Similarly, at TIGR, all sequencing reads available from dbEST and those produced inhouse were merged to form 62,808 tentative human consensus sequences (THCs), with full length transcript sequences being established in some cases (Adams *et al.*, 1995). Another approach has been used in the

construction of UniGene (Boguski and Schuler, 1995; Schuler *et al.*, 1996). In this case, ESTs and full-length mRNAs from characterised genes were organised into clusters most likely to represent distinct genes. The remaining ESTs were then screened against each other to determine those most likely to be derived from the same gene. Currently there are 62,421 UniGene clusters, that most likely represent up to two thirds of all human genes.

Until recently, only a limited number of ESTs had been mapped to specific chromosomal regions (Polymeropoulos *et al.*, 1993; Murakawa *et al.*, 1994; Berry *et al.*, 1995; Houlgatte *et al.*, 1995; Evans *et al.*, 1996). Following the generation of UniGene clusters, an international consortium was established to begin mapping each cluster using radiation hybrid panels and YAC maps, culminating in the mapping of 16,354 distinct clusters in 1996 (Schuler *et al.*, 1996) and more than 30,000 clusters in the latest release (Deloukas *et al.*, 1998). The complete mapping of all UniGene clusters will enable the construction of a transcript map that represents the majority of human genes.

1.2.3.7 Identification of Genes by Sequence Homology

Sequence comparison methods such as BLAST (Altschul *et al.*, 1990; 1997) can be used to scan dbEST and the non-redundant databases at NCBI to identify cDNA clones or gene families displaying homology to stretches of genomic sequence that are now becoming available. A direct comparison between a stretch of genomic DNA and a homologous cDNA will immediately yield the intron/exon structure of the gene and aid in the identification of alternatively spliced forms of some genes. The sequences of genes from other species can also be used to identify homologous cDNA clones in dbEST based on the conservation of functionally significant regions of the genome during evolution and this has been applied to the identification of the human gene homologues of *Drosophila*, *C. elegans*, *S. cerevisiae*, and

bacterial genes (Banfi *et al.*, 1996; Bassett *et al.*, 1997; Mushegian *et al.*, 1997; 1998).

1.3 Mapping of Genes Associated With Disease

1.3.1 Positional Cloning of Disease Genes

Historically, the cloning and molecular analysis of disease genes was based on information regarding the protein product, such as its amino acid sequence, availability of antibodies, or its function. Unfortunately, for the vast majority of disease genes, sufficient functional information is simply not available. With the generation of genetic maps of increasing resolution, it was now possible to map inherited monogenic and some multigenic disorders to specific chromosomal regions as a first step towards the cloning of the gene, a term initially referred to as "reverse genetics" or positional cloning (Orkin, 1986). Linkage analysis of genetic markers in families segregating the disorder can assign an initial map position for the responsible gene(s). Physically mapped cloned DNA contigs established between the flanking genetic markers could then subsequently be used to identify transcribed sequences that lie within this region. This approach has so far been successful in the identification of a large number of genes associated with many monogenic disorders (reviewed by Collins, 1995*a*).

1.3.2 Positional Candidate Cloning

The detailed maps of transcribed sequences being developed also provide an important tool for disease gene identification. When a new disease gene is mapped to a particular chromosomal region, it is now possible to access a list of all cDNA clones and genes assigned to that region. These transcripts are then considered candidates for that particular disease, with a direct comparison between the functions of these candidates (if known) to the features of the disease being made, leading to the identification of the affected gene. This process combines the

positional and functional cloning approaches into one, and has been referred to as positional candidate cloning (Ballabio, 1993).

1.4 Complex Disorders:Identification of Genes Involved in Cancer

While classical genetic linkage analysis is useful for the mapping of inherited single-gene disorders, it often fails to provide mapping information for more complex diseases such as epilepsy, heart disease, and cancer, where many genes can play a role in the phenotype, and environmental factors usually contribute to the disease. In these cases, alternative approaches need to be taken to determine the location of the genes that play a role in the aetiology of these disorders. The identification of genes associated with tumourigenesis relies on information that can be obtained not only through the mapping of familial cancer syndromes but through the cytogenetic analysis of tumour cells, and the finding of associated non-random cytogenetic alterations.

Heritable tendencies to cancer have been demonstrated in only about 1% of all cancer cases (Lasko *et al.*, 1991) and for most of these, predisposition is inherited as an autosomal dominant trait with detection dependant on the occurrence of a rare tumour, or in the case of a common cancer, a shifted age distribution. Cytogenetic techniques applied to leukaemias, lymphomas, and some solid tumours have led to the identification of specific chromosomal abnormalities unique to particular tumours (Mitelman *et al.*, 1997). Some of the changes seen in solid tumours include numerical aberrations, translocations, gene amplifications and interstitial and terminal deletions. These cytogenetic aberrations often give an indication as to the location of the genes involved, and in the case of retinoblastoma, deletions and non-random translocations in affected individuals have assisted in the identification and subsequent cloning

of the gene (Yunis and Ramsay, 1978; Strong *et al.*, 1981; Balaban *et al.*, 1982).

For a normal somatic cell to evolve into a metastatic tumour it requires changes at the cellular level, such as immortalisation, loss of contact inhibition and invasive growth capacity, and changes at the tissue level, such as evasion of host immune responses and growth restraints imposed by surrounding cells, and the formation of a blood supply for the growing tumour. While these processes require the coordinated induction and control of a large number of genes, only that cell which has been able to obtain all the necessary functions will eventually form a malignant foci.

Molecular genetic studies of colorectal carcinoma have provided substantial evidence that the generation of malignancy requires the sequential accumulation of a number of genetic changes within the same epithelial stem cell of the colon (Fearon and Vogelstein, 1990), with gene alterations encompassing both inactivating and activating mechanisms. Three broad classes of genes have been defined:

- **Tumour suppressor genes** – these genes inhibit cell proliferation and both alleles must be inactivated to have an effect on the cell.
- **Oncogenes** – these genes positively promote cell proliferation. Normal non-mutant versions are termed proto-oncogenes with the mutant versions excessively or inappropriately active.
- **Mutator genes** – these genes maintain the integrity of the genome. Loss of both alleles can result in an instability of the genome increasing the overall mutation rate.

For a normal colonic epithelial cell to become a benign adenoma, progress to intermediate and

late adenomas, and finally become a malignant cell, mutations in genes belonging to these three groups are required (Fearon and Vogelstein, 1990). Constitutional loss of the *APC* tumour suppressor gene at 5q21 is sufficient to initiate the formation of benign polyps. The development of intermediate and late adenomas has been linked to the activation of the *KRAS* oncogene and about 50% of late adenomas show a mutation in the *DCC* tumour suppressor gene on 18q. Also colorectal cancers, but not adenomas, have a high frequency of mutations in the *TP53* gene. Finally, mutations in the mutator genes, *MSH2* and *MLH1*, have been identified which most likely increase the overall mutation rate making each individual transition more likely.

1.4.1 Recessive Mutations in Cancer: The Retinoblastoma Paradigm

Tumour suppressor genes were first identified in the childhood cancer retinoblastoma. Both inherited and sporadic forms of this cancer exist, with the genetic form inherited as a highly penetrant autosomal dominant trait, which was mapped to chromosome 13q14 by genetic linkage analysis (Sparkes *et al.*, 1983). Cytogenetic analysis also showed deletions of this chromosomal region in patients (Yunis and Ramsay, 1978) and in tumours from patients without constitutional deletions (Balaban *et al.*, 1982).

Due to these cytogenetic findings, it was possible that the inherited retinoblastoma mutation was a recessive allele with inactivation of the normal allele required for its expression in tumourigenesis. Coupled with the observation that bilateral retinoblastoma was characteristic of the inherited disease and occurred at an early age, whereas unilateral retinoblastoma was characteristic of the sporadic form and occurred at a later age, led to the hypothesis by Knudson in 1971, that the tumour arises from two mutational steps. With this proposition, familial cancers would result from an inherited germline mutation of a gene suppressing the

growth of cells (tumour suppressor gene), such that all cells would carry this mutation. A second mutation or "hit" in any cell therefore resulted in the manifestation of the recessive mutation leading to cancer. The fact that only one more "hit" produces a cancerous cell meant that individuals with an inherited pre-disposition to the disease had an earlier age of onset and often bilateral tumours. In contrast, sporadic cases tended to be in one eye and later in onset because two "hits" were needed to the genes in the same cell.

This hypothesis was confirmed in 1983, when Cavenee and colleagues used RFLPs at 13q14 to type DNA isolated from blood and tumour samples taken from the same affected individuals. They noted that in several cases the constitutional DNA from lymphocytes was heterozygous for some markers but the tumour cells appeared homozygous for the same markers. The apparent reduction to homozygosity (or loss of heterozygosity, LOH) through the loss of one allele of these markers was suggested to be the second "hit" which was removing the remaining functional copy of the retinoblastoma gene in these individuals. The analysis of tumours in familial cases showed that the chromosome from the unaffected parent was in each instance the one eliminated from the tumour. A number of mechanisms were proposed including mitotic recombination, mitotic non-disjunction with loss of the wild-type allele or reduplication of the mutant allele, and gene conversion, deletion or mutation.

The cloning of the retinoblastoma (*RB*) gene (Friend *et al.*, 1986) and subsequent studies have established that it is a tumour suppressor gene acting through the inhibition of the G_1/S progression in the cell cycle (reviewed by Weinberg, 1995).

1.4.2 Loss of Heterozygosity and Other Tumour Types

Loss of heterozygosity analysis is basically a refined version of cytogenetic studies with this

method able to detect much smaller genetic alterations in tumour cells as compared to cytogenetically visible rearrangements. The use of a large number of genetic markers, as provided by the comprehensive genetic linkage maps available for all chromosomes, to analyse paired samples of DNA from normal and tumour tissue, has identified regions of the genome displaying LOH which are often specific to tumour types (reviewed in Lasko *et al.*, 1991). Reproducible LOH is therefore likely to indicate the presence of tumour suppressor genes important in tumour pathogenesis.

In addition to retinoblastoma, studies of other cancers have supported the model that LOH is a specific event in the pathogenesis of cancer. In Von Hippel-Lindau (VHL) syndrome both sporadic and inherited cases of the syndrome show LOH for the short arm of chromosome 3 (Bergerheim *et al.*, 1989; King *et al.*, 1986). Somatic translocations involving 3p in sporadic tumours, and genetic linkage to the same region in affected families has also been observed (Cohen *et al.*, 1979; Seizinger *et al.*, 1988). Similarly, in colorectal carcinoma, inherited forms of the disease have been mapped to the long arm of chromosome 5 (Bodmer *et al.*, 1987; Leppert *et al.*, 1987) while LOH at 5q has been reported in both the familial and sporadic versions of the disease (Sasaki *et al.*, 1989) and the *APC* gene, mapping to this region, has been shown to be involved (Groden *et al.*, 1991). Other examples include the *TP53* and *NF2* genes (reviewed in Lanfrancone *et al.*, 1994) which firmly establishes the fact that a general mechanism in human cancer is the inactivation of tumour suppressor genes by LOH.

1.5 Breast Cancer

Breast cancer is the most common malignancy seen in women affecting approximately 10% of females in the Western world. The route to breast cancer is not as well mapped as that of colon

cancer due to the histological stages of breast cancer development being largely unknown. It is known however, that breast cancer is derived from the epithelial lining of terminal mammary ducts or lobuli (Russo and Russo, 1991). The most common histological type of invasive carcinoma is the infiltrating ductal carcinoma which constitutes approximately 90% of all tumours, while the two most important non-invasive tumours are ductal carcinoma *in situ* (DCIS) and lobular carcinoma *in situ* (LCIS). Hormonal influences, such as those exerted by oestrogen, are believed to be important because of the marked increase in breast cancer incidence in post-menopausal women, but the initial steps in breast cancer development probably occur before the onset of menopause (Devilee and Cornelisse, 1994). As with colon carcinoma, it is believed that a number of genes need to become involved in a stepwise progression during breast tumourigenesis.

1.5.1 Cytogenetic Studies of Breast Cancer

Although breast carcinomas are among the most frequent malignant tumours, there is difficulty in discerning any characteristic cytogenetic abnormality due to clonal heterogeneity and the difficulty in obtaining high quality metaphases from these tumours. Therefore, less than 300 breast tumours have been karyotyped to date (Devilee and Cornelisse, 1994) with no characteristic cytogenetic alteration observed (Dutrillaux *et al.*, 1990). Overall, the most frequent changes seem to be in chromosome number, including trisomies of 7 and 18 and monosomies of 6, 8, 11, 13, 16, 17, 22, and X (reviewed by Devilee and Cornelisse, 1994). In addition, breakpoints of structural abnormalities have been shown to cluster to several regions and have included 1p22-q11, 3p11, 6p11-13, 7p11-q11, 8p11-q11, 16q, and 19q13 (Thompson *et al.*, 1993). Also, various rearrangements involving either 1q or 16q or both have been observed as the sole cytogenetic abnormality (Dutrillaux *et al.*, 1990) and interestingly, alterations in chromosome 1 and 16 have also been seen in several cases of *in situ* carcinoma

(Pandis *et al.*, 1992; Nielsen *et al.*, 1989).

1.5.2 Familial Breast Cancer

Certain women appear to be at an increased risk of developing breast cancer. The existence of genes responsible for an inherited pre-disposition to breast cancer was suggested before the turn of the century by Paul Broca, with segregation analysis many years later showing that 5 to 10% of all breast cancers were due to at least two autosomal dominant susceptibility genes. Generally, women carrying a mutation in a susceptibility gene develop breast cancer at a younger age compared to the general population, often have bilateral breast tumours, and are at an increased risk of developing cancers in other organs, particularly carcinoma of the ovary.

Genetic linkage analysis on families where the onset of breast cancer occurred before the age of 46 was successful in mapping the first susceptibility gene, *BRCA1*, to chromosome 17q21 (Hall *et al.*, 1990). However, of the 23 families studied, only 7 (40%) showed linkage to the D17S74 VNTR marker, indicating that further susceptibility genes were involved in the remaining families. Additional studies in 214 families established that mutations in *BRCA1* are responsible for predisposition in nearly all families with both breast and ovarian cancers, but only responsible for approximately 45% of those with breast cancer alone (Easton *et al.*, 1993). In a study involving families predisposed to breast cancer with at least one case of male breast cancer, linkage to *BRCA1* was not found (Stratton *et al.*, 1994) again indicating the existence of additional susceptibility loci. In 1994, linkage analysis was successful in mapping the *BRCA2* gene to chromosome 13q12-q13 (Wooster *et al.*, 1994) with this gene conferring a higher incidence of male breast cancer and a lower incidence of ovarian cancer when compared to *BRCA1*. In the analysis of approximately 200 families with at least 4 affected members, around 50% of families had convincing linkage to *BRCA1*, around 30% had linkage to *BRCA2*

and the remaining 20% were not linked to either of these loci (Szabo and King, 1995). The existence of a third locus has therefore been suggested (Serova *et al.*, 1997), with evidence of a gene on chromosome 8p proposed (Kerangueven *et al.*, 1995), however this requires further confirmation.

Analysis of breast cancers from both *BRCA1* and *BRCA2* families has shown that LOH of 17q and 13p markers respectively, is common (Smith *et al.*, 1992; Collins *et al.*, 1995*c*). Furthermore, the allele lost is always the one that carries the wild type gene with the mutant allele being retained in the tumours. This suggests in both cases tumour predisposition is due to loss of a normal *BRCA* gene, and provides evidence that both *BRCA1* and *BRCA2* are tumour suppressor genes.

In 1994, the *BRCA1* gene was positionally cloned and shown to encode a protein consisting of 1,863 amino acids (Miki *et al.*, 1994). Nearly 100 different mutations in this gene have been identified without evidence of mutation clustering, with the most common mutations leading to truncation of the protein product (Castilla *et al.*, 1994; Friedman *et al.*, 1994; Simard *et al.*, 1994; Shattuck-Eidens *et al.*, 1995; Struewing *et al.*, 1995). Although most of the coding region shows no sequence similarity to previously identified genes, the presence of a RING zinc finger domain at the amino terminus of the protein suggests it may be involved in mediating protein-protein interactions (Saurin *et al.*, 1996). The finding that the *BRCA1* protein co-localises, co-immunoprecipitates, and forms a complex with the *HsRad51* protein (Scully *et al.*, 1997), suggests that it may participate in *Rad51* functions. These functions primarily concern the double-strand-break repair pathway (Shinohara *et al.*, 1992; Baumann *et al.*, 1996), with the *Saccharomyces cerevisiae* homologue, *ScRad51*, shown to be essential for mitotic and meiotic recombination (Rockmill *et al.*, 1995).

The *BRCA2* gene was cloned in 1995 and encodes a protein consisting of 3,418 amino acids and shows no significant homology to any known protein (Wooster *et al.*, 1995). Although fewer mutations in this gene have been reported to date than with *BRCA1*, again most appear to give rise to a truncated protein product (Couch *et al.*, 1996*a*; Neuhausen *et al.*, 1996; Phelan *et al.*, 1996; Tavtigian *et al.*, 1996). Although these two genes are not significantly related in sequence, both appear to be coordinately regulated during proliferation and differentiation in mammary epithelial cells (Rajan *et al.*, 1996). Also, they have both been shown to be highly expressed in rapidly proliferating cells, with the expression peaking at the G_1/S boundary of the cell cycle, before DNA synthesis (Rajan *et al.*, 1996). This expression profile, together with the common breast cancer phenotype, suggests that these genes may be acting in the same biological pathway.

Recent studies of the *BRCA2* gene have shown that it too binds to the *Rad51* gene (Sharan *et al.*, 1997). *Brca2* knockout mice have also shown early embryonic lethality similar to that seen in *Rad51* and *Brca1* knockout mice. Also, hypersensitivity to irradiation has been observed in *Brca2* mutated cancer cells (Sharan *et al.*, 1997; Abbott *et al.*, 1998) suggesting, as with *BRCA1*, this gene may be involved in the repair of DNA breaks, thereby controlling cell cycle progression.

Additional inherited breast cancer syndromes exist, however they are rare. Inherited mutations in the *TP53* gene have been identified in individuals with Li-Fraumeni syndrome (Malkin *et al.*, 1990), a familial cancer resulting in epithelial neoplasms occurring at multiple sites including the breast. Similarly, germline mutations in the *MMAC1/PTEN* gene involved in Cowden's disease (Steck *et al.*, 1997; Liaw *et al.*, 1997) and the ataxia telangiectasia (*AT*) gene (Easton, 1994) have been shown to confer an increased risk of developing breast cancer, among other

clinical manifestations, but together account for only a small percentage of families with an inherited predisposition to breast cancer.

Somatic mutations in the *TP53* gene have been shown to occur in a high percentage of individuals with sporadic breast cancer (see 1.5.3). However, although LOH has been observed at the *BRCA1* and *BRCA2* loci at a frequency of 30 to 40% in sporadic cases (Cleton-Jansen *et al.*, 1995; Saito *et al.*, 1993), there is virtually no sign of somatic mutations in the retained allele of these two genes in sporadic cancers (Futreal *et al.*, 1994; Miki *et al.*, 1996). Possible explanations for this have been proposed (Kinzler and Vogelstein, 1997), which take into account the possible role of these genes in the regulation of cell growth. For example, the *TP53* gene may be a "gatekeeper" which directly regulates growth of cells, whereas the *BRCA* genes may be "caretakers", with inactivation of these two genes leading to an instability of the genome, rather than promoting tumour initiation directly. For normal individuals to develop breast cancer they would therefore have to acquire at least four somatic mutations in the same breast epithelial cell, two mutations in a caretaker gene, followed by mutations in both alleles of a gatekeeper gene, which may account for the low number of sporadic cases showing complete loss of either the *BRCA1* or *BRCA2* gene.

1.5.3 Other Genes Involved in Breast Cancer: Prognostic Implications

Although numerous oncogenes involved in cancer have been identified, only a small number appear to play a part in the development of breast cancer. *ERBB2* at 17q12, a gene encoding a tyrosine kinase growth factor receptor, has been shown to be amplified in about 20 to 25% of breast tumours (Slamon *et al.*, 1989). However, examination of tumours that have metastasised to the lymph nodes (node positive tumours), shows this number increases to 35%, and amplification was correlated with a poorer prognosis and relapse (Gullick *et al.*, 1991).

However, a number of subsequent studies have been unable to confirm this result (reviewed in van de Vijver, 1993).

Amplification of the proto-oncogene *MYC*, located on chromosome 8q24, has been shown to occur in about 20 to 30% of primary breast tumours. Further, in 80% of tumours this gene is overexpressed without amplification (reviewed by El-Ashry and Lippman, 1994). *MYC* has been shown to positively and negatively regulate the expression of a number of genes in apoptosis and cell cycle progression (Ryan and Bernie, 1996). Amplification of *MYC* has also been associated with a poor patient prognosis (Varley *et al.*, 1987; Tsuda *et al.*, 1989).

The chromosome region 11q13 is amplified in 15 to 20% of breast tumours (Lammie and Peters, 1991). Amplification of this region has been associated with a poor prognosis, lymph node involvement, and oestrogen and progesterone receptor positivity (Devilee and Cornelisse, 1994). Many genes are found to be co-amplified in this region, however the cyclin D1 (*CCND1*) gene appears to be the target of amplification (reviewed in Brenner and Aldaz, 1997). This gene is a direct regulator of the cell cycle and is overexpressed in 45% of breast carcinomas (Bartkova *et al.*, 1994).

The technique of comparative genome hybridisation (CGH) has identified another region commonly amplified in breast carcinomas, the long arm of chromosome 20 (Kallioniemi *et al.*, 1994). CGH is a technique aimed at detecting amplified and/or deleted regions of DNA in tumours (Kallioniemi *et al.*, 1992). Tumour and normal DNA are labelled with different fluorochromes, mixed in equal proportion, and hybridised to normal metaphase chromosomes. The hybridisation pattern of tumour and normal DNA along each chromosome is then analysed by examination of the ratio of the fluorochrome labels. Studies using CGH have found gains

and amplifications of the 20q13 region in 12-18% of primary breast tumours and 40% of cell lines (Kallioniemi *et al.*, 1994; Muleris *et al.*, 1994). In addition, 20q13 amplifications by CGH correlated with poor prognosis in breast cancers with associated lymph node metastasis (Isola *et al.*, 1995), and was the most amplified region in breast tumours that did not involve a previously known oncogene. This region of amplification has since been reduced to a 1.5 Mb segment at 20q13.2 (Tanner *et al.*, 1994). So far a number of genes mapping to this location have been excluded as the putative breast cancer oncogene (Tanner *et al.*, 1994) and further work is needed to identify more candidate genes.

Although the *RB* gene is primarily responsible for retinoblastoma, LOH for this gene has also been found in approximately 36% of breast tumours, with mutations that inactivate the remaining copy being identified (Horowitz *et al.*, 1989). In contrast to *ERBB2*, *MYC*, and *CCND1*, which are oncogenes, *RB* gene is a tumour suppressor gene which acts as a negative regulator of the cell cycle through the interaction with a variety of cellular proteins, most notably the *E2F* family of transcription factors.

The *TP53* gene is the single most commonly mutated tumour suppressor gene in cancer and maps to 17p13.1, a chromosomal site that shows frequent LOH in breast tumours (see Figure 1.2). Furthermore, it has been shown to be inactivated in a large percentage of sporadic breast cancers (Coles *et al.*, 1992). The gene is induced in response to DNA damage and functions to prevent the continued proliferation of genetically impaired cells through the induction of cell cycle arrest or apoptosis (reviewed in Hansen and Oren, 1997). The wild-type p53 protein is a sequence specific transcription factor that is able to activate genes possessing p53 consensus sites. The domain responsible for specific DNA binding was located to the central "core" of the protein which encompasses the majority of residues mutated in human cancer (Pavletich *et al.*,

1993).

1.5.4 Sporadic Breast Cancer and Loss of Heterozygosity

The use of both RFLP and STRP markers has identified numerous regions of allelic imbalance in breast cancer suggesting the presence of tumour suppressor genes in these regions. Figure 1.2 indicates that data compiled from more than 30 studies reveals the loss of DNA from at least 11 chromosome arms at a frequency of more than 25%, with regions such as 16q and 17p affected in more than 50% of tumours (Deville and Cornelisse, 1994; Brenner and Aldaz, 1995). Furthermore, some of these regions are known to harbour tumour suppressor genes shown to be mutated in individuals with both sporadic (*TP53* and *RB* genes) and familial (T*P53*, *RB*, *BRCA1,* and *BRCA2* genes) forms of breast cancer.

Although the consistent loss of certain regions of the genome has been identified, it is difficult to determine the relevance of these losses to the tumourigenic process. In most of the studies, tumours analysed were at a relatively advanced stage of progression, suggesting that some losses may be due to the general genomic instability characteristic of tumour cells. To address this problem, and to determine which tumour suppressor loci are involved in the initiation stages of breast carcinogenesis, LOH studies have been undertaken on pre-invasive ductal carcinomas (DCIS). One study examined 61 samples of DCIS with 56 polymorphic markers covering 39 autosomal arms (Radford *et al.*, 1995). They identified significant LOH for loci on 8p (18.7%), 13q (18%), 16q (28.6%), 17p (37.5%) and 17q (15.9%). These regions also show high rates of LOH in invasive cancer. In another study, allelotype profiles of 23 DCIS samples were compared with that of 29 invasive ductal breast tumours (Aldaz *et al.*, 1995). A total of 20 loci were analysed, and loss of genetic material from chromosome arms 1p, 3p, 3q, 6p, 16p, 18p, 18q, 22q, and possibly 6q and 11p, appeared to be late events in breast cancer

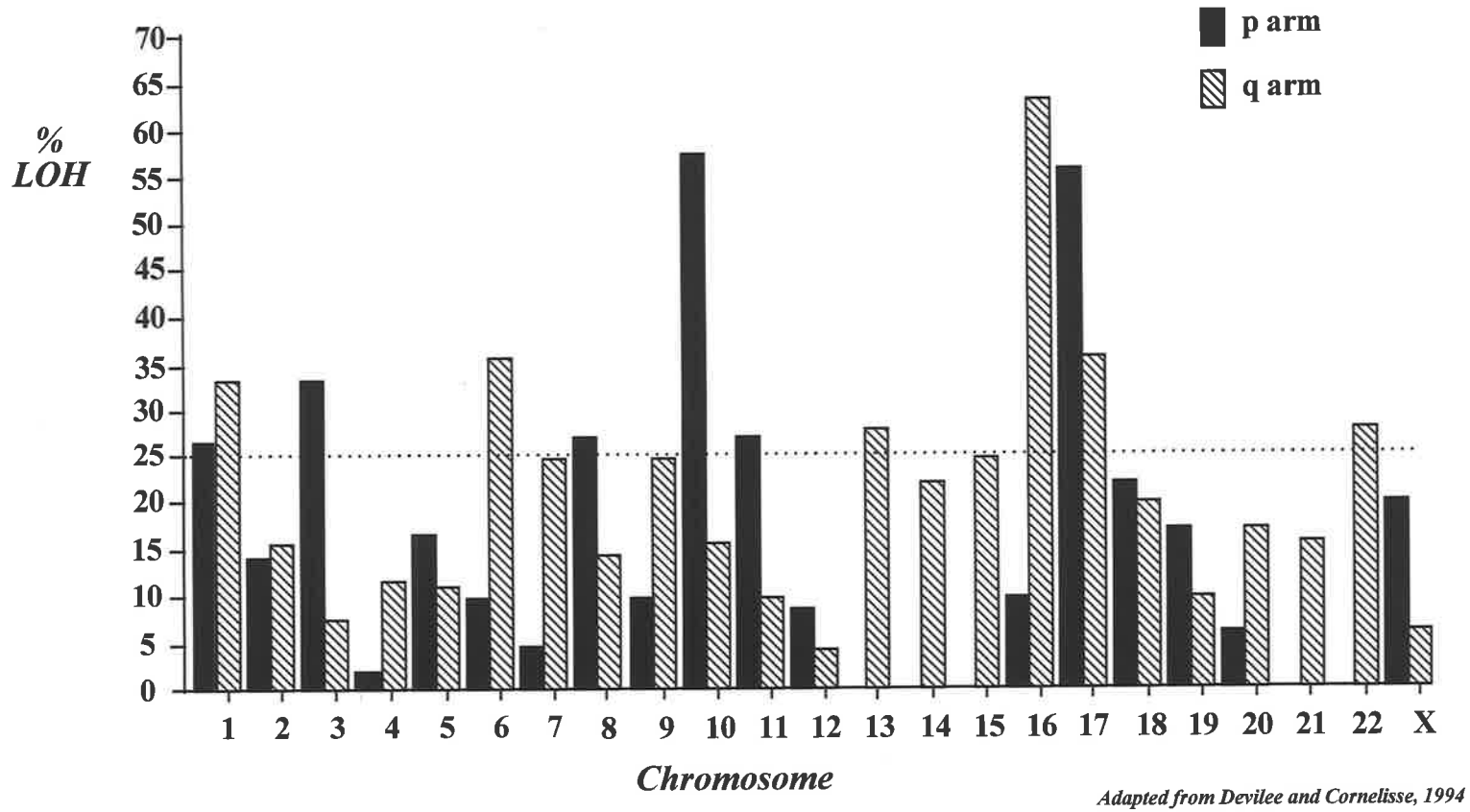


Figure 1.2: Allelotype of breast cancer. All observed allelic loss or gain was reported as loss of heterozygosity (LOH). The broken line represents a 25% cut-off.

progression due to a low percentage of allele loss in the DCIS samples. In contrast, allele losses affecting chromosome arms 7p, 16q, 17p, and 17q, appeared to be early abnormalities as they were observed in 25-30% of DCIS lesions. The common occurrence of allelic losses affecting the chromosome 16q arm in both invasive and non-invasive breast cancer, coupled with the occurrence of cytogenetic structural abnormalities between 16q and chromosome 1 observed in direct chromosome analysis of sporadic breast cancer tumours (Pandis *et al.*, 1992), suggests the presence of a tumour suppressor gene(s) on this chromosomal arm which is critically involved in the early development of a large proportion of breast cancers.

1.5.5 Loss of Heterozygosity and Chromosome 16q

Evidence for LOH on 16q has also been identified with comparable frequencies for other carcinomas. These include prostate cancer (Carter *et al.*, 1990), hepatocellular carcinoma (Zhang *et al.*, 1990; Sakai *et al.*, 1992), central nervous system primitive neuroectodermal tumours (Thomas and Raffel, 1991), Wilms' tumour (Maw *et al.*, 1992), and ovarian cancer (Sato *et al.*, 1991*b*).

To localise the position of the breast cancer tumour suppressor gene(s) on the long arm of chromosome 16, many groups have performed detailed LOH studies using markers that saturate this chromosome arm. Initial studies made use of RFLP markers, which at the time were quite limited in number and represented a low-resolution genetic linkage map. One study examined 219 sporadic primary breast tumours, of which 193 were invasive ductal carcinomas, with 7 RFLP markers on the long arm of chromosome 16 (Sato *et al.*, 1991*a*). Their findings indicated that 51% (78/153) of tumours displayed LOH for at least one marker on this chromosomal arm, with a common region of deletion detected in the chromosomal segment distal to the *HP* gene locus, and proximal to D16S157 (Figure 1.3). It was also noted that in

the group of tumours with a highly malignant phenotype (ie lymph node metastasis, large tumour size), LOH involving markers on other chromosomes also occurred. This would indicate that accumulation of genetic alterations might contribute to tumour development and/or progression in primary breast cancer.

In another study by Tsuda *et al.*, LOH analysis of 234 breast tumours, using 6 RFLP markers from 16q, established that 55% (127/230) of tumours had LOH on this chromosomal arm (Tsuda *et al.*, 1994). This high incidence of LOH was seen not only in large well-advanced tumours, but also in lower grade cancers, suggesting the presence of a tumour suppressor gene playing a role in the early development of breast cancer. A more detailed deletion map of 78 of these tumours (38/78 displayed 16q LOH), using an additional 12 RFLP markers, indicated that while 23 tumours were suggested to have partial allele loss, 15 tumours appeared to have lost the whole 16q arm. From the deletion map established, LOH was most frequent in the 16q24.2-qter region between D16S43 or D16S155, and qter (Figure 1.3). Eight cases comprised a group that showed LOH between 16cen-q22.1 but not at 16q24.2-qter (Figure 1.3). This group consisted of mainly advanced tumours displaying aggressive phenotypes, suggesting that an additional gene to the one proposed at 16q24.2-qter may be present on 16q, and this gene is involved in the formation of biologically aggressive phenotypes of breast cancer.

With the identification of microsatellite repeats and their use in the construction of detailed linkage maps of the human genome, many new markers became available for LOH analysis of breast tumours. In 1994, a study combining both traditional RFLP markers and new STRP markers mapping to 16q, were used in LOH analysis of 79 sporadic cases of breast cancer (Cleton-Jansen *et al.*, 1994). A total of twenty, 16q markers were chosen, with 63% (50/79) of

tumours showing LOH for at least one marker. As with previous studies, a number of tumours (20%) appeared to have lost the entire long arm of chromosome 16 rather than defined regions of the chromosome, further reducing the sample size for narrowing down targets of LOH. However, two common regions of deletion were noted in the remaining tumours. The first region at 16q24.3, was defined by 7 of these tumours which displayed loss for one or more of the last 3 telomeric markers only (restricted loss), with one tumour indicating this loss to be between *APRT* and D16S303 (Figure 1.3). The second region of restricted loss was identified by 7 tumours that showed interstitial loss of 16q markers, with one tumour defining this region to be at 16q22.1, between D16S398 and D16S301. These two regions correlate well with those identified by Tsuda *et al.* (1994), suggesting the presence of two tumour suppressor genes. However, the results of Sato *et al.* (1991*a*) indicate that a third region at 16q22.2-q24.2 may exist.

Another study examined 150 sporadic primary breast tumours with four 16q markers and established that 101 (67%) showed allelic imbalance for at least 1 marker (Skirnisdottir *et al.*, 1995). Results from analysis of a further five 16q markers on these 101 tumours, indicated that 48 (47%) showed LOH for all informative markers, with the remaining 53 tumours displaying partial or interstitial deletions. When the LOH pattern of tumours displaying restricted loss only is analysed, two common regions of allele loss are identified, with both regions overlapping those previously reported (Tsuda *et al.*, 1994; Cleton-Jansen *et al.*, 1994). The first region lies distal to D16S413, located at 16q24.3, and the second region lies between D16S260 and D16S265, encompassing the 16q22.1 region (Figure 1.3).

Dorion-Bonnet *et al.* (1995) conducted an allelic imbalance study of nine 16q markers in 46 primary breast carcinomas, again indicating a high incidence (65%) of LOH for this

chromosomal arm in breast cancer. While most of these tumours appeared to have lost the majority of the long arm, the three regions previously identified displaying restricted LOH overlapped with regions lost in the remaining tumours examined in this present study (Figure 1.3). A later study by this group (Driouch *et al.*, 1997) examined LOH of 16q in 24 breast cancer metastasis. Results again indicated the same three regions previously identified to be involved (Figure 1.3).

A large study of 210 sporadic cases of breast cancer by Iida *et al.* (1997) using 14 microsatellite markers on 16q, again showed LOH for the long arm of chromosome 16 is a common and significant occurrence in breast cancer, with 67% of tumours showing loss for at least 1 marker. Two target regions were identified, one region located at 16q23.2-24.1 between D16S512 and D16S515 was defined by 7 tumours showing restricted LOH, while the second region located at 16q24.3 between D16S413 and D16S303 was defined by 3 tumours (Figure 1.3). The first region overlaps that previously described by Driouch *et al.* (1997) and Sato *et al.* (1991a), with the second region agreeing with that detected by all previous studies mentioned.

Detailed LOH studies of ductal carcinoma *in situ* (DCIS) of the breast have also been done. Chen *et al.* (1996*b*), examined 35 tumour samples with 20 chromosome 16q microsatellite markers, and identified that 31 (89%) had LOH of 16q. This higher incidence of allele loss than previously observed was likely due to the use of very precise tissue microdisection of tumour material from stromal contaminants, and the use of highly informative markers. Most tumours appeared to display random loss (complex loss) of chromosome 16q markers, however results indicated that again, three regions on this chromosomal arm might be involved in breast tumourigenesis (Figure 1.3).

Preliminary LOH analysis of chromosome 16q in prostate cancer has also been done. One study examined 48 cases of primary or metastatic prostate cancers with 17 markers from chromosome 16q and established that 42% showed LOH for at least one marker (Suzuki *et al.*, 1996). Their results indicated that three commonly deleted regions were identified, each overlapping with those seen in breast carcinoma (Figure 1.3). Another study (Latil *et al.*, 1997), was able to again identify 3 regions of restricted loss, however only two of these corresponded to those observed in breast cancer (Figure 1.3). Finally, a study by Godfrey *et al.* (1997) also indicated that the regions at 16q24.3 and 16q22 were commonly deleted in prostate carcinomas.

A consistent finding in all LOH studies in breast cancer is that allelic imbalance on chromosome 16q occurs irrespectively of differences in clinico-pathological parameters. No correlation could be observed with indicators such as lymph node status, differences in clinical stage, tumour size, histology, age at diagnosis, ploidy, or family history of breast cancer. However, a weak but statistically significant correlation between allelic imbalance on 16q and a positive oestrogen receptor content was identified by Cleton-Jansen *et al.* (1994), although this was not observed in other studies. Also Skirnisdottir *et al.* (1995) were able to establish a correlation with a high progesterone receptor content and low S-phase fraction, with patients showing a low S-phase fraction having about a 19% higher survival rate than patients with high S-phase fractions. This again was not seen in other studies. However, in prostate cancer, it has been established that LOH at the 16q24.1 region is significantly associated with a clinically aggressive behaviour of the disease, metastatic disease and a higher grade of tumour (Elo *et al.*, 1997).

Figure 1.3 provides a summary of all LOH studies involving markers on the long arm of

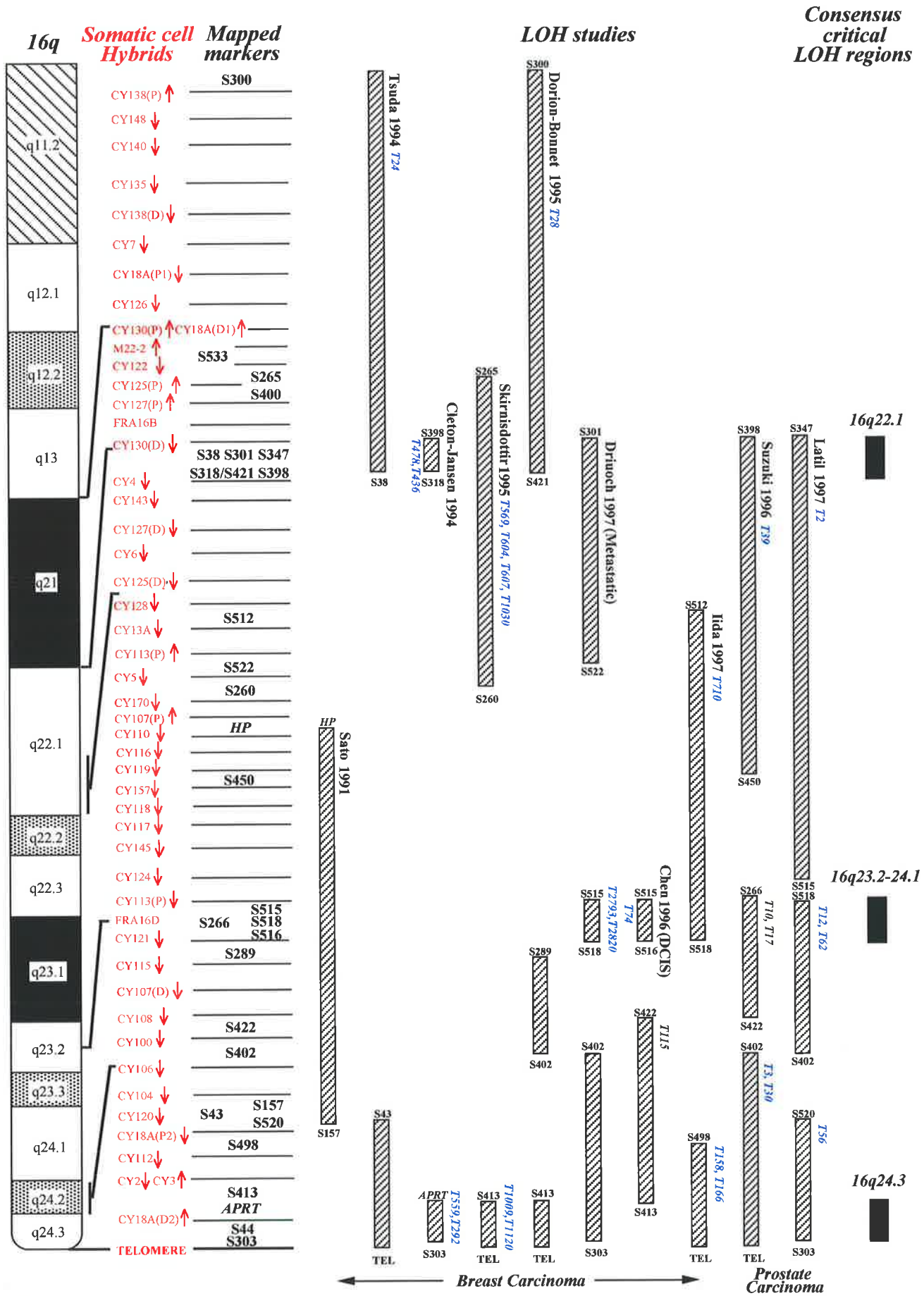


Figure 1.3: Summary of loss of heterozygosity (LOH) studies involving polymorphic markers mapping to the long arm of chromosome 16 in sporadic breast and prostate cancer. Vertical hatched bars indicate regions of DNA loss based on the analysis of the tumour samples used in each study. Those markers defining the centromeric and telomeric boundaries for loss are indicated above and below each hatched bar respectively. Those tumours defining the critical region of loss in each study are indicated in blue, otherwise conclusions of regions of loss were drawn from the analysis of tumours displaying complex deletions. The consensus critical LOH regions for this chromosomal arm based on the overlaps of the individual studies are shown in the filled boxes on the right of the figure. Three overlapping regions are defined: 16q22.1, 16q23.2-24.1, and 16q24.3.

chromosome 16 in both breast and prostate cancer. The markers used in the LOH analysis have been physically mapped to the chromosome using a somatic cell hybrid panel (Callen *et al.*, 1995), which allows integration of genetic and physical distances between markers. A striking consistency is observed in the majority of these studies. The appearance of three regions of consistent restricted LOH (16q22.1, 16q23.2-24.1, and 16q24.3) provides evidence that possibly three independent tumour suppressor genes may reside adjacent to the markers used for the LOH analysis in these regions playing a role in the genesis of these two carcinomas.

When interpreting results from LOH studies, care is needed. A number of problems can confuse the comparison of the genotypes in normal and tumour DNA from an affected individual. In the majority of tumour biopsies, a significant amount on non-malignant cells can also be found which may obscure any LOH that may be present in the tumour cells. It has been shown that tumour samples with >50% malignant cells give reliable results in an LOH screening (Devilee and Cornelisse, 1994; Dorion-Bonnet *et al.*, 1995). In the majority of LOH studies discussed above, the level of normal cell contamination for each tumour was determined histologically, such that only those samples containing at least 50% of malignant cells were used for the subsequent LOH analysis. Another approach to avoid normal cell contamination has been the use of very precise tissue microdissection, which improved the separation of tumour material from normal stromal cells (Chen *et al.*, 1996*b*). In this study, 89% of DCIS tumours were shown to have allele losses in one or more informative loci examined from the long arm of chromosome 16. This is the highest incidence of LOH that has been reported and may reflect the purity of the samples analysed. Another problem inherent to LOH analysis is that in most studies, the tumours analysed were of the invasive type and/or advanced stages of progression. It is therefore possible that not all allele losses are causative factors of tumourigenesis but may be the result of the general genomic instability characteristic

of tumours in advanced stages. This problem has been overcome by Aldaz *et al.* (1995) from the comparative allelotyping of both pre-invasive ductal carcinomas (DCIS) and invasive tumours. A third problem often arises due to the lack of informativeness of many of the polymorphic markers used in the LOH studies, particularly the RFLPs. This often leads to regions of the genome for which data are not available and prevents definition of a region of LOH. Despite these difficulties, LOH has been shown to be the mechanism responsible for the inactivation of known tumour suppressor genes (Smith *et al.*, 1992; Collins *et al.*, 1995*c*; Berx *et al.*, 1995, 1996). Since consistent LOH of defined regions on the long arm of chromosome 16 have been identified in breast and other carcinomas, it is concluded that tumour suppressor genes residing at these sites may also be the targets for LOH in these carcinomas.

1.5.6 Candidate Genes for 16q Loss of Heterozygosity

The epithelial cadherin gene (*CDH1*) maps within the region of LOH at 16q22.1 (Mansouri *et al.*, 1988). This gene is important for the maintenance of cell-cell adhesion in epithelial tissues. Evidence suggests that a loss of function in *CDH1* and/or one or more collaborating proteins contributes to increased proliferation, invasion, and metastasis in breast cancer and other solid tumours (Ilyas and Tomlinson, 1997). In addition, mutations in *CDH1* have been detected in breast cancer cell lines (Pierceall *et al.*, 1995; Hiraguri *et al.*, 1998), and in 27 of 48 infiltrating lobular breast tumours studied (Berx *et al.*, 1995, 1996), with the majority of mutations occurring in combination with LOH of the wild-type locus. However, no mutations in this gene have been identified in the most common sub-type of breast cancer, ductal carcinoma (Kashiwaba *et al.*, 1995), which have been the majority of samples analysed in reported LOH studies. Therefore other genes present in this region of q22.1 may be candidate tumour suppressor genes.

The evolutionary conserved DNA binding protein *CTCF*, which also maps to 16q22.1 (Filippova *et al.*, 1998), may be a tumour suppressor gene due to its role as one of the major factors regulating vertebrate *MYC* expression (Klenova *et al.*, 1993; Filippova *et al.*, 1996). Given dysregulated *MYC* expression is a common occurrence in the development of breast carcinomas (1.5.3), mutations affecting any regulator of *MYC* expression, including *CTCF*, may play a role in breast tumourigenesis. Recently, two breast cancer cell lines have been shown to harbour genomic *CTCF* rearrangements, and one of four fresh breast tumour samples of unknown histological type also exhibited a tumour specific genomic rearrangement of *CTCF* exons (Filippova *et al.*, 1998). Additional evidence that this gene plays a major role in breast tumourigenesis needs to be obtained.

A candidate gene mapping to the 16q23.2-24.1 LOH region is the breast cancer anti-oestrogen resistance (*BCAR1*) gene (Driouch *et al.*, 1997). It has been shown that defective retroviruses may be responsible for the loss of oestrogen receptor expression in human cells *in vitro* via insertional mutagenesis into the *BCAR1* gene (Dorssers *et al.*, 1993). Given that oestrogen receptor negativity has been associated with a poor prognosis in breast cancer, this gene may well be a target for LOH at this chromosomal site. However, this has yet to be confirmed.

At the onset of this project, the LOH analysis of additional markers mapping to 16q24.3 and additional tumour samples to those reported by Cleton-Jansen *et al.* (1994) had been achieved. Figure 1.4 summarises the results of these studies indicating those tumours that appear to display restricted loss of markers mapping to this region. The results suggest that the smallest region of LOH is between D16S3026 and D16S303, occurring in tumours BT559 and BT410 (Cleton-Jansen, personal communication). Genes located in this interval at the start of the project included *BBC1*, *CMAR*, *DPEP1*, and *MC1R*.

TUMOUR	559	410	819	413	465	378	912	757	477	545	358	367	549	919	589	309	204	645	666	753	408
16q22.1	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
D16S289	R		N	R	R	R	N	N	N	N	N	R	R	R	R	L	L		N	N	N
D16S449	N		R	R	N		R		R		R	R	R	R	L	L	L		L		N
D16S305	R	R	R	R	R	R	R	N	N	R	N	L	L	L	N	L	L	L	N	N	L
D16S1320	N	R	R	R	R	R	R	R	R	L	L		L	N	L	L	L	L	L		L
D16S413	N	R	R	L	L	L	N	L	L	L	N	N	N	L		L	L	N	L	L	
D16S2621	R		N	N	N	L	L	L	L		L	N	N	L	L	L	L	L	N	L	L
D16S7	R	R		L	L	L				N	L	L	L		L	L	N		N		L
D16S3023	R	R	N	L	N	L	L	N	L	N	L	L	N					N		L	
APRT	R	N	N	N	N	N			N			N	N	N		N	L		N		N
D16S3121	N	R			L	N	N	N	N	L	N	L	N		N			L		L	
D16S3026	R	R		N	N	N	N	N	N		L	L	N							L	
CMAR	N	L		N	N	N	L	L	L		N	N	N	N	N	N	N	N	L	L	N
DPEP1	L				N										L	N	L				
D16S303	L	N	L		N	R	N	L	L	L	R	N	L	L	N	N	L		L	N	N

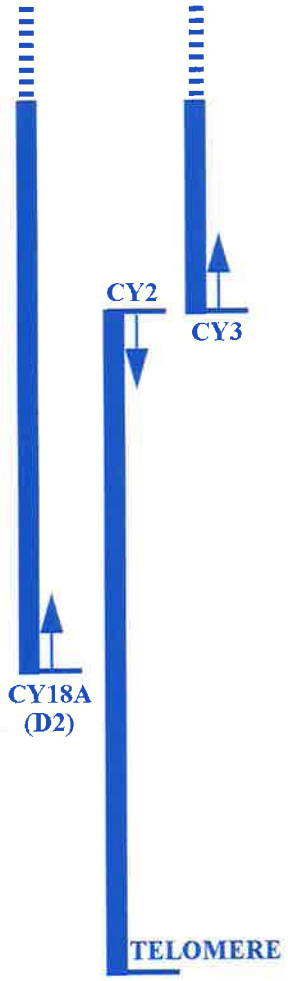


Figure 1.4: Chart showing the results of further loss of heterozygosity (LOH) analysis on DNA from breast tumours with restricted loss at 16q24.3 as described in Cleton-Jansen *et al.* (1994). Markers used are listed in the left hand column, with their physical map location, as determined by somatic cell hybrid mapping, indicated by blue columns on the right hand side. Individual tumour sample numbers are indicated across the top. **N**: Not Informative; **R**: Retention; **L**: Loss; Blank: Not Done. In each case, markers not done or not informative were scored as lost or retained based on the result of the nearest centromeric marker. This assisted in defining the smallest region of overlap and hence LOH.

The breast basic conserved (*BBC1*) gene displays decreased mRNA expression in malignant breast tumours compared to benign tumours (Adams *et al.*, 1992) and shows homology to the human ribosomal protein L13. However, the L13 protein has been shown to contain a DNA binding motif characteristic of transcription factors (Tsurugi and Mitsui, 1991), suggesting *BBC1* is not a candidate tumour suppressor gene. Mutation analysis of this gene was recently performed in samples of breast tumour DNA displaying restricted LOH for 16q24.3, but no mutations were identified in this gene in these samples (Moerland *et al.*, 1997).

The cellular matrix adhesion regulator (*CMAR*), was originally isolated from a colorectal cell line library, and shown to enhance binding to the extracellular matrix component, collagen type 1 (Pullman and Bodmer, 1992). Loss of *CMAR* activity, and hence cell attachment, could therefore possibly be an early step towards tumour invasion and metastasis. However, the recent cloning of the spastic paraplegia gene, *SPG7* (Casari *et al.*, 1998), has shown that *CMAR* is the 3' untranslated part of this gene and is therefore unlikely to play a role in breast carcinogenesis.

The renal dipeptidase (*DPEP1*) gene was originally isolated from a kidney cDNA library by subtractive hybridisation with Wilms' tumour mRNA (Austruy *et al.*, 1993). Although suggested to be a potential tumour suppressor gene, its role in tumourigenesis is yet to be determined. Finally, the melanocyte stimulating hormone receptor (*MC1R*) gene has been shown to be involved in the regulation of pigmentation phenotype (Valverde *et al.*, 1995) and is therefore not a likely candidate gene.

1.6 Project Aims

The refined 16q LOH studies in breast cancers performed by a collaborating group in Leiden, has defined a critical region at 16q24.3. This interval is bounded proximally by D16S3026 and distally by D16S303. These loci are both located within the region defined by the somatic cell hybrid breakpoint CY18A(D2) and the telomere of the long arm (Callen *et al.*, 1995). The close physical proximity of these two markers suggests that a positional cloning strategy to the identification of the proposed tumour suppressor gene is a realistic approach.

Due to the lack of cloned DNA located in this interval, the first aim of the project is to construct a detailed physical map between D16S303 and D16S3121. This map will be based on overlapping cosmids identified from the screening of a human flow-sorted chromosome 16 specific library with markers located in the critical region, which will assist in the subsequent identification of novel coding sequences. Selected cosmids spanning this region will be used as templates for exon trapping experiments to isolate transcribed sequences. Additional transcript data will be obtained from the analysis of ESTs mapping to 16q24.3 as part of the Human Gene Map construction at NCBI, and dbEST database screening of partial cosmid sequence. Trapped exons and cDNA clones will be assembled into transcription units that will be characterised further. This will involve the isolation of the full-length sequence of the corresponding gene and establishing an expression pattern for the gene. Sequence homology searches of characterised genes may then provide a possible function for the associated protein. Finally, candidates will be screened by mutation analysis of DNA from tumours defining the critical region of LOH at 16q24.3 (see Figure 1.4) to determine if any of these genes is a tumour suppressor altered in breast carcinomas.

Finally, the use of exon trapping to construct a transcription map of a segment of chromosome 17p13 will be explored. This region has been shown to contain a gene responsible for cystinosis by genetic linkage analysis, and a collaboration has been established to positionally clone the gene responsible. The collaborating laboratory has established a physical map within the critical region based on cosmid clones. This thesis describes the use of exon trapping experiments with these cosmids to identify the cystinosis gene.

Since portions of the research described in this thesis are collaborative projects, the work contributed by other individuals will be acknowledged where appropriate in the text.

Chapter 2

Materials

and

Methods

Table of Contents

Page

2.1 Materials **58**

2.1.1 Fine Chemicals and Compounds and Their Suppliers 58

2.1.2 Enzymes 60

2.1.3 Electrophoresis 61

2.1.4 Antibiotics 61

2.1.5 Bacterial Strains 61

2.1.6 Media 62

2.1.6.1 Liquid Media 62

2.1.6.2 Solid Media 63

2.1.7 DNA Vectors 63

2.1.8 DNA Size Markers 63

2.1.9 Radio-Chemicals 63

2.1.10 Miscellaneous Materials and Kits 64

2.1.11 Buffers and Solutions 64

2.2 Methods **66**

2.2.1 DNA Isolation 67

2.2.1.1 Cosmid DNA 67

2.2.1.2 Isolation of PAC and BAC DNA 68

2.2.1.3 Plasmid DNA 68

2.2.1.4 Genomic DNA 69

2.2.2 Preparation of Bacterial Glycerol Stocks 70

2.2.3 Preparation of Sheared Human Placental DNA and Salmon Sperm DNA 70

2.2.4 Isolation of Total RNA 70

2.2.4.1 RNA Isolation from Fresh Tissue 70

2.2.4.2 RNA Isolation from Cell Lines 72

2.2.5 Oligonucleotide Primer Design 72

2.2.6 Quantitation of DNA, RNA, and Oligonucleotide Primers 72

2.2.7 Restriction Digests 73

2.2.7.1 Digestion of Genomic DNA 73

2.2.7.2 Digestion of Cosmid, BAC, and PAC DNA 73

2.2.7.3 Digestion of Plasmid DNA 73

2.2.8 Agarose Gel Electrophoresis 74

2.2.9 DNA Size Markers 74

2.2.10 Southern Blotting 74
2.2.10.1 Transfer in 10X SSC 74
2.2.10.2 Transfer in 0.4 M NaOH 75
2.2.11 Northern Blotting 75
2.2.11.1 RNA Electrophoresis 75
2.2.11.2 Transfer of RNA to Membranes 76
2.2.12 ^{32}P Radio-Isotype Labelling of DNA 76
2.2.12.1 Labelling DNA Fragments in Solution 76
2.2.12.2 Labelling DNA Fragments in Agarose 77
2.2.12.3 Pre-Reassociation of Labelled Probes 77
2.2.13 Hybridisation of Membranes 78
2.2.13.1 Southern Hybridisations 78
2.2.13.2 Northern Hybridisations 78
2.2.13.3 Washing of Membranes 79
2.2.13.4 Stripping of Membranes for Re-Use 79
2.2.14 Polymerase Chain Reaction (PCR) 80
2.2.14.1 Standard PCR Reactions 80
2.2.14.2 Colony PCR Reactions 80
2.2.15 Reverse Transcriptase PCR (RT-PCR) 81
2.2.15.1 Reverse Transcription 81
2.2.15.2 PCR of Reverse Transcribed RNA 82
2.2.16 Purification of DNA Fragments 82
2.2.16.1 QIAquick Purification of PCR Fragments and Radio-Labelled Probes 82
2.2.16.2 Prep-A-Gene Purification of DNA Fragments in Agarose 83
2.2.17 Sub-Cloning of DNA Fragments 84
2.2.17.1 DNA Ligations 84
2.2.17.2 Ligating PCR Products into the pGEM-T Vector 84
2.2.17.3 Preparation of Competent Bacterial Cells 84
2.2.17.4 Transformation of Competent Bacterial Cells 85
2.2.17.5 Preparation of Colony Master Plates and Colony Lifting 85
2.2.18 DNA Sequencing 86
2.2.18.1 DyeTerminator Cycle Sequencing 86
2.2.18.2 DyePrimer Cycle Sequencing 87
2.2.19 Fluorescence *in situ* Hybridisation (FISH) 88
2.2.20 5' Rapid Amplification of cDNA Ends (5' RACE) 88
2.2.21 PCR Formatting of the Fetal Brain cDNA Library 89
2.2.21.1 Lambda Phage Plating 89
2.2.21.2 Lambda Phage Scraping 90
2.2.21.3 Pooling of Phage Scrapes 90

2.1 Materials

2.1.1 Fine Chemicals and Compounds and Their Suppliers

Ammonium sulphate	Ajax Chemicals, Australia
Bactoagar	Sigma Chemical Co., USA
Bactotryptone	Difco Laboratories, USA
Boric acid	Ajax
BSA (bovine serum albumin-pentax fraction V)	Sigma
Chloroform	BDH Lab Supplies
DAPI (diamidino phenylindole dihydrochloride)	Sigma
DEPC (diethylpyrocarbonate)	BDH
Deoxy-nucleotide triphosphates (dNTPs)	Pharmacia Biotech
Dextran sulphate	Pharmacia or Promega
N,N-dimethyl formamide	Sigma
DMEM (Dulbecco's modified Eagles medium)	Trace Biosciences
DMSO (dimethylsulphoxide)	Sigma
DTT (dithiothreitol)	Sigma
EDTA (ethylenediaminetetracetic acid; $Na_2EDTA.2H_2O$)	Ajax
Ethanol (99.5% v/v)	BDH
FCS (fetal calf serum)	Trace Biosciences
Ficoll (Type 400)	Sigma
Formamide (deionised before use)	Fluka Chemika
Glucose	Ajax
Glutamine	Trace Biosciences
Glycerol	Ajax

Human placental DNA	Sigma
IPTG (isopropylthio-β-D-galactosidase)	Progen
Isoamyl alcohol (IAA)	Ajax
Isopropanol	Ajax
N-lauroylsarcosine (Sarkosyl)	Sigma
LipofectACE	Gibco-BRL
Magnesium chloride ($MgCl_2.6H_2O$)	Ajax
Magnesium sulphate ($MgSO_4.7H_2O$)	Ajax
Maltose	BDH
βMe (β-mercaptoethanol)	BDH
Mixed bed resin (20-50 mesh)	BioRad, USA
MOPS (3-[N-morpholino]propane sulphonic acid)	Sigma
Opti MEM-1 (powder sachet)	Gibco-BRL
Paraffin oil	Ajax
PEG (polyethylene glycol) 3350	Sigma
Phenol	Wako Ind., Japan
Polyvinylpyrolidone (PVP-40)	Sigma
Potassium dihydrogen orthophosphate (KH_2PO_4)	Ajax
Propidium iodide	Sigma
Salmon sperm DNA	Calbiochem
Sodium acetate	Ajax
Sodium bicarbonate ($NaHCO_3$)	Astra Pharm., Australia
Sodium chloride	Ajax
tri-Sodium citrate	Ajax
Sodium dihydrogen orthophosphate ($NaH_2PO_4.2H_2O$)	Ajax

SDS (sodium dodecyl sulphate)	BDH
di-Sodium hydrogen orthophosphate ($Na_2HPO_4.7H_2O$)	Ajax
Sodium hydroxide	Ajax
Sucrose	Ajax
Tris-base	Boehringer Mannheim
Tris-HCl	Boehringer Mannheim
Triton X-100	Ajax
Yeast extract	Gibco-BRL

2.1.2 Enzymes

All restriction enzymes were purchased from either New England Biolabs (Beverly, Massachusetts, USA) or Progen (Brisbane, Queensland, Australia). Each enzyme was supplied with the appropriate digestion buffer and bovine serum albumin (BSA) if required. All other enzymes not part of kits are listed below along with their suppliers.

CIAP (Calf intestinal alkaline phosphatase)	Boehringer Mannheim
E.coli DNA polymerase I (Klenow fragment)	New England Biolabs
Proteinase K	Merck
RNaseH	Promega
RNAsin	Promega
Superscript^{TM} RNase H$^-$ reverse transcriptase	Gibco-BRL
T4 DNA ligase	Progen
T4 Polynucleotide kinase	Pharmacia
Taq DNA polymerase	Gibco-BRL
X-gal (5-Bromo-4-chloro-3-indoyl-β-D galactosidase)	Progen

2.1.3 Electrophoresis

Acrylamide (40%)	BioRad
Agarose: NA grade	Progen or Pharmacia
APS (Ammonium persulphate)	BioRad
Bis (2% Bis solution)	BioRad
Bromophenol blue	BDH
Ethidium bromide	Sigma
Formaldehyde (37%)	Ajax
MDE (2X gel solution)	FMC bioproducts, USA
TEMED (*N,N,N',N'*-tetramethylethylenediamine)	BioRad
Xylene cyanol	BDH

2.1.4 Antibiotics

Ampicillin	Sigma
Benzyl penicillin	CSL
Chloramphenicol	Sigma
Kanamycin	Progen
Tetracycline	Sigma

2.1.5 Bacterial Strains

All strains were streaked for single colonies on L-Agar plates containing an appropriate antibiotic (see below) from stocks kept at –70°C in L-Broth + 15% glycerol.

E.coli XL1-Blue MRF': Δ(*mcrA*)183 Δ(*mcrCB-hsdSMR-mrr*)173 *endA1 supE44 thi-1 relA1 lac* [F' *proAB lacI*^{q}*ZΔM15* Tn*10* (Tet^{r})]^{c}. This strain was a host for recombinant plasmids and

was purchased from Stratagene. XL1-Blue cells were streaked for single colonies on L-Tetracycline plates then propagated overnight at 37°C in L-Broth supplemented with tetracycline (15 μg/ml).

E.coli Y1090r-: Δ(*lac*)*U169* Δ(*lon*)*? AraD139 strA supF mcrA trpC22::*Tn*10* (Tet^r) [pMC9 Amp^r Tet^r] *mcrB hsdR.*(Note: pMC9 is pBR322 with $lacI^q$ inserted). This strain was a host for recombinant λgt11 bacteriophage and was supplied with the Clontech 5' Stretch fetal brain cDNA library. These cells were streaked for single colonies on L-Tetracycline plates and subsequently propagated overnight at 37°C in L-Broth supplemented with 10 mM $MgSO_4$ and 0.2% (v/v) maltose.

2.1.6 Media

All liquid media was prepared with double distilled water and was sterilised by autoclaving. In the cases where liquid media was prepared for the pouring of solid media plates, antibiotics were added once the autoclaved media had cooled to a temperature of approximately 50°C.

2.1.6.1 Liquid media

L-Broth (Luria-Bertaini Broth): 1% (w/v) Bactotryptone; 0.5% (w/v) yeast extract; 1% (w/v) NaCl; pH to 7.5 with NaOH.

OMI (pH7.1): 1 sachet of Opti MEM-1, 28 ml $NaHCO_3$, 80 ml FCS, 55 nM βMe, 1 ml benzyl penicillin, *per* litre.

Supplemented DMEM (pH7.0): 13 g DMEM, 44 ml $NaHCO_3$, 4 mM glutamine, 1 ml benzyl penicillin, 100 ml FCS, *per* litre.

TSS: L-Broth; 10% (w/v) PEG 3350; 5% (v/v) DMSO; 50 mM $MgCl_2$.

2.1.6.2 Solid Media

L-Agar: L-Broth; 1.5% (w/v) Bactoagar

L-Ampicillin: L-Broth; 1.5% (w/v) Bactoagar; 100 μg/ml ampicillin.

L-Chloramphenicol: L-Broth; 1.5% (w/v) Bactoagar; 34 μg/ml chloramphenicol.

L-Kanamycin: L-Broth; 1.5% (w/v) Bactoagar; 50 μg/ml kanamycin.

L-Tetracycline: L-Broth; 1.5% (w/v) Bactoagar; 15 μg/ml tetracycline.

Top Agar: L-Broth; 0.7% Bactoagar.

2.1.7 DNA Vectors

pSPL3B-CAM was used for exon trapping from cosmid DNA. This vector was kindly provided by Dr. Tim Connors and Dr. Greg Landes (USA).

Puc19 was used for the subcloning of cosmid DNA fragments and was purchased from New England Biolabs.

2.1.8 DNA Size Markers

SPP1 (*Eco*RI restricted)	Progen or Bresatec
Puc19 (*Hpa*II restricted)	Bresatec
Drigest III (λcI857 Sam 7 *Hin*dIII/ ϕX-174 *Hae*III restricted)	Pharmacia

2.1.9 Radio-Chemicals

[α-^{32}P]dCTP (3,000 Ci/mmol)	Amersham
[γ-^{32}P]dATP (5,000 Ci/mmol)	Amersham

2.1.10 Miscellaneous Materials and Kits

ABI Prism™ DyePrimer Cycle sequencing kit	Perkin Elmer
ABI Prism™ DyeTerminator Cycle sequencing kit	Perkin Elmer
BigDye Terminator Cycle sequencing kit	Perkin Elmer
Breast cancer cell lines	ATCC, USA
CloneAMP pAMP10 kit	Gibco-BRL
ExpressHYB solution	Clontech
Genescreen Plus Nylon membranes	Dupont
halfTERM DyeTerminator sequencing reagent	GenPack
Labelling mix-dCTP	Pharmacia
Multiple tissue Northern blot (Catalogue # 7760-1)	Clontech
Oligo-dT_{12-18}	Gibco-BRL
pGEM-T cloning kit	Promega
Prep-A-Gene purification kit	BioRad
Qiagen Plasmid mini kit	Qiagen
QIAquick purification kit	Qiagen
5' RACE kit	Gibco-BRL
Random hexamers	Pharmacia or Perkin Elmer
R*Tth* reverse transcriptase RNA PCR kit	Perkin Elmer
Trizol	Gibco-BRL
3MM Whatman paper	Lab Supply

2.1.11 Buffers and Solutions

All solutions were prepared with sterile double distilled water followed by autoclaving at 120°C for 15 to 30 minutes. Solutions provided with kits are not mentioned. Buffers and

solutions routinely used in this study were as follows:

10X Agarose loading buffer: 100 mM Tris-HCl (pH8.0); 200 mM EDTA (pH8.0); 2.0% (w/v) sarkosyl; 15% (w/v) ficoll 400; 0.1% bromophenol blue; 0.1% xylene cyanol.

1X pAMP10 annealing buffer: 20 mM Tris-HCl (pH8.4); 50 mM KCl; 1.5 mM $MgCl_2$.

Cell lysis buffer: 0.32 M sucrose; 10 mM Tris-HCl; 5 mM $MgCl_2$; 1% (v/v) Triton X-100 (pH7.5).

Colony denaturing solution: 1.5 mM NaCl; 0.5 M NaOH.

Colony neutralising solution: 3 M NaCl; 0.5 M Tris-HCl.

De-ionised formamide: 2 g of mixed bed resin/50 ml formamide. Mix for ~1 hour then filter.

100X Denhart's solution: 2% (w/v) ficoll 400; 2% (w/v) polyvinylpirolidone; 2% (w/v) BSA.

10X dNTP labelling solution: 1 mM labelling mix-dCTP.

Filter denaturing solution: 0.5 M NaOH.

Filter neutralising solution: 0.2 M Tris-HCl; 2X SSC.

5X First strand buffer (supplied with Superscript reverse transcriptase): 250 mM Tris-HCl (pH8.3); 375 mM KCl; 15 mM $MgCl_2$.

Formamide loading buffer: 50% (v/v) glycerol; 1 mM EDTA (pH8.0); 0.25% (w/v) bromophenol blue; 0.25 % (w/v) xylene cyanol.

Gel denaturing solution: 2.5 M NaCl; 0.5 M NaOH.

Gel neutralising solution: 1.5 M NaCl; 0.5 M Tris-HCl (pH7.5).

10X Labelling buffer: 0.5 M Tris-HCl (pH7.5); 0.1 M $MgCl_2$; 10 mM DTT; 0.5 mg/ml BSA; 0.05 A_{260}/µl random hexamers (Pharmacia).

10X MOPS buffer: 0.2 M MOPS; 50 mM NaAc; 1 mM EDTA.

Phenol: buffered with Tris-HCl (pH7.4).

PBS (Phosphate buffered saline): 138 mM NaCl; 2.7 mM KCl; 8.1 mM $Na_2HPO_4.7H_2O$; 1.2 mM KH_2PO_4 (pH7.4).

10X PCR buffer (supplied with *Taq* DNA polymerase): 200 mM Tris-HCl (pH8.0); 500 mM KCl.

2X PCR mix: 33 mM $(NH_4)_2SO_4$, 133 mM Tris-HCl (pH8.8); 20 mM βME (added fresh each

time); 0.013 mM EDTA; 0.34 mg/ml BSA; 20% (v/v) DMSO; 0.4 mM dNTPs.

3X Proteinase K buffer: 10 mM NaCl; 10 mM Tris-HCl; 10 mM EDTA (pH8.0).

RNA hydrolising solution: 50 mM HaOH; 1.5 M NaCl.

10X RNA loading dye: 50% glycerol; 1 mM EDTA (pH8.0); 0.25% bromophenol blue; 0.25% xylene cyanol.

RNA neutralising solution: 0.5 M Tris (pH7.4); 1.5 mM NaCl.

SM buffer: 50 mM Tris-HCl; 100 mM NaCl; 8 mM $MgSO_4$; 0.01% (v/v) gelatin (pH7.5).

3 M Sodium acetate pH5.2 (NaAc): 24.6 g sodium acetate/100ml; pH 5.2 with acetic acid.

Southern hybridisation solution: 50% (v/v) de-ionised formamide; 5X SSPE; 2% SDS; 1X Denhart's; 10% (w/v) dextran sulphate; 100 µg/ml denatured salmon sperm DNA.

20X SSC: 3 M NaCl; 0.3 M tri-sodium citrate (pH7.0).

20X SSPE: 3.6 M NaCl; 0.2 M $NaH_2PO_4.2H_2O$; 0.02 M EDTA.

1X TBE: 90 mM Tris-base; 90 mM boric acid; 2.5 mM EDTA (pH8.0).

TE: 10 mM Tris-HCl (pH7.5); 0.1 mM EDTA.

TKM: 10 mM Tris-base; 10 mM KCl; 1 mM $MgCl_2$.

X-gal: 50 mM (2% w/v) solution, using N,N,dimethyl-formamide as diluent.

2.2 Methods

Most of the methods described below are those routinely used in the Department of Cytogenetics and Molecular Genetics at the Women's and Children's Hospital (WCH), Adelaide. Most procedures are based on those presented in *Molecular Cloning: A Laboratory Manual* (Sambrook *et al.*, 1989) unless otherwise specified. Only methods used in common throughout the project are presented here. Those procedures specific for individual chapters are mentioned in detail in their corresponding methods section.

2.2.1 DNA Isolation

All DNA preparations except for genomic DNA isolation were based on the alkaline lysis procedure adapted from Sambrook *et al.* (1989). DNA was subsequently purified using Qiagen columns with slight modifications, in some cases (specified below), to manufacturers conditions (Qiagen Plasmid Mini Handbook, 1995). All buffers and columns were supplied with each Qiagen kit.

2.2.1.1 Cosmid DNA

All cosmids were received as bacterial stabs and were streaked for single colonies on L-Kanamycin plates. A single colony was grown overnight at 37°C in 200 ml of L-Broth and cultures were spun down for 15 minutes at 4,000 rpm the next day. DNA was isolated using a Qiagen Tip-20 column following minor changes of the manufacturers methods. This involved resuspending the bacterial pellet in a total of 2 ml of buffer P1, followed by the addition of 2 ml of buffer P2. After careful mixing, the resuspended pellet was left at room temperature for 5 minutes. Two ml of P3 was then added, followed by further gentle mixing and a 10 minute incubation on ice. This mixture was spun at 13,000 rpm, the supernatant transferred to a fresh tube, and spun for a further 10 minutes. During the second spin a Qiagen Tip-20 column was equilibrated with 1 ml of QBT buffer. The supernatant was applied to the equilibrated column and allowed to drain completely through. The column was subsequently washed with 4 ml of buffer QC, the DNA eluted into a 1.5 ml eppendorf tube with 0.8 ml of buffer QF, and precipitated with 0.56 ml isopropanol. After mixing, the solution was spun at 13,000 rpm for 30 minutes. The resultant DNA pellet was washed with 1 ml of 70% ethanol, spun for a further 5 minutes, and allowed to air-dry for 20 minutes. The DNA was resuspended in 15 µl of sterile water and quantitated by spectrophotometry (2.2.6).

2.2.1.2 Isolation of PAC and BAC DNA

PAC and BAC clones received as bacterial stabs were streaked for single colonies on L-Kanamycin or L-Chloramphenicol plates respectively. PAC clones were initially grown overnight at 37°C in 10 ml of L-Broth plus kanamycin (50 μg/ml). The next day, 200 ml of L-Broth plus kanamycin (50 μg/ml) was seeded with 6.7 ml of this culture and grown for 1.5 hours at 37°C. After addition of IPTG to a final concentration of 0.5 mM, the cells were grown a further 5 hours and then spun down at 4,000 rpm. Supernatants were removed and cell pellets were left overnight at -20°C. BAC clones were grown overnight at 37°C in 200 ml of L-Broth plus chloramphenicol (34 μg/ml). Cells were pelleted the next day at 4,000 rpm and the supernatant discarded. Both PAC and BAC DNA was isolated using Qiagen Tip-100 columns using the same procedure as cosmid DNA isolation with adjustments to volumes being made. These included the use of 4 ml of P1, P2 and P3 buffers per DNA isolation; the Tip-100 columns were equilibrated with 4 ml QBT buffer, and washed with 10 ml buffer QC; DNA was eluted into a 10 ml sterile tube with 5 ml of QF buffer, then aliquoted into 6 eppendorf tubes. DNA was precipitated with 0.7 volumes of isopropanol, and pellets were washed, as per cosmid DNA preparation. Individual dried DNA pellets were resuspended in 5 μl of sterile water, then pooled into one tube and quantified by spectrophotometry (2.2.6). Care was taken to resuspend the DNA pellets with wide bore pipette tips to prevent shearing of the DNA.

2.2.1.3 Plasmid DNA

Single colonies were grown overnight at 37°C in 20 ml L-Broth plus ampicillin (100 μg/ml). Plasmid DNA isolations were performed using Qiagen Tip-20 columns with no modifications being made to manufacturers conditions (Qiagen Plasmid Mini Handbook, 1995). DNA pellets

were resuspended in 15 µl of sterile water and quantitated by spectrophotometry (2.2.6).

2.2.1.4 Genomic DNA

DNA was isolated from peripheral blood lymphocytes, mouse/human somatic cell hybrids, and breast cancer cell lines by procedures adapted from Wyman and White (1980). All DNA preparations were kindly performed by either Shirley Richardson or Jean Spence (Department of Cytogenetics and Molecular Genetics, WCH). All cell lines were maintained by Sharon Lane and Cathy Derwas (Department of Cytogenetics, WCH).

Cell lysis buffer was added to either a blood sample or to a pellet of cells obtained from a cell line culture to a volume of 30 ml. The tube was left on ice for 30 minutes, then spun at 3,500 rpm for 15 minutes at 4°C. With aspiration, 20 ml of the supernatant was removed, followed by the addition of a further 20 ml of cell lysis buffer, and a repeat centrifugation. The supernatant was discarded, and 3.25 ml of 3X Proteinase K buffer, 500 µl of 10% (w/v) SDS and 200 µl of Proteinase K (at 10 mg/ml) was added and mixed with the pellet. After an overnight incubation at 37°C with mixing, 5 ml of phenol was added to the solution, followed by a 15 minute incubation at room temperature on a rotating wheel. The solution was spun at 3,000 rpm for 10 minutes, the aqueous layer removed to a clean tube, and phenol added to a volume of 10 ml. After a 10 minute incubation at room temperature, the mix was centrifuged for 10 minutes, and the aqueous layer removed to a clean tube. Chloroform was added to a volume of 10 ml, and the sample mixed and re-centrifuged. DNA was precipitated by mixing with 300 µl of 3 M NaAc (pH5.2) and 10 ml of 100% ethanol. DNA pellets were washed with 70% ethanol, air-dried, resuspended in 100 µl of TE, and quantitated by spectrophotometry (2.2.6).

2.2.2 Preparation of Bacterial Glycerol Stocks

All bacterial clones were maintained as glycerol stocks that were prepared from a 10 ml aliquot of an overnight culture, which was being used for DNA isolation purposes. The 10 ml culture aliquot was spun at 3,000 rpm and the pellet resuspended in 1 ml of L-Broth containing 15% glycerol. These stocks were kept at -70°C until more DNA was required, whereby a 5 µl aliquot was used to streak for single colonies.

2.2.3 Preparation of Sheared Human Placental DNA and Salmon Sperm DNA

Desiccated human placental DNA and salmon sperm DNA was purchased commercially and reconstituted in sterile TE to a concentration of 5 mg/ml and 20 mg/ml respectively. Aliquots of 1 ml were transferred to 1.5 ml eppendorf tubes and incubated at 100°C for between 6 to 20 hours. Periodically, 1 µl aliquots were taken and analysed on 1.5% agarose gels until the average sized fragments were below 700 bp. The concentration of the samples were then analysed by spectrophotometry and adjusted to their starting values (2.2.6).

2.2.4 Isolation of Total RNA

The isolation of RNA from tissues and breast cancer cell lines was done under RNAse-free conditions. This consisted of using solutions that had been incubated at room temperature in 0.2% (v/v) DEPC, then autoclaved, using Gilson pipettors that had been cleaned with RNAseZAP (Ambion), wearing gloves at all times, and using filtered Gilson pipette tips. RNA was isolated using a procedure based on that of Chomczynski and Sacchi (1987) using the Trizol reagent. Isolated RNA was kept at -70°C whilst not in use.

2.2.4.1 RNA Isolation from Fresh Tissue

RNA was isolated from a number of tissues from a 20 week old male human fetus. Access to

these samples was kindly granted by Dr. Roger Byard (Department of Histopathology, WCH). Selected tissues were frozen in liquid nitrogen and stored at -70°C until RNA was required. For every 100 mg of frozen tissue, 1 ml of Trizol reagent was used. An appropriate amount of tissue (usually 100 to 500 mg) was removed from the frozen stock and placed in a sterile, mortar dish containing a small amount of liquid nitrogen to keep the sample frozen. The tissue was ground to a fine powder with a sterile pestle. Liquid nitrogen was periodically added to the tissue to prevent thawing. The mortar and pestle had previously been washed thoroughly with ethanol, rinsed with DEPC-treated water, then heat treated for 4 hours at 160-180°C to inactivate any RNases. The ground tissue was then transferred to a 10 ml tube containing the appropriate amount of Trizol. The powder was resuspended by gentle mixing and transferred to a sterile 5 ml glass homogeniser (Lab Supply). The solution was homogenised until lumps of powder were no longer visible, transferred to a 1.5 ml eppendorf tube, and left at room temperature for 5 minutes. One-tenth the volume of chloroform was added and the tube was shaken for 25 seconds and placed on ice for a further 5 minutes. The sample was then spun at 12,000 rpm at 4°C for 15 minutes, and the supernatant was transferred to a new tube. To this, an equal volume of isopropanol was added, and the tube incubated at 4°C for 15 minutes. Following centrifugation at 4°C for 15 minutes, the RNA pellet was washed in 800 µl of 75% ethanol and the pellet allowed to air-dry for 15 minutes. The RNA was resuspended in 50 µl of DEPC-treated water, and pooled into one tube if the original tissue homogenate was transferred into more than one eppendorf. A second precipitation step involving the addition of one-twentieth the volume of 4 M NaCl and 2 volumes of 100% ethanol to the resuspended RNA was performed. This was incubated at -20°C for 1 hour then spun for 15 minutes at 4°C. The RNA pellet was again washed with 800 µl of 75% ethanol, air-dried for 20 minutes, resuspended in DEPC-treated water, and quantitated by spectrophotometry (2.2.6).

2.2.4.2 RNA Isolation from Cell Lines

RNA isolation from breast cancer cell lines followed the same procedure as that for RNA isolation from tissue sources. Typically, $1X10^7$ cells obtained from cell culture were washed in PBS and spun down at 1,200 rpm for 10 minutes. The cells were then resuspended in 1 ml of Trizol and procedures identical to those described in 2.2.4.1 were subsequently followed. In all RNA isolations, the integrity of the RNA preparation was tested by running a small aliquot on a 0.8% agarose gel.

2.2.5 Oligonucleotide Primer Design

Where possible, oligonucleotide primers for PCR and sequencing were designed to contain approximately 50% G-C content, and an annealing temperature of 60°C, calculated by 2x(A+T)+4x(G+C) (Suggs *et al.*, 1981). Ideally, primers were not to possess runs of identical bases and not to contain four contiguous base pairs of inter-strand or intra-strand complementarity. In addition, primers were designed such that G-C, C-C, G-G, or C-G bases were present at their terminal 3' ends wherever possible. All oligonucleotides were purchased commercially from either Gibco-BRL or Bresatec (Adelaide, Australia).

2.2.6 Quantitation of DNA, RNA and Oligonucleotide Primers

Samples were diluted in water and their absorbance was measured with a deuterium lamp at 260nm. The absorbance was multiplied by the dilution factor and a conversion factor particular to the type of DNA being quantitated. Relevant conversion factors were 50 for double stranded DNA, 40 for RNA, and 33 for oligonucleotides, which gave a concentration in µg/ml. Spectrophotometers used were a Pharmacia Biotech Ultrospec 3000 or a CECIL CE-2020.

2.2.7 Restriction Digests

All digests were performed overnight at 37°C unless otherwise specified by the manufacturer. BSA at a final concentration of 100 μg/ml was included in the digest if recommended.

2.2.7.1 Digestion of Genomic DNA

Ten micrograms of genomic DNA from normal individuals, somatic cell hybrids or breast cancer cell lines was digested with 20 units of the appropriate restriction enzyme in a 50 μl volume. The next day, a 5 μl aliquot was run on a 0.8% agarose gel (2.2.8) in order to confirm complete digestion. The remaining 45 μl was run overnight (2.2.8).

2.2.7.2 Digestion of Cosmid, BAC, and PAC DNA

For a standard digest, 250 ng of cosmid DNA and 500 ng BAC or PAC DNA was digested in a 15 μl volume with 5 units of each restriction enzyme. This amount was sufficient for one lane on an agarose gel. Digestion reactions were scaled up based on how many lanes were required and whether the digest was for the purpose of the isolation of specific restriction fragments (see 3.2.4.1, 3.2.4.4, and 3.2.5.1). For double digests, a restriction enzyme buffer was used which was compatible for both enzymes (NEB catalogue, 1995).

2.2.7.3 Digestion of Plasmid DNA

Plasmid DNA was digested mainly for the purposes of insert isolation from cDNA clones purchased from Genome Systems. Five micrograms of plasmid DNA was digested in a 50 μl volume and 20 units of the appropriate restriction enzyme. Inserts were subsequently purified from the agarose gel using the Prep-A-Gene purification system (2.2.16.2).

2.2.8 Agarose Gel Electrophoresis

Electrophoresis of genomic and cell line DNA was usually performed in 0.8% (w/v) agarose gels and run at 18 mA per gel, overnight. The analysis of cosmid, plasmid, BAC or PAC DNA digests were usually performed in 0.8% (w/v) agarose gels, and run at 100 volts. The analysis of PCR products was performed in 1.5 to 2.5 % (w/v) agarose gels depending on the size of the expected products to be analysed, and run at 120 volts. All electrophoresis was carried out in 1X TBE buffer. Agarose loading buffer was added to each DNA sample (1X final concentration) prior to loading. DNA was visualised under UV light after staining gels in 0.02% ethidium bromide solution for 30 minutes.

2.2.9 DNA Size Markers

Depending on the size of the DNA fragments to be analysed and the relative concentration of the agarose gel used, dictated what molecular weight markers were run. Generally *Eco*RI digested SPP-1 phage markers (size range 0.5 to 8.5 kb), Puc19 DNA digested with *Hpa*II (size range 60 to 500 bp), or DrigestIII markers (size range 72 bp to 23 kb) were used.

2.2.10 Southern Blotting

2.2.10.1 Transfer in 10X SSC

This method was adapted from that originally described by Southern (1975) and was reserved for the transfer of genomic DNA, including breast cancer cell line DNA and somatic cell hybrid DNA. The agarose gel was soaked in 500 ml of gel denaturing solution for 60 minutes, followed by 500 ml of gel neutralising solution for a further 60 minutes. Prior to the completion of the gel neutralisation step, a piece of GeneScreen Plus nylon membrane was cut to size and soaked in water for 10 minutes followed by a 15 minute soak in 10X SSC. The gel was placed face down on a blotting tray containing 10X SSC, and overlayed with the pre-

treated membrane. Air bubbles were carefully removed and 2 pieces of 3MM Whatman paper soaked in 10X SSC were placed on the membrane, followed by three dry sheets. A wad of paper towels was placed over this and left to transfer overnight. The next day, the position of the wells was marked on the membrane, which was then soaked for 1 minute in filter denaturing solution followed by a 2 minute soak in filter neutralising solution. The membrane was then left to dry overnight or baked in the microwave oven for 5 minutes at half power.

2.2.10.2 Transfer in 0.4 M NaOH

This method was adapted from the method of Reed and Mann (1985) and was used for the transfer of digested cosmid, plasmid, BAC or PAC DNA. The gel to be blotted was stained in ethidium bromide, photographed, and the DNA size markers were stabbed into the gel with a needle and ink. GeneScreen Plus membranes cut to size were soaked in water for 5 minutes, followed by 0.4 M NaOH for another 5 minutes. The gel was blotted in 0.4 M NaOH overnight without any pre-treatment as described above except that the blotting tray contained 0.4 M NaOH. The next day, the position of the wells and DNA size markers was marked and the membrane was treated in filter neutralisation solution for 2 minutes. Membranes were then air-dried or dried in the microwave as above.

2.2.11 Northern Blotting

All procedures were carried out according to procedures adapted from Sambrook *et al.* (1989) using RNAse free conditions as described in 2.2.4.

2.2.11.1 RNA Electrophoresis

A solution containing 1% (w/v) agarose (146 ml) in 1X TBE was prepared. After dissolving the agarose by boiling, 20 ml of 10X MOPS buffer and 34 ml of formaldehyde were added,

and the gel was poured. This step was performed in a fume hood. Each sample was prepared for electrophoresis by the addition of 20 µg of total RNA to 2 µl of 10X MOPS buffer, 3.5 µl of formaldehyde, 10 µl of de-ionised formamide, and 2 µl of 10X RNA loading dye in a 20 µl final volume. Samples were incubated at 65°C for 15 minutes, loaded, and run at 75 volts in 1X MOPS buffer. Some samples were loaded twice so that following gel electrophoresis, this portion of the gel could be cut away from the main gel and stained with ethidium bromide. This would allow visualisation of the 28S and 18S ribosomal bands for size markers and to check the integrity of the RNA samples following electrophoresis.

2.2.11.2 Transfer of RNA to Membranes

The gel to be blotted was rinsed in DEPC-treated water for 20 minutes, followed by a 30 minute incubation in RNA hydrolising solution, a 20 minute incubation in RNA neutralising solution, and a final incubation for 40 minutes in 20X SSC. During this last step, the pre-cut Genescreen membrane was incubated in DEPC-treated water for 10 minutes followed by a 15 minute incubation in 20X SSC. The gel was then blotted in 20X SSC overnight as described in 2.2.10.1. The following day, the membrane was rinsed in 20X SSC for 1 minute, and allowed to dry at room temperature.

2.2.12 ^{32}P Radio-Isotype Labelling of DNA

Double-stranded DNA fragments were labelled with [α-^{32}P]dCTP using a modified procedure to that described by Feinberg and Vogelstein (1983). Reactions were scaled up or down accordingly from that shown below.

2.2.12.1 Labelling DNA Fragments in Solution

In a standard labelling reaction, 50 ng of double stranded DNA was made up to 34 µl with

sterile water, mixed, and incubated at 100°C for 5 minutes. After spinning down the contents, 5 µl of 10X dNTP labelling solution, 5 µl of 10X labelling buffer, 5 µl of [α-^{32}P]dCTP (50 µCi), and 5 units of *E.coli* DNA polymerase I were added. This was incubated at 37°C for 30 minutes. The reaction was stopped with the addition of 1 µl of 0.5 M EDTA, and if required, un-incorporated ^{32}P was removed using QIAquick columns (2.2.16.1). DNA probes that required blocking of repetitive sequences were then pre-reassociated (2.2.12.3). Probes not requiring pre-blocking were heat denatured at 100°C for 5 minutes before addition to pre-hybridised filters (2.2.13.1).

2.2.12.2 Labelling DNA Fragments in Agarose

The DNA fragment contained in agarose was heated at 100°C for 2 minutes. An aliquot of approximately 50 ng was then removed and transferred to a tube with sterile water such that the volume was 34 µl. This solution was heated at 100°C for 5 minutes and the same steps as in 2.2.12.1 were subsequently followed.

2.2.12.3 Pre-Reassociation of Labelled Probes

Probes suspected to contain human repetitive elements or probes to be used for Northern hybridisations had their repetitive elements blocked, before hybridisation with the membrane, with sheared denatured human placental DNA (2.2.3). The procedure used was based on that of Sealy *et al.* (1985). To the labelled probe (50 µl reaction), 100 µl of human placental DNA (5 mg/ml), and 50 µl of 20X SSC were added. This was incubated at 100°C for 10 minutes, ice for 1 minute, followed by an incubation at 65°C for at least 1 hour. The probe could then be added directly to the pre-hybridised membranes (2.2.13.1).

2.2.13 Hybridisation of Membranes

All hybridisations, including Northern blot and high-density grid hybridisations were performed in a HYBAID orbital midi oven. High-density grid hybridisation procedures are described in detail in chapter 3.

2.2.13.1 Southern Hybridisations

Southern filters were hybridised based on a modification of Brown (1993). Membranes were placed in glass bottles and pre-wet in 5X SSC for 1 minute. This solution was replaced with either 10 ml of Southern hybridisation solution (for small bottles) or 15 ml (for large bottles) and filters were pre-hybridised at 42°C for at least 2 hours. Once the DNA probe had been labelled and either denatured (2.2.12.1) or pre-reassociated (2.2.12.3), it was added directly to the pre-hybridised filters.

2.2.13.2 Northern Hybridisations

Northern membranes, both commercially purchased and produced in the lab, were hybridised according to protocols supplied with the Clontech multiple tissue blots (User manual, PT1200-1). Membranes were pre-soaked in 5X SSC for 1 minute, then pre-hybridised at 65°C for at least 2 hours in 10 ml of ExpressHyb solution containing denatured salmon sperm DNA (100 μg/ml). After the DNA probe had been labelled (2.2.12), cleaned (2.2.16.1), and pre-reassociated (2.2.12.3), the pre-hybridisation solution was removed from the membrane, and replaced with a fresh 10 ml of ExpressHyb solution containing the probe and salmon sperm DNA (100 μg/ml). Hybridisation proceeded overnight at 65°C. Commercial Northern blots were hybridised with the control β-Actin probe supplied with the membrane to test for the integrity and loading of the RNA samples.

2.2.13.3 Washing of Membranes

Southern membranes were washed based on procedures presented in Sambrook *et al.* (1989). Filters were treated sequentially at 42°C for 10 minutes in solutions containing 2X SSC and 1% SDS. If high background was still evident, the filters were washed for another 20 minutes at 42°C in 2X SSC and 1% SDS (heated to 65°C first), followed by another 20 minute wash in 0.1X SSC and 1% SDS at 65°C if needed.

Northern membranes were washed according to manufacturers specifications (User manual, PT1200-1). Filters were first rinsed in 200 ml of a solution containing 2X SSC and 0.05% SDS at room temperature for 1 minute at a time. Counts were monitored after each rinse. Following this, four 10 minute washes at room temperature in the same solution were performed with constant monitoring after each wash. If the counts were still high (>50 cpm), the membranes were washed in a solution containing 0.1X SSC and 0.1% SDS for 5 minutes at a time at 65°C. All washed membranes were exposed for the appropriate time to X-OmatK XK-1 Kodak diagnostic film either at room temperature or –70°C.

2.2.13.4 Stripping of Membranes for Re-Use

Southern membranes were stripped of radio-labelled probe by incubating the washed filters at 45°C for 30 minutes in a solution containing 0.4 M NaOH. The filters were then transferred to a solution containing 0.1% SDS; 0.1X SSC; 0.2 M Tris-HCl (pH7.5), and incubated a further 15 minutes. Northern membranes were stripped of radio-labelled probe by the addition of a solution of 0.5% (w/v) SDS that had been heated to 100°C, followed by an incubation at room temperature until cool.

2.2.14 Polymerase Chain Reaction (PCR)

All PCR reactions were performed using *Taq* DNA polymerase in one of two buffer systems; 2X PCR mix (modification of Kogan *et al.* 1987) or 10X PCR buffer (supplied by Gibco-BRL). In each case 1.5 mM $MgCl_2$ was used unless high background or no PCR products were obtained with a control template. If this occurred, the PCR conditions were optimised by varying the concentration of $MgCl_2$ used in the reaction. All PCR reactions were set up on ice, and were performed in 0.5 ml PCR tubes using a thermal cycler 480 (Perkin Elmer Cetus) or a PCR express thermal cycler (Hybaid). All colony PCRs were performed in 0.2 ml PCR tubes using an FTS-960 thermal sequencer (Corbett Research).

2.2.14.1 Standard PCR Reactions

Routine PCR reactions were performed in 10 µl volumes using 1 µl of 10X PCR buffer, 0.2 µl of 10 mM dNTPs, 0.3 µl of 50 mM $MgCl_2$, 75 ng of each primer, 0.25 units of *Taq* DNA polymerase, and template DNA. Template DNA amounts were 100 ng for genomic and cell line DNA, 10 ng for BAC and PAC DNA, and 1 ng for plasmid DNA. All reactions were overlayed with a drop of paraffin oil and incubated at 94°C for 2 minutes, followed by 35 cycles of 94°C for 30 seconds; 60°C for 1 minute; 72°C for 2 minutes, followed by a final elongation step of 72°C for 7 minutes. Each reaction was then run on an agarose gel of the appropriate percentage (2.2.8). When the cloning of PCR products was anticipated, the original reaction was scaled up to 20 µl. Half was then examined on an agarose gel while the remainder was kept for sub-cloning into the pGEM-T vector (2.2.17.2).

2.2.14.2 Colony PCR Reactions

A single colony was picked with a sterile disposable pipette tip, streaked onto a grid position

on a fresh L-Agar plate containing the appropriate antibiotic (master plate), and the tip placed into a PCR tube containing a drop of paraffin oil. The tubes were left for 10 minutes at room temperature during which the PCR reaction mix was prepared. This mix consisted of 5 µl of 2X PCR mix, 0.3 µl of 50 mM $MgCl_2$, 0.2 µl of each primer (30 ng of each), 3.6 µl of sterile water, and 0.2 µl of *Taq* DNA polymerase (0.25 units). The pipette tips were removed from the PCR tubes and the tubes were incubated at 99°C for 10 minutes. The samples were then held at 80°C, and 10 µl of the prepared PCR reaction mix was added to each tube below the level of the oil. The tubes were incubated at 95°C for 2 minutes followed by 35 cycles of 95°C for 15 seconds; 60°C for 30 seconds; 72°C for 2 minutes, and a final extension step of 7 minutes at 72°C. In general, 5 µl of each reaction was examined on a 2.5% agarose gel.

2.2.15 Reverse Transcriptase PCR (RT-PCR)

2.2.15.1 Reverse Transcription

All procedures were adapted from those supplied with the Superscript enzyme. All reactions were carried out on ice at all times. Three micrograms of total RNA or 100 ng of $polyA^+$ mRNA was added to either 50 pmoles of random hexamers (Perkin Elmer) or 100 pmoles of oligo dT_{12-18}. The volume was made up to 11.5 µl with the addition of DEPC-treated water, mixed, and incubated at 65°C for 5 minutes. Following a 1 minute incubation on ice, the contents of the tube were spun down briefly, and 4 µl of 5X 1st strand buffer, 2 µl of 0.1 M DTT, 1 µl of 10 mM dNTPs, and 0.5 µl (20 units) of RNAsin were added. This was incubated at 42°C for 2 minutes, followed by the addition of 1 µl (200 units) of Superscript reverse transcriptase, and a further incubation at 42°C for 30 minutes. The reaction was terminated by incubating at 70°C for 10 minutes and samples were kept at -20°C until needed. A control reaction was included for each individual experiment where the reverse transcriptase enzyme

was omitted. This was to test for genomic contamination present within the RNA template that may be seen following the PCR step (2.2.15.2). All polyA$^+$ mRNA samples were kindly provided by Dr. Jozef Gecz (Department of Cytogenetics and Molecular Genetics, WCH).

2.2.15.2 PCR of Reverse Transcribed RNA

Two microliters of the reverse transcription reaction was combined with 2 µl of 10X PCR buffer, 0.4 µl of 10 mM dNTPs, 0.6 µl of 50 mM $MgCl_2$, 1 µl of each primer (150 ng of each), and the volume made up to 19 µl with sterile water. After the addition of 1 µl of *Taq* DNA polymerase (0.5 units) to all tubes, including those tubes without reverse transcriptase, one drop of paraffin oil was added, and the tubes were incubated at 94°C for 2 minutes. The samples were then incubated at 94°C for 30 seconds; 60°C for 1 minute; 72°C for 2 minutes for 35 cycles, followed by a final elongation step at 72°C for 7 minutes. Ten microlitre aliquots of all RT-PCR products were analysed on 2.5% (w/v) agarose gels while the remaining aliquot for many reactions was kept for subcloning into the pGEM-T vector (2.2.17.2) or for direct purification using QIAquick columns (2.2.16.1) for sequencing purposes. A positive control PCR to test the success of the cDNA synthesis was done for each reverse transcription reaction with primers to the esterase D (*ESTD*) housekeeping gene (GenBank Accession number M13450). These primers give rise to a 452 bp amplicon from cDNA template only. Primer sequences are listed in Table 2.1.

2.2.16 Purification of DNA Fragments

2.2.16.1 QIAquick Purification of PCR Fragments and Radio-Labelled Probes

PCR products were cleaned according to manufacturers protocols using QIAquick columns (Qiagen) and buffers which were supplied with the kit. Briefly, the remaining PCR reaction

was removed from the paraffin oil and transferred to a clean tube. Each sample was mixed with a 5X volume of buffer PB. This mixture was added to a purification column, and spun for 1 minute at 13,000 rpm. The collection tube was drained and 750 µl of buffer PE was then added, followed by a 1 minute spin at 13,000 rpm. The collection tube was again drained, and the column spun again. The DNA was eluted by the addition of 50 µl of sterile water to the column, followed by centrifugation at 13,000 rpm for 1 minute with a clean collection tube. The purified DNA was quantitated by spectrophotometry (2.2.6).

Radio-labelled probes were purified from un-incorporated nucleotides using the same procedure as described above. However, only 700 µl of buffer PE was used to reduce the chance of contaminating the centrifuge from overfilling of the spin column.

2.2.16.2 Prep-A-Gene Purification of DNA Fragments in Agarose

Restriction fragments excised from agarose gels were purified from the agarose using the Prep-A-Gene purification system (Bio-Rad) according to manufacturers conditions. Where possible, fragments to be isolated were run in 0.8% (w/v) gels, and trimmed of excess agarose once cut from the gel. Five microlitres of Prep-A-Gene matrix was used per microgram of DNA excised. Three gel slice volumes of binding buffer were added to the excised band, followed by a 10 minute incubation at 50°C to dissolve the agarose. After addition of the required amount of matrix, the tube was vortexed briefly, and incubated at room temperature for 10 minutes on a rotating wheel. The tube was then spun for 30 seconds, and the supernatant removed. The pellet was rinsed and resuspended in binding buffer equivalent to 25X the amount of added matrix, and spun a further 30 seconds. The pellet was then washed twice with a 25X matrix volume of wash buffer. The pellet was then air-dried for 10 minutes at room temperature, resuspended in at least 1 pellet volume of sterile water, and incubated at 42°C for 5 minutes.

The tube was then spun for 30 seconds and the supernatant transferred to a fresh tube. The above step was repeated and the supernatants pooled. The concentration of the DNA eluted was then determined by spectrophotometry (2.2.6).

2.2.17 Sub-Cloning of DNA Fragments

2.2.17.1 DNA Ligations

All ligations were performed in a volume of 10 µl at 15°C for 3 to 16 hours. Typical ligations involved 1:1 molar ratios of the vector and the insert to be subcloned based on methods described in Sambrook *et al.* (1989). In most reactions 50 ng of vector DNA was used. Ligations included 5 units of T4 DNA ligase (MBI) using the buffer supplied with the enzyme.

2.2.17.2 Ligating PCR Products into the pGEM-T Vector

The pGEM-T vector system (Promega) was employed to clone DNA fragments amplified by PCR (except for the PCR products generated during exon trapping). This system takes advantage of the addition of a single adenosine residue by thermostable polymerases to the 3' end of all duplex molecules. The A overhangs are ligated to single thymidine overhangs present in the pGEM-T vector. In most cases, 1 µl of the completed PCR reaction was added to 1 µl of the pGEM-T vector, 1 µl of 10X ligation buffer, and 1µl of T4 DNA ligase in a 10 µl volume. Ligations were performed overnight at 15°C.

2.2.17.3 Preparation of Competent Bacterial Cells

Bacterial cells were made competent based on a modification of Chung *et al.* (1989). XL-1 Blue cells were streaked for single colonies on an L-Tetracycline plate from a 5 µl aliquot of a glycerol stock. A single colony was grown overnight in 10 ml of L-Broth plus tetracycline (15

μg/ml) at 37°C. The next day, 50 ml of L-Broth plus tetracycline (15 μg/ml) was seeded with 1 ml of the overnight culture and grown until an A_{600} of 0.3 was reached (approximately 1.5 to 2 hours). The cells were then transferred to a 50 ml tube and spun at 2,200 rpm for 10 minutes. The supernatant was removed and the cells were carefully resuspended in 5 ml of ice cold TSS media. The cells were left on ice for 10 minutes before being transformed (2.2.17.4). Any left over cells were transferred to ice-cold eppendorf tubes, and frozen in liquid nitrogen. Frozen cells were then stored at -70°C until required.

2.2.17.4 Transformation of Competent Bacterial Cells

For this procedure, a modified version of the technique described in Chung *et al.* (1989) was used. For each reaction, 100 μl of competent XL-1 Blue cells were thawed on ice and 5 μl of the ligation reaction was added. Following gentle mixing, the cells were placed at 4°C for 30 minutes. During this step, 880 μl of L-Broth was placed in a 10 ml tube and glucose was added to a concentration of 20 mM. The cells and DNA mix were then added, and the tube was incubated at 37°C for 1 hour. For vectors with blue/white colour selection, an aliquot of 200 μl of transformed cells, 40 μl of X-Gal (20 mg/ml) and 20 μl of 200 mM IPTG were then spread on an L-Ampicillin plate, and incubated overnight at 37°C. Otherwise, 200 μl of transformed cells alone was spread onto an L-Ampicillin plate.

2.2.17.5 Preparation of Colony Master Plates and Colony Lifting

Colonies representing potential recombinant molecules were picked and streaked onto a gridded master plate. Following an overnight incubation at 37°C, inserts of gridded clones were subsequently amplified by colony PCR (2.2.14.2) or transferred to nylon membranes for screening purposes. The procedure to transfer bacterial clones to membranes was based on a

modified version to that described by Grunstein and Hogness (1975). Master plates containing colonies for transfer were first placed at 4°C until chilled. Agar plates containing individually gridded bacterial colonies were overlayed with a membrane and markers used to orient the filters were stabbed into the edge of the membrane using a needle and ink. This was left for 5 minutes. The membrane was removed and placed colony side up on a sheet of Whatman 3MM paper soaked in colony denaturing solution and left for 7 minutes. The membrane was then transferred to a sheet of Whatman 3MM paper soaked in colony neutralisation solution and left for 5 minutes. This step was repeated on a separate sheet of Whatman paper, and the membranes were then soaked in 500 ml of 2X SSC for 5 minutes, air-dried for 10 minutes, and baked at 65°C for 1 hour in an oven. In some instances duplicate colony lifts were done in which a second membrane was placed onto the agar plate once the first one had been removed. All steps subsequent to this were as stated above.

2.2.18 DNA Sequencing

The sequencing of both double stranded plasmid DNA and PCR product DNA was achieved with cycle sequencing using reagents supplied with the ABI Prism DyeTerminator or DyePrimer (-21M13 forward and M13RP1) Cycle Sequencing Kit. All samples were electrophoresed on an Applied Biosystems Model 373A DNA sequencer by Dr. Jozef Gecz or Dr. Julie Nancarrow (Department of Cytogenetics and Molecular Genetics, WCH), or on the same machine when it was relocated to the Australian Genome Research Facility in Brisbane.

2.2.18.1 DyeTerminator Cycle Sequencing

For each sequencing reaction, either 180 ng of purified PCR product DNA or 500 ng of plasmid DNA was mixed with 20 ng of the desired primer, 4 µl of the DyeTerminator mix, 4 µl of halfTERM Dye Terminator Sequencing Reagent, and made up to a volume of 20 µl with

sterile water. After the addition of a drop of paraffin oil, the tubes were incubated for 25 cycles at 96°C for 30 seconds; 50°C for 15 seconds; 60°C for 4 minutes. The sample was then purified using procedures provided with the kit (Revision A, 1995). This involved removal of the sample from the oil and transfer to an eppendorf tube containing 2 µl of 3 M NaAc (pH5.2) and 50 µl of 95% ethanol. Following a 10 minute incubation on ice, the tubes were spun at 13,000 rpm for 30 minutes, and the supernatant was discarded. The pellet was then washed with 250 µl of 70% ethanol, spun for a further 5 minutes and the ethanol removed. The pellet was allowed to dry for 10 minutes at room temperature, and then kept at 4°C in the dark until run on a sequencing gel.

2.2.18.2 DyePrimer Cycle Sequencing

For the sequencing of each template, four separate sequencing reactions were needed. To the first two tubes, one containing 4 µl of dd-ATP mix and the other containing 4 µl of dd-CTP mix, 1 µl (200 ng) of DNA template was added. To the remaining two tubes, one containing 8 µl of dd-GTP mix and the other containing 8 µl of dd-TTP mix, 2 µl (400 ng) of DNA template was added. After the addition of a drop of paraffin oil, the tubes were incubated for 15 cycles of 95°C for 30 seconds; 55°C for 30 seconds; 70°C for 1 minute, followed by a further 15 cycles of 95°C for 30 seconds; 70°C for 1 minute. Samples were then purified using procedures provided with the kit (Revision B, 1995). All four samples from each template were removed from the oil and combined in a tube containing 80 µl of 95% ethanol. Following a 10 minute incubation on ice, the tubes were spun at 13,000 rpm for 30 minutes then the supernatant was removed. The pellet was washed with 250 µl of 70% ethanol, spun for 5 minutes, and air-dried for 10 minutes after removal of the supernatant. Samples were subsequently kept at 4°C in the dark until run on a sequencing gel.

2.2.19 Fluorescence *in situ* Hybridisation (FISH)

All FISH analysis was kindly performed by Elizabeth Baker, Erica Woollatt, or Helen Eyre (Department of Cytogenetics, WCH). Briefly, whole cosmid or plasmid DNA was nick translated with biotin-14-dATP and hybridised *in situ* at a final concentration of 20 ng/μl to metaphases from 2 normal males. The FISH method had been modified from that previously described (Callen *et al.*, 1990*a*). Chromosomes were stained before analysis with both propidium iodide (as counter-stain) and DAPI (for chromosome identification). Images of metaphase preparations were captured by a cooled CCD camera using the CytoVision Ultra image collection and enhancement system (Applied Imaging Int. Ltd.). FISH signals and the DAPI banding pattern were merged for figure preparation.

2.2.20 5' Rapid Amplification of cDNA Ends (5'RACE)

The cloning of the 5' ends of transcripts was aided by the use of the 5'RACE kit (Gibco-BRL). All components were supplied with the kit except the gene specific primers (GSP) which are listed in the relevant chapters that used this technique. Briefly, 1 μg of total RNA or 100 ng of polyA$^+$ mRNA was reverse transcribed with 2.5 pmoles of GSP1 at 42°C for 50 minutes, followed by a 15 minute incubation at 70°C. After the addition of 1 μl of RNAse mix, the tubes were left for a further 30 minutes at 37°C. The first strand cDNA was purified using Glass Max spin columns supplied with the kit. The cDNA was eluted from the columns with 50 μl of sterile water preheated to 65°C. Ten microlitres of the purified sample was then added to 65 μl of DEPC-treated water, 5 μl of 5X tailing buffer, and 2.5 μl of 2 mM dCTP and incubated for 3 minutes at 94°C followed by a 1 minute incubation on ice. After the addition of 1 μl of the TdT mix, the tubes were incubated at 37°C for 10 minutes. The TdT was then heat inactivated for 10 minutes at 65°C. The tailed cDNA was subjected to PCR using a GSP2

primer and the deoxyinosine containing anchor primer (AAP) specific for the dC tail. This was followed by a nested PCR using a GSP3 primer and a primer homologous to the AAP primer (AUAP). Negative control reactions excluding reverse transcriptase were included for each reaction. Ten microlitres of the PCR products were analysed on 1.5% (w/v) agarose gels, and the remaining sample was cloned into the pGEM-T vector (2.2.17.2) and subsequently sequenced (2.2.18.2) using vector specific primers. AAP and AUAP primer sequences are listed in Table 2.1.

2.2.21 PCR Formatting of the Fetal Brain cDNA Library

A random-primed and polyA-primed human fetal brain 5'-Stretch Plus cDNA phage library (Clontech) was formatted for rapid screening using PCR. This involved plating a 1.5 fold representation of all clones in the library on one hundred 15 cm L-Agar plates (15,000 phage per plate), transferring the amplified clones from each plate to a 10 ml tube, and taking an aliquot from each tube to produce 40 pools. These pools were subsequently used for templates in PCR reactions with gene specific primers or a combination of vector and gene specific primers.

2.2.21.1 Lambda Phage Plating

Y1090 *E. coli* cells were first grown in 10 ml of L-broth, 100 µl of 1 M $MgSO_4$, and 100 µl of 20% maltose, overnight at 37°C. The following day, 100 eppendorf tubes were set up, such that each contained 500 µl of cells combined with 15,000 plaque forming units of phage (pfu) from a dilution of the cDNA library (diluted in SM buffer). These tubes were incubated at 37°C for 30 minutes. During this period, 8 ml of Top Agar was aliquoted into 100, 10 ml tubes and left at 50°C. For each of the 100 tubes, the mixture of cells and phage were then added to the Top Agar, mixed by inversion, then poured onto a pre-warmed 15 cm L-Agar plate. After

the Top Agar had set, the plates were incubated overnight at 37°C.

2.2.21.2 Lambda Phage Scraping

To each of the 100 plates, 10 ml of L-Broth was added, and the Top Agar (containing the amplified phage) was scraped away with a sterile scalpel blade, along with the L-Broth, into a 10 ml tube. These tubes were spun to remove the agar and the supernatant transferred to fresh 10 ml tubes. These tubes were kept at 4°C.

2.2.21.3 Pooling of Phage Scrapes

The 100 tubes were reduced to 40 pools that could be used for subsequent PCR reactions. The 100 tubes were split into 4 groups (1-25, 26-50, 51-75, 76-100) and for the first group 200 µl from tubes 1 to 5 were combined in a 1.5 ml eppendorf and called pool A1, tubes 6-10 combined to form pool A2, etc up to A5. Then 200 µl from tubes 1, 6, 11, 16, and 21 were combined to form pool A6, tubes 2, 7, 12, 17, and 22 combined to form pool A7, etc up to A10. This was duplicated for tubes 26-50 (pools B1-B10), 51-75 (pools C1-C10), and 76-100 (pools D1-D10). This produced 40 pools, which were subsequently used as templates in PCR reactions (5.2.2). The results presented in this thesis used these pools specifically for the isolation of 5' end sequences from transcripts being characterised (chapter 5). Subsequent PCR products were subcloned into pGEM-T as described in this chapter. This pooled library has also been used to isolate cDNA clones corresponding to trapped exons identified in chapter 4, however these results are not discussed in this thesis.

TABLE 2.1

Sequences of Primers of General Use

Primer	Sequence (5'-3')
T3	ATT AAC CCT CAC TAA AGG GA
T7	TAA TAC GAC TCA CTA TAG GG
D16S3121F	CAT GTT GTA CAT CGT GAT GC
D16S3121R	AGC TTT TAT TTC CCA GGG GT
D16S3026F	CTC CCT GAG CAA CAA ACA CC
D16S3026R	GGT CAT TTA TAT GCG CCT GA
AAP	GGC CAC GCG TCG ACT AGT ACG GGI IGG GII GGG IIG
AUAP	GGC CAC GCG TCG ACT AGT AC
SA5	CTA GAA CTA GTG GAT CTC CAG G
SD5	CCC TCG AGG TCG ACC CAG C
SA2	ATC TCA GTG GTA TTT GTG AGC
SD6	TCT GAG TCA CCT GGA CAA CCT
dUSA4	CUA CUA CUA CUA CAC CTG AGG AGT GAA TTG GTC G
dUSD2	CUA CUA CUA CUA GTG AAC TGC ACT GTG ACA AGC TGC
PucF	CAC GAC GTT GTA AAA CGA CGG CCA GT
PucR	TGT GAG CGG ATA ACA ATT TCA CAC AGG A
ESTDF	GGA GCT TCC CCA ACT CAT AAA TGC C (bases 423-447)
ESTDR	GCA TGA TGT CTG ATG TGG TCA GTA A (bases 875-851)

Note. The numbers in brackets after the esterase D (*ESTD*) primers refer to the positions of the primers with respect to the GenBank accession number, M13450, for this gene. I: inosine residues; U: uracil residues.

Chapter 3

Physical Mapping of Chromosome 16q24.3

Table of Contents

Page

3.1 Introduction **92**

3.1.1 Chromosome 16q24.3 and Fanconi Anaemia 93

3.2 Methods **95**

3.2.1 Chromosome 16 High-Density Cosmid Library Screening 95

3.2.2 Probes used for Initial Cosmid Identification 96

3.2.3 Human Specific PAC and BAC Library Hybridisations 96

3.2.4 Isolation of Cosmid Ends 97

3.2.4.1 Digestion of Cosmid DNA 97

3.2.4.2 Labelling of T3 and T7 Oligonucleotides 97

3.2.4.3 Hybridisation and Washing 98

3.2.4.4 Cosmid End Purification 98

3.2.5 Isolation of BAC and PAC Ends 98

3.2.5.1 DNA Digestion 98

3.2.5.2 Digest Re-Ligation and Transformation 99

3.2.6 Mapping of Microsatellites D16S3121 and D16S3026 99

3.2.6.1 PCR Mapping 99

3.2.6.2 Southern Mapping 100

3.3 Results **100**

3.3.1 Initial Identification of Cosmids from 16q24.3 100

3.3.2 Restriction Enzyme Analysis and Cosmid Walking 103

3.3.3 BAC and PAC High-Density Grid Screening 109

3.3.4 Establishment of Contig Orientation 110

3.3.5 Physical Map Coverage 112

3.4 Discussion **113**

3.4.1 Physical Map Integrity 117

3.4.2 Integration of Physical and Genetic Maps 118

3.1 Introduction

To identify new candidate breast cancer tumour suppressor genes mapping to the critical LOH interval at 16q24.3, a clone based physical map contig covering this region was needed. At the start of this project the most commonly used and readily accessible cloned genomic DNA fragments were contained in either lambda, cosmid or YAC vectors. During the construction of a whole-chromosome 16 physical map, clones from a number of YAC libraries were incorporated into the map (Doggett *et al.*, 1995). These included clones from a flow-sorted chromosome 16-specific YAC library (McCormick *et al.*, 1993), from the CEPH Mark I and MegaYAC libraries (Albersten *et al.*, 1990; Bellanne-Chantelot *et al.*, 1992) and from a half-telomere YAC library (Riethman *et al.*, 1989). Detailed STS and Southern analysis of YAC clones mapping at 16q24.3 established that very few were localised between the somatic cell hybrid breakpoint CY18A(D2) and the long arm telomere (unpublished data). However, those that were located in this region gave inconsistent mapping results and were suspected to be rearranged or deleted. Coupled with the fact that YAC clones make poor sequencing substrates, the difficulty in isolating the cloned human DNA, and an independent aim of our group to sequence this telomeric region, a physical map based on cosmid clones was the preferred option.

A flow-sorted chromsome 16 specific cosmid library had previously been constructed (Longmire *et al.*, 1993), with individual cosmid clones gridded in high-density arrays onto nylon membranes. These filters collectively contained ~15,000 clones representing an approximately 5.5 fold coverage of chromosome 16. Individual cosmids mapping to the critical region were identified by the subsequent hybridisation of these membranes with 9 markers identified by this and previous studies to map to the region (3.2.2).

The strategy to align overlapping cosmid clones was based on their STS content and restriction endonuclease digestion pattern. Those clones extending furthest within each initial contig would then be used to walk along the chromosome by the hybridisation of the ends of these cosmids back to the high-density cosmid grids. This process would continue until all initial contigs were linked and therefore the region defining the location of the breast cancer tumour suppressor gene would be contained within the map. Individual cosmid clones representing a minimum tiling path in the contig will then be used for the identification of transcribed sequences by exon trapping (chapter 4), and for the sequencing efforts to follow in the future.

3.1.1 Chromosome 16q24.3 and Fanconi Anaemia

During the early stages of the project, the gene responsible for Fanconi anemia complementation group A (*FAA*) had been mapped to chromosome 16q24.3 on the basis of classical linkage analysis and allelic association in a South African founder population (Pronk *et al.*, 1995). This study indicated that the *FAA* gene was located distal to the 16q24.3 marker D16S498, a region overlapping with the critical LOH region observed in breast cancer.

In 1928, Fanconi described anemia in three brothers who had a range of clinical manifestations, including microencephaly, genital hypoplasia and hyperreflexia. It is now known that the associated disease has the principal characteristic of progressive bone marrow failure afflicting children at an early age. Affected individuals are also predisposed to cancer, possibly explained by the fact that Fanconi anemia (FA) cells show increased chromosome breakage and hypersensitivity to DNA cross-linking agents. The disease has therefore been included among the group of DNA repair disorders.

Fanconi anemia is inherited as an autosomal recessive disorder and is rare, affecting

approximately 1/350,000 individuals, with a carrier frequency of 1/300 (Digweed, 1993). The clinical heterogeneity in FA is supported by an observed genetic heterogeneity, implying the role of a number of genes in this disorder. Previously, patients were placed into groups based on the complementation of the sensitivity of lymphoblastoid cells to cross-linking agents by utilising somatic cell fusion studies between different patient cell lines. These studies identified 8 complementation groups (FA-A through FA-H) each thought to correspond to a specific gene defect (Joenje *et al.*, 1997). As there was no phenotypic differences between patients in all complementation groups, it was suggested that the FA proteins may function as a large multi-protein complex, or may belong to the same metabolic pathway.

The *FAA* and *FAC* genes account for 75-80% of patients with *FAA* the most common (Buchwald, 1995). The *FAC* gene has been cloned (Strathdee *et al.*, 1992), maps to chromosome 9q22.3, and encodes a highly hydrophobic protein of 558 amino acids that shows no homology to known genes. A location for the *FAD* gene has been identified at 3p21-26 (Whitney *et al.*, 1995), but the gene has not yet been cloned. The *FAG* gene has been mapped to 9p13 (Saar *et al.*, 1998), and subsequent complementation studies have been successful in isolating the gene responsible (de Winter *et al.*, 1998). *FAG* is identical to the previously isolated *XRCC9* gene (Liu *et al.*, 1997), which shows no homology to *FAC* or to any other known genes.

The original linkage mapping data for *FAA* has been refined through the analysis of more families with additional microsatellite markers at 16q24.3. Significant linkage disequilibrium was observed with the three most telomeric markers (D16S3121, D16S3026, and D16S303), with a recombinant in one family indicating a location distal to D16S3121 (The FAB consortium, 1996-Appendix A1). This region precisely overlaps with that seen in LOH studies

of breast cancer, and prompted the formation of a collaboration between the group at the Women's and Children's Hospital (WCH) in Adelaide and those groups focussed on the cloning of the *FAA* gene. This collaborative group has termed itself the FAB (Fanconi anemia/Breast cancer) consortium, with the construction of a physical and transcription map aiming not only towards the cloning of a breast cancer tumour suppressor gene, but also to the cloning of the gene responsible for *FAA*.

3.2 Methods

3.2.1 Chromosome 16 High-Density Cosmid Library Screening

Cosmids were identified from a flow-sorted chromosome 16 genomic cosmid library. Chromosome 16 was sorted from the mouse/human somatic cell hybrid CY18, which contains this chromosome as the only human DNA (Callen, 1986), and *Sau*3A partially digested CY18 DNA was ligated into the *Bam*HI cloning site of the cosmid sCOS-1 vector (Longmire *et al.*, 1993). This cloning site is flanked on either side by RNA promoters (T3 and T7) which enable identification of insert ends (3.2.4). The library consists of a set of 10 membranes containing high density gridded (~15,000) clones representing approximately 5.5X coverage of chromosome 16. All grids were hybridised and washed using methods described in Longmire *et al.* (1993). Briefly, the 10 filters were pre-hybridised in 2 large bottles for at least 2 hours in 20 ml of a solution containing 6X SSC; 10 mM EDTA (pH8.0); 10X Denhardt's; 1% SDS and 100 μg/ml denatured fragmented salmon sperm DNA at 65°C. Overnight hybridisations with ^{32}P labelled probes (2.2.12) were performed in 20 ml of fresh hybridisation solution at 65°C. Filters were washed sequentially in solutions of 2X SSC; 0.1% SDS (rinse at room temperature), 2X SSC; 0.1% SDS (room temperature for 15 minutes), 0.1X SSC; 0.1% SDS (room temperature for 15 minutes), and 0.1X SSC; 0.1% SDS (twice for 30 minutes at 50°C if

needed). Membranes were exposed at -70°C for between 1 to 7 days.

3.2.2 Probes used for Initial Cosmid Identification

Initial markers used for cosmid grid screening were those known to be located between the somatic cell hybrid breakpoints CY18A(D2) and the long arm telomere (Callen *et al.*, 1995). These are indicated in Table 3.1 and included three genes, *CMAR*, *DPEP1*, and *MC1R*; the microsatellite marker D16S303; an end fragment from the cosmid 317E5, which contains the *BBC1* gene; and four cDNA clones, yc81e09, yh09a04, D16S532E, and ScDNA-C113. The IMAGE consortium cDNA clone, yc81e09, was obtained through screening an arrayed normalised infant brain oligo-dT primed cDNA library (Soares *et al.*, 1994), with the insert from cDNA clone ScDNA–A55 (performed by Dr. Greg Lennon at the Lawrence Livermore National Laboratory, USA). Both the ScDNA-A55 and ScDNA-C113 clones were originally isolated from a hexamer primed heteronuclear cDNA library constructed from the mouse/human somatic cell hybrid CY18 (Whitmore *et al.*, 1994). The IMAGE cDNA clone yh09a04 was identified from direct cDNA selection of the cosmid 37B2 which was previously shown to map between the CY18A(D2) breakpoint and the 16q telomere (Apostolou, 1997). The EST, D16S532E, was also mapped to the same region. Subsequent to these initial screenings, restriction fragments representing the ends of cosmids were used to identify additional overlapping clones (see Table 3.2).

3.2.3 Human Specific PAC and BAC Library Hybridisations

Four sets of either BAC or PAC library filters were used for screening purposes. One set of PAC library filters and the Human Release II BAC library filters were obtained from Genome Systems (St. Louis, Missouri, USA) while the RPCI-4 PAC and RPCI-11 (Segment 3) BAC library filter sets were obtained from BACPAC resources (Rosewell Park Cancer Institute,

New York, USA).

All library filters were hybridised and washed according to manufacturers recommendations. Initially, membranes were individually pre-hybridised in large glass bottles for at least 2 hours in 20 ml of 6X SSC; 0.5% SDS; 5X Denhardt's; 100 µg/ml denatured salmon sperm DNA at 65°C. Overnight hybridisations with ^{32}P labelled probes were performed at 65°C in 20 ml of a solution containing 6X SSC; 0.5% SDS; 100 µg/ml denatured salmon sperm DNA. Filters were washed sequentially in solutions of 2X SSC; 0.5% SDS (room temperature 5 minutes), 2X SSC; 0.1% SDS (room temperature 15 minutes) and 0.1X SSC; 0.5% SDS (37°C 1 hour if needed).

3.2.4 Isolation of Cosmid Ends

3.2.4.1 Digestion of Cosmid DNA

Cosmid DNA was isolated for digestion as described in 2.2.7.2. DNA (500ng) was double digested with *Not*I in combination with a range of other enzymes, including *Hind*III, *Bam*HI, *Xba*I, *Hinc*II, *Bgl*II, *Mlu*I, and *Eco*RI. Digests were divided in half, run in 0.8% (w/v) agarose gels in tandem, such that after transferring to nylon membranes (2.2.10.2) 2 filters would be generated, one of which to screen with the T3 oligo, the other to be probed with T7 (primer sequences shown in Table 2.1).

3.2.4.2 Labelling of T3 and T7 Oligonucleotides

The T3 and T7 oligos were 5' end labelled with 5 ul of [γ-^{32}P]dATP (50 µCi) using 2 µl of T4 polynucleotide kinase (20 units) as described by Chaconas and van de Sande (1980). Reactions were performed in a 50 µl volume using 300 ng of either the T3 or T7 oligo (Promega), and

incubated at 37°C for 60 minutes.

3.2.4.3 Hybridisation and Washing

Membranes were pre-hybridised as described in 2.2.13.1. The labelled T3 and T7 probes were then added separately to the pre-hybridised membranes and incubated overnight at 42°C. The next day the membranes were washed twice for 5 minutes at room temperature in a solution of 5X SSC; 0.1% SDS. Further washes at room temperature in 2X SSC; 0.1% SDS were performed until membranes were at ~5 to 10 counts per minute.

3.2.4.4 Cosmid End Purification

Cosmid DNA (5 μg) was double digested with 30 units of each enzyme as described in 2.2.7.2 in a 50 μl reaction volume. The enzyme combination used was that determined in 3.2.4.3. The restriction fragment representing the end of the cosmid was excised from a 0.8% agarose gel and the DNA was purified from the agarose using the Prep-A-Gene purification kit (2.2.16.2). To determine whether the T3 or T7 end of a particular cosmid extended furthest within the contig, the ends were hybridised to the overlapping cosmids.

3.2.5 Isolation of BAC and PAC Ends

BAC and PAC ends were isolated based on methods described by the manufacturers of the BAC and PAC high-density membranes (Genome Systems PAC manual).

3.2.5.1 DNA Digestion

BAC or PAC DNA (1 μg) was digested in a 20 μl reaction volume with a restriction enzyme that did not cut within the vector. This would remove all of the internal fragments from the genomic insert such that the two extremities of the insert could be religated. For PAC clones,

*Sac*I was sufficient, while for BAC clones *Sac*II, *Mlu*I, *Nhe*I, or *Xcm*I were used. Following an overnight digestion, enzymes were heat denatured at 65°C for 20 minutes.

3.2.5.2 Digest Re-Ligation and Transformation

In a 20 µl volume, 2 µl of the overnight digest was re-ligated at 15°C for 4 hours, then transformed into XL-1 blue cells as described in 2.2.17.4. Recombinant clones produced result from the re-ligation of the extreme ends of the PAC or BAC clone. DNA was subsequently isolated from a recombinant clone using methods described in 2.2.1.2. In both vectors, a *Not*I site flanks each end of the cloning site, such that a double digest involving *Not*I and the restriction enzyme chosen to originally digest the DNA clone, will result in the excision of each end of the PAC or BAC insert. This end was subsequently excised and purified from an agarose gel (2.2.16.2) and used in grid screens as described in 3.2.3.

3.2.6 Mapping of Microsatellites D16S3121 and D16S3026

3.2.6.1 PCR Mapping

A standard PCR was performed as described in 2.2.14.1 using primers from D16S3121 and D16S3026. Total human DNA and CY18 somatic cell hybrid DNA (containing chromosome 16 as its only human complement) were positive controls, while the hybrid cell line A9 (mouse background hybrid from which all mouse/human hybrids have been produced) and a no DNA template reaction were negative controls. Cosmids representing a minimum tiling path in the established contigs were chosen for mapping purposes. PCR product sizes for D16S3121 and D16S3026 were 71 to 87 bp and 197 to 213 bp respectively. Primer sequences are shown in Table 2.1.

3.2.6.2 Southern Mapping

Cosmid DNA corresponding to clones shown to be positive for D16S3121 and D16S3026 by PCR was digested and transferred to nylon membranes as described in 2.2.10.2. Membranes were probed with ^{32}P labelled purified PCR products generated from the amplification of human genomic DNA with the D16S3121 and D16S3026 primers, which allowed mapping to specific restriction fragments. Methods used are described in 2.2.12.1, 2.2.13.1, and 2.2.13.3.

3.3 Results

3.3.1 Initial Identification of Cosmids from 16q24.3

A total of nine separate markers mapping between the somatic cell hybrid breakpoint CY18A(D2) and the 16q telomere were initially used to identify cosmid clones derived from this region through screening of the high-density cosmid filters. A representative autoradiograph of cosmid grid screening is shown in Figure 3.1A. Clones corresponding to positive hybridisation signals were obtained from Los Alamos National Laboratories (LANL) and colony master grids were prepared from isolated single colonies (2.2.17.5). The probe used for the initial identification of these cosmids was then used to screen for any false positive clones before DNA was isolated for restriction enzyme analysis. Table 3.1 indicates the cosmids detected, and subsequently confirmed as true positives, for each of the initial markers used. In general, the majority of clones displaying hybridisation signals on the high-density grids were later confirmed to be true positives.

As can be seen from Table 3.1, the markers yc81e09 (ScDNA–A55), *MC1R*, and D16S532E all identified the same set of cosmid clones, indicating their close physical proximity. Similarly, cosmids identified by *CMAR* and an end fragment of the cosmid 317E5 could be grouped into

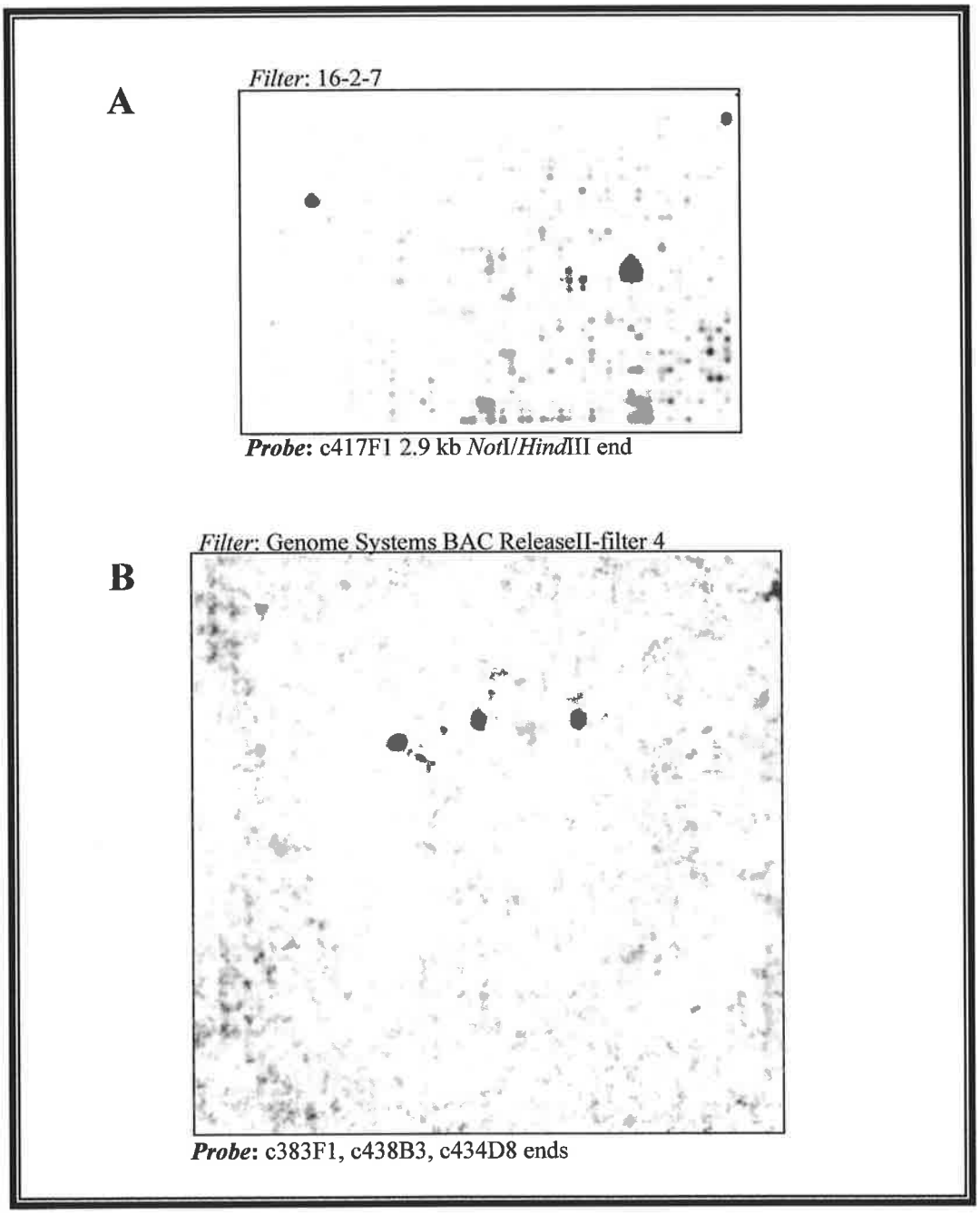


Figure 3.1: Representative autoradiographs of typical results seen from the screening of (**A**) cosmid high-density and (**B**) BAC high-density membranes. Each cosmid filter contains 1,536 individual clones with 10 filters comprising the set. Each BAC filter contains 18,432 clones, each one represented twice within a filter to avoid false positive signals. The BAC filter above shows the identification of three clones with each signal seen as two spots on a lighter exposure. Both the Genome Systems BAC and PAC filter sets consist of 6 and 7 membranes respectively, while the BACPAC resources filter sets comprise of 6 filters each.

TABLE 3.1

Cosmids Identified from Initial Marker Screening

Marker	Cosmids Identified		
yc81e09	360D7 416C2 416C4	416C1 374E6 377F1	416B6 336A8 337G8
MC1R	374E6 336A8	337G8 416B6	416C1 377F1
D16S532E	377F1 336A8	337G8 374E6	416C1
yh09a04	352A12 309E9	361H2	361G2
317E5 end (*BBC1*)	317E5 408G10 342G5	376F9 410H4 325G2	383H6 432B2 384F2
CMAR	432B2	384F2	
DPEP1	444C11 444D11	435H5 435H6	324F5
ScDNA-C113	434D8 396B4	372C1 348F3	361E4
D16S303	425E4		

Note. All initial high-density cosmid grid screenings were performed by Dr. Sinoula Apostolou. All subsequent confirmation of true positives was performed by the candidate. Markers in green type were of close physical proximity based on the cosmids identified (shared cosmids are indicated in green type), as too were markers in red (shared cosmids indicated in red type). All other markers identified unique sets of cosmids which are represented in black type.

another set, however the remaining markers identified unique sets of cosmids. Overall, six groups of cosmid clones served as nucleation points for walking. Subsequent to these initial screenings, the Cadherin 15 (*CDH15*) gene was mapped to the critical region (CY18A(D2) to 16qter) by Dr. Gabriel Kremmidiotis (Kremmidiotis *et al.*, 1998). He was successful in identifying a single cosmid clone (448H4) containing this gene which was also used for contig assembly and walking.

3.3.2 Restriction Enzyme Analysis and Cosmid Walking

The degree of overlap between cosmids within each of the initial six groups of clones was determined by restriction enzyme analysis, however due to the resolution of the gel system used, restriction fragments of less than 600 bp were not detected and therefore not included in the restriction map. The sCOS-1 cosmid vector has a *Not*I site flanking each end of the cloning site allowing for excision of the insert. Double digests were therefore performed involving *Not*I, in combination with a range of other restriction enzymes, including *Eco*RI, *Bam*HI, *Eag*I, *Xba*I, and *Hind*III. The enzyme combination chosen for each group of cosmids was the one producing a range of restriction fragments that allowed easy visualisation of cosmid overlaps within each contig. This kept the number of hybridisations to confirm overlaps to a minimum. Fragments of unique size within each contig were excised from agarose gels, labelled with ^{32}P, and used as a probe to detect overlapping restriction fragments (Figure 3.2).

The isolation of restriction fragments representing the extreme ends of cosmids within each of the initial six groups was essential for the subsequent identification of overlapping clones to extend each contig through re-screening of the cosmid grids. Figure 3.3 shows an example of the identification of the T3 and T7 ends of the cosmid 431F1. Three criteria were used to decide which restriction fragment was chosen to represent the end of each cosmid. The first

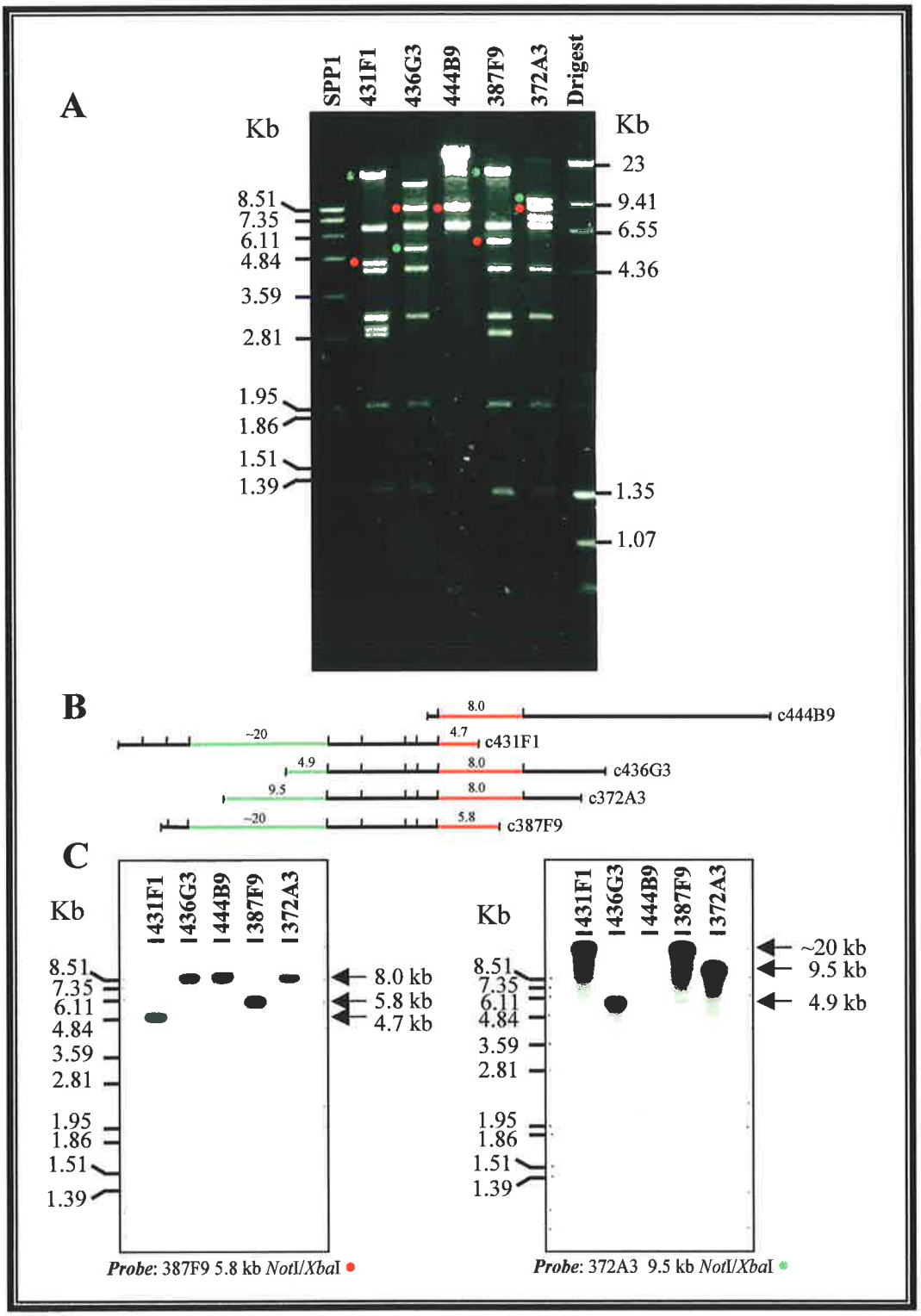


Figure 3.2: Confirmation of overlaps between cosmids identified from grid screening with the cosmid 431F1 T3 end. Most clones shared common *NotI*/*Xba*I restriction fragments (**A** and **B**), however those of differing sizes were used as hybridisation probes to determine overlapping bands (**C**). For example, the 387F9 5.8 kb fragment hybridised to the 8 kb band of cosmids 436G3, 444B9, and 372A3, and the 4.7 kb fragment of 431F1 (indicated by the red dots and fragments). The green dots and fragments represent the results of probing with the 9.5 kb fragment of cosmid 372A3.

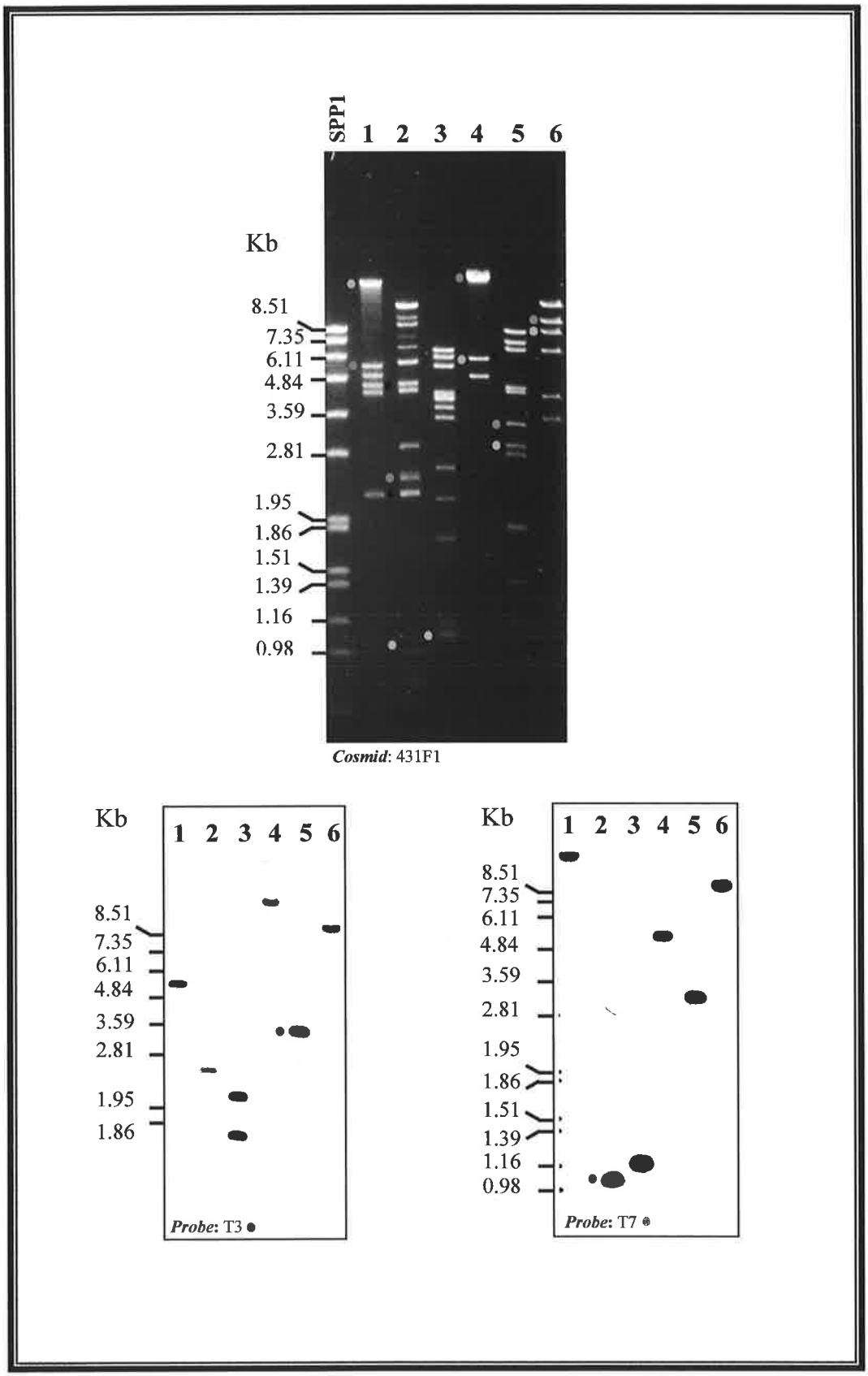


Figure 3.3: Identification of the T3 and T7 ends of the cosmid 431F1. The clone was digested with **1.** *Not*I/*Hind*III, **2.** *Not*I/*Bam*HI, **3.** *Not*I/*Bgl*II, **4.** *Not*I/*Mlu*I, **5.** *Not*I/*Hinc*II, or **6.** *Not*I/*Eco*RI. Southern membranes were probed with either T3 (red dots) or T7 (green dots) 5'-end labelled probes. The restriction fragment subsequently chosen was ideally 1 to 2 kb in size and ran independently of other fragments from the same digest. The digest eventually chosen to allow the isolation of the T3 end was *Not*I/*Hinc*II, while for the T7 end, a *Not*I/*Bam*HI digest was used. Both end fragments are illustrated by the blue dots on the corresponding autoradiographs.

was the size of the end fragment, which ideally was 1 to 2 kb in size to prevent high background on subsequent grid screenings. Secondly, the fragment should not migrate with similar sized bands on an agarose gel. This would make it difficult to cut out the correct band without contamination from similar sized fragments. The third criteria was the presence of specific binding of the T3 or T7 probes such that a single restriction fragment per digest was detected by each probe. If the probe bound to more than one restriction fragment, that particular digest was ignored. This is clearly seen in the binding of the T3 oligo probe to the *Not*I/*Bgl*II digest of cosmid 431F1 in Figure 3.3. Similarly, the *Not*I/*Bam*HI T3 end of this cosmid was also ignored because the 2.6 kb fragment was seen to co-migrate with a 2.5 kb fragment. In some instances, the restriction mapping of certain clones was sufficient to identify a suitably sized fragment representing the extreme end of the contig without the need to screen with T3 and T7.

Figure 3.4 shows the identification of the T3 or T7 ends of cosmid 431F1, demonstrating which extends the furthest in the contig and would be suitable for subsequent re-screening of the grids for walking. As can be seen, the 1 kb *Not*I/*Bam*HI T7 end detected not only the 431F1 cosmid, but also hybridised to cosmids 352A12 and 309E9, suggesting this end does not extend the furthest within this contig. However, the 3.2 kb *Not*I/*Hinc*II T3 end only hybridised to 431F1, demonstrating that this end was the correct one to use for further grid-screening.

Table 3.2 shows the cosmids identified from all end probe screenings. From the results of cosmid grid screens, clones were only ordered when the probe used to screen the grids also detected the cosmid from which it was derived. All six initial contigs were successfully extended and linked except for the contig represented by the ScDNA-C113 probe and the

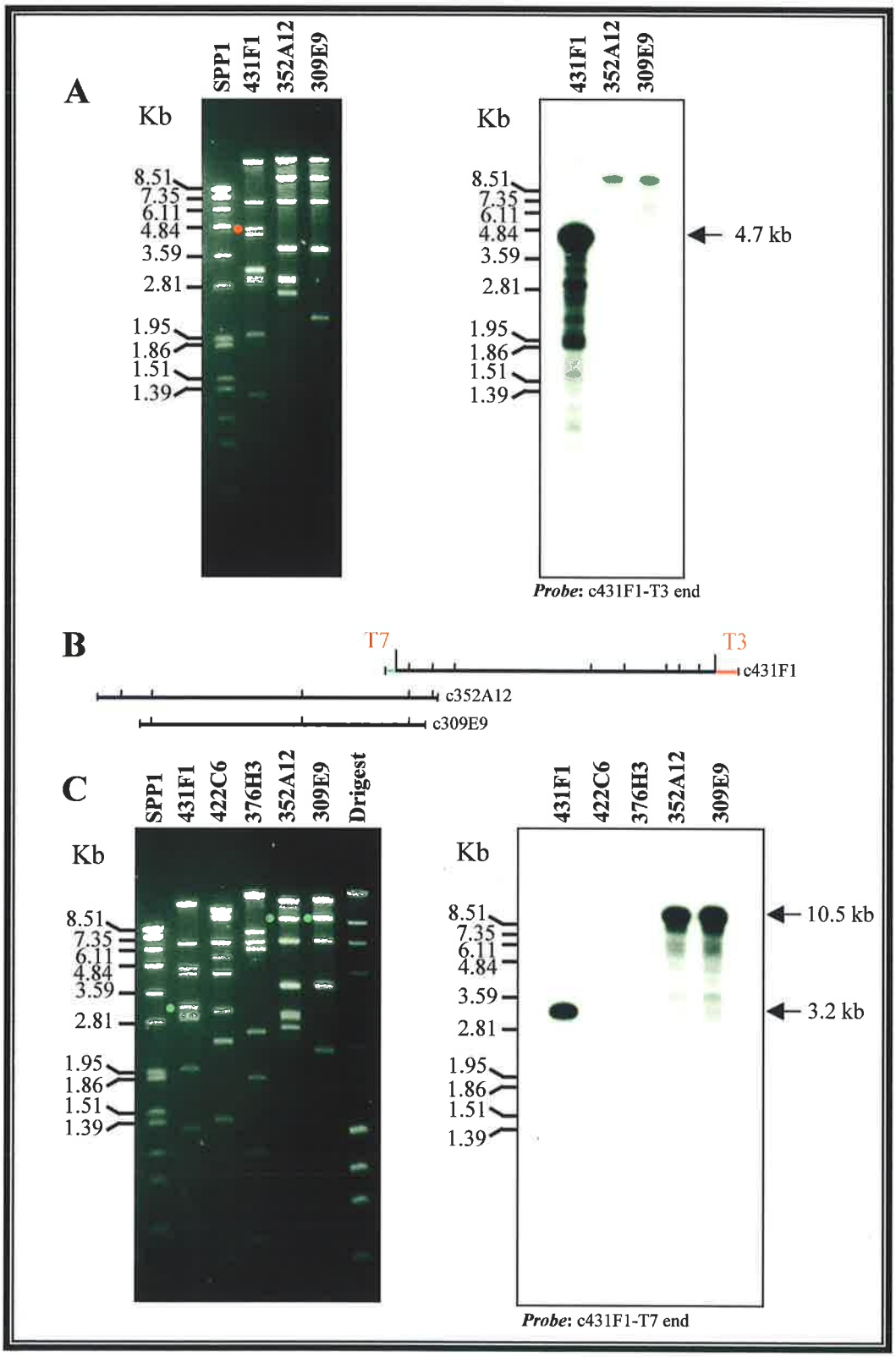


Figure 3.4: Determining whether the T3 or T7 end of the cosmid 431F1 extends furthest from within the contig established in (**B**). The hybridisation of the T3 end seen in (**A**), shows that the end fragment only detects a band originating from c431F1 (red dot) and not the overlapping cosmids. In (**C**), the T7 end probe hybridised not only to c431F1, but also to the overlapping cosmids 309E9 and 352A12 (represented by green dots). The end of choice to re-screen the cosmid grids was therefore the T3 end of c431F1.

TABLE 3.2

Cosmids, PACs and BACs Identified from End Fragment Walking

End Probe	Clones Identified			End Probe	Clones Identified		
c377F1	c377F1 c318C5	c377B2	c377B3	c334G10	c334G10 c321D3	c321G5	c438B3
c318C5	c318C5	c403B8	c393B2	c438B3	c438B3		
c403B8	403B8	393B2	c301F3	c425E4 T3	c425E4	c369E1	
c360D7	c354E12	c360D7		c369E1	c369E1		
c354E12	c354E12 c398E10	c360D7 c415C4	c397G1 c417F1	c425E4 T7	c425E4 c439G8	c308A9	c316C1
c417F1	c417F1 c402C2	c346H2 c367G6	c346H1 c367G7	c439G8	c447C5 c441C3 c429B5 c439G8	c420H1 c420G1 c360D2 c344G2	c413D8 c316C1 c308A9
c402C2	c402C2	**c444B9**					
c352A12	c352A12	c431F1		c429B5	c378H6 c429B5	c367H3	c344G2
c431F1	c431F1 c387F9	c436G3 **c444B9**	c372A3	c367H3	c367H3	c372B12	c329B4
c361H2	c361H2	c378G9		c372B12	c372B12 **c301F3** c364C4	c364B4 c432E8	c340H3 c341H12
c378G9	**c378G9**						
c435H6	c435H6	c371D5	c421E4	c348F3	c348F3 c434E1	c408G1 c365A7	c412H12
c371D5	c371D5	c344H1					
c344H1	c344H1	**c378G9**	404A4	c365A7	c365A7	c308G1	
c444C11	c444C11	c340B6	398B4	c308G1	c308G1 c408C8	c383F1	c386E7
c340B6	c340B6 **c365H1**	c317E5 c410H4	c383G9	c383F1	c383F1		
c408G10	c408G10 c383G9	c422A5 **c365H1**	c383F9	c434D8	c434D8	c396B4	
c383H6	c383H6	c427B11	c447A5	c438B3 c434D8 c383F1	p74O15	p156I14	**p168P16**
c427B11	c427B11	c358D12					
c358D12	c358D12	c335F8		c383F1 c448H4	**b561E17**		
c335F8	c335F8	c393C9		b561E17	b202C4	b276J3	b2K12
c393C9	c393C9	c334G10		b561E17	p692I10	p754F23	

Note. End clone hybridisations of the high-density cosmid/BAC/PAC filters listed in red were performed by the candidate. All other screenings were done by Dr. Sinoula Apostolou or Dr. Rachel Gibson. Verification of true positive clones with colony grid screening was performed by the candidate. Clones in green were responsible for linking the initial six mini-contigs and subsequent joining of the cosmid containing *CDH15*. c: cosmid; b: BAC; p: PAC.

448H4 cosmid identified by the *CDH15* gene.

3.3.3 BAC and PAC High-Density Grid Screening

Repeated screening of the high-density cosmid filters failed to identify overlapping clones that would extend from either end of the 448H4 cosmid, the ScDNA-C113 contig, or the established main contig. The availability of both human BAC and PAC high-density filters later in the project allowed the screening of an alternatively cloned source of genomic material to aid in closing these gaps. Screening of the PAC filters with the end fragments of the cosmids 438B3, 434D8, and 383F1 identified a single clone positive for the 434D8 and 438B3 probes, thus enabling the linkage of the C113 contig to the main contig. Screening of the BAC filters with end fragments to the 448H4 cosmid and the 383F1 end was successful in identifying one clone positive for all three probes, resulting in the linking of all contigs. Extension of the resultant contig was subsequently achieved through the further screening of either the PAC or BAC grids with clone ends, allowing for quicker walks due to the larger insert sizes as compared to cosmids. Figure 3.1B shows a representative autoradiograph from a PAC filter hybridisation and Table 3.2 shows the clones identified from such BAC and PAC filter screenings.

In some instances cosmid end probes did not hybridise to the expected clones on the cosmid grids which hindered the linking of contigs. This is highlighted with the end of c444C11, which only identified itself and c340B6, when it should have identified c365H1 also. In contrast, a walk proceeding towards this region from the opposite side, with the end of c408G10, identified a number of clones including c365H1, but not the expected c340B6. Both the c340B6 and c365H1 clones were later found to be almost identical, yet additional walks were carried out before contig linkage was identified. This may have been due to human error in the

interpretation of the grid screen results, or over-use of the membranes. Another possibility is that during high-density grid production, a variable amount of DNA from different cosmid clones may have been bound to each membrane. In addition, variations between filter sets may also have existed.

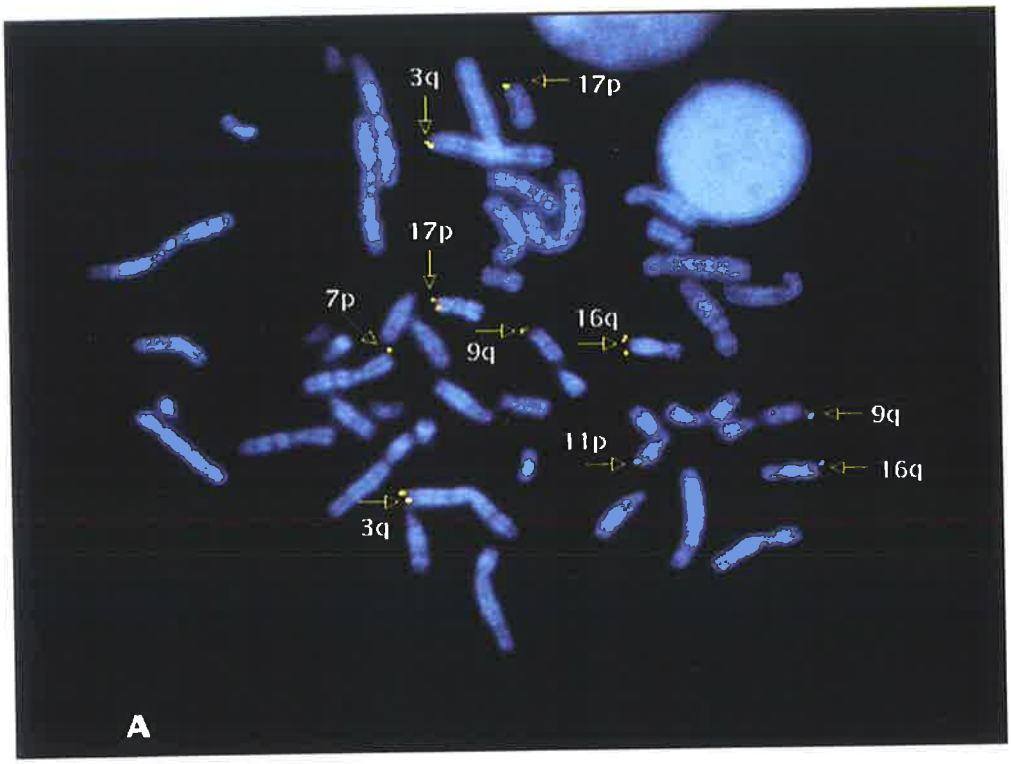
Only a single cosmid (c438A4) did not show a restriction enzyme digestion pattern consistent with overlapping clones and was therefore omitted from further analysis and not incorporated into the physical map. Through hybridisation of individual *Eco*RI fragments, the extent of the deletion in c438A4 was shown to be ~7 kb.

3.3.4 Establishment of Contig Orientation

As the microsatellite D16S303 was known to be the most telomeric marker in the 16q24.3 region (Callen *et al.*, 1995), fluorescence *in situ* hybridisation (FISH) to normal metaphase chromosomes using whole cosmids mapping in the vicinity of this marker, was used to define the telomeric limit for the contig. The cosmid 369E1 showed clear fluorescent signals at the telomere of the long arm of chromosome 16. However, this probe also gave clear signal at the telomeres of chromosomal arms 3q, 7p, 9q, 11p, and 17p (Figure 3.5A). Conversely, the cosmid 439G8, which mapped proximal to D16S303, gave fluorescent signals only at 16qter with no consistent signal detected at other telomeres (Figure 3.5B). Further work by Ms. Karen Lower (Department of Cytogenetics and Molecular Genetics, WCH) showed an identical result when the 439G8 and 369E1 cosmid ends immediately adjacent to the D16S303 marker were used as FISH probes to metaphase chromosomes. These results enabled us to establish the microsatellite marker D16S303 as the boundary of the transition from euchromatin to the subtelomeric repeats, providing a telomeric limit to the contig.

Figure 3.5

FISH analysis of metaphase chromosomes using cosmid probes. (**A**) Cosmid probe 369E1, located distal to D16S303, reveals hybridisation signals at the telomere of the long arm of chromosome 16 and at the telomeres of chromosomal arms 3q, 7p, 9q, 11p, and 17p. This indicates the presence of subtelomeric repeats within c369E1 providing an orientation and telomeric limit for the clone contig. (**B**) Cosmid probe 439G8, located immediately proximal to D16S303, gives fluorescent signals at 16q24.3 only, suggesting this marker forms the boundary between the subtelomeric repeats and euchromatin of chromosome 16q.


3q
17p
17p
7p
9q
16q
9q
11p
16q
3q
A


16q
16q
B

3.3.5 Physical Map Coverage

The proximal extent of the contig was determined by the isolation of the ends of PAC 754F23, the most proximal clone in the contig. When the T7 end was used as a hybridisation probe on Southern blots made from somatic cell hybrid DNA, the hybrid CY18A was negative for this probe, indicating the contig does not extend to the hybrid breakpoint CY18A(D2).

To confirm that the contig contained the microsatellite markers D16S3121 and D16S3026, both PCR using individual cosmid clones, and hybridisation to cosmid digests were performed (Table 3.3).

TABLE 3.3

Mapping of D16S3121 and D16S3026 to the Clone Contig

Template	PCR Results		Southern Hybridisation	
	D16S3121	D16S3026	D16S3121	D16S3026
Human	+	+		
A9	-	-		
CY18	+	+		
No Template	-	-		
c438B3	+	+	+	+
c321D3	+	+	+	+
c335F8	-	-		

Note. CY18 is a mouse/human somatic cell hybrid containing chromosome 16 as its only human complement. Hybrid A9 is the mouse background from which all somatic cell hybrids have been made. Blank spaces represent templates not used for Southern hybridisation.

This established that both markers were contained within the contig, and further, D16S3026 mapped to a 9.5 kb *Eco*RI fragment present in the overlapping cosmids 438B3 and 334G10, while D16S3121 was also present in these 2 cosmids (Figure 3.6). This suggested the distance between these 2 markers was a maximum of 15 kb. The contig therefore contained both of the

markers (D16S303 and D16S3026) defining the minimum LOH region seen in breast cancer and the region most likely to contain the *FAA* gene.

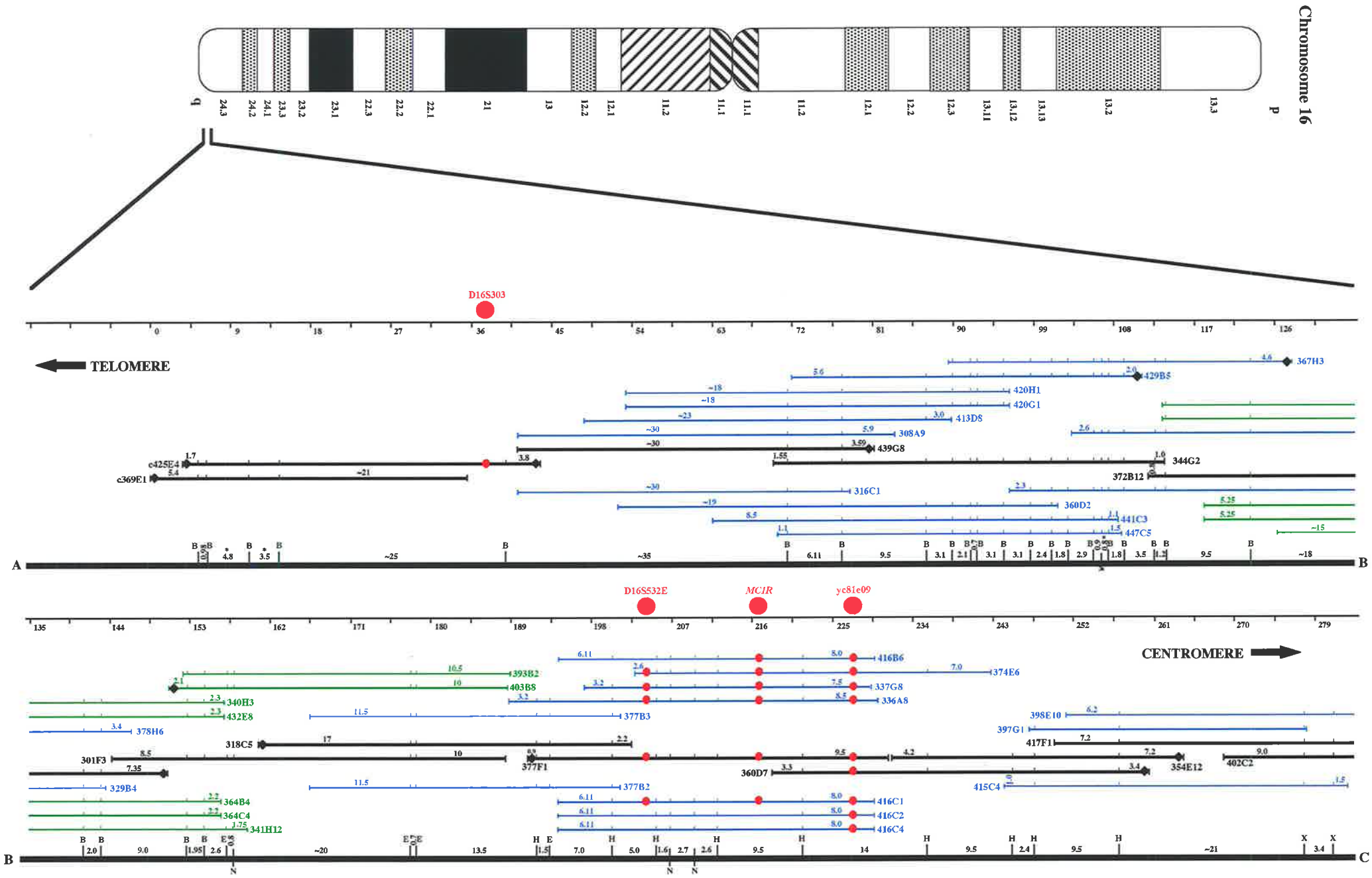
In total, 35 cosmid walks were needed with the identification of 103 cosmid clones, 35 of which form a minimum tiling path. In addition, 3 PAC screens and 2 BAC screens identified 5 PACs and 4 BACs, of which 3 contribute to the minimum tiling path. Together, these clones enabled the construction of a comprehensive physical map extending approximately 1.1 Mb from the telomere of chromosome 16q (Figure 3.6). Based on the nature of the libraries used, a five to six fold coverage of the region was expected to be achieved. While this was true for most of the contig, there were areas of both under- and over-representation. It was particularly important to confirm overlaps in the regions of under-representation and in regions where overlapping cosmids did not share common restriction fragments (overlap of 425E4 and 439G8, and overlap of 344H1 and 371D5 are examples).

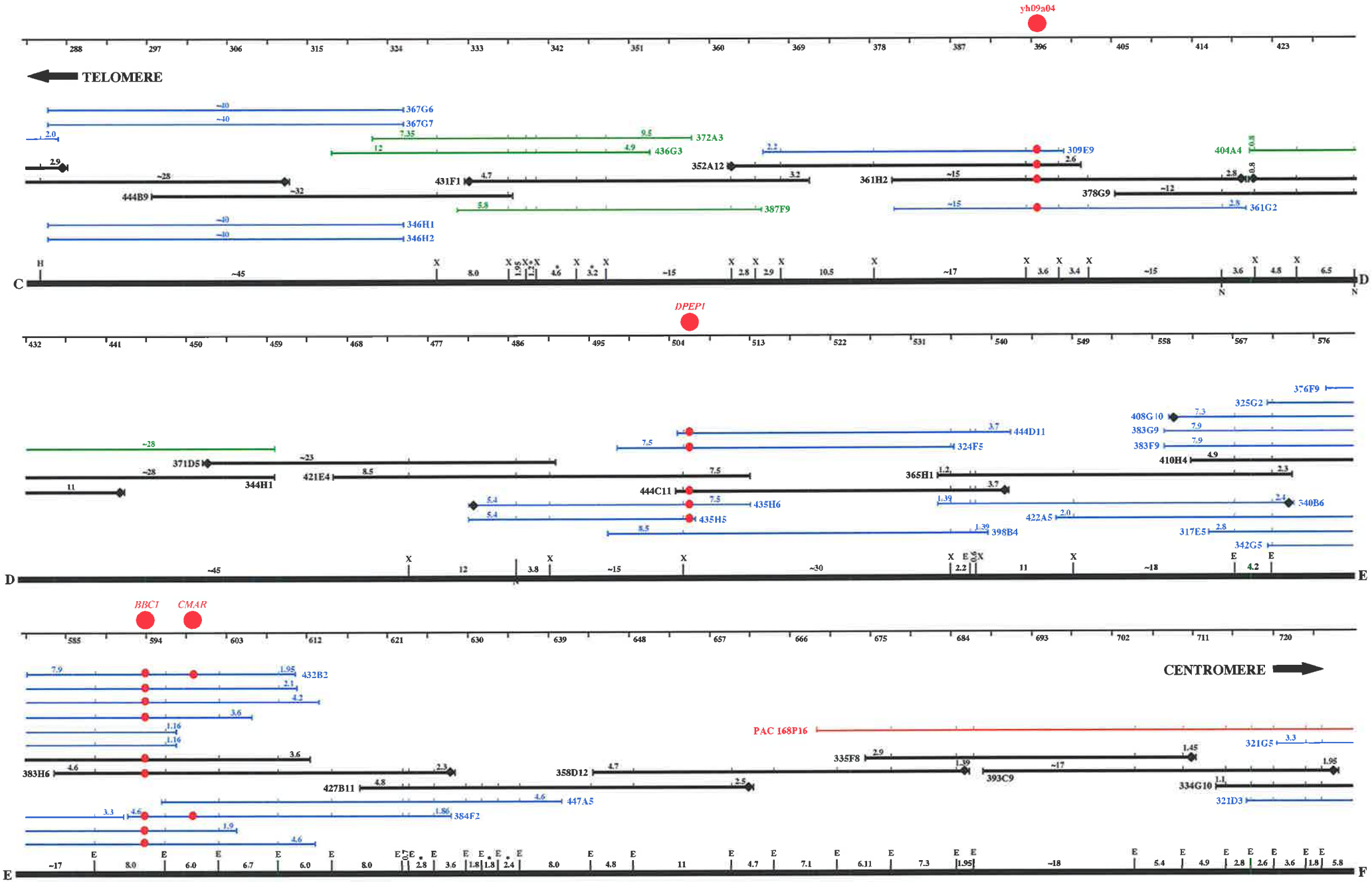
3.4 Discussion

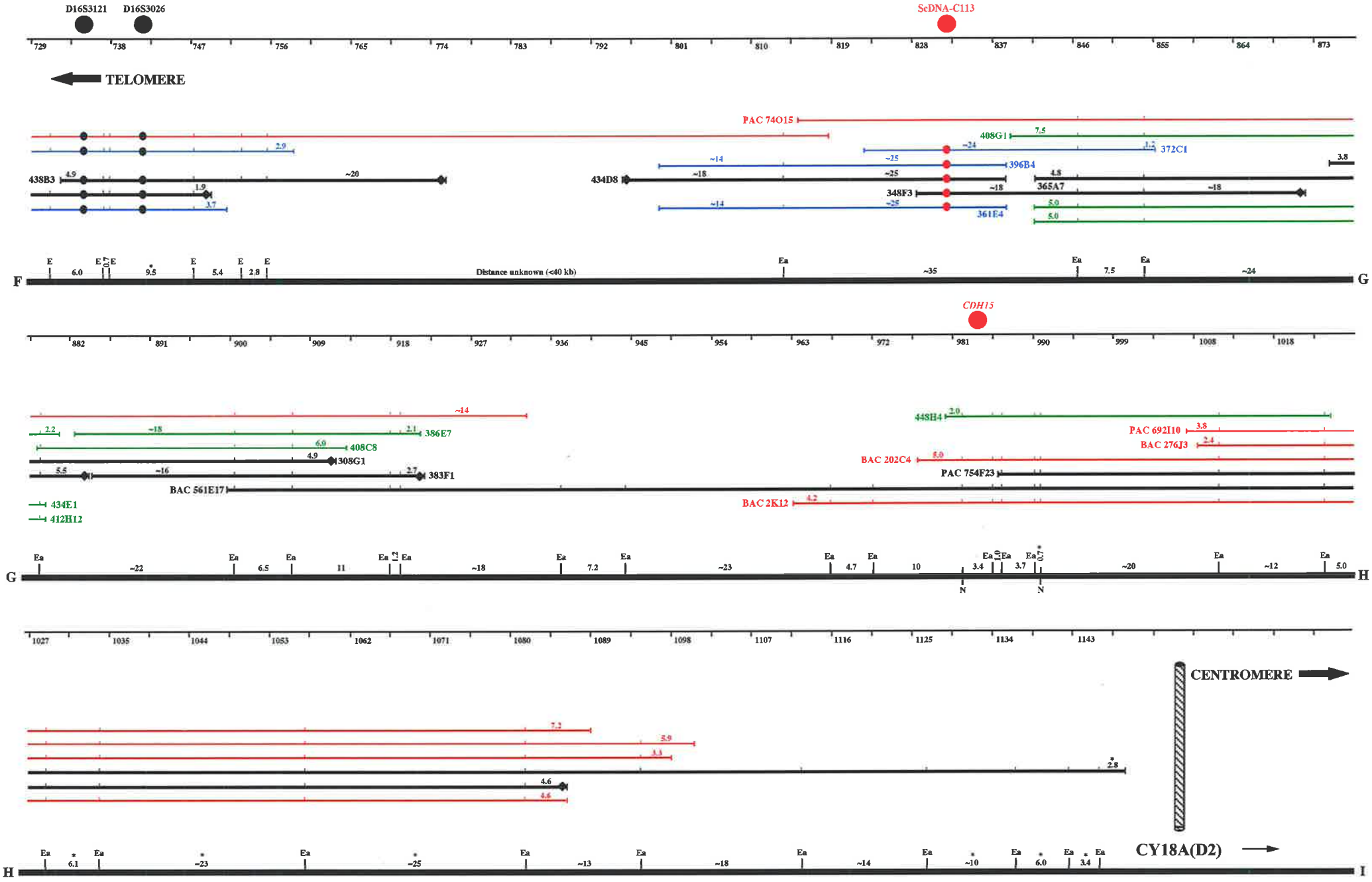
The construction of comprehensive clone based physical maps of chromosomal regions known to contain disease causing genes, assists in the subsequent identification and cloning of such genes. The chromosomal region defined by the genetic markers D16S303 and D16S3026, which lies between the mouse/human somatic cell hybrid breakpoint CY18A(D2) and the long arm telomere, has been shown by linkage analysis to contain a gene responsible for *FAA*, and through LOH studies, has also been suggested to harbour a tumour suppressor gene playing a crucial role in the development of primary breast cancer. A 1.1 Mb high-resolution clone-based physical map has successfully been constructed across this cytogenetic interval, and incorporates both the D16S303 and D16S3026 markers.

Figure 3.6

Physical map of the restricted region of loss of heterozygosity (LOH) at 16q24.3 seen in breast cancer. The map for the ~1.1 Mb contig is divided into 8 segments spanning 3 pages and extends from **A** (telomeric end) to **I** (centomeric end). At the top of each segment is a scale bar indicating distance in kilobases, while the bottom line with vertical bars indicates restriction enzyme sites. Ea: *Eag*I; E: *Eco*RI; X: *Xba*I; N: *Not*I; H: *Hind*III; B: *Bam*HI. *Not*I sites are indicated below this line. Within each restriction fragment a size in kilobases is indicated, while the sizes of cosmid end fragments are indicated above each clone. Restriction fragment sizes with an asterisk indicate the order of this fragment is not known with respect to the adjacent telomeric fragment. Clones representing a minimum tiling path in the contig are indicated by thicker black horizontal lines, as compared to other clones. Overlapping cosmids are represented by blue horizontal lines and overlapping BACs and PACs by red lines. Green cosmids represent those clones characterised by Ms. Joanna Crawford, Dr. Gabriel Kremmidiotis, or Dr. Ram Seshadri. Red closed circles represent the locations of the markers used to initially identify cosmid clones mapping to the 16q24.3 region. The closed black circles indicate the location of the D16S3121 and D16S3026 microsatellite markers which define the centromeric limit for the minimum region of LOH. Black diamonds indicate the ends of clones used to re-screen the cosmid, BAC, and PAC high-density filters for contig extension. The CY18A(D2) breakpoint is indicated by the vertical hashed bar in segment 8. The distance of this breakpoint to the end of PAC754F23 is not known.


Chromosome 16
q
p
24.3
24.2
24.1
23.3
23.2
23.1
22.3
22.2
22.1
21
13
12.2
12.1
11.2
11.1
11.1
11.2
12.1
12.2
12.3
13.11
13.12
13.13
13.2
13.3
D16S303
TELOMERE
CENTROMERE
A
B
C
D16S532E
MC1R
yc81e09
367H3
429B5
420H1
420G1
413D8
308A9
439G8
344G2
372B12
c425E4
c369E1
316C1
360D2
441C3
447C5
416B6
374E6
337G8
336A8
393B2
403B8
340H3
432E8
377B3
378H6
318C5
301F3
377F1
360D7
398E10
397G1
417F1
354E12
402C2
415C4
329B4
377B2
364B4
364C4
341H12
416C1
416C2
416C4


yh09a04
TELOMERE
367G6
367G7
372A3
436G3
309E9
404A4
352A12
431F1
361H2
444B9
378G9
387F9
361G2
346H1
346H2
C
D
DPEP1
376F9
325G2
408G10
383G9
383F9
410H4
444D11
324F5
371D5
344H1
421E4
365H1
444C11
435H6
435H5
340B6
422A5
317E5
342G5
398B4
E
BBC1
CMAR
CENTROMERE
432B2
PAC 168P16
321G5
335F8
383H6
358D12
427B11
393C9
334G10
321D3
447A5
384F2
F


D16S3121
D16S3026
ScDNA-C113
729
738
747
756
765
774
783
792
801
810
819
828
837
846
855
864
873
TELOMERE
PAC 74O15
408G1
372C1
396B4
438B3
434D8
348F3
365A7
361E4
Distance unknown (<40 kb)
F
G
CDH15
882
891
900
909
918
927
936
945
954
963
972
981
990
999
1008
1018
386E7
408C8
308G1
383F1
BAC 561E17
434E1
412H12
44SH4
PAC 692I10
BAC 276J3
BAC 202C4
PAC 754F23
BAC 2K12
H
1027
1035
1044
1053
1062
1071
1080
1089
1098
1107
1116
1125
1134
1143
CENTROMERE
CY18A(D2)
I

3.4.1 Physical Map Integrity

The flow-sorted chromosome 16 cosmid library used in the map construction proved to be an extremely reliable source of cloned material and has been shown to be effective in other mapping studies (Giles *et al.*, 1997; Sood *et al.*, 1997). Only three gaps were encountered which could not be closed through successive cosmid grid screenings, however each of these gaps was covered through the identification of either BAC or PAC clones. Similar gaps in contig construction have also been reported by Sood *et al.* (1997). In this instance, gaps were also successfully covered by the isolation of a BAC clone, which suggests that low copy number may be important in propagating some genomic regions using bacterial systems.

Cosmids have been shown to be relatively stable cloning vehicles, a fact supported by this map which established that only one clone (c438A4) showed an altered restriction enzyme digestion pattern relative to other overlapping cosmids. This region, containing both the D16S3121 and D16S3026 markers, had several stable cosmids including c321G5, c321D3, c438B3 and c334G10, as well as one PAC clone (PAC 168P16). The 438A4 clone was later shown to contain a ~7 kb deletion within an *Eco*RI restriction fragment and was therefore omitted from the physical map.

In most regions of the map, the expected density of cosmid clones was achieved. However regions of under-representation were seen, particularly proximal walks from the initial 317E5 contig. In this instance, a PAC clone (PAC 168P16) was identified that spanned this region of low cosmid representation, and was shown to contain *Eco*RI fragments identical to those seen in the overlapping cosmid clones. Also, an individual cosmid clone (c427B11) from this region was used as a FISH probe to metaphase chromosomes and as a hybridisation probe on Southern membranes containing somatic cell hybrid DNA, to confirm it mapped to the correct

region. Although the under-representation of cosmid clones in particular regions may reflect the nature of the DNA content of that region, other factors such as the probe used for the cosmid walking need to be considered. Probes that are high in repetitive elements, such as *Alu* sequences, may have very little unique sequence left to bind once the repeats have been quenched through pre-reassociation. Another factor was that cosmid high-density grids that had been used frequently showed reduced signals increasing the likelihood that positive clones were not detected. In addition, the distribution of *Sau*3A sites within regions of low cosmid representation may be a contributing factor. A lack of these sites may lead to restriction fragments too large to clone such that this region will not be represented as a cosmid clone.

3.4.2 Integration of Physical and Genetic Maps

The microsatellite markers D16S3121 and D16S3026 were mapped to the same cosmid clone by PCR and Southern hybridisation and are a maximium of 15 kb apart. A comparison with the genetic linkage data of Dib *et al* (1996) however, indicated that the likely order of genetic markers was determined to be S413-S3026-S3023-S3121-TEL, with a sex-averaged distance of ~2.3 cM between D16S3026 and D16S3023, and 0.1 cM between D16S3023 and D16S3121. Previous physical mapping studies (Callen *et al.*, 1995) had determined that D16S3023 and D16S413 map to the cytogenetic interval proximal to the D16S3121 and D16S3026 markers, thus highlighting an error in the genetic linkage data. Based on the established contig, along with previous physical mapping data, the most likely order of markers is S413-S3023-S3026-S3121-TEL. A reversal of the order of S3026 and S3023 on the linkage map of Dib *et al.* (1996) would therefore indicate a sex-averaged distance of 0.1 cM between D16S3026 and D16S3121, consistent with the results established from the contig construction.

A comparison between the genetic and physical distances between D16S303 and D16S3121

could not be made as D16S303 is not a Genethon marker and has therefore not been included in the comprehensive linkage maps established by these groups. It does appear however, that D16S303 is the most telomeric marker on the long arm of chromosome 16 (Kozman *et al.*, 1993). FISH to metaphase chromosomes using a cosmid mapping immediately telomeric to this AC repeat, identified DNA homology with the telomeres of a number of other chromosomes, indicating the presence of subtelomeric repeats within this cosmid. In contrast, FISH with a cosmid immediately telomeric to D16S303 showed unique signal at 16q24.3 only. This suggests the marker lies at the boundary between the subtelomeric repeats of the long arm of chromosome 16 and the start of the euchromatin. Analysis of the short arm of chromosome 16 has also indicated that an AC repeat may border the divergence of the telomeric region to the euchromatic region (Wilkie and Higgs, 1992). The presence of genes at a high density starting just a few kilobases proximal to the subtelomeric repeats has been identified at this and other sites (van Deutekom *et al.*, 1996; Flint *et al.*, 1997).

The depth of coverage across the majority of the established contig now provides an essential resource for the sequencing of this chromosomal region and for the identification of additional microsatellite markers that may assist in the further refinement of the minimum LOH region seen in breast cancer. Given the few genes and ESTs mapped to the contig, the clones also provide templates for the identification of new transcripts that lie within this region, which will be candidates for the *FAA* gene and a tumour suppressor gene involved in breast cancer. Chapter 4 describes the use of exon trapping specifically for this purpose, together with the analysis of partial cosmid sequence and cDNA database screening.

Chapter 4

Identification of Transcribed Sequences at 16q24.3

Table of Contents

	Page
4.1 Introduction	**120**
4.2 Methods	**123**
4.2.1 Digestion of Cosmid DNA	123
4.2.2 Phenol/Chloroform Extraction of Digested DNA	123
4.2.3 Preparation of the Exon Trapping Vector	124
4.2.4 Subcloning of Cosmids into the pSPL3B-CAM Vector	124
4.2.4.1 Ligations	124
4.2.4.2 Transformations	124
4.2.4.3 Cell Scraping and DNA Isolation	125
4.2.5 Preparation of COS-7 Cells	125
4.2.6 Transfection of COS-7 Cells	125
4.2.7 Isolation of Cytoplasmic RNA from COS-7 Cells	126
4.2.8 Reverse Transcription	127
4.2.9 Primary PCR Amplification	127
4.2.10 Secondary PCR Amplification	128
4.2.11 Uracil DNA Glycosylase Cloning of Secondary PCR Products	128
4.2.12 Exon Confirmation	129
4.2.12.1 Colony PCR	129
4.2.12.2 DNA Isolation and Sequence Analysis	129
4.2.12.3 Physical Mapping of Trapped Products	130
4.2.12.4 Database Homology Searches	130
4.2.13 Analysis of the Human Gene Map	130
4.2.14 Single-Pass Cosmid Sequencing	131
4.3 Results	**131**
4.3.1 Exon Trapping	131
4.3.2 Analyses of Trapped Exon Sequences	136
4.3.3 Cloning of the *FAA* gene	141
4.3.3.1 Identification of cDNA Clones Homologous to Trapped Exons	141
4.3.3.2 Identification of Full-Length cDNA Sequence	142
4.3.3.3 Mutation Analysis	143
4.3.4 Analysis of the Human Gene Map	144
4.3.5 Analysis of Single-Pass Cosmid Sequencing	146
4.4 Discussion	**147**
4.4.1 Exon Trapping	147
4.4.2 Comparison of Transcript Identification Methods	153
4.4.3 Candidate Breast Cancer Tumour Suppressor Genes	154

4.1 Introduction

At the start of this phase of the project international efforts in the physical mapping of cDNA clones (ESTs) were not yet underway. This necessitated the identification of transcribed sequences by other means. At this time, the two most commonly used procedures for the identification of transcribed sequences within relatively large genomic intervals were direct cDNA selection and exon amplification (discussed in detail in 1.2.3.4 and 1.2.3.5). The latter procedure was chosen because it is not limited by the tissue or developmental expression of genes. Instead, the expression of cloned genomic DNAs is driven by a promoter in tissue culture cells, which allows for coding sequences to be identified without prior knowledge of their expression profile. This therefore circumvents the need to analyse multiple cDNA libraries as with cDNA selection. Exon trapping using the pSPL3B vector relies on the presence of a functional splice acceptor and donor site flanking an exon. As a result, the terminal 5' and 3' exons of a gene are not selected, and similarly, transcripts that contain less than three exons will not be identified. However, as a general procedure for identifying coding sequences within large genomic regions, this technique has been shown to be very effective (Burn *et al.*, 1996; Chen *et al.*, 1996*a*; Couch *et al.*, 1996*b*; Church *et al.*, 1997).

The construction of a detailed physical map of chromosome 16q24.3, between the markers D16S3026 and D16S303, encompasses the region most likely to contain the *FAA* gene and a breast cancer tumour suppressor gene. The majority of this map is comprised of cosmids, clones forming a minimum tiling path in the contig were therefore used as substrates in exon trapping experiments to help identify new transcripts mapping to the critical region. The outline of the exon trapping procedure is shown in Figure 4.1 and described in detail in the methods section. However a modified version of the previously reported pSPL3B vector was

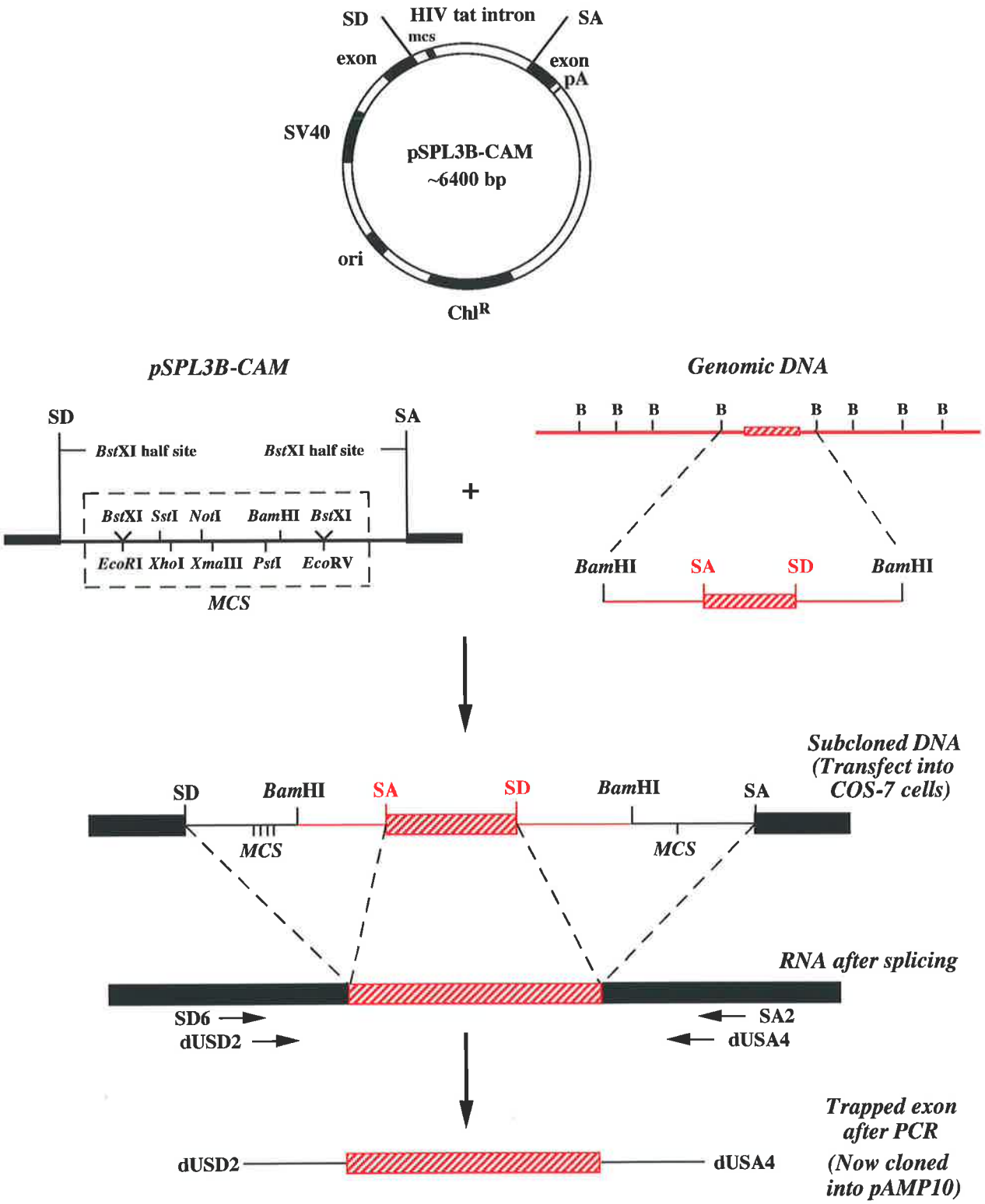


Figure 4.1: Schematic diagram of the exon trapping procedure. All vector sequences are in black while human genomic DNA sequences are in red. The pSPL3B-CAM vector contains a multiple cloning site within an intron of the HIV-1 *tat* gene. This intron is bordered by exons of the rabbit B-globin gene which provide a splice donor (SD) and splice acceptor (SA) site. Cosmid DNA is digested at specific restriction sites (eg *Bam*HI) such that a single exon (red box) may be contained within a *Bam*HI fragment. The exon is flanked by a SA and SD site. After subcloning cosmid DNA into pSPL3B-CAM, DNA is isolated and transfected into COS-7 cells. Cytoplasmic RNA is isolated for RNA-based PCR analysis. After generation of cDNA using a vector specific primer, a first round of PCR is performed using the outside primers SD6 and SA2. A second round of PCR is then performed using the nested primers dUSD2 and dUSA4. These primers allow uracil DNA glycosylase cloning of the PCR products. Trapped exons are recognised following gel electrophoresis of the PCR products.

used (Burn *et al.*, 1995). This new vector, pSPL3B-CAM, has had the ampicillin resistance gene replaced with a gene conferring chloramphenicol resistance. This assists in the initial step of sub-cloning cosmid DNA restriction enzyme fragments into the trapping vector by allowing selection against contaminating re-ligated ampicillin resistant cosmid vector clones. It has also been noted that increasing the complexity of the DNA used for each trapping experiment dramatically decreases the exon recovery (Yaspo *et al.*, 1995). For this reason, the maximum number of cosmids used in any one trapping experiment was three, with the majority of traps involving a single cosmid.

To determine the authenticity of the products trapped, a comparison of the sequences of these products can be made to nucleotide sequences present in the database of ESTs (dbEST). Significant homology to an EST is often sufficient initial evidence for mapping the corresponding cDNA clone to the region from where the exon was trapped. Otherwise, expression of the exons needs to be confirmed, by for example, linking to adjacent exons by RT-PCR, or screening cDNA libraries. Some of these procedures will be discussed in detail in chapters 5 and 6.

The publication of the UniGene collection and the development of the Human Gene Map (Schuler *et al.*, 1996), coupled with single-pass sequencing of selected cosmids from within the physical map contig (performed by Los Alamos National Laboratories) during the course of the exon trapping experiments, provided alternative ways to identify cDNA clones mapping to the cosmid contig. The main objective of all these methods was to allow the construction of a transcription map of the 16q24.3 region which would provide a source of candidate genes that may play a role in *FAA* and breast cancer.

4.2 Methods

A number of procedures were only performed in this chapter and chapter 7, and are therefore mentioned in detail in these chapters. More general techniques are described in detail in chapter 2, however slight modifications are referred to when appropriate. Exon trapping procedures were adapted from those used by Church *et al.* (1994).

4.2.1 Digestion of Cosmid DNA

Five micrograms of cosmid DNA was double digested in a volume of 50 µl with *Bam*HI and *Bgl*II, or with *Pst*I alone, in the presence of BSA at a concentration of 100 µg/ml. Digestion was at 37°C overnight using 15 units of *Bam*HI and *Pst*I enzymes and 30 units of *Bgl*II (the double digest was performed in the restriction buffer specific for *Bam*HI which was not optimal for *Bgl*II, therefore double the amount of *Bgl*II enzyme was used).

4.2.2 Phenol/Chloroform Extraction of Digested DNA

An aliquot of 250 ng from the overnight digest was kept aside for later analysis. A further 150 µl of sterile water was added to the remaining digest and mixed. An equal volume of phenol was then added to each digest and mixed. After centrifugation at 13,000 rpm for 7 minutes, the aqueous layer was transferred to a fresh eppendorf tube. An equal volume of chloroform:IAA (24:1) was added, the solution mixed, and again the samples were spun at 13,000 rpm for 7 minutes, with the aqueous layer transferred to a clean eppendorf. The DNA was precipitated by the addition of one-tenth the volume of 3 M NaAc (pH 5.2) and 2 volumes of 100% ethanol. After incubating at -80°C for 1 hour, the samples were spun at 13,000 rpm for 30 minutes. DNA pellets were washed with 1 ml of 70% ethanol, air-dried for 15 minutes, and resuspended in 10 µl of sterile water. An aliquot of 2 µl was run on a 0.8% agarose gel with

the original digest aliquot being run alongside to determine the efficiency of the purification procedure.

4.2.3 Preparation of the Exon Trapping Vector

A total of 10 µg of the exon trapping vector pSPL3B-CAM was digested in a 50 µl volume overnight at 37°C with 15 units of either *Pst*I or *Bam*HI. The following day, 50 µl of water was added to the digests and they were cleaned with QIAquick columns according to manufacturers instructions (2.2.16.1). The cleaned vector digests were then treated with 2.5 units of calf intestinal alkaline phosphatase (CIAP) in a volume of 100 µl using the buffer supplied. The dephosphorylation reactions were carried out at 37°C for 1 hour. Following this, the samples were again cleaned by QIAquick columns and the DNA quantitated by spectrophotometry (2.2.6).

4.2.4 Subcloning of Cosmids into the pSPL3B-CAM Vector

4.2.4.1 Ligations

Digested and dephosphorylated pSPL3B-CAM DNA was ligated to digested cosmid DNA essentially as described in 2.2.17.1. Reactions involved the use of 50 ng of vector DNA in combination with 100 ng of cosmid DNA. An important control was the ligation of 50 ng of vector with no cosmid DNA present.

4.2.4.2 Transformations

The method was the same as that described in 2.2.17.4. However, instead of plating out 200 µl of the transformed cells, the cells were spun down and the pellet resuspended in 100 µl of L-Broth. The entire 100 µl was then plated onto an L-Chloramphenicol plate and incubated at

37°C overnight. No X-Gal or IPTG were spread with the cells. Subcloning of the cosmids into the trapping vector was deemed successful if the number of colonies on the experimental plate exceeded the number of colonies on the vector-only ligation control plate by 10 fold.

4.2.4.3 Cell Scraping and DNA Isolation

Three millilitres of L-Broth plus chloramphenicol (34 μg/ml) was aliquoted onto each successful ligation plate, and the cells were scraped off into the solution with a sterile glass spreader. The solution was then transferred to a sterile 10 ml tube and DNA was isolated using a Qiagen Tip-20 (2.2.1.3).

4.2.5 Preparation of COS-7 Cells

The COS-7 cells were grown in supplemented DMEM media. The cells were passaged one day prior to transfection by placing $8X10^5$ to $16X10^5$ cells into 10 ml of supplemented DMEM in 75 cm^2 tissue culture flasks. This resulted in 40 to 60% confluence the next day.

4.2.6 Transfection of COS-7 Cells

Flasks were inspected for correct confluence and pSPL3B-CAM subcloned cosmid DNA was transfected into the cells using the LIPOFECTACE reagent. DNA from 1 to 3 subcloned cosmids was added to any one flask of cells per transfection. Briefly, in a sterile 10 ml tube, 500 μl of Opti-MEM 1 (OMI) medium was added to 1 μg of each subcloned cosmid to be transfected. In a 1.5 ml eppendorf tube, 500 μl of OMI was added to 30 μl of LipofectACE and left at room temperature for 5 minutes. This was then added to the 10 ml tube above and incubated for a further 10 minutes at room temperature. The supplemented DMEM media was removed from the COS-7 cells and 10 ml of OMI was added to each flask. The cells were then incubated for 5 minutes at 37°C in a 5% CO_2 incubator. Four millilitres of OMI was then

added to each 10 ml tube containing the DNA/LipofectACE/OMI mix. The 10 ml of OMI media from each flask of cells was removed, and the combined media above was then added to the appropriate flask of cells. The cells were then incubated overnight at 37°C in a 5% CO_2 incubator. The next day, the lipid-DNA complexes were poured off from the COS-7 cells, and 10 ml of supplemented DMEM was added. The cells were incubated for a further 48 hours at 37°C in a 5% CO_2 incubator.

4.2.7 Isolation of Cytoplasmic RNA from COS-7 Cells

All RNA isolation procedures were done using RNAse free conditions as described in 2.2.4. The supplemented DMEM was removed from each flask and 10 ml of ice-cold PBS was added and the flask was left on ice. The PBS was then removed and the wash step was repeated 2 more times. Ten millilitres of PBS was then added to the washed cells and they were removed from the culture flask using a sterile plastic scraper (Costar). The cells were then transferred into a sterile 10 ml tube, which was placed on ice. The cells were spun at 1,200 rpm for 5 minutes and the pellet was resuspended in 400 µl of TKM and 1 µl of RNAsin by gently tapping the tube. After a 5 minute incubation on ice, 20 µl of 10% Triton X-100 was then added, followed by a further 5 minute incubation on ice. The mix was then spun at 1,200 rpm for 5 minutes and the supernatant transferred to a 1.5 ml eppendorf tube on ice. Thirty microlitres of 5% SDS was then added. This mix was phenol/chloroform extracted twice as described in 4.2.2. Following this, the RNA was precipitated by the addition of 15 µl of 5 M NaCl and 900 µl of ethanol. This was held at -70°C for 30 minutes. The precipitated RNA was spun at 13,000 rpm for 20 minutes and the pellet was washed in 1 ml of 75% ethanol. The pellet was air-dried for 10 to 20 minutes and resuspended in 50 µl of DEPC-treated water. The RNA was quantitated by spectrophotometry and a sample run on a 0.8% agarose gel to check

the integrity of the RNA isolated.

4.2.8 Reverse Transcription

Reverse transcription of RNA isolated from COS-7 cells was as described in 2.2.15.1. However, the primer used for reverse transcription was the pSPL3B-CAM specific primer, SA2. The sequence of this primer is shown in Table 2.1. One microlitre of a 20 μM stock of this primer was used. After a 30 minute incubation at 42°C with the Superscript enzyme, the samples were heated to 55°C for 10 minutes. One microlitre (2 units) of RNAseH was then added and the tubes were incubated for a further 10 minutes at 55°C. Products were stored at -20°C whilst not in use.

4.2.9 Primary PCR Amplification

Eight microlitres of the reverse transcription reaction were incubated with 4 μl of 10X PCR buffer, 0.8 μl of 10 mM dNTPs, 1.2 μl of 50 mM $MgCl_2$, 2 μl each of a 20 μM stock of oligonucleotides SA2 and SD6 (primer sequences are shown in Table 2.1), and made up to a volume of 39.5 μl with sterile water. The components were mixed and a drop of paraffin oil was added. After heating the tubes at 94°C for 5 minutes, the tubes were held at 80°C while 0.5 μl (2.5 units) of *Taq* DNA polymerase was added. The tubes were then incubated at 94°C for 1 minute, 60°C for 1 minute, 72°C for 5 minutes for 6 cycles, followed by a 10 minute incubation at 72°C. To prevent the amplification of false-positive (vector-only) products in subsequent PCR reactions, 2.5 μl (25 units) of *Bst*XI was added to the PCR reaction and incubated overnight at 55°C. The following day, a further 0.5 μl (5 units) of *Bst*XI was added to each tube and incubated for a further 2 hours at 55°C.

4.2.10 Secondary PCR Amplification

Five microlitres of the first PCR reaction were mixed with 4.5 µl of 10X PCR buffer, 1 µl of 10 mM dNTPs, 1.5 µl of 50 mM $MgCl_2$, 1 µl each of a 20 µM stock of oligonucleotides dUSD2 and dUSA4, and the reactions made up to a volume of 49.5 µl with water. The contents were mixed and one drop of paraffin oil was added. After heating the tubes at 94°C for 5 minutes, the tubes were held at 80°C while 0.5 µl (2.5 units) of *Taq* DNA polymerase was added. The tubes were then incubated at 94°C for 1 minute, 60°C for 1 minute, 72°C for 3 minutes for 30 cycles, followed by a 10 minute incubation at 72°C. Five microlitres were then run on a 2.5% agarose gel to examine the potential number of trapped exons amplified.

Oligonucleotides SD2 and SA4 were modified to include deoxy-UMP residues at their 5' ends (see Table 2.1). This was to allow cloning of PCR products using the CloneAMP pAMP10 kit.

4.2.11 Uracil DNA Glycosylase Cloning of Secondary PCR Products

After secondary amplification, the PCR products contain the dUMP-containing sequence at their 5' termini. Treatment with uracil DNA glycosylase renders dUMP residues abasic, and unable to base-pair, resulting in 3' protruding termini which can then be ligated to complementary pAMP10 vector ends.

Two microlitres of the secondary PCR reaction were mixed with 2 µl of pAMP10 cloning vector, 15 µl of 1X pAMP10 annealing buffer, and 1 µl (1 unit) of uracil DNA glycosylase. The contents were mixed and incubated at 37°C for 30 minutes. Five microlitres of the reaction was then transformed into 100 µl of competent XL-1 Blue cells (2.2.17.4). One-tenth of the transformation was plated on L-Ampicillin plates and incubated overnight at 37°C.

4.2.12 Exon Confirmation

Colony PCR using pAMP10 specific primers was used to determine the size of the products trapped. Clones were grouped according to size and a representative clone from each group was chosen for further analysis. All clones with an insert size of 409 bp (determined by colony PCR) were omitted from further analysis as they represented products generated from vector-only clones containing no insert.

4.2.12.1 Colony PCR

Recombinant colonies were picked with a sterile pipette tip, streaked on a master plate for later use, and the tip placed in a PCR tube containing a drop of oil. The colony PCR procedure was identical to that described in 2.2.14.2, the only variation was the number of colonies analysed and the primers used in the PCR. For an exon trap involving 1 cosmid, 24 colonies were analysed, for a 2 cosmid trap, up to 90 colonies were analysed, and for a 3 cosmid trap, up to 135 colonies were analysed. The primers used for the colony PCR were PucF and PucR with their sequence shown in Table 2.1.

4.2.12.2 DNA Isolation and Sequence Analysis

Clones selected for further analysis were grown in 20 ml of L-Broth plus ampicillin (100 μg/ml) and DNA was isolated using Qiagen Tip-20 columns (2.2.1.3). The trapped products were sequenced using DyePrimer F and R sequencing kits (2.2.18.2). As the amplified trapped products were cloned non-directionally into pAMP10, the DNA sequence obtained was aligned in the sense orientation, and the vector sequences were removed. This left sequence corresponding to the trapped product alone.

4.2.12.3 Physical Mapping of Trapped Products

The inserts of trapped products were removed from their vector with a *Not*I/*Sal*I double digest. The digests were run on a 2.5% agarose gel and inserts were cut directly out of the agarose, labelled with ^{32}P, and used as probes on Southern blots containing cosmid DNA from which the product was originally trapped.

4.2.12.4 Database Homology Searches

The BLASTN program (Altschul *et al.*, 1997) was used to search for nucleotide sequence homology between the trapped products and sequences in the GenBank non-redundant and EST databases (http://www.ncbi.nlm.nih.gov/index.html). In the majority of cases, the critical P value indicating significant homology was taken to be $\leq 10^{-5}$ (1.0e-5). To reveal any open reading frames within the sequences of the trapped products, the Applied Biosystems SeqEd (version 1.0.3) software was used.

4.2.13 Analysis of the Human Gene Map

The Human Gene Map published in October 1996 was analysed with the microsatellite marker D16S520 and the telomere of the long arm of chromosome 16 defining the genetic interval to be searched (http://www.ncbi.nlm.nih.gov/index.html). The D16S520 marker has been physically mapped (Callen *et al.*, 1995), and lies 2 cytogenetic intervals proximal to the *FAA* and critical LOH region (see Figure 1.3). Identified cDNA clones not homologous to exons already trapped or to cDNA clones identified through cosmid sequencing, were mapped by Southern hybridisation to DNA from cosmids within the contig or to somatic cell hybrid DNA (2.2.13.1). In some cases, PCR amplification from somatic cell hybrid DNA using primers corresponding to the cDNA clone was done (2.2.14.1). The availability of the GeneMap'98 (http://www.ncbi.nlm.nih.gov/genemap/) towards the end of this study provided additional

mapped cDNAs, however they are not included in the final results.

4.2.14 Single-Pass Cosmid Sequencing

Nineteen selected cosmid clones from the 16q24.3 contig were partially sequenced by a shotgun approach at the Los Alamos National Laboratory essentially as described in Venter *et al* (1996). These clones are indicated in Figure 4.4. In the majority of cases, between 15 to 20 kb of non-contiguous sequence was obtained, and was used to search for homology to ESTs present in dbEST using the BLAST algorithm. Those ESTs not already identified through exon trapping and Human Gene Map searches were analysed further. Corresponding cDNA clones were mapped by Southern hybridisation to the cosmid contig if they appeared to be legitimate cDNAs.

4.3 Results

4.3.1 Exon Trapping

In total, 26 cosmids representing a minimum tiling path in the physical map contig of 16q24.3, were chosen for exon trapping experiments, with the results summarised in Table 4.1 and the cosmids highlighted in Figure 4.4. A vital step in the initial stages of the procedure was the complete sub-cloning of all *Pst*I or *Bam*HI/*Bgl*II restiction fragments from each selected cosmid into the exon trapping vector. Subsequent steps in the procedure were not performed until the majority of the cosmid was shown to be subcloned (Figure 4.2). For each transfection of COS-7 cells, a maximum of 3 separate sub-cloned cosmids was used, however in most cases, a single cosmid only was used. A total of 891 individual clones were analysed by colony PCR from the 26 cosmids used for exon trapping, with 529 (59%) containing an insert of size equal to that of a "vector-only" control product. These were subsequently omitted from further

TABLE 4.1

Results of Exon Trapping Experiments

Number of cosmids examined	26
Clones analysed by colony PCR	891
Vector only, no insert clones	529 (59%)
Number of trapped clones with inserts	362
Number of redundant clones	266
Unique trapped clones	96
Repetitive clones	9
E.coli contamination	1
Aberrant splicing of vector	15
Total trapped exons	71
No homology to existing sequences[a]	36
Significant homology to ESTs[a]	28
Homology to known chromosome 16 genes[b]	7

[a] 2 of these clones were generated from mis-priming of the dUSA4 oligo
[b] 2 of these clones were derived from cryptic splicing events

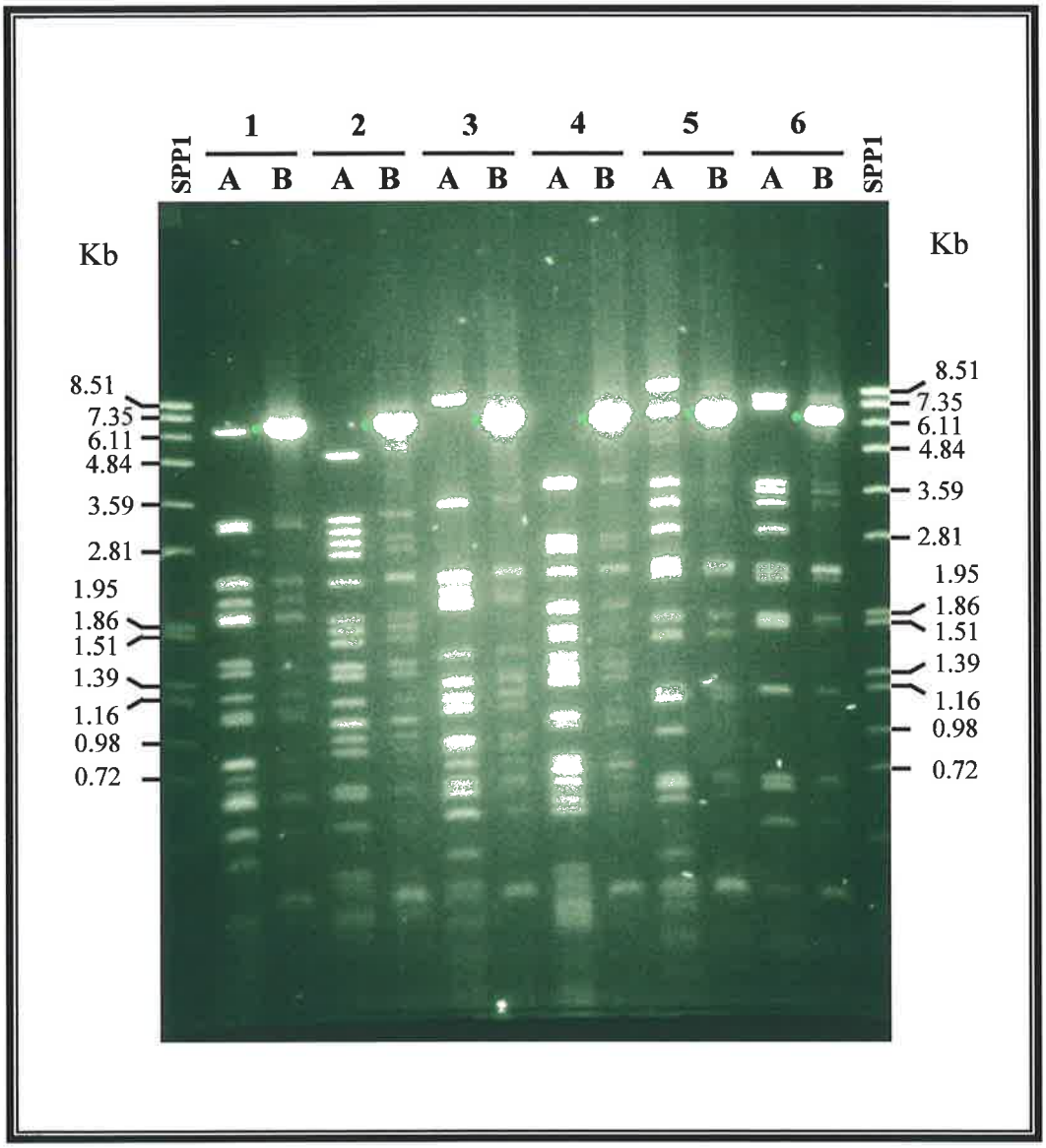


***Figure 4.2*:** Subcloning of *Pst*I digested cosmids into the pSPL3B-CAM exon trapping vector. In each case lane **A** represents the original cosmid digested with *Pst*I, while lane **B** represents a *Pst*I digest of the same cosmid that has been subcloned into the trapping vector. In most cases, each restriction fragment is represented by a pSPL3B-CAM subclone, indicating success of the initial stages of the trapping procedure. Cosmid **1**: 431F1; **2**: 354E12; **3**:427B11; **4**: 410H4; **5**; 421E4; **6**: 371D5. Green dots represent the linearised exon trapping vector.

analysis. For each individual trapping experiment, colony PCR products with size greater than that of the "vector-only" product were grouped according to size and one clone from each group was analysed further. This resulted in 96 trapped products that were further characterised. Figure 4.3 shows an example of the products generated during the amplification stages of the trapping procedure and the subsequent colony PCR results observed from the sub-cloning of these products.

From the sequence results of these 96 selected clones, 9 showed homology to repetitive sequences such as *Alu* elements and one clone showed homology to *E. coli* DNA, and were therefore eliminated from further analysis. Of the remaining 86 trapped products, 15 appeared to be the result of aberrant splicing originating from within the trapping vector alone, generating products containing vector sequence only. After elimination of these vector-only aberrantly generated products, a total of 71 unique trapped exons remained. Of these, 7 trapped products showed homology to previously characterised chromosome 16 genes (see 4.4.1 and Table 4.2), with 3 of these belonging to the *PISSLRE* gene and 2 to the *PRSM1* gene. Of the remaining 2, one clone was the result of an aberrant splicing event in the *PRSM1* gene while the remaining clone showed homology to *PISSLRE* and appeared to result from the splicing events occurring as a result of co-ligation of 2 non-contiguous cosmid restriction fragments into the pSPL3B-CAM vector (Table 4.2).

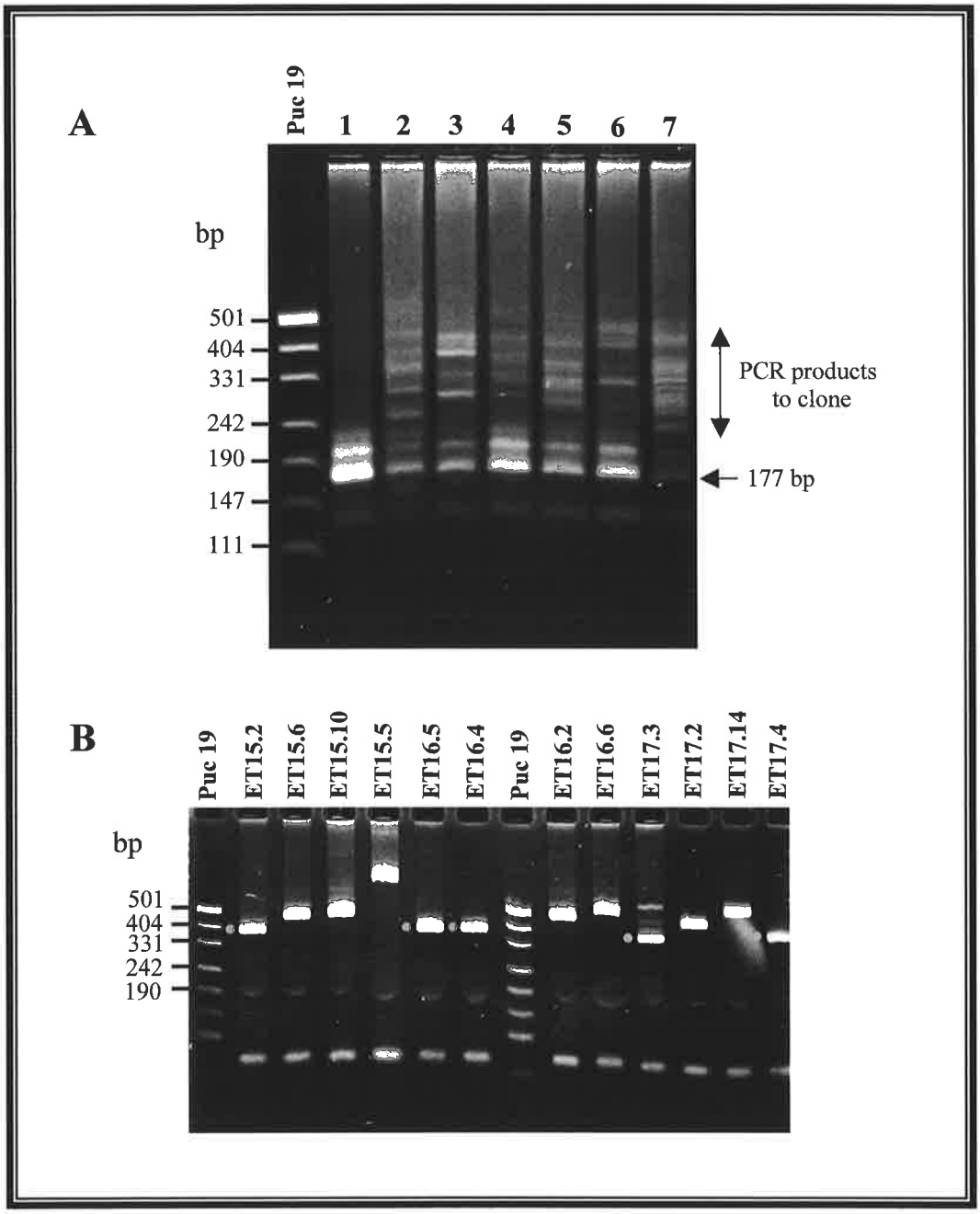


Figure 4.3: (**A**) Secondary PCR products generated from the use of the dUSD2 and dUSA4 primers on templates derived from **1**: pSPL3B-CAM vector only control; **2**: c344H1; **3**: c402C2; **4**: c417F1; **5**: c361H2; **6**: c444B9; **7**: c360D7. Vector only clones give rise to a 177 bp PCR product as indicated with an arrow. Products greater in size than this in the cosmid lanes 2-7 were viewed as potential trapped exons and were subsequently cloned and analysed by colony PCR. The source of the second band seen in the vector only control lane is unknown. (**B**) Colony PCR products obtained from the amplification of subcloned products from (**A**) with the PucF and PucR primer set. Clones containing vector only sequence give rise to a 409 bp product. Clones with a size greater than this were subsequently pooled based on their size and a representative clone from each group was analysed further. Bands highlighted in green represent vector-only containing clones.

TABLE 4.2

Trapped Exons Identical to Portions of Chromosome 16 Genes

Clone	Size	BLASTN
ET32.100	72bp	X78342 H. sapiens PISSLRE mRNA 100% nt 281-352
ET32.136	93bp	X78342 H. sapiens PISSLRE mRNA 100% nt 353-445
ET32.26	60bp	X78342 H. sapiens PISSLRE mRNA 100% nt 729-788
ET33.5	98bp	U58048 PRSM1 mRNA 100% nt 262-359
ET33.47	147bp	U58048 PRSM1 mRNA 100% nt 360-505
ET33.82[a]	119bp	U58048 PRSM1 mRNA 100% nt 567-635
ET32.160[b]	346bp	X78342 H. sapiens PISSLRE mRNA 100% nt 1,240-1,460

Note. In the BLAST homology column the following information is included: GenBank accession number, gene name, percent identity, and region of nucleotide (nt) homology. GenBank accession numbers have been obtained for each clone and are represented by AF039814-AF038818, AF039822, and AF039824. ET: trapped exon, followed by the clone number.
[a]Aberrant splicing.
[b]Co-ligation of non-contiguous cosmid fragments.

Twenty eight trapped exons displayed significant homology to sequences in dbEST (Table 4.3), however two of these clones appeared to be generated from the mis-priming of the dUSA4 oligo during the exon amplification stage, resulting in truncated clones. Finally, a total of 36 trapped products showed no homology to database sequences (Table 4.4), with 2 clones again arising from dUSA4 mis-priming. However, these four mis-primed clones could still have been generated from correctly spliced exons. All 71 trapped products, which included the two aberrantly spliced clones, were mapped back to their cosmid of origin by Southern analysis and were integrated into the physical map as seen in Figure 4.4.

4.3.2 Analysis of Trapped Exon Sequences

As can be seen from Table 4.3, a number of individual trapped exons showed homology to the

TABLE 4.3

Trapped Exons Showing Homology to Database Sequences (March 1998)

Clone	Size	ORF	BLASTN
ET27.12[a]	122bp	+2	dbEST: AA063452 (zm05d03.s1) 100% nt 337-394 1.5e-16
ET27.11	87bp	0,+2	nr: U19859 (M. mus growth arrest cDNA) 88% nt 143-226 1.3e-19 dbEST: AA019846 (ze60g08.s1) 96.5% nt 10-94 4.8e-28
ET26.8	124bp	+2	nr:D83703 (peroxisome assembly factor-2) 85% nt 2229-2338 6.0e-5 dbEST: AA350737 98% nt 170-233 1.5e-18
ET26.1[b]	151bp	+1,+2	dbEST: W22606 (Human retina cDNA) 100% nt 80-230 1.9e-55
ET26.3[a,b]	135bp	+2	dbEST: W22606 (Human retina cDNA) 100% nt 142-232 2.9e-30
ET6	134bp	+1,+2	nr: X96707 (Mus.musculus def-8-mRNA) 86.5% nt 88-194 7.5e-22 dbEST: N35521 (yx61a06.r1) 100% nt 27-160 1.6e-48
ET7	35bp	0,+1,+2	dbEST: N35521 (yx61a06.r1) 100% nt 1-26 0.0045
ET17.23	160bp	0	dbEST: AA009201 (mh02d02.r1) 89% nt 238-397 1.2e-46
ET6.14	93bp	+1,+2	dbEST: AA009201 (mh02d02.r1) 89% nt 39-131 1.2e-23 dbEST: AA424369 (zv90e08.r1) 93% nt 446-538 1.3e-26
ET17.9	93bp	0,+1,+2	dbEST: AA424369 (zv90e08.r1) 99% nt 352-445 4.9e-29
ET17.2	66bp	0,+1	dbEST: AA424369 (zv90e08.r1) 100% nt 1-37 2.9e-7 dbEST: AA078773 (zm21b08.r1) 100% nt 281-346 6.9e-20
ET17.14	164bp	0	dbEST: AA078773 (zm21b08.r1) 100% nt 4-86 4.3e-25
ET14.22[b]	68bp	0,+1,+2	dbEST: R06205 (ye94g05.r1) 100% nt 87-154 9.2e-21
ET13.6[b]	203bp	0,+2	dbEST: R06205 (ye94g05.r1) 100% nt 1-154 8.1e-59

Note. In the BLAST homology column the following information is included: database, GenBank accession number (clone description), percent identity, region of nucleotide (nt) homology, and P value. ORF: open reading frames. All sequences of trapped exons represent the sense strand and were examined for ORF. 0: Frame 1; +1: Frame 2; +2: Frame 3. GenBank accession numbers have been obtained for each clone and are represented by AF039785-AF039798. ET: trapped exon, followed by the clone number.

[a] Homology is in the opposite orientation.

[b] Overlapping clone.

TABLE 4.3 (Cont.)

Trapped Exons Showing Homology to Database Sequences (March 1998)

Clone	Size	ORF	BLASTN
ET13.2	230bp	0,+2	dbEST: AA064599 (zm13c09.r1) 97% nt 1-149 7.4e-51
ET15.24	109bp	+1,+2	dbEST: R07213 (yf14a03.r1) 100% nt 87-194 8.0e-34
ET19	258bp	+1	dbEST: R07213 (yf14a03.r1) 97% nt 1-86 9.5e-26
ET15.22	131bp	0,+3	dbEST: R07213 (yf14a03.r1) 98% nt 195-256 3e-22
ET15.10	104bp	+1	dbEST: R61618 (yh09a04.r1) 100% nt 143-219 1.6e-24
ET15.5	466bp	none	dbEST: R07162 (yf14a03.s1) 99% nt 7-204 1e-105 dbEST: R61618 (yh09a04.r1) 97% nt 380-419 8e-11
ET3.5	93bp	0,+1,+2	dbEST: Z39757 (clone c-1ih04) 98% nt 30-122 1.3e-32
ET2.1	158bp	0	nr: U83246 (human copine I mRNA) 65% nt 1018-1149 3.3e-11 dbEST: F05508 (clone c-0dc06) 77% nt 70-214 1.3e-28
ET3.25	276bp	0	nr: U83246 (human copine I mRNA) 54% nt 694-861 0.044 dbEST: Z41974 (clone c-04f05) 65% nt 15-170 3.9e-15
ET2.8	103bp	0	dbEST: C15128 (human fetal brain cDNA) 90% nt 110-212 5.1e-32
ET2.5	229bp	+1	dbEST: AA023715 (mh78a10.r1) 89% nt 108-336 7.1e-73 dbEST: D63198 (Hum505B01B) 96% nt 176-300 1.1e-43
ET24.35	85bp	0	dbEST: AA027336 (ze97g11.s1) 100% nt 49-130 2.6e-25
ET33.6[c]	191bp	none	nr: AC002479 (human BAC clone) 90% nt 3905-4000 2.3e-10 dbEST: W81593 (zd88g02.r1) 77% nt 165-248 1.9e-11
ET34.1[c]	109bp	0	dbEST: AA023715 (mh78a10.r1) 86% nt 371-449 5.3e-17

Note. In the BLAST homology column the following information is included: database, GenBank accession number (clone description), percent identity, region of nucleotide (nt) homology, and P value. ORF: open reading frames. All sequences of trapped exons represent the sense strand and were examined for ORF. 0: Frame 1; +1: Frame 2; +2: Frame 3. GenBank accession numbers have been obtained for each clone and are represented by AF039799-AF039807, AF039812, AF039813, and AF039819-AF039821. ET: trapped exon, followed by the clone number.

[c] Due to mis-priming of the dUSA4 primer during exon amplification.

TABLE 4.4

Trapped Exons with no Homology to Database Sequences

Clone	Size	ORF	Clone	Size	ORF
ET5.7	43bp	+1,+2	ET22.13	130bp	0,+1,+2
ET27.40[a]	107bp	+1,+2	ET22.19	63bp	0,+2
ET27.32	68bp	0	ET22.22[a]	70bp	+2
ET27.9	50bp	0,+1,+2	ET22.1	66bp	0,+1
ET18	114bp	0	ET25.45	98bp	+1
ET13.9	113bp	+2	ET3.26	248bp	+1
ET16.6	125bp	0	ET13.20	239bp	+1,+2
ET6.1	113bp	+1,+2	ET24.28	68bp	+1
ET6.6	134bp	+1,+2	ET23.23[b]	49bp	0,+2
ET6.4	200bp	+1	ET34.76[b]	223bp	+1
ET2	74bp	+1,+2	ET5.2[a]	246bp	none
ET1	114bp	+2	ET15.9	153bp	none
ET15.6	84bp	+1	ET33.9	125bp	none
ET32.68	133bp	+2	ET24.42	87bp	none
ET32.46	115bp	0,+1	ET24.17	152bp	none
ET4.19	72bp	0	ET22.24[a]	168bp	none
ET2.37	41bp	+2	ET25.41	82bp	none
ET34.87	103bp	0	ET25.8	74bp	none

Note. ORF: open reading frames. All sequences of trapped exons represent the sense strand and were examined for ORF. 0: Frame 1; +1: Frame 2; +2: Frame 3. GenBank accession numbers have been obtained for each clone and are represented by AF039754-AF039778, AF039808-AF039811, and AF039823.

[a] Overlapping clones.

[b] Mis-priming of the dUSA4 primer during exon amplification.

same ESTs. This enabled some of the exons to be placed into tentative transcription units based on their homology to identical cDNAs. The first example is the group of exons comprising ET17.23/ET6.14/ET17.9/ET17.2/ET17.14 that could be linked together based on three overlapping cDNA clones (Genome Systems clones mh02d02, zv90e08, and zm21b08). In addition, ET17.9 and ET6.14 appear to be adjacent exons based on the homology observed with the zv90e08 cDNA clone. The trapped exons ET6 and ET7 could also be linked based on a common homology to the cDNA clone yx61a06. Again, these trapped products appear to be adjacent exons of this cDNA.

Both ET2.1 and ET3.25 showed 65 and 54% homology, respectively, to the human copine I mRNA (unpublished reference in GenBank) and to a number of EST clones in dbEST. One of these cDNAs (yl70c06) belongs to a contig of sequences present in the TIGR database (http://www.tigr.org/tigr_home/tdb/), THC126737. A clone from this contig (ym60b12) however, could not be mapped to the cosmid contig at 16q24.3 by Southern hybridisations. Using this clone as a probe for FISH to metaphase chromosomes showed signals at both 3q21 and 16q24.3 (data of Elizabeth Baker). This suggests that the cDNA clones belonging to THC126737 are part of the copine I gene, which maps to chromosome 3q21, whereas exons ET2.1 and ET3.25 are part of a closely related gene at 16q24.3.

There were a number of examples of exons overlapping with each other. The trapped exon ET14.22 was contained within the last 68 base pairs of ET13.6, suggesting the latter trapped product contained at least 2 contiguous exons of a transcript simultaneously trapped from a single genomic cosmid fragment. Similarly, the 70 base pair exon ET22.22 is contained within the 168 base pair exon ET22.24. There were two instances where clones overlapped each other but were trapped from opposite strands of DNA (ET26.1/ET26.3 and ET5.2/ET27.40),

and one instance where two non-overlapping trapped products identified a common cDNA clone, however again, one appeared to be trapped from the opposite strand (ET2.5 and ET34.1). The remaining trapped products presented in Table 4.3 identified unique EST homology and therefore probably represent individual genes. In total, 15 separate novel transcripts can be identified based on EST homologies in dbEST, which map to the cosmid contig.

Table 4.4 indicates that 36 trapped products did not show homology to sequences in available databases. These may represent exons that belong to the 5' portions of the transcripts identified above, or may themselves belong to genes not yet represented by ESTs in dbEST. Given that 28 of them have open reading frames (ORF), the chances that they do represent real exons is encouraging. Those clones without ORFs, are likely to result from splicing events occurring within non-coding genomic DNA, or may represent non-coding exons present in the 3' or 5' untranslated regions of genes.

4.3.3 Cloning of the *FAA* Gene

4.3.3.1 Identification of cDNA Clones Homologous to Trapped Exons

Another large group of trapped products that could be linked based on their EST homologies was ET15.24/ET19/ET15.22/ET15.10 and ET15.5. The cDNA clones that these exons had in common were yf14a03, and yh09a04, the latter of these being one of the initial EST markers used for physical map construction. The yh09a04 clone was originally identified by direct cDNA selection with a cosmid overlapping with c361H2, and subsequently sequenced by Dr. Sinoula Apostolou. She also sequenced the yf14a03 clone and together showed that these cDNAs contained 2,140 base pairs. From sequence alignments, the trapped exon ET19 overlapped with the 5' end of this combined sequence by 87 base pairs, thus extending it a

further 171 bases. The trapped exon ET15.24 was also contained within the yf14a03 portion of the combined cDNA sequence, and appeared to be the adjacent 3' exon to ET19 based on sequence homology, while ET15.5 showed 100% sequence homology towards the 3' end. The large size of this trapped exon suggests that it may consist of a number of adjacent exons, although it may be a cryptic product as it did not possess an open reading frame. The remaining trapped exons showing homology to these two cDNA clones did appear to be generated from aberrant splicing events. ET15.22 began with complete nucleotide homology to the cDNA sequence established, but then the sequence diverged completely after the first 60 base pairs, while ET15.10 was trapped from the anti-sense strand of this transcript.

4.3.3.2 Identification of Full-Length cDNA Sequence

The results of work described in this section and 4.3.3.3 were performed by other members of the FAB consortium who were delegated exons from the region surrounding yf14a03 and yh09a04 for transcript characterisation. The ultimate view was to perform DNA mutation analysis of *FAA* patients and of breast tumour material. As a consequence, in most instances, raw data is not shown but can be obtained from the publication arising from this work (The FAB consortium, 1996-Appendix A1).

From Southern mapping of the trapped exons belonging to the yh09a04/yf14a03 sequence, it can be seen that the transcript they are part of extends 5' towards the 16q telomere, with a total of 5 exons (ET1/ET2/ET6.4/ET6.6 and ET6.1) trapped from the 431F1 cosmid, mapping to this adjacent 5' region (Figure 4.4). These exons were ultimately used to extend the transcripts sequence by reverse transcribed PCR (RT-PCR) and by screening cDNA libraries for larger clones. Sequencing of these overlapping cDNA clones revealed an open reading frame of 4,365 base pairs that would encode a protein of 1,455 amino acids. Northern blot

analysis indicated that the gene is expressed in a variety of tissues, including kidney, heart, brain, skeletal muscle, and liver. Although transcripts of several sizes were detected, the most prominent was 4.7 kb in length, a size consistent with the length of the cDNA sequence obtained for the gene. Searches of nucleotide and protein databases using BLAST failed to reveal strong homologies with genes of known function.

4.3.3.3 Mutation Analysis

Evidence that this cDNA may be the *FAA* gene was obtained when a patient of Italian origin was shown to be partially deleted when a region of the cDNA was amplified by RT-PCR. This deletion of 274 nucleotides was found to create a premature stop codon with the deletion removing ET6.4 and ET2. In total, four intragenic deletions have been identified from this candidate cDNA in FA patients, three of which produce truncated proteins. This provides strong evidence that the sequence isolated is the *FAA* gene. Further evidence has been provided by the co-incident cloning of the same gene by functional complementation and the identification of over 70 different additional mutations (Lo Ten Foe *et al.*, 1996; Levran *et al.*, 1997; Wijker *et al.*, 1998).

From the complete sequence and intron/exon structure of *FAA* (Ianzano *et al.*, 1997-Appendix A2), a total of 10 trapped products could be associated with the gene. ET6.1, ET6.6, ET2, ET1, and ET15.24 corresponded to exon 7, 14, 21, 22, and 33 respectively, while ET19 represented exons 31 and 32. ET6.4 contained exons 18, 19, and 20, however these exons had been spliced together in the wrong order, ET15.5 constituted a truncated version of the 3' terminal exon, and the remaining exons, ET15.22 and ET15.10, were also the result of aberrant splicing.

4.3.4 Analysis of the Human Gene Map

In an effort to increase the density of the transcript map in the critical region of LOH seen in breast cancer, ESTs from the available Human Gene Map (October 1996 version) were analysed. Table 4.5 displays a summary of the UniGene clusters that have been mapped to the genetic interval defined by the D16S520 marker and the long arm telomere of chromosome 16. A total of 23 unique clusters were identified with an additional 4 mapped ESTs that did not have an associated UniGene cluster number. Nine of these clusters belonged to previously characterised genes mapped to this chromosome, with 6 of them localised to the clone contig established in the CY18A(D2)-qter region. These are indicated on Figure 4.4. The *MTG16* and *GALNS* genes map to the cytogenetic interval immediately proximal to the CY18A(D2)-16qter interval, while the *ICSBP1* gene has not been mapped to the somatic cell hybrid panel of chromosome 16 and was not identified by exon trapping.

Representative cDNA clones from the Hs11173 and Hs14642 UniGene clusters were mapped to the cosmid contig by Southern hybridisation (Figure 4.4), however they were not identified by exon trapping. A further 2 clusters (Hs26975 and Hs10238) mapped to the established contig also, but were located outside of the region that was exon trapped, and another four clusters (Hs7081, Hs11968, Hs31443, and Hs79077) were physically mapped to the adjacent centromeric hybrid interval (CY2/CY3-CY18A(D2)) by Mr. Jason Powell (Department of Cytogenetics and Molecular Genetics, WCH). A total of 9 UniGene clusters contained cDNA clones for which physical mapping data has not yet been established with none of them showing homology to ESTs identified through exon trapping. Only one cluster (Hs7130) from the 23 contained the same set of cDNA clones that was identified by the trapped exons ET2.1 and ET3.25.

TABLE 4.5

Analysis of UniGene ESTs Mapped Between D16S520 and 16qter (August 1998)

Mapped EST	UniGene cluster	Comments
D38554	Hs57302	Homologous to the *PRSM1* gene[a]
X65634	Hs2823	Homologous to the *MC1R* gene[a]
U06088	Hs32293	Homologous to the *GALNS* gene[b]
WI-1435	Hs119436	Homologous to the *BBC1* gene[a]
SHGC-9877	Hs2286	Homologous to the *ICSBP1* gene[c]
WI-17119	Hs86297	Homologous to the *FAA* gene[a]
stSG8033	Hs77313	Homologous to the *PISSLRE* gene[a]
stSG8665	Hs110099	Homologous to the *MTG16* gene[b]
D29107	Hs78497	Homologous to the *CMAR* gene[a]
stSG2568	Hs7130	ET2.1 and ET3.25 also show homology to these ESTs
a006Q31	Hs11173	Maps to the physical map contig established in 16q24.3[d]
stSG9210	Hs14642	Maps to the physical map contig established in 16q24.3[d]
WI-13072	Hs26975	Maps to the contig but not in region that was exon trapped
SGC33145	Hs10238	Maps to the contig but not in region that was exon trapped
SGC36958	Hs7081	Maps to the CY2/CY3-CY18A(D2) interval[e]
WI-12410	Hs11968	Maps to the CY2/CY3-CY18A(D2) interval[e]
WI-16080	Hs31443	Maps to the CY2/CY3-CY18A(D2) interval[e]
KIAA0233	Hs79077	Maps to the CY2/CY3-CY18A(D2) interval[e]
WI-11796	Hs10172	Not physically mapped
WI-15838	Hs7970	Not physically mapped
WI-16844	none	Not physically mapped
WI-18377	Hs40530	Not physically mapped
stSG4762	Hs81505	Not physically mapped
A008S19	none	Not physically mapped
A008U15	none	Not physically mapped
WI-16952	none	Not physically mapped
SGC32044	Hs106326	Not physically mapped

Note. [a] maps to the established clone contig at 16q24.3; [b]maps to the CY2/CY3-CY18A(D2) cytogenetic interval; [c] physical map location is unknown; [d] the cDNA clones within this UniGene cluster were not identified by exon trapping; [e] this region is proximal to that which was exon trapped (mapped by Mr. Jason Powell).

4.3.5 Analysis of Single-Pass Cosmid Sequencing

Table 4.6 describes the database homologies observed through BLASTN analysis of cosmid sequence that were not identified through a search of the Human Gene Map or from exon trapping.

TABLE 4.6

BLAST Analysis of Cosmid Partial Sequence Data

Cosmid	Homologous EST	Cosmid	Homologous EST
344G2	AA099020 (zn45h01.s1) [a]	344H1	T49146 (yb09d10.s1)
377F1	U47634 (Human beta-tubulin class III isotype mRNA).	410H4	H06994 (yl81b07.r1)
360D7	AA035673 (ze25b01.r1)	383H6	H05447 (yl81b07.s1)
417F1	AA234782 (zr78b01.r1) [a]	358D12	H57222 (yr08f01.r1) [a]
361H2	R01230 (ye81b05.s1)		

Note. Nineteen cosmids were partially sequenced but only those clones with homology to ESTs not identified by exon trapping or Human Gene Map searches are displayed. BLAST analysis was performed by Dr. Norman Doggett, Dr. David Callen and the candidate.

[a] Clones considered low priority and omitted from further characterisation as detailed analysis of corresponding cDNA clones indicated they were likely to be genomic contaminants. The rest were physically mapped to the cosmid contig using cDNA inserts corresponding to the relevant EST.

A total of 8 additional ESTs were discovered, however, the ye81b05 cDNA clone and ESTs corresponding to the beta-tubulin class III isotype mRNA do belong to UniGene clusters that have both been mapped between D16S422 and the 16q telomere on the gene map. The D16S422 marker has been physically mapped proximal to D16S520 and therefore may explain why it was not identified in the previous gene map search. All cosmids sequenced identified many ESTs present in dbEST, except the 444B9 cosmid sequence, which failed to detect any EST matches. Three of the 8 new ESTs identified appeared to be genomic contaminants as polyadenylation signals were not seen at their 3' ends, and no additional cDNA clones had homology matches with them. In addition, the yl81b07 cDNA was found to map to the same

*Eco*RI restriction fragment as the cDNA clone identified by ET2.8 and therefore was omitted from further analysis as they probably represented the same transcript. This left only 4 new ESTs that were subsequently mapped by Southern hybridisation back to the clone contig and were incorporated into the transcript map (Figure 4.4).

4.4 Discussion

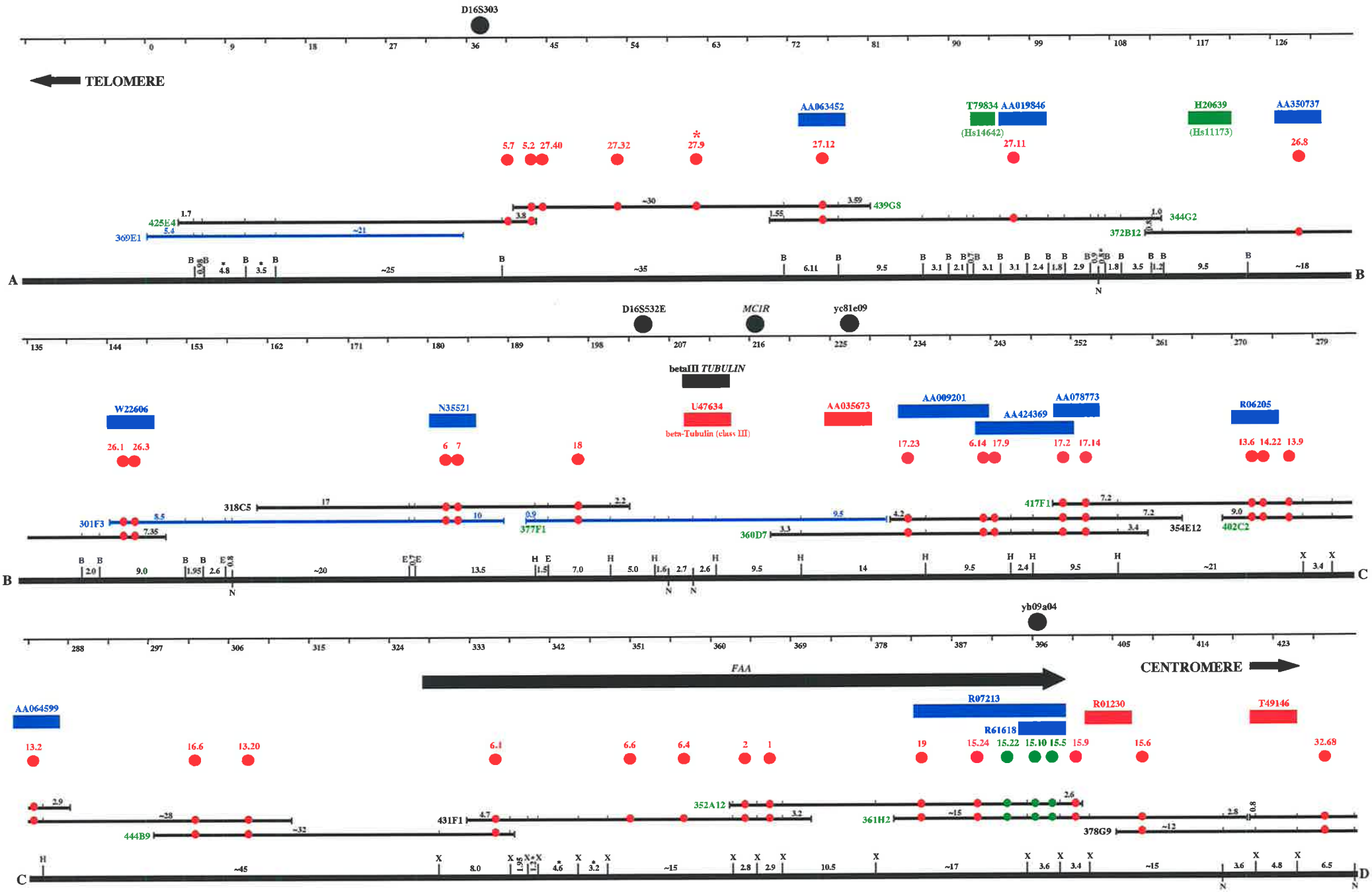
4.4.1 Exon Trapping

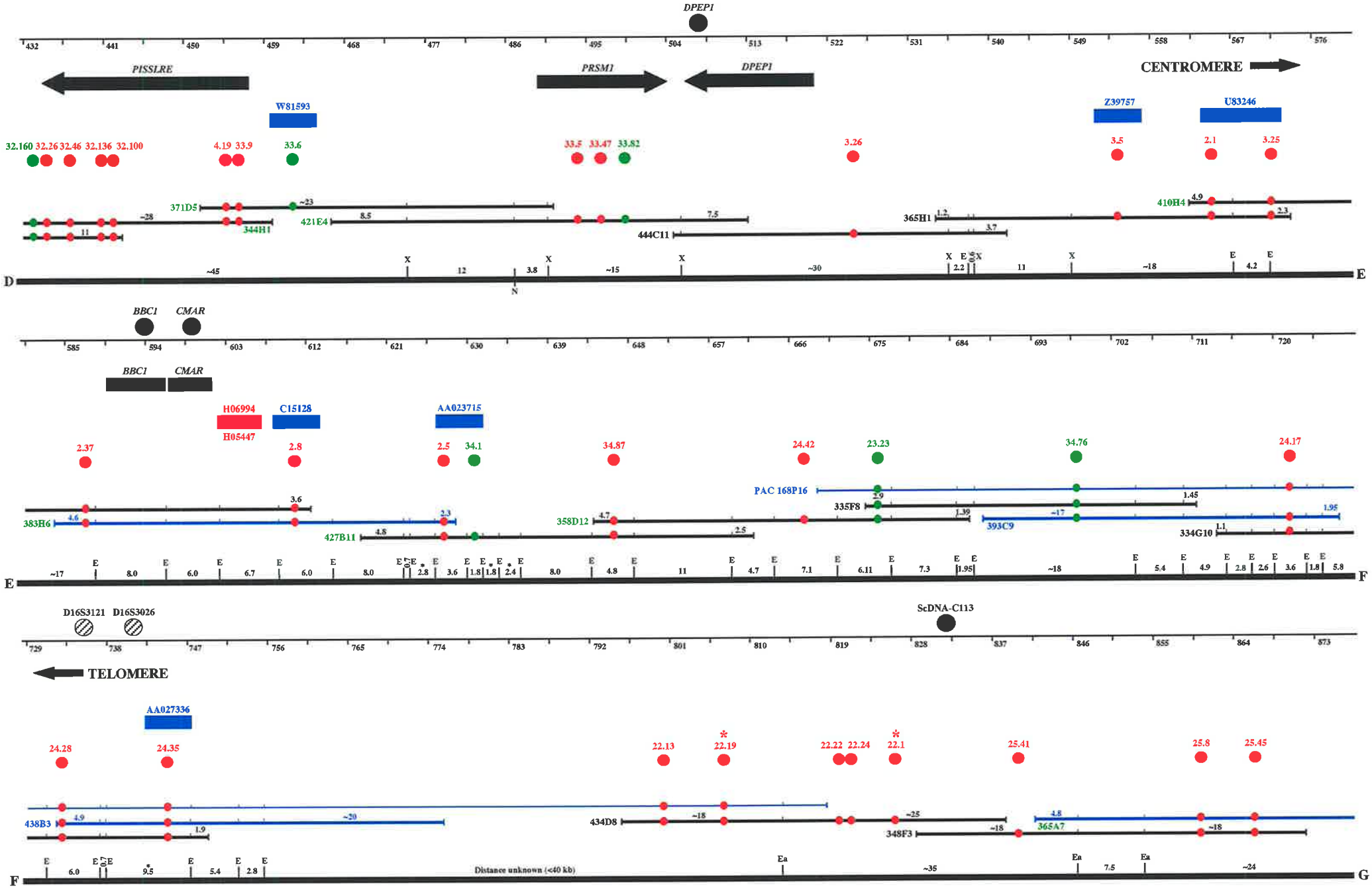
In total, 71 different exons (including aberrantly generated clones) were identified from the 26 cosmids used in the trapping procedure. This represents an average of 2.7 traps per cosmid or 1 exon every 13 kb. Similar successes with this procedure have been reported (Burn *et al.*, 1996; Chen *et al.*, 1996*a*; Church *et al.*, 1997). To maximise the number of unique exons trapped it is important to minimise the complexity of the pSPL3B-CAM recombinant clones introduced into the COS-7 cells. In most cases, trapping from 1 cosmid at a time provided the maximum return of trapped exons. Due to the size of the region being analysed, a single clone representing a particular size group from each trapping experiment was chosen for further analysis. Given more time, a detailed analysis of additional clones within each size group would undoubtedly reveal the presence of different exons possessing a similar size to the one chosen, which would further increase the effectiveness of the exon trapping procedure (this was found in later studies, see chapter 7).

A high proportion (59%) of clones analysed by colony PCR appeared to be "vector-only" clones. These would be generated if the cosmid genomic insert did not have a functional exon present (an exon with an attached 5' splice donor and 3' splice acceptor site). The *Bst*XI digestion of the trapping procedure (4.2.9) is designed to eliminate such clones, however it

Figure 4.4

Integrated physical and transcription map of the restricted region of loss of heterozygosity (LOH) at 16q24.3 seen in breast cancer. The map is divided into 6 segments which span 2 pages and extends from **A** (telomeric end) to **G** (centromeric end). Only the region that was exon trapped is indicated. Clones representing a minimum tiling path in the contig are shown, with cosmids used for exon trapping indicated by black horizontal lines. The clone number of cosmids used as templates for partial genomic sequencing at Los Alamos National Laboratories is indicated in green type. Black circles above the scale bars indicate the markers mapping to the region, while the hatched black circles show the location of the D16S3121 and D16S3026 markers which define the centromeric limit of the LOH region. The position of exons trapped from the region are indicated by red circles, while exons generated from aberrant splicing events are represented by green circles. EST clones identified from database homologies with trapped exons are shown by blue boxes. ESTs homologous to partial cosmid sequence are represented as red boxes, while ESTs from mapped UniGene clusters are shown as green boxes. Characterised genes, including *FAA,* are indicated by black boxes, with their orientation, if known, shown in a 5' to 3' direction (as indicated by arrows). Red asterisks above certain trapped exons indicate that the order of these clones could not be determined with respect to exons immediately distal to them. For each EST clone shown, the GenBank accession number is indicated above the appropriately coloured box.


D16S303
TELOMERE
AA063452
T79834 AA019846
(Hs14642)
H20639
(Hs11173)
AA350737
5.7 5.2 27.40
27.32
27.9
27.12
27.11
26.8
425E4
369E1
439G8
344G2
372B12
A
B
D16S532E
MCIR
yc81e09
betaIII TUBULIN
W22606
N35521
U47634
beta-Tubulin (class III)
AA035673
AA009201
AA424369
AA078773
R06205
26.1 26.3
6 7
18
17.23
6.14 17.9
17.2 17.14
13.6 14.22 13.9
301F3
318C5
377F1
360D7
417F1
354E12
402C2
C
yh09a04
FAA
CENTROMERE
AA064599
R07213
R61618
R01230
T49146
13.2
16.6
13.20
6.1
6.6
6.4
2 1
19
15.24 15.22 15.10 15.5 15.9
15.6
32.68
444B9
431F1
352A12
361H2
378G9
D


DPEP1
PISSLRE
PRSM1
DPEP1
CENTROMERE
W81593
Z39757
U83246
32.160 32.26 32.46 32.136 32.100
4.19 33.9
33.6
33.5 33.47 33.82
3.26
3.5
2.1
3.25
371D5
344H1
421E4
444C11
365H1
410H4
D
E
BBC1
CMAR
H06994
H05447
C15128
AA023715
2.37
2.8
2.5
34.1
34.87
24.42
23.23
34.76
24.17
PAC 168P16
335F8
383H6
358D12
427B11
393C9
334G10
E
F
D16S3121
D16S3026
ScDNA-C113
TELOMERE
AA027336
24.28
24.35
22.13
22.19
22.22 22.24
22.1
25.41
25.8
25.45
438B3
434D8
348F3
365A7
F
G
Distance unknown (<40 kb)

appears that this step is not efficient. This, coupled with the fact that a high percentage of cosmid fragments cloned into the trapping vector probably do not contain a functional exon, may account for the large proportion of contaminating clones. The procedure could be improved through the purification of primary PCR products before subsequent digestion with *Bst*XI. These clones however did not interfere with the overall effectiveness of the trapping procedure, as they could be simply eliminated by colony PCR before further characterisation.

A total of two exons were trapped from the *PRSM1* gene, which had been previously mapped incorrectly between the somatic cell hybrid breakpoints CY2/CY3 and CY18A(D2) at 16q24.3 (Scott *et al.*, 1996), and two exons were identified from the *PISSLRE* gene, which had also been previously mapped to chromosome 16q24 (Bullrich *et al.*, 1995). This study therefore provided refined localisations for these two genes. Both the *CMAR* and *MC1R* genes are intronless, explaining why no exons were trapped from them, however the *DPEP1* gene has 10 exons, none of which were identified through exon trapping. Similarly, no exons were trapped from the *BBC1* gene. This may be a reflection of the distribution of *Bst*XI, *Pst*I, *Bam*HI, and *Bgl*II sites within these genes, resulting in either large restriction fragments not ligating into the pSPL3B-CAM vector efficiently, or small restriction fragments that lack intact exons. Similar gaps in transcription maps have been encountered (Brody *et al.*, 1995; Burn *et al.*, 1996) suggesting that exon trapping should be coupled with other gene identification methods for the completion of a comprehensive transcription map. However, even using a combination of five methods in a 600 kb region around the *BRCA1* locus, Brody and co-workers (1995) were unable to identify clones for at least 2 known genes.

A total of 10 exons (represented by 7 independent traps) were isolated from the *FAA* gene that aided in its initial identification and subsequent genomic characterisation. This gene has been

shown to consist of 43 exons (Ianzano *et al.*, 1997) indicating that the trapping procedure was successful in cloning approximately 25% of its exons. Given that the gene spans about 80 kb of genomic sequence, a total of 6 exons would be expected to be identified based on the average number of exons trapped per cosmid. The trapped exon ET6.1 corresponded to exon 7 of *FAA*, and was the most 5' exon identified. This product therefore provided essential information for the subsequent identification of the 5' sequence of the gene by RT-PCR. This is often a difficult and time consuming task in gene characterisation, especially in larger genes whose corresponding cDNA clones may only extend a maximum of 3 kb towards the 5' end in most conventional libraries. Since FA patients are at a higher risk of developing cancers, and their cells show an increased sensitivity to DNA crosslinking agents, the *FAA* gene could be the target for LOH at 16q24.3. A recent study has screened this gene for mutations by SSCP analysis in 19 breast tumours with specific LOH at 16q24.3 (Cleton-Jansen *et al.*, 1999). However, no tumour-specific mutations were identified, indicating this gene is not involved in breast carcinogenesis.

There were a number of instances of exons being trapped as a result of the incorrect usage of splice sites, either in the trapping vector or in the cloned genomic cosmid DNA. This has not been reported in other studies and may be a reflection of the number of products characterised from each trapping experiment with the result that these rarer clones were being analysed. The availability of the genomic structure of the *FAA* gene has assisted in identifying how some of these trapped products were generated. The first example is ET15.22 whose first 60 base pairs is identical to exon 34 of *FAA*, with the remainder of the sequence identical to intron 34. From the analysis of sequence within this intron it appears that the 5' splice donor site of intron 34 has been ignored with an alternative site being used 72 base pairs downstream of the correct site. This results in the generation of a change in reading frame for the protein and is most

likely not a true alternative splicing event. The score of this aberrant splice site based on its identity to consensus donor sites (Shapiro and Senapathy, 1987) was 75%, compared to a score of 83% for the correct splice site. The similarity of sequences at these 2 regions reflects the complex and extremely sensitive mechanism of mRNA processing. The second example was ET15.10, which was the result of the use of aberrant splice donor and acceptor sites within intron 40 and exon 41 on the opposite strand of the FAA gene. ET15.5 used a cryptic splice donor site within the 3' UTR of the gene as a result of the absence of splice donor sites at 3' terminal exons, and finally ET6.4 was the result of co-ligation of non-contiguous fragments of cosmid DNA into the trapping vector. This led to exons 18, 19, and 20 being spliced together in the wrong order. Detailed information regarding the associated genes of other trapped exons is not yet available, but presumably some of these exons were also generated by similar mechanisms.

At least 14 potential new transcripts have been identified by database homologies of trapped products to EST sequences that can now be considered candidates for a breast cancer tumour suppressor gene. This does not include those trapped exons that had no database homology, which may represent as yet unsequenced transcripts or genes expressed in defined developmental stages or tissues. Eight of the 14 potential new transcripts were homologous to UniGene clusters, with only one of these 8 placed on the NCBI Human Gene Map. Thus exon trapping has succeeded in mapping 7 UniGene clusters to chromosome 16q24.3. Of the remaining 6 new transcripts that did not have associated UniGene clusters, all but one overlapped with additional cDNA clones. This indicates that construction of these clusters is not 100% efficient. An alternative would be to examine the tentative human consensus (THC) clusters constructed at The Institute of Genomic Research (see 1.2.3.6) with these remaining transcripts.

4.4.2 Comparison of Transcript Identification Methods

From analysis of the 1996 NCBI Human Gene Map, 27 UniGene clusters and singleton ESTs have been mapped between the D16S520-qter region, with only nine of these not yet physically mapped to the somatic cell hybrid panel. Of these nine, seven are located proximal to either D16S3121 or D16S3026 by genetic linkage, suggesting that they may not be located within the region defining the minimum LOH seen in breast cancer. All nine clones will be physically mapped in the immediate future. A total of 17 clusters are either physically located outside of the critical LOH region or represent previously characterised chromosome 16 genes. Only 2 new transcripts were identified, mapping to the critical LOH region, that were not found with exon trapping.

Analysis of single-pass cosmid sequence helped identify 4 new cDNA clones not identified by exon trapping or searches of the Human Gene Map. Together, these techniques of transcript identification have allowed the isolation of at least 20 new putative genes, each of which needs to be characterised further to determine if it has a role in breast tumourigenesis. By far, the most efficient method was the exon trapping procedure, although analysis of the complete sequence of the entire cosmid contig, if available, would serve as the ultimate resource for gene isolation. Combined with the 7 previously characterised genes, at least 27 transcripts can be placed on the transcript map between D16S3026 and D16S303, a region of approximately 750 kb, giving a density of 1 gene per 28 kb. Given that an estimated 50,000 to 100,000 genes exist in the $3X10^9$ human genome, there is expected to be on average 1 gene every 30 to 60 kb, suggesting that a large percentage of the genes located in this region may already have been identified. However it is still to be determined if some of these transcripts relate to the same gene or if they are actually transcribed.

4.4.3 Candidate Breast Cancer Tumour Suppressor Genes

Through the development of the transcript map of the critical region of LOH, three previously characterised genes have now been located to this region. The first of these is the *PISSLRE* gene, which belongs to a family of cyclin-dependant protein serine/threonine kinases that are critical for cell cycle progression. Dominant-negative constructs of the gene have been shown to halt cell cycle progression in G_2-M, and when overexpressed in U2OS cells (osteosarcoma cell line with an inactive Rb protein), suppresses growth (Li *et al.*, 1995). Although tumour suppressor gene inactivation involves loss of function mutations, the *PISSLRE* gene may still play a role in breast cancer due to its role in the cell cycle. Recent SSCP mutation analysis of this gene however, has failed to identify nucleotide changes specific for breast tumour DNA, which suggests it is not involved in breast carcinogenesis (Crawford *et al.*, manuscript submitted). The second is the *PRSM1* gene, a metallopeptidase of the zincin superfamily. These genes play a major role in many biological processes including pattern formation, growth factor activation, and the synthesis of collagen which forms the fibrous scaffold of the extracellular matrix of tissues (Scott *et al.*, 1996). A role in cancer metastasis could therefore be predicted, through the degradation of extracellular matrix promoting tumour invasion. The third characterised gene to be identified is the class III isotype beta tubulin gene. Tubulin is an alpha/beta heterodimer and is the subunit protein of microtubules. Both the alpha and beta units exist in many isotypes each encoded by different genes, with some appearing to have no functional significance while others perform specific functions such as formation of particular organelles and interactions with specific proteins. The beta tubulin class III isotype has been shown to be expressed exclusively in neuronal tissues (Draberova *et al.*, 1998) and therefore is not a likely candidate breast cancer tumour suppressor gene.

To determine which of the remaining novel transcripts identified may be involved in breast

carcinogenesis, clues to the functions must be derived from the homology that may be observed between them and genes already characterised. One of the more interesting trapped exon homologies was to a mouse growth arrest-specific cDNA clone. The mouse sequence was 88% identical with ET27.11, suggesting that this was the human homologue. Growth arrest-specific genes were first identified from a mouse NIH 3T3 subtracted cDNA library enriched for RNA sequences preferentially expressed during growth arrest (Schneider *et al.*, 1988). The possibility that the transcript identified by this trapped exon is a tumour suppressor gene is investigated in detail in chapter 5. Further characterisation of trapped exons that do not show homology to presently known genes, and therefore there is no immediate clue to their function, will be described in chapter 6.

Chapter 5

Characterisation of the GAS11 and C16orf3 Genes

Table of Contents

	Page
5.1 Introduction	**156**
5.2 Methods	**160**
5.2.1 *GAS11* Sequence Assembly	161
5.2.2 Isolation of 5' End Sequences	161
5.2.3 Database Homology Searches	164
5.2.4 Northern Hybridisations	164
5.2.5 Genomic Structure Characterisation of *GAS11*	164
5.2.5.1 Direct Cosmid Sequencing	166
5.2.5.2 PCR Amplification of cDNA and Genomic DNA	166
5.2.5.3 Sequencing c344G2 BamHI Subcloned DNA	166
5.2.6 PCR	167
5.2.7 RT-PCR	167
5.2.8 R*Tth* PCR	168
5.2.9 3' RACE	168
5.2.10 Mutation Analysis	169
5.2.10.1 Single Stranded Conformation Polymorphism (SSCP) Analysis	169
5.2.10.2 Screening for Homozygous Deletions	169
5.2.10.3 Exon Skipping	169
5.3 Results	**171**
5.3.1 *GAS11* Sequence Determination	171
5.3.1.1 Sequencing cDNA Clones	171
5.3.1.2 5' Sequence Isolation	173
5.3.2 *GAS11* Northern Analysis	173
5.3.3 *GAS11* Gene Structure	178
5.3.4 Comparison of *GAS11* and its Mouse Homologue	178
5.3.5 Analysis of zb57b10	180
5.3.6 Amino Acid Identity	182
5.3.7 Sequence Determination and Characterisation of *C16orf3*	182
5.3.7.1 5' Sequence Identification of C16orf3	183
5.3.7.2 Allele Frequencies of C16orf3	186

5.3.8 Analysis of ET27.12 186
5.3.8.1 Orientation of ET27.12 186
5.3.8.2 Linking ET27.12 to C16orf3 188
5.3.8.3 3' RACE 188
5.3.9 Mutation Analysis of *GAS11* 189
5.3.9.1 SSCP Analysis 189
5.3.9.2 Homozygous Deletions 189
5.3.9.3 Exon Skipping 191
5.3.10 Mutation Analysis of *C16orf3* 191

5.4 Discussion **191**
5.4.1 *GAS11* 191
5.4.2 *C16orf3* 194

5.1 Introduction

During exon amplification from cosmids mapping to the minimum region of LOH at 16q24.3, dbEST homology searches identified a trapped exon with significant homology to a mouse transcript described as a growth arrest-specific (*GAS*) gene. As their name implies, these genes were first identified and isolated in mice based on their preferential expression when cell division in culture was stopped by serum deprivation or growth to confluence (Schneider *et al.*, 1988). Their isolation was achieved through subtractive hybridisation, where a cDNA library specific for the arrested stage was pre-hybridised with a 10-fold excess of mRNA obtained from NIH 3T3 cells that were actively growing in nutrient rich medium. The resultant library led to the identification of six cDNA clones (*Gas1-6*) which represented genes specifically expressed during the quiescent phase of the cell cycle (G_0). These genes were likely to be functionally distinct rather than representing a family of related proteins. In principle, *Gas* genes could encode products that cause arrested cell growth and are therefore likely tumour suppressor gene candidates.

Studies of the mouse *Gas1* gene have shown it codes for a membrane-associated protein that accumulates at the surface of cells following growth arrest and is then down regulated during the G_0-S phase transition. Down regulation of its expression has been correlated with an increase in the expression of the *c-myc* gene following the addition of serum to quiescent NIH 3T3 cells, and similarly, as cells enter G_0 following serum deprivation, *Gas1* mRNA levels increase while those of *c-myc* decrease (Del Sal *et al.*, 1992). In addition, the down-regulation of *Gas1* was also observed in NIH 3T3 cells that had been transfected *in vitro* with activated *h-ras* and *k-ras* oncogenes when compared to untransfected cells (Cairo *et al.*, 1992). Taken together, the membrane protein encoded by *Gas1* may suppress growth with its down-

regulation resulting in cell cycle progression. The human homologue of this gene has been isolated and shares 82% amino acid identity with mouse *Gas1*, with the domains typical for an integral membrane protein being conserved (Del Sal *et al.*, 1994). Again, the up regulation of *GAS1* was observed in human fibroblasts arrested for growth by either serum deprivation or contact inhibition, while the levels decreased significantly during exponential growth. The finding that this gene maps to human chromosome 9q21.3-q22 (Evdokiou *et al.*, 1993), a region commonly deleted, or otherwise rearranged in certain myeloid malignancies (Sreekantaiah *et al.*, 1989), coupled with its ability to inhibit DNA synthesis, suggest *GAS1* may be a potential tumour suppressor gene.

The expression of *Gas2* mRNA is also abundant during growth arrest, but is subsequently down regulated upon reintroduction into the cell cycle by serum stimulation or dilution (Schneider *et al.*, 1988). The gene encodes a cytoplasmic protein that colocalises with the microfilament system, and upon serum stimulation of quiescent cells, is hyperphosphorylated and becomes more abundant at the cell border (Brancolini and Schneider, 1994). An important post-translational modification of the *Gas2* protein is observed during apoptosis, with antibodies specific for the carboxy- or amino-terminal ends of the protein indicating that the carboxy-terminal end is removed. Subsequent overexpression of the carboxy-terminal truncated forms gave rise to dramatic alterations in the actin cytoskeleton and in cell shape, changes which are normally encountered during cell death by apoptosis (Brancolini *et al.*, 1995). A consensus binding site for the interleukin-1β-converting enzyme (ICE)-like protease, a gene which is activated during apoptosis (Kumar, 1995), exists at the carboxy-terminal end of the *Gas2* gene, providing further evidence for a possible regulatory role of *Gas2* in microfilament dynamics, during both cell cycle and apoptosis. The human *GAS2* gene has subsequently been isolated and the amino acid sequence has been shown to be extremely

conserved between the two species, with only 8/314 residues different, 7 of which represented conservative substitutions. The human homologue has also been shown to share functional homology to murine *Gas2*, with a similar pattern of tissue distribution, similar intracellular localisation, and identical proteolytic processing during apoptosis (Collavin *et al.*, 1998). This gene has been mapped to human chromosome 11p14.3-p15.2, a region demonstrated to contain tumour suppressor loci using subchromosomal fragments (Koi *et al.*, 1993), with deletions or rearrangements of the short arm of chromosome 11 also observed in human tumours (Mitelman *et al.*, 1997). In addition, a related human gene, *GAR22*, maps to a region on chromosome 22 which shows LOH in a variety of tumours (Zucman-Rossi *et al.*, 1996). This suggests that *GAS2* may be a candidate tumour suppressor gene that is part of a gene family, with inactivation of these genes disrupting the apoptotic response of certain cells to specific signals, possibly allowing cells with aberrations to continue dividing and become cancerous.

The *Gas3* gene codes for a 144 amino acid protein that consists of three putative transmembrane domains, with *in vitro* translation products of *Gas3* mRNA showing it associates with the detergent TX-114, indicating it is a true integral membrane protein (Manfioletti *et al.*, 1990). The protein of the rat homologue, *SR13*, has been localised to the fully-differentiated and quiescent Schwann cells that constitute the myelin sheath of sciatic nerves, and following nerve injury, *SR13* mRNA is rapidly down regulated, coinciding with the time of Schwann cell proliferation (Welcher *et al.*, 1991). *PMP22*, the human homologue to these genes, maps to chromosome 17p11.2 and has been shown to be duplicated in Charcot-Marie-Tooth disease type 1A (CMT1), a disease which is characterised by distal muscle wasting and weakness, severely slowed nerve conduction and demyelination in nerve biopsies. Altered expression of the *PMP22* gene may therefore affect Schwann cell proliferation giving

rise to these clinically observed features.

The biological function of both the *Gas4* and *Gas5* genes is presently unclear. However the *Gas6* gene and its human homologue share similar patterns of expression and have been shown to possess a high degree of homology with each other, and significant similarity to human protein S, a negative coregulator in the blood coagulation pathway, suggesting it is a member of the same family (Manfioletti *et al.*, 1993). The product of *Gas6* has been shown to be the ligand for the Sky receptor tyrosine kinase and can stimulate tyrosine phosphorylation of this receptor (Ohashi *et al.*, 1995). It is also a ligand for the Axl tyrosine kinase adhesion receptor which is closely related to Sky (Stitt *et al.*, 1995; Varnum *et al.*, 1995). Receptor tyrosine kinases play a major role in signal transduction pathways that lead to cell proliferation and differentiation suggesting *Gas6* may play an important role in these processes.

Subsequent to the initial identification of *Gas1-Gas6* by Schneider *et al* (1988), three additional *Gas* genes have been detected (Brenner *et al.*, 1989; Lih *et al.*, 1996). One of these genes has since been shown to encode the platelet-derived growth factor α-receptor protein, PDGFαR. Analysis showed that synthesis of mRNA encoding PDGFαR in NIH 3T3 cells was induced by growth arrest, confirming that it is a *Gas* gene (Lih *et al.*, 1996). Expression of the PDGFαR gene product during G_0 may therefore lead to an accumulation of this protein on the surface of growth arrested cells, consequently advancing the cell cycle. This shows that not all products of *Gas* genes may have a negative effect on the growth of cells but some may act in conjunction with other growth factors to promote cell cycling.

The mouse *Gas* gene identified from exon trapping does not seem to correspond to any of the previously characterised genes and is an unpublished reference in GenBank. The human

homologue of this gene will therefore be referred to as *GAS11*, named in numerical order to previous *GAS* genes present in GDB. Based on the functional characterisation of other *GAS* genes, this new transcript was considered a potential breast cancer tumour suppressor candidate. To determine if *GAS11* has a possible role in breast carcinogenesis, the gene must be screened for mutations in breast tumour material isolated from individuals showing restricted LOH at 16q24.3. The following steps will be undertaken to allow this screening to proceed. Firstly, Northern blot analysis using probes based on partial gene sequences will determine the size of the messenger RNA encoded by *GAS11*. The complete nucleotide sequence of the gene will then be obtained from the sequencing of overlapping cDNA clones and use of the 5' RACE procedure that aids in the isolation of the 5' end of genes. The determination of the full-length sequence is essential for the prediction of the protein product of the transcript, which may provide functional information about the gene. The genomic organisation of *GAS11* will then be determined in order to amplify each exon of the gene to allow screening for DNA mutations in affected individuals using single stranded conformation polymorphism (SSCP) analysis. Possible gross deletions of *GAS11* in breast cancer tumours will also be determined by Southern analysis of breast cancer cell line DNA by hybridisation with a cDNA probe. Exon skipping is another possible mechanism of mutation that will also be explored in patient material and cell lines. The characterisation and mutation analysis of another transcript, *C16orf3* (chromosome 16 open reading frame 3), that overlaps with *GAS11*, will also be presented in this chapter.

5.2 Methods

Only those procedures performed specifically in this chapter are mentioned in detail. More general techniques are described in detail in chapter 2, however slight modifications are

referred to when appropriate in the text.

5.2.1 *GAS11* Sequence Assembly

Clones showing homology to the trapped exons ET27.11 (clone zb57b10) and ET27.12 (clones yg15d07 and zm05d03) were purchased from Genome Systems and DNA was isolated as described in 2.2.1.3. Inserts were excised from their respective vectors and used as probes on Southern blots of digested cosmid DNA to verify their physical map location with respect to the cosmid contig. These methods are described in detail in chapter 2. Sequencing of cDNA clones was performed by DyeTerminator and DyePrimer chemistry (2.2.18) using the primers listed in Table 5.1.

5.2.2 Isolation of 5' End Sequences

5' RACE was performed as described in 2.2.20. For *GAS11*, fetal skeletal muscle total RNA was used as a template with *GAS11* gene specific primers (GasGSP) as listed in Table 5.2. Two separate GasGSP3 primers were used, GasGSP3A, located in exon 2, and GasGSP3B, present within the extra 109 bp belonging to intron 2 of *GAS11*. For *C16orf3*, polyA$^+$ mRNA isolated from fetal skeletal muscle and total RNA isolated from fetal heart tissue were used as templates. *C16orf3* gene specific primers (C16GSP) are listed in Table 5.2.

PCR screening of the fetal brain cDNA pools was also used to extend the *GAS11* sequence. Standard PCR reactions were performed (2.2.14.1) using 2 µl of each phage pool as a template in 20 µl reactions. However, before the addition of the *Taq* DNA polymerase, the tubes containing all components, were overlayed with a drop of paraffin oil, and incubated at 96°C for 5 minutes. The tubes were then incubated at 80°C while the enzyme was added to each reaction. Following this, all tubes were incubated at 94°C for 4 minutes, then 35 cycles of

TABLE 5.1

Sequences of Primers used for Transcript Sequence Assembly

Primer name	Primer sequence (5' – 3')	Primer position
***GAS11* Sequence Identification**		
T3	ATT AAC CCT CAC TAA AGG GA	Vector
T7	TAA TAC GAC TCA CTA TAG GG	Vector
M13F	TGT AAA ACG ACG GCC AGT	Vector
M13R	CAG GAA ACA GCT ATG ACC	Vector
W1	TCT TGG GAG ATC ACA CGG AG	bases 307-325
W2	TGC ATC TCG CTC ATC TCC TC	bases 1003-984
W3	AGG AGA TGA GCG AGA TGC AG	bases 985-1004
R1	TGT GAT CGG ACA GAC GCT GG	bases 1502-1521
R2	TCA CCT CCG ACC ACA GTC G	bases 1813-1831
R3	GAC TCA AGA ACA TCC TCC AG	bases 1357-1338
R4	CTG CCA GTC TTC GCT CTA AC	bases 2114-2133
AA1	GTC CCT AAA CTG CTC TGG AAC	bases 3005-2985
G1	ACA TGC GGG CAC TGA AGG TG	bases 544-563
27.12A	CTA CAC AGA CCC AGA TGG TC	bases 2838-2819
27.12B	CTC CAC ATG TGG GTA GAA GC	bases 2717-2736
***GAS11* Sequence Confirmation**		
MgasF	TGA CCT GCT GCG TAC ATA TG	bases 1432-1451 (U19859)
MgasR	TCA TTG CTG ACG GGT AAG TC	bases 1681-1662 (U19859)
G17	GAA GAA AGG CAA AGC CAA AGG	bases 143-163
G8	AGC TGC CTC CGT GTG ATC TC	bases 331-312
G6	TCA GGT GAG GCA CTG AGC	within intron 2
***C16orf3* Mapping, Sequence Assembly, Allele Frequency Determination**		
C16F	ACA ACG CAG TGT TCA ATC CA	bases 871-890
C16R	TTT TAC ATG CAG GCA AAC CA	bases 997-978
C16A	TCT GTT TGG CTC ACA TGG TG	bases 414-433
C16B	TGG CTC AAC GTG CGA TGT GC	bases 650-669

Note. The primer positions are based on the sequence of the *GAS11* (Accession number AF050079) and *C16orf3* genes (Accession number AF050081-large allele) presented in Figure 5.3 and 5.7. Vector refers to the vector that the *GAS11* cDNA insert was cloned into. Numbers in brackets refer to the GenBank accession number of the sequence for which the primers were designed.

TABLE 5.2

Sequences of Primers used for Transcript End-Sequence Identification

Primer name	Primer sequence (5' – 3')	Primer position
***GAS11* 5' Sequence Identification**		
GasGSP1	GCT TGT ACA CCT TGA TCT CC	bases 420-401
GasGSP2	AGC TGC CTC CGT GTG ATC TC	bases 331-312
GasGSP3A	TGC TCC TTG CTC ATG TCC TC	bases 211-192
GasGSP3B	TCA GGT GAG GCA CTG AGC	within intron 2
λGT11R	TTG ACA CCA GAC CAA CTG GTA ATG	Vector
W4	CCT TTG GCT TTG CCT TTC TTC	bases 163-143
***C16orf3* 5' Sequence Identification**		
C16GSP1(C16R)	TTT TAC ATG CAG GCA AAC CA	bases 997-978
C16GSP2	AGT GGC TAA GCA GAG AGG TG	bases 962-943
C16GSP3	GGA AAG ATG TCT GGA GCT GC	bases 921-902
ET27.12 3' RACE		
27.12GSP1[a]	CTA CAC AGA CCC AGA TGG TC	bases 2838-2819
27.12GSP2	AGA TGA AGG CAG GCA CCT AC	bases 2805-2786
27.12GSP3	TCG ACA CTG AGG GCT ACT AG	bases 2775-2756
AP1	CCA TCC TAA TAC GAC TCA CTA TAG GGC	Nested primer
AP2	ACT CAC TAT AGG GCT CGA GCG GC	Nested primer

Note. The primer positions are based on the sequence of the *GAS11* (Accession number AF050079) and *C16orf3* genes (Accession number AF050081-large allele) presented in Figure 5.3 and 5.7. Vector refers to the vector that the fetal brain cDNA library was cloned into. [a] This primer was biotinylated at its 5' end.

94°C for 1 minute; 60°C for 2 minutes; 72°C for 3 minutes, with a final incubation of 72°C for 7 minutes. Primers used for the PCR were a primer specific for the λgt11 vector cloning site (λgt11R), and a primer to the 5' end of zb57b10 (W4), which are shown in Table 5.2. All products were analysed on 2.5% agarose gels with aliquots subsequently cloned into the pGEM-T vector and sequenced.

5.2.3 Database Homology Searches

The BLASTN and BLASTP programs (Altschul *et al.*, 1997) were used to search for nucleotide sequence homology and amino acid homology respectively, between *GAS11* and *C16orf3* and sequences deposited in the GenBank non-redundant and EST databases (http://www.ncbi.nlm.nih.gov/index.html). In the majority of cases, the critical P value indicating significant homology was taken to be $\leq 10^{-5}$ (1.0e-5).

5.2.4 Northern Hybridisations

Commercial Northern blot filters (Clontech) were probed sequentially with the purified insert of the cDNA clone zb57b10, which represented the *GAS11* gene, and a PCR product generated from primers designed to the cDNA clone representing *C16orf3* (yd83e07), as described in 5.2.6. A Northern blot produced from breast cancer cell line RNA (2.2.11) was probed with the insert of the cDNA clone yg15d07, representing the *GAS11* gene. This filter was not probed with the β-Actin control as it was still being used for other purposes not reported here.

5.2.5 Genomic Structure Characterisation of *GAS11*

Three methods were chosen as described below, with all primer sequences listed in Table 5.3.

TABLE 5.3

Sequences of Primers used for the Identification of *GAS11* Gene Structure

Exon	Primer name	Primer sequence (5' – 3')	Method of identification
1 3' SD	GN	GGG CGC CCT GGT CCC GGA GTC	Seq. 2.4 kb *Bam*HI fragment
2 5' SA	G18	TGC TCC TTG CTC ATG TCC TC	Seq. 3.1 kb *Bam*HI fragment
2 3' SD	G17	GAA GAA AGG CAA AGC CAA AGG	Seq. 3.1 kb *Bam*HI fragment
3 5' SA	PUCR	TGT GAG CGG ATA ACA ATT TCA CAC AGG A	Seq. 0.7 kb *Bam*HI fragment
3 3' SD	W1	TCT TGG GAG ATC ACA CGG AG	Genomic PCR
4 5' SA	G9A	TCA CTG GCC AGC TCC TGC TC	Genomic PCR
4 3' SD	G9	TGA CAG AGA TGA AGG CTG AG	Seq. 2.1 kb *Bam*HI fragment
5 5' SA	GC	CTC GAA CTT GCC TCT CAA AAT C	Seq. 3.1 kb *Bam*HI fragment
5 3' SD	GB	CAC ACC GAG GAG ATC ACC AG	Seq. 3.1 kb *Bam*HI fragment
6 5' SA	G10	CGA GTT CGT CCC TCA GCA TC	Direct cosmid seq.
6 3' SD	G11	GAT GCT GAG GGA CGA ACT CG	Genomic PCR
7 5' SA	W2	TGC ATC TCG CTC ATC TCC TC	Genomic PCR
7 3' SD	W3	AGG AGA TGA GCG AGA TGC AG	Genomic PCR
8 5' SA	N1	GTG AAT CGC TGC TCT AAC AC	Genomic PCR
8 3' SD	G14	GTG TTA GAG CAG CGA TTC AC	Seq. 9.5 kb *Bam*HI fragment
9 5' SA	GM	ACT GAA GAC TCT TGC CAC AG	Seq. 9.5 kb *Bam*HI fragment
9 3' SD	G21	TCT AAC CTG GAC CCT GCA GC	Seq. 9.5 kb *Bam*HI fragment
10 5' SA	GL	TGG TGA ACC GTT GCT CTA AC	Seq. 9.5 kb *Bam*HI fragment
10 3' SD	G15	AGT CGA AGA ACA GCA CCA TC	Genomic PCR
11 5' SA	G9R	TTC CTA GGC TGG TTC TCT GG	Genomic PCR

Note. SA: splice acceptor site; SD: splice donor site.

5.2.5.1 Direct Cosmid Sequencing

The entire *GAS11* gene lies within the cosmid 344G2 and as such, DNA from this cosmid was chosen as a template for direct sequencing using primers designed to the sequence of the gene. This method has been described in Ianzano *et al.* (1997). Ten micrograms of cosmid DNA was first denatured in 2 M NaOH, 2 mM EDTA (pH 8.0) in a 100 µl volume for 30 minutes at 37°C. The DNA was then precipitated with 10 µl of 3 M NaAc (pH 5.2), and 250 µl of 100% ethanol at –70°C for 30 minutes. Following a 30 minute spin at 13,000 rpm, the pellet was washed in 1 ml of 70% ethanol, respun for 5 minutes, and allowed to air-dry for 10 minutes. The DNA was resuspended in 10 µl of sterile water and a standard DyeTerminator cycle sequencing reaction was performed (2.2.18.1), however 4 µl of halfTERM mix was substituted for another 4 µl of the Dye Terminator mix. Samples were then cleaned as described in 2.2.18.1.

5.2.5.2 PCR Amplification of cDNA and Genomic DNA

Primers specific for the *GAS11* coding sequence were designed 100 to 150 bp apart and used in PCR reactions on human genomic DNA or c344G2 DNA. PCR products generated were compared to the products obtained when cDNA was used as a template in the PCR reaction. If the products were larger with genomic DNA as template than cDNA, they were subsequently purified (2.2.16.1) or cloned into pGEM-T and sequenced (2.2.17.2).

5.2.5.3 Sequencing c344G2 BamHI Subcloned DNA

Individual *Bam*HI fragments from cosmid 344G2 were purified from agarose gels (2.2.16.2) and subcloned into the Puc19 vector (2.2.17). Each subclone then served as a template for DyeTerminator sequencing reactions using primers specific for the cDNA sequence of the

gene. In each of the three procedures mentioned, the divergence of the cDNA sequence to genomic sequence was taken to indicate the exon to intron transition.

5.2.6 PCR

All PCR reactions were performed essentially as described in 2.2.14.1. Amplification of the 3' end of the mouse *Gas11* gene was achieved using 200 ng of mouse A9 cell line DNA as a template with the MgasF and MgasR primers (Table 5.1). The product was cloned into the pGEM-T vector and sequenced.

Amplification of the *C16orf3* gene to generate a product for Northern analysis was achieved using the C16A and C16R primers (Table 5.1) with 50 ng of c344G2 DNA as a template. Reaction volumes were scaled up to 50 µl and a 45 µl aliquot of the PCR reaction was purified using QIAquick columns (2.2.16.1).

The frequency of the large and small alleles of *C16orf3* were determined using PCR amplification of DNA isolated from normal Caucasian individuals using the C16R and C16B primers (Table 5.1).

5.2.7 RT-PCR

All RT-PCR reactions were performed as described in 2.2.15. The sequences of all primers used are shown in Table 5.1. In the cases where products were to be sequenced, an aliquot of the reaction was subcloned into the pGEM-T vector and sequenced using vector specific primers.

5.2.8 R*Tth* PCR

This technique allowed reverse transcription to occur at 70°C, due to the reverse transcriptase activity of *Thermus thermophilus* DNA polymerase in the presence of Mn^{++} ions. This enabled a gene specific primer to be used to initiate cDNA synthesis, and facilitated the denaturing of any secondary structure of the RNA. All reaction buffers were supplied with the r*Tth* kit. In a 20 µl reaction, 2 µl of 10X RT buffer, 2 µl of $MnCl_2$, 1.6 µl of 10 mM dNTPs, 100 ng of gene specific primer 1, 250 ng of $polyA^+$ fetal brain mRNA, and 2.5 units of r*Tth* polymerase, were mixed and incubated at 70°C for 25 minutes. Following this, 8 µl of 10X chelating buffer, 6 µl of 25 mM $MgCl_2$, and 100 ng of gene specific primer 2 were added to the initial 20 µl reaction and made up to a volume of 100 µl with sterile water. After addition of two drops of paraffin oil, the tubes were incubated at 95°C for 2 minutes, followed by 35 cycles of 95°C for 30 seconds; 60°C for 30 seconds; 72°C for 30 seconds, with a final elongation step of 72°C for 7 minutes. Following this, 20 µl of each sample was electrophoresed through a 2.5% (w/v) agarose gel. A control reaction was performed for each individual experiment where the r*Tth* enzyme was omitted from the initial reverse transcription step but was added for the PCR step. This established whether PCR products were derived from genomic contamination from within the RNA template. A positive control provided with the kit was also used. All primer sequences used are shown in Table 5.1.

5.2.9 3' RACE

A procedure adapted from Gecz *et al.* (1997) was used with primers specific for ET27.12. First strand cDNA was synthesised from $polyA^+$ fetal brain mRNA using an anchored oligo-dT primer, 5'-CCATCCTAATACGACTCACTATAGGGCT-CGAGCGGC$(T)_{18}$VN-3', in which V was any of the four A, C, G, and T nucleotides and N was C, G, or A. Second strand cDNA

was synthesised in a linear PCR using a biotinylated gene specific primer (27.12GSP1). ET27.12 specific cDNAs were then captured on streptavidin-coated magnetic beads (Dynal), and the 3' end was subsequently amplified using the nested primers 27.12GSP2 and AP1, followed by an additional nested PCR using the primers 27.12GSP3 and AP2. Resulting PCR products were then cloned into the pGEM-T vector and sequenced. All primer sequences are listed in Table 5.2.

5.2.10 Mutation Analysis

5.2.10.1 Single Stranded Conformation Polymorphism (SSCP) Analysis

Other members of the FAB consortium performed SSCP mutation analysis on the *GAS11* and *C16orf3* genes. Individual laboratory variations in procedures employed exist, but were essentially as described in chapter 6 (6.2.5). However, only a subset of tumours from the same panel of breast cancer samples as described in 6.2.6 were used (see Table 5.7), with the LOH status of the majority of them shown in Figure 1.4. Primers used for the SSCP analysis of both these genes are listed in Table 5.4.

5.2.10.2 Screening for Homozygous Deletions

DNA from selected breast cancer cell lines (T47D, MDA134/1, MDA231/51, MCF7, and MB157) and normal male and female individuals was digested with *Pst*I, *Bam*HI, *Hind*III, or *Eco*RI in separate tubes (2.2.7.1). After Southern transfer of the electrophoresed digests, membranes were hybridised with the insert to the cDNA clone ze60g08 that contains the entire *GAS11* coding sequence and 3' UTR.

5.2.10.3 Exon Skipping

Selected primers listed in Table 5.1 and 5.3 were used in RT-PCR reactions on RNA isolated

TABLE 5.4

Oligonucleotide Primers for SSCP Analysis of *GAS11* and *C16orf3*

Primer set	Nucleotide sequences (5'-3')	Size (bp)
GAS11		
Exon 1	CGT GGA GCT TAG GTT CCA GC TTC CCA GGG TCT CCG TGT AC	368
Exon 2	AAT CAG AGC CAG CCA AGA AG TAG GCA CCC TGC GTG GTG	216
Exon 3	TCC TCC CTC TTT CTT CCT TG CTC TGT GGT GAC CAG GTT AG	340
Exon 4	CAG GTG AGC AGA TGG TAC TC AAC AAC CTG TCA GCC TCC AG	280
Exon 5	GAG TTA GGG AAT ATG GAG GTG CCT GAG TTA ACA AGC TCT GG	257
Exon 6	TCC ATG TTC TGT AGC CGT AG TGC AGC CCA CAT TTT AGG AC	334
Exon 7	CTT TGG TTC TGC CTG CTG AG TGA TGC TCC TCC CTC CTG AG	288
Exon 8	CTT GGA TGA CTG TGA ATC TTG TGG TGA ACC GTT GCT CTA AC	203
Exon 9	CAT CTT CCC AGC GAG ATG TC ACT GAA GAC TCT TGC CAC AG	340
Exon 10	ATC AAA CCA GGC TGC TAT TC TTG CAT TTA GAA CAG TCT GG	247
Exon 11	CTT TAT CCT TCT GCA CCG TG TTC CTA GGC TGG TTC TCT GG	307
C16orf3		
Set 1	TCT GTT TGG CTC ACA TGG TG GCA CTG CAC ATC GCA CGT TG	261
Set 2	TGG ACC CTG AAG AAG CCT TC TTT TAC ATG CAG GCA AAC CA	352[a], 376[b]

[a] Size of product generated from the small allele
[b] Size of product generated from the large allele

from selected breast cancer cell lines to test for the presence of exon skipping in the *GAS11* gene. The breast cancer cell lines examined were BT20, MDA-453, ZR-75, BT-474, BT-549, ZR-75-30, UACC893, MDA-468, and MDA-231. The normal breast epithelial cell line, HBL-100, and fetal skeletal muscle RNA were positive controls. A total of 4 overlapping primer sets were used, G9R/W3, N1/G1, G10/W1, and G9A/GN with products being run on 2.5% agarose gels. In addition Dr. Anne-Marie Cleton-Jansen generated RT-PCR products from the same panel of tumours as used for SSCP analysis using overlapping primer sets to the *GAS11* coding sequence.

5.3 Results

5.3.1 *GAS11* Sequence Determination

5.3.1.1 Sequencing cDNA Clones

Based on the results of chapter 4, the trapped exons ET27.11 and ET27.12 are situated approximately 20 kb from each other and were therefore suspected to be part of the same transcript. This possibility was strengthened when detailed analysis of dbEST BLAST results indicated that the cDNA clone showing homology to ET27.11 (zb57b10) shared the same 3' origin as the cDNA clone with homology to ET27.12 (zm05d03), however they had different 5' end sequences. An additional cDNA clone, yg15d07, also shared the same 3' and 5' ends as zm05d03 but appeared to extend the sequence 5' by 86 bp, and for this reason this clone was chosen to be sequenced, along with zb57b10. The full sequence of these two clones established that they were both significantly homologous to a mouse growth arrest-specific cDNA (GenBank accession U19859) again confirming that both ET27.11 and ET27.12 were part of the same transcript. Figure 5.1A shows the extent of overlap between the yg15d07 and zb57b10 cDNA clones, indicating that although they were both homologous to the mouse *Gas*

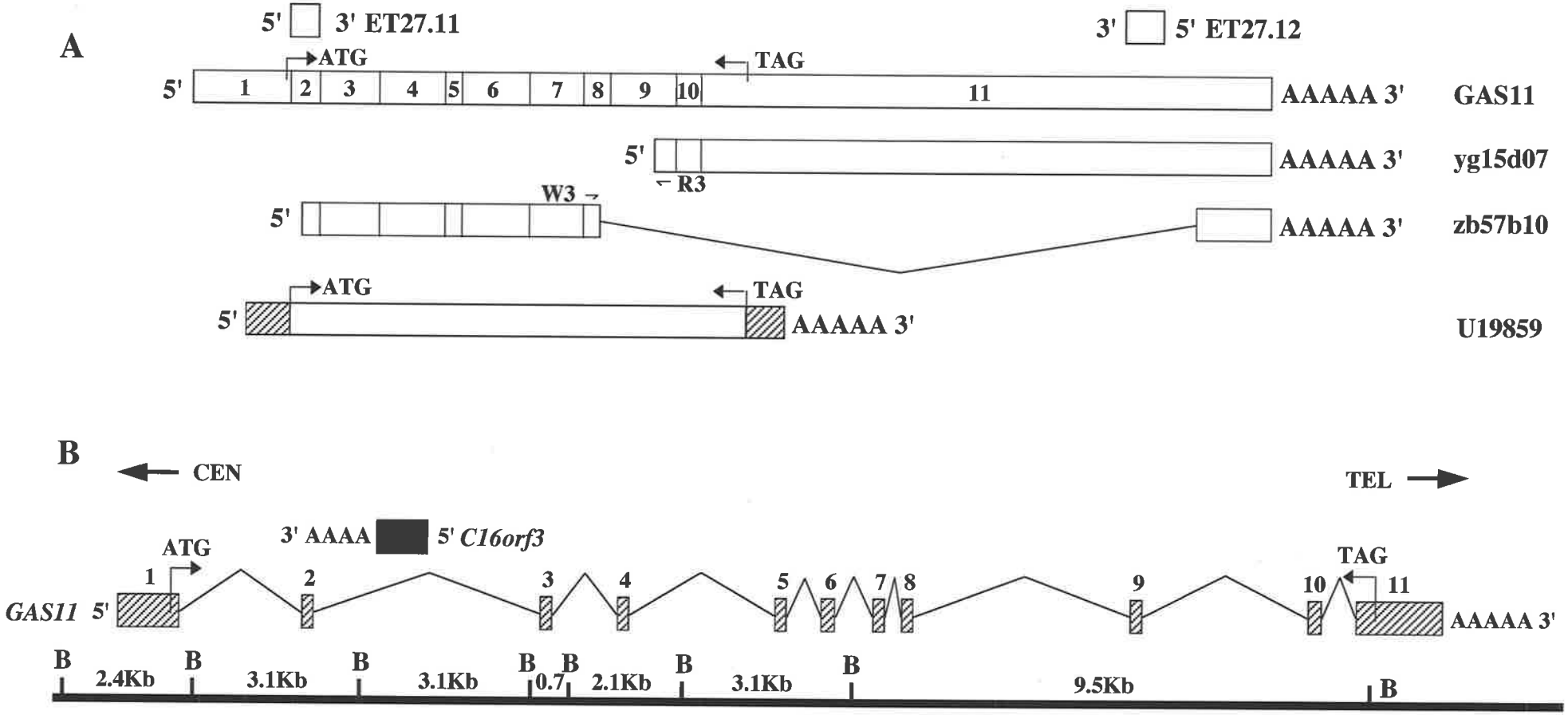


Figure 5.1 : (A) Alignment of the human and mouse (U19859) *GAS11* genes with the cDNA clones yg15d07 and zb57b10, and the trapped exons ET27.11 and ET27.12. Hatched boxes represent the 5' and 3' untranslated regions (UTR) of the mouse *Gas* gene, which are completely different in sequence composition from the human gene. The trapped exon ET27.11 represents exon 2 of *GAS11*, whereas ET27.12 was trapped from the noncoding strand within the 3' UTR. The cDNA clone zb57b10 is missing bases 1,100-2,951 of *GAS11*, which includes part of exon 8, all of exons 9 and 10, and most of the 3' UTR except the last 235 bp. This cDNA clone was joined to the yg15d07 clone by RT-PCR using the primers W3 and R3 (see 5.3.1.1). (B) A partial restriction map of the 16q24.3 region encompassing the *GAS11* and *C16orf3* genes. B: *Bam*HI. The *C16orf3* gene (black box) lies within intron 2 of the *GAS11* gene and is transcribed in the opposite orientation. The distribution of *GAS11* exons across the genomic region is shown by hatched boxes.

U19859 sequence in different regions, they did not overlap other than the last 221 bp at their 3' ends. However, RT-PCR experiments using the W3 and R3 primers on fetal brain RNA were successful in amplifying a product which when sequenced was able to join these two cDNA clones to form one complete sequence.

5.3.1.2 5' Sequence Isolation

Analysis of the cDNA sequence obtained identified an open reading frame (ORF) of 1,431 bp extending from the start of the sequence, followed by a large 3' untranslated region (UTR) of 1,630 bp. The zb57b10 clone only contained up to base number 974 of the ORF which was then attached to the last 235 bp of the 3' UTR, indicating this clone may be a splice variant of the gene. However, based on the homology to the mouse gene, it appeared that the 5' methionine start codon had not been obtained. The application of 5' RACE and PCR screening of the fetal brain cDNA pools, extended the 5' sequence by 71 bp and 57 bp respectively (Figure 5.2A and B). Within the extended 5' sequence an in-frame methionine start codon was identified which as well as being highly homologous to the KOZAK consensus sequence (Kozak, 1991) was identical to the start codon used in the mouse gene. This indicated that *GAS11* has an ORF of 1,434 bp that codes for a protein consisting of 478 amino acids (Figure 5.3). Two polyadenylation sites were identified within the 3' UTR of the gene (http://lita.itba.mi.cnr.it/~webgene/wwwHC_polya.html), the first of these at position 3,155 (TATAAA) and the second at position 1,668 (ATGAAA), both of which do not conform to the most commonly observed signal (AATAAA).

5.3.2 *GAS11* Northern Analysis

The insert of the cDNA clone zb57b10 identified 2 major bands of approximately 3.4 kb and 1.8 kb in all tissues of a commercial Northern blot (Figure 5.4.A). In the heart, placenta, liver,

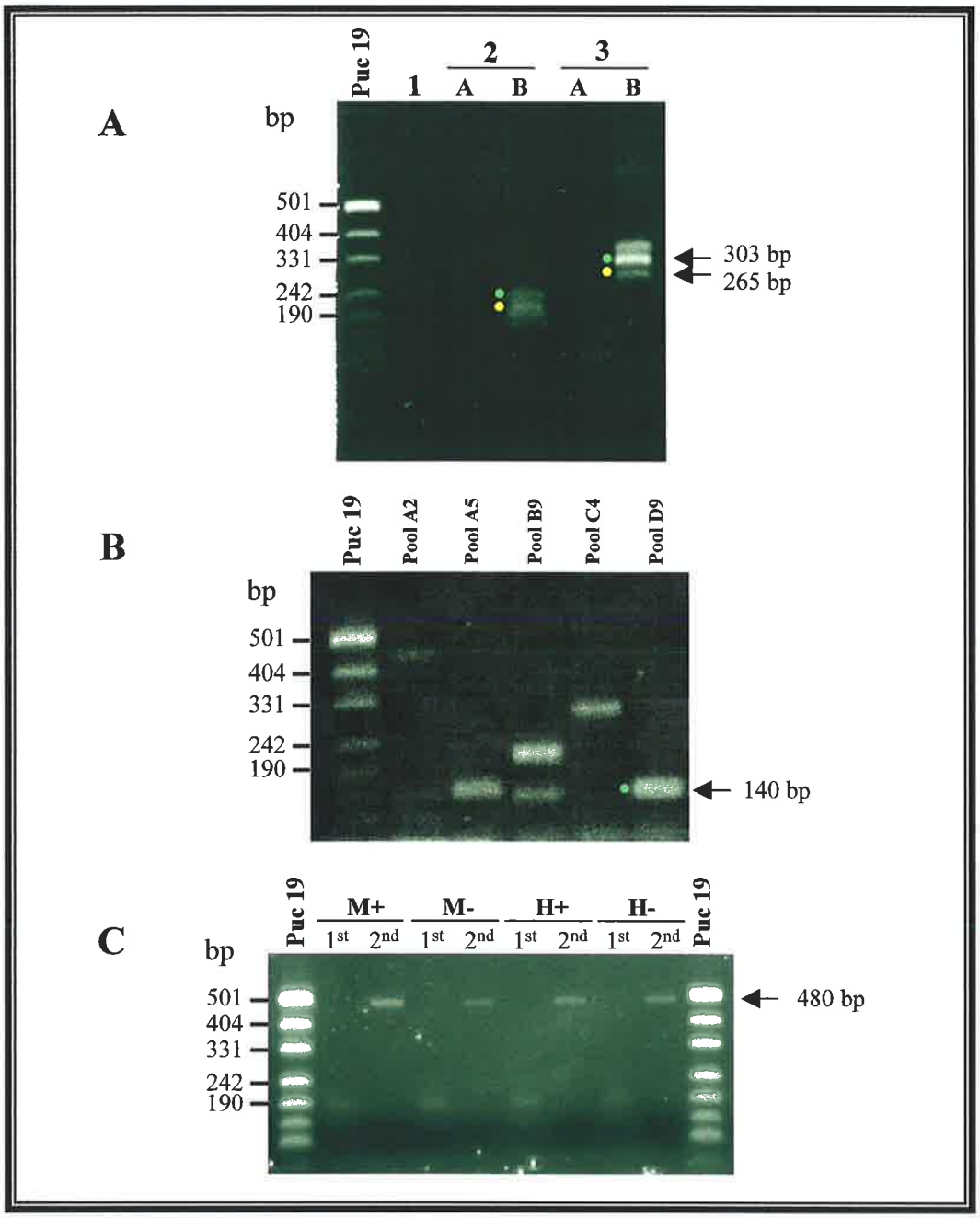
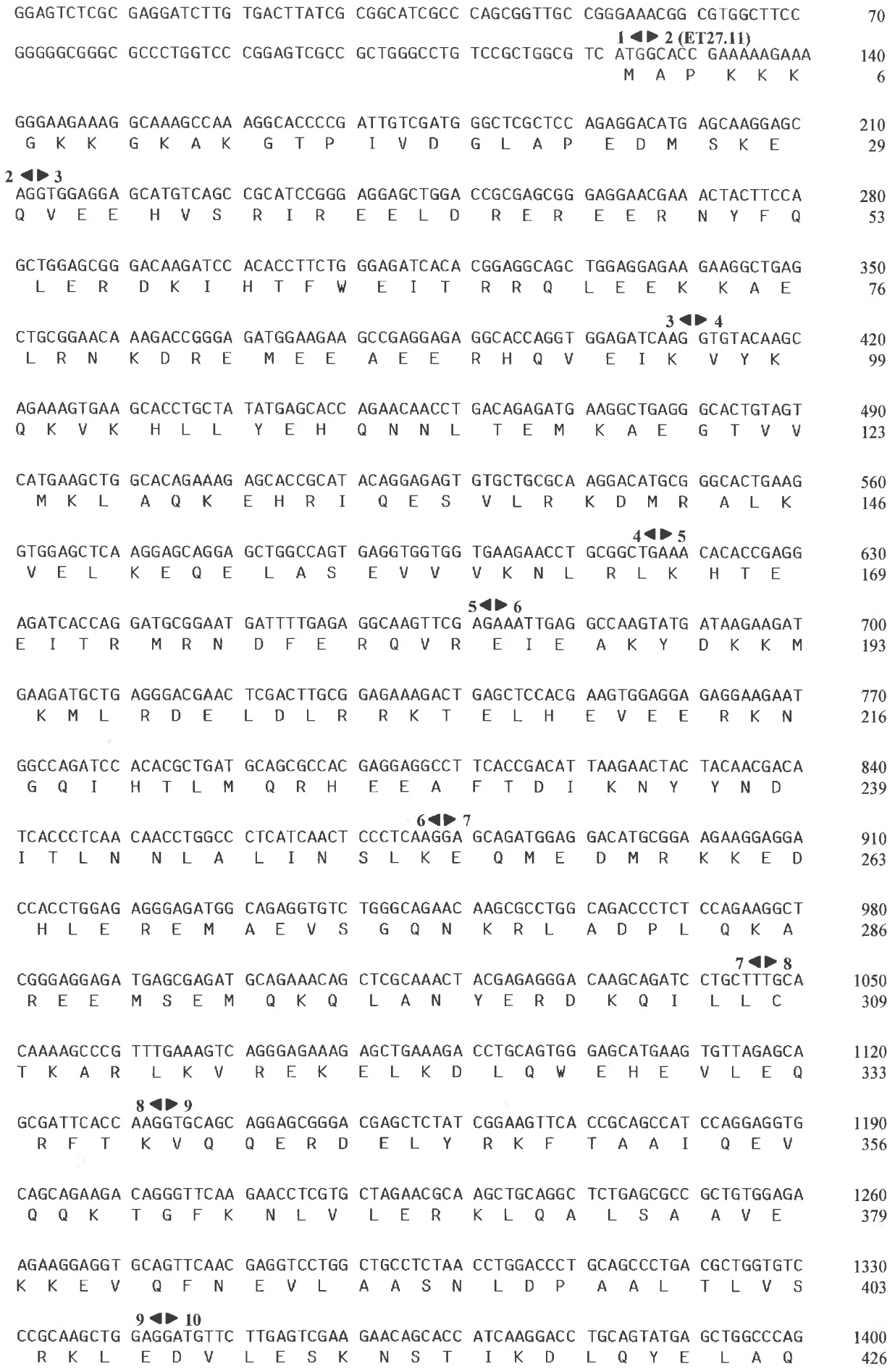


Figure 5.2: Results of 5' sequence extension for *GAS11* and *C16orf3*. (**A**) 5' RACE results for *GAS11*. Lane **1**: 10 ul sample of first PCR with GasGSP2/AAP; **2**: 10 ul sample of second PCR with GasGSP3A/AUAP; **3**: 10 ul sample of second PCR with GasGSP3B/AUAP. Lane **A** in **2** and **3** represents a control reaction where reverse transcriptase was omitted in the cDNA synthesis step. Two products (green and yellow dots) were cloned and sequenced from lane **3B**. The 303 bp fragment extended the *GAS11* sequence by 71 bp. (**B**) Fetal brain cDNA pool screening with a *GAS11* primer and a vector specific primer. Pools shown were those initially identified (from all 40 pools) to amplify PCR products. Sequencing of the 140 bp fragment in pool D9 (green dot) extended the *GAS11* sequence by 57 bp. (**C**) 5' RACE results for *C16orf3*. **M**: fetal muscle RNA with (+) and without (-) reverse transcriptase; **H**: fetal heart RNA with (+) and without (-) reverse transcriptase. **1st** refers to the first PCR with C16GSP2/AAP, whereas **2nd** refers to the second nested PCR with C16GSP3/AUAP. The 480 bp band seen in each of the **2nd** lanes was cloned and sequenced and shown to be derived from the self priming of the AUAP oligo.

Figure 5.3

Nucleotide and deduced amino acid sequence of the *GAS11* gene beginning from the proposed transcription start site. The gene sequence extends over two pages. Numbers in the right hand column indicate the nucleotide and amino acid positions. The translation start site at position 123 and the stop codon at position 1,557 are indicated by green type. The two polyadenylation signals are underlined. Horizontal arrowheads are located above the first and last nucleotide of a given exon. Each exon is represented by a red number with the trapped exon ET27.11 shown in brackets. Nucleotides in red type indicate the position of the trapped exon ET27.12 which was trapped from the opposite strand.


GGAGTCTCGC GAGGATCTTG TGACTTATCG CGGCATCGCC CAGCGGTTGC CGGGAAACGG CGTGGCTTCC 70
1 ◀▶ 2 (ET27.11)
GGGGGCGGGC GCCCTGGTCC CGGAGTCGCC GCTGGGCCTG TCCGCTGGCG TC ATGGCACC GAAAAAGAAA 140
M A P K K K 6
GGGAAGAAAG GCAAAGCCAA AGGCACCCCG ATTGTCGATG GGCTCGCTCC AGAGGACATG AGCAAGGAGC 210
G K K G K A K G T P I V D G L A P E D M S K E 29
2 ◀▶ 3
AGGTGGAGGA GCATGTCAGC CGCATCCGGG AGGAGCTGGA CCGCGAGCGG GAGGAACGAA ACTACTTCCA 280
Q V E E H V S R I R E E L D R E R E E R N Y F Q 53
GCTGGAGCGG GACAAGATCC ACACCTTCTG GGAGATCACA CGGAGGCAGC TGGAGGAGAA GAAGGCTGAG 350
L E R D K I H T F W E I T R R Q L E E K K A E 76
3 ◀▶ 4
CTGCGGAACA AAGACCGGGA GATGGAAGAA GCCGAGGAGA GGCACCAGGT GGAGATCAAG GTGTACAAGC 420
L R N K D R E M E E A E E R H Q V E I K V Y K 99
AGAAAGTGAA GCACCTGCTA TATGAGCACC AGAACAACCT GACAGAGATG AAGGCTGAGG GCACTGTAGT 490
Q K V K H L L Y E H Q N N L T E M K A E G T V V 123
CATGAAGCTG GCACAGAAAG AGCACCGCAT ACAGGAGAGT GTGCTGCGCA AGGACATGCG GGCACTGAAG 560
M K L A Q K E H R I Q E S V L R K D M R A L K 146
4 ◀▶ 5
GTGGAGCTCA AGGAGCAGGA GCTGGCCAGT GAGGTGGTGG TGAAGAACCT GCGGCTGAAA CACACCGAGG 630
V E L K E Q E L A S E V V V K N L R L K H T E 169
5 ◀▶ 6
AGATCACCAG GATGCGGAAT GATTTTGAGA GGCAAGTTCG AGAAATTGAG GCCAAGTATG ATAAGAAGAT 700
E I T R M R N D F E R Q V R E I E A K Y D K K M 193
GAAGATGCTG AGGGACGAAC TCGACTTGCG GAGAAAGACT GAGCTCCACG AAGTGGAGGA GAGGAAGAAT 770
K M L R D E L D L R R K T E L H E V E E R K N 216
GGCCAGATCC ACACGCTGAT GCAGCGCCAC GAGGAGGCCT TCACCGACAT TAAGAACTAC TACAACGACA 840
G Q I H T L M Q R H E E A F T D I K N Y Y N D 239
6 ◀▶ 7
TCACCCTCAA CAACCTGGCC CTCATCAACT CCCTCAAGGA GCAGATGGAG GACATGCGGA AGAAGGAGGA 910
I T L N N L A L I N S L K E Q M E D M R K K E D 263
CCACCTGGAG AGGGAGATGG CAGAGGTGTC TGGGCAGAAC AAGCGCCTGG CAGACCCTCT CCAGAAGGCT 980
H L E R E M A E V S G Q N K R L A D P L Q K A 286
7 ◀▶ 8
CGGGAGGAGA TGAGCGAGAT GCAGAAACAG CTCGCAAACT ACGAGAGGGA CAAGCAGATC CTGCTTTGCA 1050
R E E M S E M Q K Q L A N Y E R D K Q I L L C 309
CAAAAGCCCG TTTGAAAGTC AGGGAGAAAG AGCTGAAAGA CCTGCAGTGG GAGCATGAAG TGTTAGAGCA 1120
T K A R L K V R E K E L K D L Q W E H E V L E Q 333
8 ◀▶ 9
GCGATTCACC AAGGTGCAGC AGGAGCGGGA CGAGCTCTAT CGGAAGTTCA CCGCAGCCAT CCAGGAGGTG 1190
R F T K V Q Q E R D E L Y R K F T A A I Q E V 356
CAGCAGAAGA CAGGGTTCAA GAACCTCGTG CTAGAACGCA AGCTGCAGGC TCTGAGCGCC GCTGTGGAGA 1260
Q Q K T G F K N L V L E R K L Q A L S A A V E 379
AGAAGGAGGT GCAGTTCAAC GAGGTCCTGG CTGCCTCTAA CCTGGACCCT GCAGCCCTGA CGCTGGTGTC 1330
K K E V Q F N E V L A A S N L D P A A L T L V S 403
9 ◀▶ 10
CCGCAAGCTG GAGGATGTTC TTGAGTCGAA GAACAGCACC ATCAAGGACC TGCAGTATGA GCTGGCCCAG 1400
R K L E D V L E S K N S T I K D L Q Y E L A Q 426

10◀▶11

```
GTCTGTAAGG CCCATAACGA CCTGCTGCGC ACGTATGAGG CAAAGCTGCT GGCCTTCGGG ATCCCTCTGG    1470
 V  C  K   A  H  N  D   L  L  R   T  Y  E   A  K  L  L   A  F  G   I  P  L     449

ACAACGTGGG CTTCAAGCCC TTGGAAACAG CTGTGATCGG ACAGACGCTG GGCCAGGGCC CCGCGGGACT    1540
D  N  V  G   F  K  P   L  E  T   A  V  I  G   Q  T  L   G  Q  G   P  A  G  L    473

GGTGGGCACC CCGACGTAGC TGCCCCCCCT GGGGGGCCAC AGCCCAGAGA ACCAGCCTAG GAACACTCGG    1610
  V  G  T   P  T  .                                                             478

GATGACACCC CTTATCACAC CAAGGACAGC AAGTTTTTTA GATTTTATCA TCAGCAAATG AAAGCTTTTC    1680

ACATGTTCTT GCCATCCTCT TTCCTGGCTC TGTGGAGGAG AACCACCTGC AGGACCCTCA CCCATGGTGT    1750

CCCTGTCGCT CCCTTCCCTG GGTGCCGCAC GTCCAGCCTG TGTCCAGGCC TACTCCCTGG TCTCACCTCC    1820

GACCACAGTC GGCGGCACCT TCTCAGAGTG CCCCGCACTC ACCTGGGGGT TGGGGCAATG CCGCGCTGTG    1890

CTGCCTGTCT TCGCGCCACT GTTGTCCCAC CGAATGGACA GCTTTGCAGG TGCTGGCACT AACTTCATTG    1960

ACACCTGAGT CACAGCTGCC CAGTGGGATT CTCCAGGGGG CCGGGACTTC CCTAGGAAGT GGTGAGCCAA    2030

TGCTCCCTGA TGAGCACAAA GCCCGCTCTG TTGAGGGCTG GGTGGGTGCA GCCAGCGTGC GGGAACGGGC    2100

AGGCAGCCTC CCGCTGCCAG TCTTCGCTCT AACTCCCTCG GTAGGTGATG TAGGACCAGG GGCACGTGGA    2170

ACTTCTGGGC CTTGCTGGTG ATGGTTAAAA CAACCTGAGA TGGAGAGGCC AGGAGAGAGT ATAAGGGGAT    2240

AGCAGCAAAC CACCTATCTG GCCCCAACAC ACCTGAGAGA ATTCAGCAGC CCAGACTGAG GGTCTGGGAT    2310

GGGGTGAACC TTCCGCACCA GAGGGACACT CCACAGAAGC CACAGCCCAG TAAGTCAGGC GCTTCTGCGG    2380

CGGCTCCAGT GTGGGGTGAG GCAGTGAGGT TAGGCCCAGA GAGCTGGAGT TGGCTCAGAT GAAAACCTCT    2450

GTCAACAAAG AGGGGATGAA TCACCCTTGG CCCAGCCTCC CCACAAAGCC TGACCCTGGG CAGGTGAGTG    2520

ACGGGTGTGT CCTCGTAGAG TCTATTGCTG CCTGGACACC TTTCTTTTGG GAGCTCAAAG CAAGTGAGCT    2590

CACCTACCTG CCACCGCCCA GGACCAGTCT GCCCACTGCC TAAATGATGC CCGGCCAGCA GGACCTGGCC    2660

TGCAGATCCC AGTGAGTCAT GAGCCTCAGC CCCCTCCAGC CCACTGGGGC TCTCAC CTCC ACATGTGGGT    2730

AGAAGCTTTC CTGCCCCCTC TTCCTCTAGT AGCCCTCAGT GTCGAAGGTG AGCTTGTAGG TGCCTGCCTT    2800

CATCTGGTCC AGGACAGTGA CCATCTGGGT CTGTGTAG CT GGGGAGAGAA TGAGGCTGCA GAGATGGGGA    2870

CCAGAAGCCC CCCACCCCAG CTTTCCTGGG TCTGCATCCC AGTGGGCCTC AGACACTGCC CTGCCACCTG    2940

TCAGACTTGG GTGAGCAGAC ACAGTGAGGC TGTTAGGTCC TGCAGTTCCA GAGCAGTCTA GGGACACCAC    3010

TGCCCTGTCT TTAGGAAATC ACAACACAGA GAAGCAAAAA GGGAAAAAGT CTCCCACAAT TTATCCCATG    3080

AGCAAGAACC ACTTTATAGC TGGCATATAT TTTTCCAGAT TTTCTCTATG CATAAGTATA TTTGTTTAAA    3150

AACTTATAAA GTGGATTATA CTATTTGTAT TGTTTT                                          3186
```

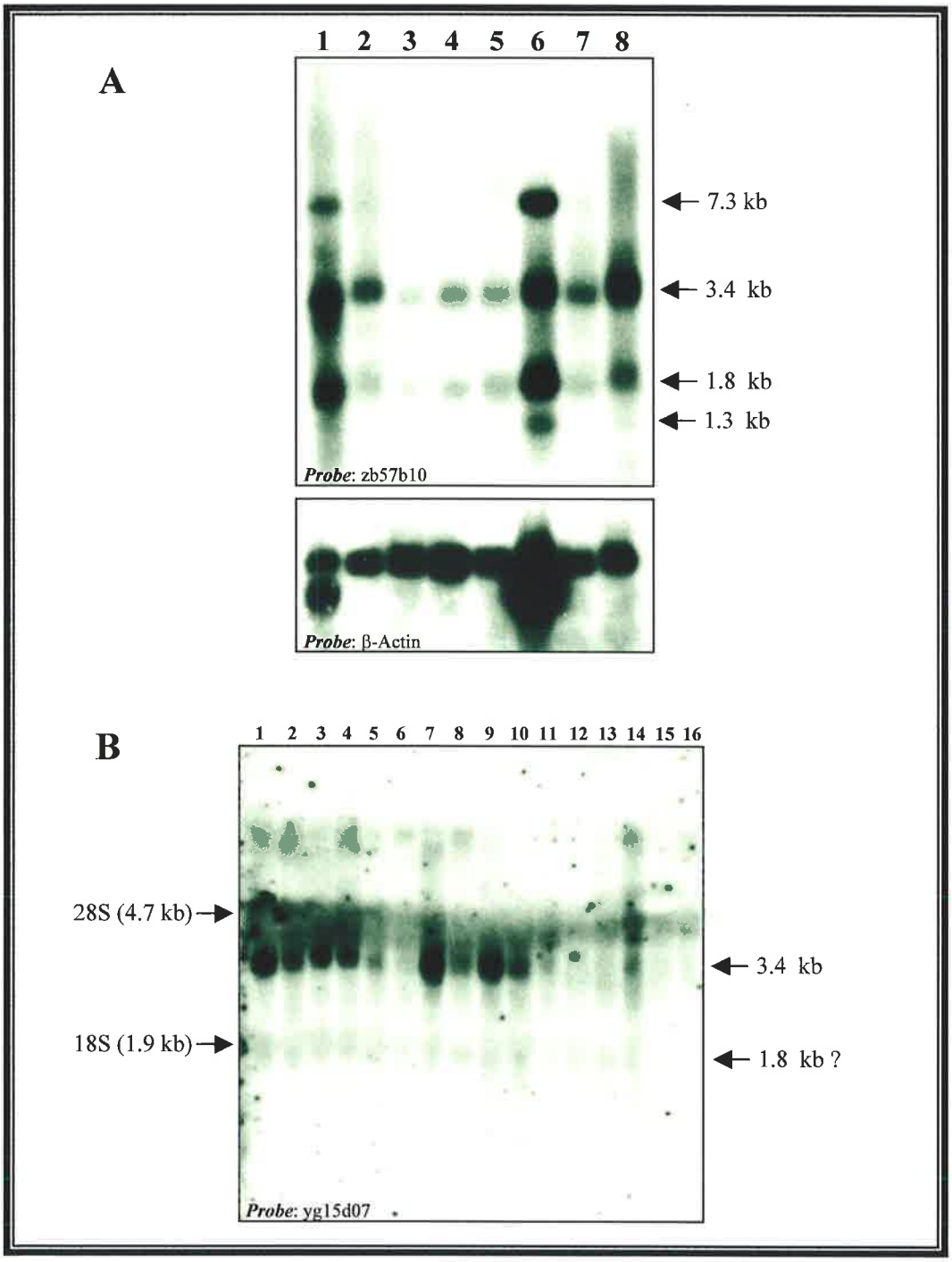


Figure 5.4: Northern analysis of the *GAS11* gene. (**A**) Hybridisation results of a commercial blot consisting of adult tissues probed with the insert to the cDNA clone zb57b10. **1**: Heart; **2**: Brain; **3**: Placenta; **4**: Lung; **5**: Liver; **6**: Skeletal muscle; **7**: Kidney; **8**: Pancreas. The control β-actin hybridisation is shown directly below. (**B**) Hybridisation of the yg15d07 cDNA insert to RNA prepared from selected breast cancer cell lines and fetal tissues. **1**: BT20; **2**: MDA453; **3**: ZR75; **4**: BT474; **5**: BT549; **6**: ZR75-30; **7**: HBL100; **8**: UACC893; **9**: MDA468; **10**: MDA231; **11**: HS578T; **12**: Liver; **13**: Spleen; **14**: Lung; **15**: Heart; **16**: Skeletal muscle. The 28S and 18S ribosomal bands are indicated on the left of the autoradiograph. It is impossible to determine if the 18S ribosomal background is masking the presence of the 1.8 kb mRNA band of the *GAS11* gene.

and skeletal muscle these bands appeared to be expressed at the same levels whereas in the remaining tissues the 3.4 kb transcript was preferentially expressed. A 7.3 kb band was also detected in the heart and skeletal muscle, and an additional band of approximately 1.3 kb was seen in the skeletal muscle alone. Expression of *GAS11* in mammary cells was confirmed by Northern analysis with RNA from selected breast cancer cell lines using the insert of the cDNA clone yg15d07. A band of approximately 3.4 kb was detected after a 10-night exposure with the 1.8 kb band possibly masked by the 18S ribosomal band co-migrating with this mRNA (Figure 5.4.B).

5.3.3 *GAS11* Gene Structure

From genomic structure characterisation, the gene consists of 11 exons separated by 10 introns that collectively cover approximately 25 kb of genomic DNA (Figure 5.1.B). Table 5.5 summarises the sequences of the splice site junctions with all splice sites obeying the ag/gt rule and relative splice site consensus strengths based on Shapiro and Senapathy (1987) indicated. Exon 1 contained the start codon that constituted the last 3 bases of this exon, while exon 11 contained the stop codon as well as the complete 3' UTR. The trapped exon ET27.11 corresponded to exon 2, however ET27.12 was trapped from the opposite strand within the 3' UTR of *GAS11* (bases 2,838 to 2,717).

5.3.4 Comparison of *GAS11* and its Mouse Homologue

Although the human and mouse genes are highly homologous within their coding regions both appeared to use different stop codons. A PCR product was generated from mouse DNA at the region corresponding to the stop codon used by the human gene. Sequencing of this product highlighted errors in the data submitted to GenBank representing the mouse gene (U19859). Six bases were incorrect and there was an extra G residue at base 1,543 of U19859. This

TABLE 5.5

Splice Sites of the *GAS11* Gene

Exon	Size (bp)	3' Splice site (intron/exon)	Consensus strength (%)	5' Splice site (exon/intron)	Consensus strength (%)	Intron size (kb)
1	124	5'UTR		TGGCGTCATG/**gt**gagcaggg	83	3.2
2	87	tccaccaa**ag**/GCACCGAAAA	84.5	CAAGGAGCAG/**gt**gagcagag	90.8	3.6
3	198	acctggcc**ag**/GTGGAGGAGC	80.4	GGAGATCAAG/**gt**gagtgggg	96.7	1.1
4	207	tgccctgc**ag**/GTGTACAAGC	92.8	CCTGCGGCTG/**gt**aggtgtgg	81.3	3.2
5	55	tcctgtgc**ag**/AAACACACCG	93	CAAGTTCGAG/**gt**cagtttcc	89.9	0.62
6	206	gtgctttc**ag**/AAATTGAGGC	79.1	CTCCCTCAAG/**gt**gctgtgcg	66.24	0.69
7	168	gtgcctgc**ag**/GAGCAGATGG	93.4	GATCCTGCTT/**gt**gagtttcc	76	0.362
8	87	tttcctcc**ag**/TGCACAAAAG	94.9	ATTCACCAAG/**gt**gagtggca	96.7	?
9	210	gtccctac**ag**/GTGCAGCAGG	92	CAAGCTGGAG/**gt**aggcccta	83.3	?
10	66	ctctcctc**ag**/GATGTTCTTG	94	GGTCTGTAAG/**gt**acggctgt	83.9	0.64
11	1777	gtctccac**ag**/GCCCATAACG	93.9	3'UTR		

Note. Consensus strengths are based on Shapiro and Senapathy, 1987. Exon sequences are shown in uppercase whereas intronic sequence is represented by lowercase letters. All splice site boundaries observed the gt/ag rule and are shown in bold.

changes the ORF such that the same stop codon is now shared between the mouse and human *GAS11* gene.

Two independent human cDNA clones (zb57b10 and ze60g08) contained an additional 109 bp of sequence between exon 2 and 3 which was not seen in the mouse *Gas11* homologue or a mouse cDNA clone (vi93h01). However, inclusion of this extra sequence as a possible additional exon disrupts the ORF of the gene such that an alternative start site down-stream would have to be used for translation. RT-PCR using primers located in exon 2 and 3 of *GAS11* (G17 and G8) failed to amplify a product containing the extra 109 bp sequence from a number of tissues (Figure 5.5.A). However, RT-PCR using a primer located within the 109 bp sequence (G6) together with the exon 2 primer did amplify products (Figure 5.5.B). A comparison with genomic sequence of the 0.7 kb *Bam*HI fragment indicated that the 109 bp was identical to intronic sequence immediately 5' of exon 3. Subsequent cloning of two of the RT-PCR amplified products followed by a sequence comparison with genomic DNA indicated the extra 109 bp was most likely derived as a result of the use of a cryptic splice acceptor site located 109 bp 5' to the true splice acceptor site of exon 3 (score of 88%). The use of an additional cryptic splice acceptor site located 14 bp 5' to this first cryptic site (score of 82.6%) appeared to give rise to the second RT-PCR product sequenced. The largest RT-PCR band seen in Figure 5.5.B was unable to be cloned and most likely represents a heteroduplex band generated from the two smaller products which differ in sequence by only 14 bp. These results indicate that transcripts containing this extra sequence must exist but may represent unspliced mRNA, splicing intermediates generated during processing of this intron, or splicing artifacts.

5.3.5 Analysis of zb57b10

Based on sequence analysis, the fetal lung cDNA clone zb57b10 was missing half of exon 8, all

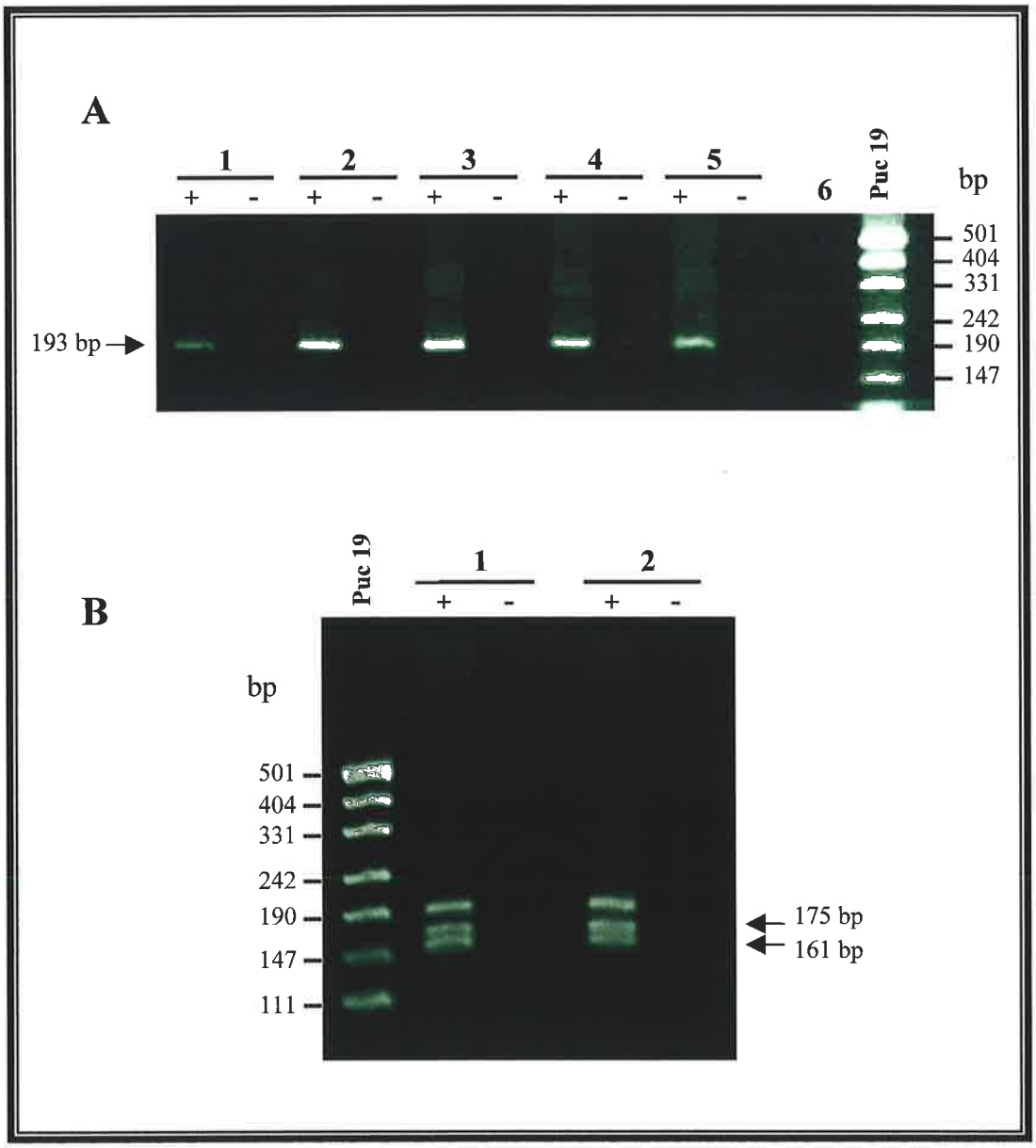


Figure 5.5: Confirmation of the extra 109 bp seen between exons 2 and 3 of *GAS11*. (**A**) RT-PCR was performed using a primer in exon 2 in combination with an exon 3 primer on fetal RNA tissues (lanes **1** to **4**) **1**: Spleen; **2**: Lung; **3**: Kidney; **4**: Skeletal muscle; **5**: breast cancer cell line BT20 RNA; **6**: genomic DNA, with (+) or without (-) reverse transcriptase. A single band of 193 bp was observed which when sequenced did not contain the extra 109 bp. (**B**) RT-PCR was performed using an exon 2 primer in combination with a primer present in the extra 109 bp region on fetal RNA tissues **1**: Brain; **2**: Skeletal muscle, with (+) or without (-) reverse transcriptase. The two smallest bands were sequenced and shown to be derived from the use of cryptic splice acceptor sites located 109 bp and 123 bp respectively 5' to the true splice acceptor site of exon 3.

of exon 9 and most of the 3' UTR including exon 10 and therefore did not appear to be derived from alternate splicing of the *GAS11* gene. To confirm the existence of this transcript, RT-PCR using primers (AA1 and W3) located either side of the region where this clone diverged in sequence from *GAS11* was initiated using fetal brain RNA. A product was not generated from this tissue. However, subsequent experiments using polyA$^+$ mRNA from skeletal muscle were successful in amplifying a RT-PCR product of the correct size from this tissue indicating that the zb57b10 cDNA clone is not an artifact.

5.3.6 Amino Acid Identity

Overall the *GAS11* gene is 87% and 96% identical with its mouse homologue (U19859) at the nucleotide and amino acid level respectively. This is a conservative estimate as the U19859 sequence has been shown to have sequencing errors elsewhere and may contain more. In contrast, the 5' and 3' UTRs in mouse and human do not show any homology. Promoter prediction (http://www-hgc.lbl.gov/projects/promoter.html) located a possible transcription start site at position –121 with respect to the *GAS11* translation start but no TATA box was located. The program PSORT (http://psort.nibb.ac.jp/) predicted that the gene product might be located in the nuclear compartment. From BLASTP analysis, significant amino acid similarity was observed to a T lymphocyte triggering factor of *Trypanosoma brucei rhodesiense* (GenBank accession AF012853) as well as weaker similarities to a portion of the non-muscle myosin heavy chain of *Drosophila melanogaster,* chicken, human, rabbit, *gallus gallus*, and rat. Other weak similarities included the human and mouse kinesin heavy chain protein and a kinesin-related protein of *Xenopus laevis*.

5.3.7 Sequence Determination and Characterisation of *C16orf3*

The cDNA clone yd83e07, had previously been mapped to the 16q24.3 region

(http://www.ncbi.nlm.nih.gov/cgi-bin/SCIENCE96/loc?stSG9210) using the GB4 radiation hybrid panel. This clone represented a new transcript which was subsequently referred to as *C16orf3* (chromosome 16 open reading frame 3). Using the same primer set (C16F and C16R), this gene was subsequently mapped by PCR to the same cosmid from which *GAS11* is transcribed (c344G2). PCR on individual *Bam*HI fragments of this cosmid indicated that *C16orf3* mapped within intron 2 of *GAS11*. The 5' and 3' sequences of yd83e07 present in GenBank were shown to be continuous with genomic DNA sequence obtained from the 3.1 kb *Bam*HI fragment and further, the ends overlapped with each other such that the resultant cDNA insert sequence completely resembled genomic DNA.

5.3.7.1 5' Sequence Identification of C16orf3

Northern blot hybridisations identified a band of approximately 1.2 kb and of weak intensity primarily in the heart, skeletal muscle, pancreas, and liver, with very faint signals detected in the other tissues (Figure 5.6). As the yd83e07 cDNA insert was only 833 bp in length, 5' RACE was performed to identify additional 5' sequence. A single band of approximately 400 bp was obtained from both tissues used, however identical products were also observed in the negative (no reverse transcriptase) controls (Figure 5.2C). Sequencing of these bands indicated that they were cloning artifacts generated from the self-priming of the AUAP primer during nested PCR.

Examination of the sequence obtained for the *C16orf3* gene (Figure 5.7) indicates an ORF of only 375 bp coding for a protein of 125 amino acids (large allele, see 5.3.7.2) that does not show homology to known gene products present in available databases. The gene appears to be transcribed in the opposite orientation to the *GAS11* gene. Promoter prediction analysis indicates a possible transcription start site at –558 with respect to the translation start site and

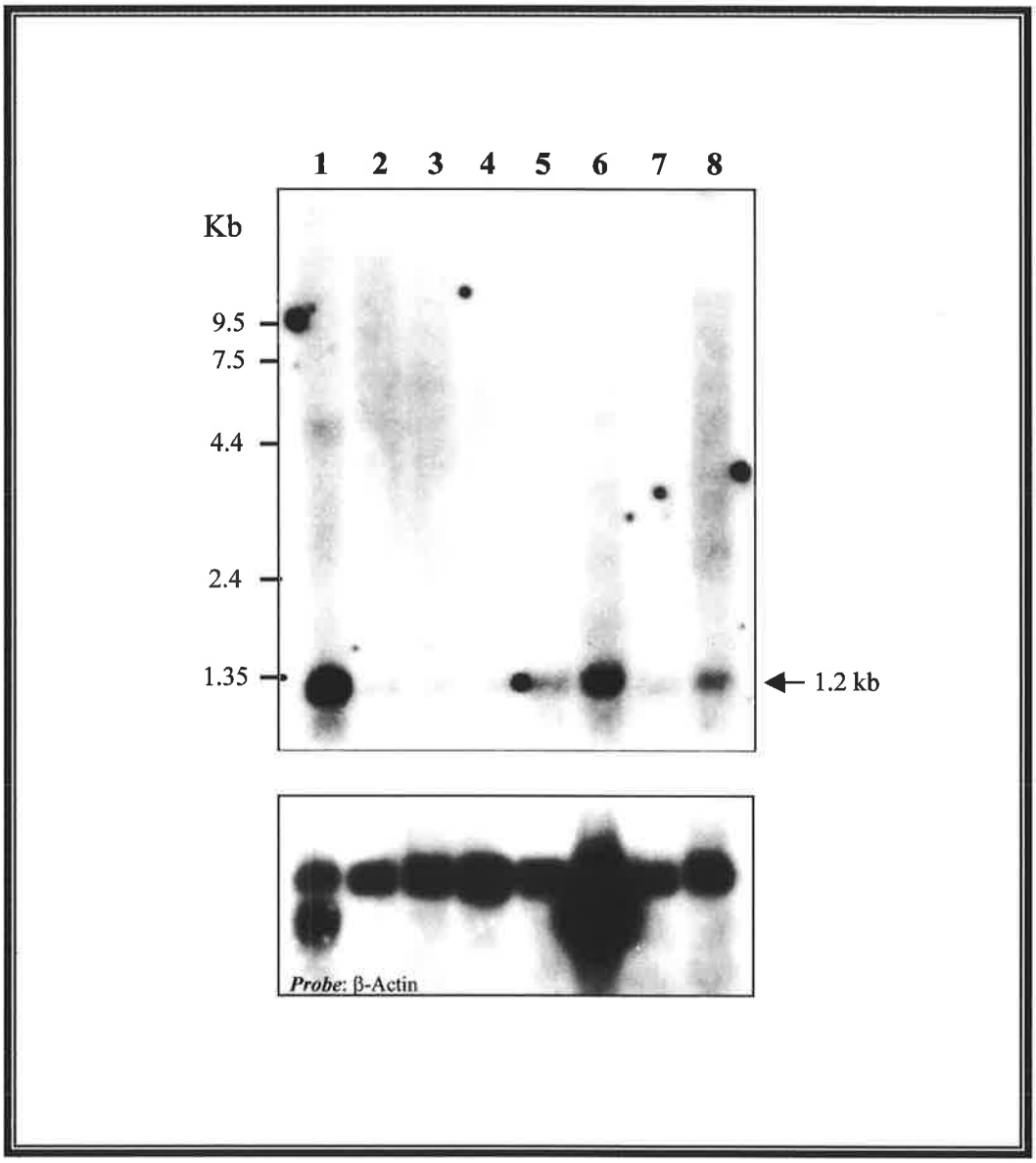


Figure 5.6: Northern analysis of the *C16orf3* gene. A single band of 1.2 kb in size was detected after a ten night exposure at -80°C. The probe used was a PCR product generated from genomic DNA using the C16R and C16A primers. A β-actin control probe is shown directly below the autoradiograph. Adult tissues represented are **1**: Heart; **2**: Brain; **3**: Placenta; **4**: Lung; **5**: Liver; **6**: Skeletal muscle; **7**: Kidney; **8**: Pancreas.

```
ACCTGCAGTC CCAGCTACTG GGCAGCCTGA AGCAGCAGGA TGGTGTGAAC CCAGGAGGTG GAGCTTGCAG TGAGC    75

CGAGGTCGCG CCACCGCACT CCAGCCTGGG CCACACAGCG AGATTCCGTC AGAATCAGTT ACTTTTCGGG CACAG   150

CCCCAGGCCA CTTACTGTGA GCCTTTTTCT TTCTCAACAC CACATTCCCC ACAGGGAAAA CACATTTCTC ACCTC   225

AAAAGAAGAC AAGACAACGA GCAAACAAGA AGGAGCAGCA GGAGGGGTTC TGAGCCGAGG ATGCCGGGCA GACAT   300

GAGGGAGACA CGCACCCCCG AATCCAACCA GTGCCTCGGC ACAACGACAA ATGTCTTCAC GTCACAGACC TTTAG   375

AAGCTCCTGG GCAGACCTGA ACCAAGGCTC CTGACTGGTC TGTTTGGCTC ACATGGTGTT GAGATTTTGC CATCA   450

CTCAATATTC AGATTTCTTA TAAATATCCA GATTTCCAGC TTCTCTTGGA AAATCAGAAA AAAACAGCAC TGAAC   525

TCCTAGGCCC ACAAGGCACT CCCCAGTGAA CAG ATGAAAC TGTCCTCTGC TGCGGGGCAG GAGTCTCCAG GTCAC  600
                                      M  K  L  S  S  A  A  G  Q  E  S  P  G  H    14

CCCCATCCCT CCCCACCTGC CTGGACCCTG AAGAAGCCTT CTGAGTCTGT GGCTCAACGT GCGATGTGCA GTGCA  675
 P  H  P  S  P  P  A  W  T  L  K  K  P  S  E  S  V  A  Q  R  A  M  C  S  A    39

AGGGCCTGCC CCGTAGCCTG CCCCGTAGGC TGCCCCGCAG CCTGCCCCGT AGGCTGCCCC ATAGCCTGCC CCGTA  750
 R  A  C  P  V  A  C  P  V  G  C  P  A  A  C  P  V  G  C  P  I  A  C  P  V    64

AGCTGCCCCG TAGCCTGCCC CGTAGGCTGC CCCGTA GGCT CCATGGCCAC TGCCCCACAA GGCCTGTCTC CACAG  825
 S  C  P  V  A  C  P  V  G  C  P  V  G  S  M  A  T  A  P  Q  G  L  S  P  Q    89

GAATGGGAAG CGGACAGGGA GACGGGCAGC AGCTCACATG CTGGGACAAC GCAGTGTTCA ATCCATTCTC CATCC  900
 E  W  E  A  D  R  E  T  G  S  S  S  H  A  G  T  T  Q  C  S  I  H  S  P  S   114

AGCAGCTCCA GACATCTTTC CAGAACACAA ACC TGACCCC ATCACCTCTC TGCTTAGCCA CTGGCTTAAA CTGCC  975
 S  S  S  R  H  L  S  R  T  Q  T                                               125

AATGGTTTGC CTGCATGTAA AATAAAGCCA TTCTTTACCA TTAAAAAA                            1023
```

Figure 5.7: Nucleotide sequence of the *C16orf3* gene. The corresponding amino acids are indicated below the nucleotide sequence. Residue numbers are indicated in the right hand column. The small allele is missing 2 copies of the 12 bp imperfect repeat situated in the open reading frame (ORF) of the gene. This repeat region is shown in red. The translation start site and the ORF stop codon are indicated by green type and the position of the poyadenylation signal is underlined.

the PSORT program indicates that the gene product may be a cytoplasmic protein.

5.3.7.2 Allele Frequencies of C16orf3

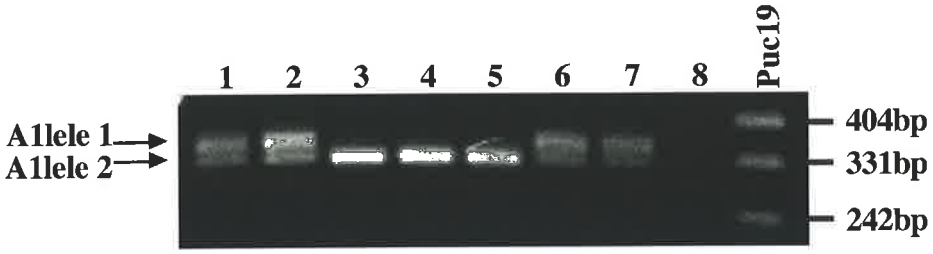
PCR amplification from human genomic DNA using primers to the 5' and 3' ends of *C16orf3* resulted in the identification of two alleles in some individuals (Figure 5.8.A). Sequencing of both these products revealed they differ in size by 24 bp due to an extra 2 copies of a 12 bp imperfect repeat motif located within the coding region of the gene (Figure 5.8.B). Analysis of 95 normal Caucasian individuals established that 39% were homozygous for the small allele, 18% were homozygous for the large allele and 43% were heterozygous. This indicates that either the small or large allele can be transcribed.

5.3.8 Analysis of ET27.12

5.3.8.1 Orientation of ET27.12

The presence of the trapped exon ET27.12 in the 3' UTR of *GAS11* suggested it might belong to an overlapping gene transcribed from the opposite strand. RT-PCR using the r*Tth* reverse transcriptase enzyme was used to determine if ET27.12 was transcribed in both orientations. The 27.12A and 27.12B primers were individually used in separate reverse transcription reactions followed by the addition of the other primer to each tube for the subsequent PCR. These experiments confirmed that ET27.12 is transcribed in the same orientation and is therefore part of the *GAS11* gene (Table 5.6). However, when tested to determine if it was transcribed in the opposite orientation, a product of the correct size was also observed, but in both experiments products in the negative control tube (no r*Tth* in the reverse transcription reaction) of the same size were also seen. It was therefore impossible to determine from these results if this exon can be transcribed in both orientations.

A



B

```
Allele 1                          TGGCTCA ACGTGCGATGTG CAGTGCAAGGGC   31
Allele 2                          TGGCTCA ACGTGCGATGTG CAGTGCAAGGGC   31

Allele 1  CTGCCCCGTAGC CTGCCCCGTAGG CTGCCCCGCAGC CTGCCCCGTAGG        79
Allele 2  CTGCCCCGTAGC                           CTGCCCCGTAGG        55

Allele 1  CTGCCCCATAGC CTGCCCCGTAAG CTGCCCCGTAGC CTGCCCCGTAGG       127
Allele 2  CTGCCCCATAGC CTGCCCCGTAAG CTGCCCCGTAGC CTGCCCCGTAGG       103

Allele 1  CTGCCCCGTAGG CTCCATGGCCAC TGCCCCACAAGG CCTGTCTCCACA       175
Allele 2  CTGCCCCGTAGG CTCCATGGCCAC TGCCCCACAAGG CCTGTCTCCACA       151

Allele 1  GGAATGGGAAGC GGACAGGGAGAC GGGCAGCAGCTC ACATGCTGGGAC       223
Allele 2  GGAATGGGAAGC GGACAGGGAGAC GGGCAGCAGCTC ACATGCTGGGAC       199

Allele 1  AACGCAGTGTTC AATCCATTCTCC ATCCAGCAGCTC CAGACATCTTTC       271
Allele 2  AACGCAGTGTTC AATCCATTCTCC ATCCAGCAGCTC CAGACATCTTTC       247

Allele 1  CAGAACACAAAC CTGACCCCATCA CCTCTCTGCTTA GCCACTGGCTTA       319
Allele 2  CAGAACACAAAC CTGACCCCATCA CCTCTCTGCTTA GCCACTGGCTTA       295

Allele 1  AACTGCCAATGG TTTGCCTGCATG TAAAA                         348
Allele 2  AACTGCCAATGG TTTGCCTGCATG TAAAA                         324
```

Figure 5.8: Identification and analysis of the large and small alleles of the *C16orf3* gene. (**A**) PCR on genomic DNA of normal individuals (lanes **1**-**7**) with primers to *C16orf3*. Two alleles were detected as indicated by arrows. Lane **8** represents a no-DNA template control. (**B**) Sequence comparison of the large (Allele 1) and small (Allele 2) alleles of *C16orf3*. The imperfect 12 bp repeat is shown in red type. Primer sequences used to amplify these products are underlined.

5.3.8.2 Linking ET27.12 to C16orf3

As *C16orf3* appeared to be transcribed in the same orientation as ET27.12 was trapped, it was possible they belonged to the same transcript and ET27.12 was a further 5' exon of *C16orf3* not identified through 5' RACE. The r*Tth* reverse transcriptase enzyme was again used to initiate cDNA synthesis using a primer specific for the *C16orf3* gene (C16R). Subsequent PCR using the 27.12A primer was unable to generate products from fetal brain polyA$^+$ mRNA (Table 5.6).

TABLE 5.6

r*Tth* RT-PCR results of ET27.12

	1st strand primer (cDNA synthesis)						
2nd primer (PCR)	**27.12A**		**27.12B**		**C16R**		**Kit control**
	+RT	**-RT**	**+RT**	**-RT**	**+RT**	**-RT**	
27.12B	+	F	NA		NA		NA
27.12A	NA		+	F	-	-	NA
Kit control	NA		NA		NA		+

Note. NA: not applicable; +RT: r*Tth* is included in the first strand synthesis; -RT: r*Tth* is omitted from first strand synthesis; +: a band of the expected size was obtained; F: faint band of the expected size; -: no bands seen on the gel.

Given that *C16orf3* shows greater expression in skeletal muscle, RNA from this source should have been used. However, RT-PCR using primers from either end of the established sequence for this gene failed to amplify a product from this tissue at all.

5.3.8.3 3' RACE

To identify a possible 3' end to the ET27.12 exon, 3' RACE was attempted. A PCR product

was generated from polyA$^+$ mRNA only, of about 230 bp (data not shown). From sequence analysis this product appeared to have been generated from the binding of the original anchored oligo-dT primer to bases 3,046 to 3,058 present in the 3' UTR of the *GAS11* gene, where the A residue constituted 10 of these 13 bases. The sequence of the product beyond the ET27.12 3' end was continuous with the sequence seen beyond this exon in the 3' UTR of *GAS11*, suggesting this product does not represent part of an overlapping gene to which ET27.12 belongs.

5.3.9 Mutation Analysis of *GAS11*

5.3.9.1 SSCP Analysis

To determine if this gene was a potential tumour suppressor gene involved in breast cancer, primers were designed adjacent to exon sequences in intronic DNA to test for the presence of mutations in breast tumour DNA. Primer sequences are listed in Table 5.4 and SSCP results for the analysis of 17 tumour and matching normal DNA samples are shown in Table 5.7. No nucleotide base changes were observed in the coding region of *GAS11*, however polymorphisms were detected in introns 2 and 10 (IVS2+39G/A; IVS10+26A/G and IVS10+44C/T).

5.3.9.2 Homozygous Deletions

A probe representing the entire coding region of *GAS11* (ze60g08 cDNA insert) indicated that the genomic interval containing this gene is polymorphic for all enzymes tested which complicated interpretation of results. However, hybridisation patterns seen in DNA from normal individuals were also observed in the breast cancer cell line samples indicating that no homozygous deletions of this region were disrupting the gene in the cell lines tested.

TABLE 5.7

Results of SSCP Analysis of *GAS11*

Tumour	Exon										
Loss 16q24.3	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
204	+/+	ND	+/+	ND	ND	+/+	+/+	ND	ND	ND	+/+
309	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	P	+/+
358	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
367	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
413	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
549	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
559	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	P	+/+
589	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
645	+/+	+/+	ND	+/+	+/+	ND	ND	+/+	+/+	+/+	ND
666	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
757	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
819	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
919	+/+	P	ND	+/+	+/+	ND	ND	+/+	+/+	+/+	ND
Complex loss											
152	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
380	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
670	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
683	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+

Note. +/+ indicates the same SSCP pattern was observed in both the tumour DNA and normal DNA respectively from the same affected individual. P indicates a polymorphism detected as a bandshift in tumour and normal DNA from the same individual. ND: not done.

5.3.9.3 Exon Skipping

Of all the breast cancer cell lines and breast tumours examined, not one gave rise to truncated RT-PCR products with all primer sets tested. This excluded exon skipping as a mutational mechanism and confirmed that there was no alternative splicing occurring in the gene.

5.3.10 Mutation Analysis of *C16orf3*

Primers designed to encompass the ORF of the *C16orf* gene (Table 5.4) were unable to identify SSCP changes specific for the breast tumour DNA samples analysed.

5.4 Discussion

5.4.1 *GAS11*

The cDNA clones belonging to the trapped exons ET27.11 and ET27.12 have been successfully linked and shown to be part of the *GAS11* gene. ET27.11 corresponded to exon 2 of this gene while ET27.12 was trapped from the antisense strand within the 3' UTR. The possibility that this trapped product belonged to an overlapping gene transcribed in the opposite orientation was explored but failed to confirm the existence of such a gene. Analysis of the nucleotide sequences surrounding this exon established that they highly resembled splice acceptor and donor consensus sequences (scores of 89.2% and 91% respectively). Therefore it was likely this exon was trapped from genomic DNA. Based on the results of Northern analysis it appears that many isoforms of this gene exist, particularly in skeletal muscle and heart. Ubiquitous expression of a 3.4 kb mRNA corresponds to the observed nucleotide sequence length obtained for the gene, while the ubiquitous 1.8 kb mRNA was most likely derived from the use of an alternative polyadenylation signal located in the 3' UTR due to the fact that exon skipping experiments failed to identify alternatively spliced *GAS11* transcripts.

The greater abundance of the 3.4 kb mRNA in certain tissues suggests the presence of elements within the 3' UTR of *GAS11* playing a role in its regulation.

The truncated cDNA clone zb57b10 corresponding to ET27.11 was not considered an alternatively spliced transcript because the point of divergence from the *GAS11* sequence was in the middle of exon 8 and within the 3' UTR which in both cases did not correspond to a splice site consensus sequence. However, RT-PCR was successful in confirming the existence of this truncated transcript in fetal skeletal muscle tissue with the size of this truncated mRNA most likely accounting for the 1.3 kb Northern band seen in skeletal muscle alone. The 7.3 kb Northern band also appears to be relatively tissue specific in its expression suggesting it may be a transcript particular to muscle tissue.

The gene consists of 11 exons that span ~25 kb of genomic DNA and codes for a 478 amino acid protein that is at least 96% identical to its mouse homologue. This high conservation has also been observed for many of the other previously characterised *GAS* genes. To determine whether the *GAS11* gene was a tumour suppressor gene involved in breast carcinogenesis, mutation analysis was done on paired tumour and normal DNA isolated from affected individuals that defined the minimum LOH region at 16q24.3. The results of SSCP analysis indicated that no nucleotide base changes were present in the tumour DNA samples tested, however polymorphisms were detected in introns 2 and 10.

The SSCP technique is not 100% efficient in detecting all base changes, with the efficiency related to the size of the DNA fragment under analysis, sequence composition of the DNA fragment, and the properties of the gel system used (Glavac and Dean, 1993; Savov *et al.*, 1992). Studies of the sensitivity of SSCP analysis have produced a range of efficiencies from

~30 to over 90% (reviewed in Jordanova *et al.*, 1997), with the highest figures reported when DNA fragments of ~150 bp were used together with a range of electrophoretic conditions. However since polymorphisms were identified in the *GAS11* SSCP screening and given the sporadic nature of the tumours that were being analysed, the chance of missing all disease causing mutations within the 17 tumour samples tested was therefore quite low.

Despite this, other methods of mutation detection were also examined and included a test for exon skipping within the gene and a search for homozygous deletions in a subset of breast cancer cell lines. Exon skipping as a result of splice site mutations or intronic deletions has been observed as a mutation mechanism for a number of genes including *TP53* in breast cancer, *RB* in hepatocellular carcinoma, and *ATP7A* in occipital horn syndrome (Farshid *et al.*, 1994; Voglino *et al.*, 1997; Qi and Byers, 1998). Homozygous deletions have also been instrumental in identifying and narrowing down regions known to contain disease genes identified from LOH studies or linkage analysis (Fearon *et al.*, 1990; Mollenhauer *et al.*, 1997; Town *et al.*, 1998). With the use of the whole *GAS11* cDNA insert, 25 kb of genomic DNA was effectively screened at once. Again these studies failed to detect aberrations in the *GAS11* transcript in breast tumour material suggesting this gene does not have a role in breast carcinogenesis.

A function for the *GAS11* gene based on its amino acid sequence and database homology searches is not immediately obvious. What is consistent, is the similarity of this gene to the non-muscle myosin heavy chain gene of a number of species, as well as similarity to the kinesin heavy chain gene. Kinesin is a member of a family of microtubule based motor proteins that perform force-generating tasks which include organelle transport and chromosome segregation (reviewed in Goldstein, 1993). The kinesin molecule consists of a tail domain at the C-

terminus, followed by a stalk domain, and finally a motor (head) domain at the N-terminus. The motor domain of this protein has been shown to have a structural similarity to the core of the catalytic domain of the actin-based motor myosin (Kull *et al.*, 1996), while the tail domain consists of a 28 amino acid long repeated motif which is also seen in the tail of myosin (Goldstein, 1993). The *GAS11* gene displays homology to this tail domain repeat motif (amino acids 58 to 442) and may therefore be structurally related to both kinesin and myosin suggesting it may also play a role in force-generating tasks within cells. The restricted expression of this gene during growth arrest could be correlated with the processes that occur during this phase of the cell cycle. Some of these include the production of ribosomes, mitochondria, and other cytoplasmic molecules which are needed for S phase and subsequent cell division. If *GAS11* is involved in force-generating tasks within cells its expression may increase during growth arrest to provide transport for new organelles being synthesised. However, given these processes occur in the cytoplasm, it does not agree with the predicted nuclear localisation for the *GAS11* protein product.

5.4.2 *C16orf3*

Through sequence analysis within intron 2 of the *GAS11* gene, it was established that the *C16orf3* gene was transcribed in the opposite orientation to *GAS11* and appeared to be intronless. This gene codes for a 125 amino acid protein that has no significant homology with known amino acid sequences or domains. However, within the coding region of the gene a polymorphism was detected such that two alleles of the gene exist differing by 8 amino acids (2 copies of a tetrapeptide repeat). Given the presence of just a single cDNA clone in dbEST representing this gene, the faint hybridisation signals seen from Northern analysis, and a failure to detect RT-PCR products from this gene in a number of fetal tissues, suggests that it is not abundantly expressed. SSCP analysis failed to detect variant bands in tumour DNA samples

when compared to matched normal samples suggesting this gene does not play a role in breast tumourigenesis.

The possibility of mutations occurring outside the coding region of both these genes was not explored in this study. Alterations in the pattern of DNA methylation have been observed as a consistent mechanism in a wide variety of cancers (Jones, 1996) due to aberrant hypermethylation of CpG islands in the 5' regulatory regions of genes. Transcriptional silencing by methylation of several tumour suppressor genes has been identified including *RB* (Sakai *et al.*, 1991), *VHL* (Herman *et al.*, 1994), and *CDH1* (Yoshiura *et al.*, 1995). Future DNA methylation studies of both *GAS11* and *C16orf3* therefore may need to be explored to provide further evidence that both of these genes are not involved in breast tumourigenesis.

Chapter 6

Characterisation of Transcription Unit 6 (T6)

Table of Contents

	Page
6.1 Introduction	**196**
6.2 Methods	**197**
6.2.1 T6 Sequence Assembly	197
6.2.2 Direct Cosmid Sequencing	199
6.2.3 Genomic Structure Characterisation	199
6.2.4 PCR	199
6.2.5 Single Stranded Conformation Polymorphism (SSCP) Analysis	199
6.2.5.1 PCR Amplification of Tumour and Normal DNA	200
6.2.5.2 SSCP Gels	203
6.2.5.2.1 10% Polyacrylamide Gels	203
6.2.5.2.2 Mutation Detection Enhancement (MDE) Gels	203
6.2.6 Patient Material used for SSCP	204
6.2.7 Exon Skipping	204
6.3 Results	**205**
6.3.1 Sequence Analysis of T6 cDNA Clones	205
6.3.2 T6 Northern Analysis and 5' Sequence Isolation	205
6.3.3 T6 Sequence Homologies	206
6.3.4 T6 Genomic Structure	210
6.3.5 Mutation Analysis	210
6.3.5.1 SSCP Results	210
6.3.5.2 Exon Skipping	214
6.3.6 Alternative T6 Transcripts	214
6.3.6.1 T6A	217
6.3.6.2 T6B	217
6.3.6.2.1 Mapping T6B	220
6.3.6.2.2 T6B Northern Analysis and Sequence Homologies	221
6.3.6.3 T6C	221
6.4 Discussion	**222**
6.4.1 T6	222
6.4.2 T6 Isoforms	223

6.1 Introduction

From the results described in chapter 4, a number of potential novel genes were identified in 16q24.3 based on the nucleotide sequence homology of trapped exons to ESTs present in dbEST. A group of five exons consisting of ET17.23, ET6.14, ET17.9, ET17.2 and ET17.14 could be grouped into a single transcription unit termed T6, based on the nucleotide homology of overlapping cDNA clones (see Table 4.3). In addition, the close physical location of these trapped exons with respect to each other provided further evidence that they belonged to the same gene (Figure 4.4). Another feature of this region was the presence of two additional cDNA clones. The first of these, yc81e09, was identified following the construction of a hn-cDNA library of human chromosome 16 (Whitmore *et al.*, 1994). The hn-cDNA clone ScDNA-A55 was used as a probe to screen a gridded fetal brain cDNA library at Lawrence Livermore laboratories, subsequently selecting yc81e09 as a homologous clone. A second additional cDNA clone located close to this group of trapped exons was ze25b01, identified from BLASTN analysis of the partial sequence of cosmid 360D7 (Table 4.6). Based on their close physical proximity, these trapped exons and cDNA clones may therefore collectively represent a single gene, which because of its location at 16q24.3, could be considered a candidate breast cancer tumour suppressor gene.

This chapter describes the determination of the relationship of the trapped exons and homologous cDNA clones located within this close physical proximity and the subsequent characterisation of the transcript. As described in the previous chapter, this involves determining the complete nucleotide sequence of the associated gene in order to obtain the amino acid sequence of the encoded protein product. This may provide functional information based on homology to previously characterised gene products present in available databases,

and therefore a prediction of its possible role in breast carcinogenesis. In addition, determination of its genomic structure and subsequent mutation analysis of the full-length transcript using an extended panel of breast tumour material from that used in the previous chapter, will determine if this gene is mutated in breast cancer.

The majority of the work presented in this chapter was performed by the candidate and Ms. Joanna Crawford with a small amount of sequence data provided by Dr. Sinoula Apostolou. Both individuals will be acknowledged for their contributions where appropriate in the text.

6.2 Methods

Only those procedures performed specifically in this chapter are mentioned in detail. More general techniques are described in chapter 2, or other chapters as indicated. However slight modifications are referred to when appropriate in the text.

6.2.1 T6 Sequence Assembly

The cDNA clones yc81e09, ze25b01, and zb17g05 were purchased from Genome Systems and DNA was isolated using Qiagen columns (2.2.1.3). All clones were initially sequenced using DyePrimer chemistry (2.2.18.2) which was specific for vector sequences adjacent to the insert cloning sites, followed by DyeTerminator sequencing (2.2.18.1) with insert specific primers. Clones yc81e09 and ze25b01 were sequenced by the candidate with confirmation provided from the sequencing of the zb17g05 clone by Ms. Joanna Crawford. Dr. Sinoula Apostolou also provided assistance in sequence confirmation. Primers used to sequence yc81e09 and ze25b01 are shown in Table 6.1. The BLASTN and BLASTP programs (Altschul *et al.*, 1997) were used to search for nucleotide sequence homology and amino acid homology respectively, between T6 and sequences deposited in the GenBank non-redundant and EST databases as

TABLE 6.1

Sequences of Primers used for T6 Sequence Identification and Confirmation

Primer name	Primer sequence (5' – 3')	Primer position
T6 Sequence Identification		
M13F*	TGT AAA ACG ACG GCC AGT	Vector
M13R*	CAG GAA ACA GCT ATG ACC	Vector
T3	ATT AAC CCT CAC TAA AGG GA	Vector
T7	TAA TAC GAC TCA CTA TAG GG	Vector
177D12F	CTC CTT CCT CGG GCC TCT CC	bases 1969-1950
177D12R	CTG CCG GCT GGA TTA CCG CAG	bases 1065-1085
177D12R1	TCC CTG AGT GTG AGC AGA GCT	bases 1379-1399
T6 5' Sequence Identification		
T6GSP1	TCT CGG TGT CAT CTC CAT CC	bases 412-393
T6GSP2	TCT GTG TTT CCA CGC TGA CC	bases 383-364
T6GSP3	CAG GGT CAT CCT CAA GAT CG	bases 298-279
AA7	AGC TCG AAG CGG TTG TTG AC	bases 266-247
T6 Sequence Confirmation		
AA2	TCC ACA TGC TTT ATT CCA GC	bases 598-579
AA3	CTT GCA TTT CGA TCT CCG TG	bases 144-163
zq3	CAG TGC TGT CCT CAA TCC TC	bases 559-540
AA9	CCT AGG GGT GGG AGG AAG C	bases 497-515
AA5	TCT GTG TTT CCA CGC TGA CC	bases 384-364

Note. The primer positions are based on the sequence of the T6 gene presented in Figure 6.2 or the T6B transcript in Figure 6.5. Primer positions are indicated in red lettering for T6B. Vector refers to the vector containing the ze25b01 and yc81e08 cDNA clones. * These primers were part of the Dye Primer sequencing kit.

described in 5.2.3.

6.2.2 Direct Cosmid Sequencing

The cosmid 360D7 contains the T6 transcript and was used as a template to sequence the 5' UTR using the BigDye Terminator sequencing kit. In a 20 µl volume, 600 ng of cosmid DNA was mixed with 30 ng of AA7 primer (Table 6.1) and 8 µl of the BigDye mix. After addition of a drop of paraffin oil, the tubes were incubated at 95°C for 5 minutes. PCR was then performed with 80 cycles of 95°C for 30 seconds; 50°C for 20 seconds; 60°C for 4 minutes. All reactions were cleaned as described in 2.2.18.1.

6.2.3 Genomic Structure Characterisation

Ms. Joanna Crawford was solely responsible for the sequencing of all exon/intron boundaries of the T6 gene using methods similar to those described in chapter 5. Primarily the two methods chosen were sequencing genomic PCR products generated from primers specific for the cDNA sequence of T6, and direct sequencing of the 360D7 cosmid with cDNA specific primers as described in 6.2.2.

6.2.4 PCR

PCR reactions were performed essentially as described in 2.2.14.1. All products were analysed on 2.5% (w/v) agarose gels. RT-PCR reactions were performed as described in 2.2.15. The sequences of all primers used are shown in Table 6.1.

6.2.5 Single Stranded Conformation Polymorphism (SSCP) Analysis

The large number of exons and tumour samples to be analysed dictated that the SSCP analysis was to be shared among three individuals. Exons 1 to 10 were analysed by Ms. Joanna

Crawford using tumour samples 204, 309, 358, 589, 645, 757, 919, 424, 438, 439, 377, 355, 555, 581, 152, 380, 670, 683, 768, and 594. Exons 11 to 18 were examined by the candidate using the same set of tumour samples as above, except sample 448 was used instead of sample 377, while Dr. Anna Savoia (Italy) tested all exons with the tumour samples 367, 413, 549, 559, 666, and 819. Exons 1 and 18 were each split into two overlapping PCR fragments due to their size. These are referred to as exon 1.1, 1.2, 18.1, and 18.2.

6.2.5.1 PCR Amplification of Tumour and Normal DNA

Exon specific sequences were amplified from tumour and control DNA samples using procedures modified from Kogan *et al.* (1987). Ten microlitre reactions were prepared which contained 5 µl of 2X PCR mix, 150 ng of each primer (Table 6.2), 30 ng of patient DNA, 0.5 units of *Taq* DNA polymerase, 0.2 µl of [α-^{32}P]dCTP and $MgCl_2$ optimal for each primer pair (see Table 6.2). After addition of a drop of paraffin oil, the tubes were incubated for 10 cycles at 94°C for 1 minute; 55°C for 1.5 minutes; 72°C for 1.5 minutes, followed by a further 25 cycles of 94°C for 1 minute; 60°C for 1.5 minutes; 72°C for 1.5 minutes, with a final extension of 72°C for 7 minutes. Primers specific for exons 1.2 and 18.2 were amplified for 35 cycles with an annealing temperature of 60°C. Following the PCR reaction, 10 µl of formamide loading buffer was added to each sample, mixed, spun down, and left at 4°C until ready to load on the appropriate gel. Prior to loading, samples were incubated at 100°C for 10 minutes, then placed on ice. Four microlitres of each sample was then immediately loaded on the appropriate gels.

Exons in which SSCP changes were detected were amplified from tumour and normal DNA from the same individual using the conditions originally used for the SSCP analysis with the

TABLE 6.2

Oligonucleotide Primers for Mutation Analysis of T6

Primer set	Nucleotide sequences (5'-3')	Size (bp)	Optimal [Mg^{++}]
T6 SSCP			
Exon 1.1	GCA CGG CGC GCA GAC ACC CA ACA TAG GAC CGA ACG ACC AC	276	0.5 mM
Exon 1.2	GTG GTC GTT CGG TCC TAT GT CTG GCA TCA CAA CCC CGC T	317	0.5 mM
Exon 2	CAC ATC TCA CTA TTC ATA TCT C GGC TGG GTG CCT TGG AAC TC	299	1.5 mM
Exon 3	CGC TTG CTG CTT TAA TTC TCA C CAC AGC CAG GCC CAA CTC AC	172	1.5 mM
Exon 4	GTA AAT ACA GTG TAA TAG GTC TC GCC ACA CCT GTG CTC CAC	232	0.5 mM
Exon 5	CTT TTG GAA TCG TCT TAG AAC GGT AAA CTA GCT TCA GGA AC	215	1.5 mM
Exon 6	CTT GGT CCC CAC ACG CCA C GGG CCA GTG CAC AGA CAA C	204	2.5 mM
Exon 7	CGG AGC TGC CCA CGT CCT TG GGG CAG GAG GCA GCA TGC CT	313	1.5 mM
Exon 8	AGC TGG GGC CTT GCT GAG TG CTG TCT TGA TAG CTT GAG GTG	215	1.5 mM
Exon 9	CAG AGG AGG GGC CGT GGT TG CCC AGC CTC ACT CGC AGG AG	222	1.5 mM
Exon 10	GCA TTC TCC TTT GTG TCG TTG GAA CAG GGG CAT CGT CAG G	227	1.5 mM
Exon11	CCA CTC TCA GCC TGA CGA TG GCG CAC AGA TAC TGA ACC AAG	246	1.5 mM
Exon12	GTC TTT ATG ACT TAA GGC TCC AC CCA GGA AGA GCA GCC CCC AC	275	0.5 mM

TABLE 6.2 (Cont.)

Oligonucleotide Primers for Mutation Analysis of T6

Primer set	Nucleotide sequences (5'-3')	Size (bp)	Optimal [Mg^{++}]
T6 SSCP			
Exon 13	CAG ACT CTC CAC CTT AGA ACA TCC CAG GCT CCG AAC CCA CA	265	1.5 mM
Exon 14	GAG GCA GGC TGT GGT CTG AAG AGG CCC CCG TGT CTG GGA TG	283	1.5 mM
Exon 15	TTG CTG TGC TTT ATC TGC TGT CAC GCA CGC CTG CAT GTG G	226	2.0 mM
Exon 16	GAC TCT GCT GGG AGA GGT AG GAG GTA CCT CTC TGG CCT GA	157	1.5mM
Exon 17	GGT CCC ACG AGT CTC AGT TC ATG GAT ATG CCC GCA CCC AC	182	1.0 mM
Exon 18.1	CTC CAC AGT GCC TCC TGC TT GGA CCA GGG AGC GAC ACA CT	354	1.5 mM
Exon 18.2	AGT CGG CCA GTT GCC TGA AG TGG GAC GGC GCT CAG TGC CT	131	1.0 mM
T6 Exon Skipping			
Set 1	CTG AGC CAG CTG GTG AAC CT GGA CCA GGG AGC GAC ACA CT	660	1.5 mM
Set 2	CTC ATC GAC CAC CTG GCC TT CGG TGG ATA TTC CTG GGT G	518	1.5 mM
Set 3	TGA GGA GTA CCA GCA GGC TC CAT GGT GAG CGC CTG CTG T	613	1.5 mM
Set 4	TCT CAT GCA AGT GGC AAA C CTG ATC CTC TTG AAA GCG GCA	552	1.5 mM
Set 5	GTG GTC GTT CGG TCC TAT GT GAG AAC GTG CTT CCT GGA G	550	1.5 mM

omission of [α-^{32}P]dCTP. PCR products were cleaned using QIAquick columns (2.2.16.1) and sequenced using DyeTerminator chemistry (2.2.18.1) with the primers originally used for the amplification.

Population screening of polymorphisms identified from SSCP analysis used the same PCR conditions as above, however 50 ng of genomic DNA was used as a template. Gel conditions were performed as for the original SSCP analysis (6.2.5.2).

6.2.5.2 SSCP Gels

To reveal any conformational differences between PCR products generated from tumour as opposed to normal tissue within affected individuals, non-denaturing gels were used. The dimensions of the gel plates used were 50 cm X 37.5 cm X 0.4 cm and two separate gel systems were used to maximise the chance of detecting a bandshift (Hayashi, 1992; Orita *et al.*, 1989).

6.2.5.2.1 10% Polyacrylamide Gels

10% SSCP gels were prepared by the addition of 24.5 ml of 40% acrylamide, 10 ml of 2% Bis, 5% glycerol, 20 ml of 5X TBE, and 130 µl each of TEMED and 25% (w/v) APS per 100 ml of gel solution. Polymerisation was allowed to proceed for at least 2 hours. The gel was then run for 24 hours at 700 volts in 1X TBE after sample loading.

6.2.5.2.2 Mutation Detection Enhancement (MDE) Gels

To increase the chance of detecting any mutation, samples were also run on MDE gels. MDE is a modified polyacrylamide based vinyl polymer, which allows for increased sensitivity

without loss of resolution. MDE gels were prepared by the addition of 30 ml of MDE solution, 9.6 ml of 1X TBE, and 100 µl each of TEMED and 25% (w/v) APS per 80 ml of gel solution. The gel was run for 24 hours at 700 volts in 0.6X TBE.

6.2.6 Patient Material used for SSCP

Tumour DNA was isolated from freshly frozen breast tissue containing at least 50% tumour cells as assessed by H and E stained sections. Corresponding normal DNA was isolated from peripheral blood lymphocytes from the same patients. All DNA samples were generously isolated and provided by one of our collaborators, Dr. Anne-Marie Cleton-Jansen (Leiden). The tumours used for this analysis were selected for LOH only at 16q24.3, loss at 16q24.3 and 16q22.1 with retention in between, whole 16q loss, and a small number with complex loss. This was to ensure that in the majority of samples tested, chromosomal band 16q24.3 was the region targeted by LOH. All LOH analysis of tumour samples was done by Dr. Anne-Marie Cleton-Jansen as previously described (Cleton-Jansen *et al.*, 1994), using at least 25 polymorphic markers spanning the whole 16q chromosome arm (see Figure 1.4 for the latest unpublished data).

6.2.7 Exon Skipping

Selected primers listed in Table 6.2 were used in RT-PCR reactions (2.2.15) on RNA isolated from breast cancer cell lines to test for the presence of exon skipping in the T6 gene. The breast cancer cell lines examined were BT20, MDA-453, ZR-75, BT-474, BT-549, ZR-75-30, UACC893, MDA-468, and MDA-231, with amplification from fetal heart RNA and the normal breast epithelial cell line, HBL-100, acting as positive controls. A total of 5 overlapping primer sets were used, 177.31/177.25, 177.34/177.17, 177.13/177.22, 177.38/177.40, and 177.1/177.8 with products being run on 2.5% agarose gels.

6.3 Results

6.3.1 Sequence Analysis of T6 cDNA Clones

Sequencing of the cDNA clone yc81e09 indicated an insert of 1,640 nucleotides which contained an open reading frame (ORF) of 1,540 bp extending from the 5' end. This sequence was used to screen dbEST which identified a total of 92 independent homologous human cDNA clones (including yc81e09) together with 51 mouse ESTs, indicating that it is a relatively abundant transcript. An overlapping cDNA, zb17g05, contained a larger insert which not only contained an extra 66 bp absent in yc81e09, but extended the sequence 5' a further 264 bp, with both additions maintaining the ORF. The 5' end was also homologous to the cDNA ze25b01 that was originally identified from sequence analysis of the cosmid 360D7. This clone provided a further 190 bp of 5' sequence and confirmed that both itself and yc81e09 most likely form part of the same transcript, as suggested by physical mapping results. Additional cDNA homologies (cDNA clone zq50d09) identified a further 60 bp of 5' sequence such that in total 2,220 bp of sequence was obtained which contained an ORF of 2,055 bp extending from the 5' end.

This sequence also contained all of the trapped exons (ET17.23, ET6.14, ET17.9, ET17.2, and ET17.14) which had been physically mapped to the same region, with ET17.9 being the adjacent 5' exon to ET6.14. These exons immediately provide a partial genomic structure to the gene.

6.3.2 T6 Northern Analysis and 5' Sequence Isolation

Ms. Joanna Crawford used the insert of the cDNA clone zb17g05, which contained most of the T6 sequence, to probe a commercial Northern blot filter. A 2.4 kb band was detected in all

tissues (data not shown). To obtain more 5' sequence for T6, 5' RACE was performed on fetal brain RNA using T6 gene specific primers (T6GSP) listed in Table 6.1. Figure 6.1A shows the results of these experiments. Two PCR amplified products of 300 bp and 334 bp were cloned and sequenced with the larger fragment extending the existing T6 sequence 5' by 27 bp. A separate 5' RACE procedure performed by Ms. Joanna Crawford also extended the T6 sequence to this point suggesting this may represent a transcription start site for the gene. In addition, analysis of homologous mouse entries in dbEST showed high homology (85%) over the entire transcript except for the sequence immediately 5' to the proposed transcription start. Taken together, these results provide evidence that the full-length sequence of T6 has been obtained. In total, the T6 transcript consists of 2,247 nucleotides, which contains an ORF of 2,007 bp encoding a protein consisting of 669 amino acids (Figure 6.2). A single polyadenylation site (TATAAA) was identified within the 3' UTR of the gene beginning at position 2,230 with the most 5' translation start site beginning at base 76. The methionine encoded by this start site does not have a strong KOZAK consensus, however the next 3' in-frame methionine is 706 bp away, suggesting the first ATG start site at base 76 is the one used.

Direct cosmid sequencing was used to obtain sequence immediately 5' to the proposed transcription start site of T6. Analysis of this sequence established that 279/370 (75%) nucleotides examined were either a G or C residue indicating that the 5' region is extremely high in G-C content. This is consistent with this region being the site for the promoter of T6.

6.3.3 T6 Sequence Homologies

BLASTP analysis of the NCBI non-redundant database with the T6 protein identified weak homology to a hypothetical protein of *S. pombe* (40% similarity over 657 amino acids), the *C. elegans* protein K07A12.1 (47% similarity over 401 amino acids), and a probable membrane

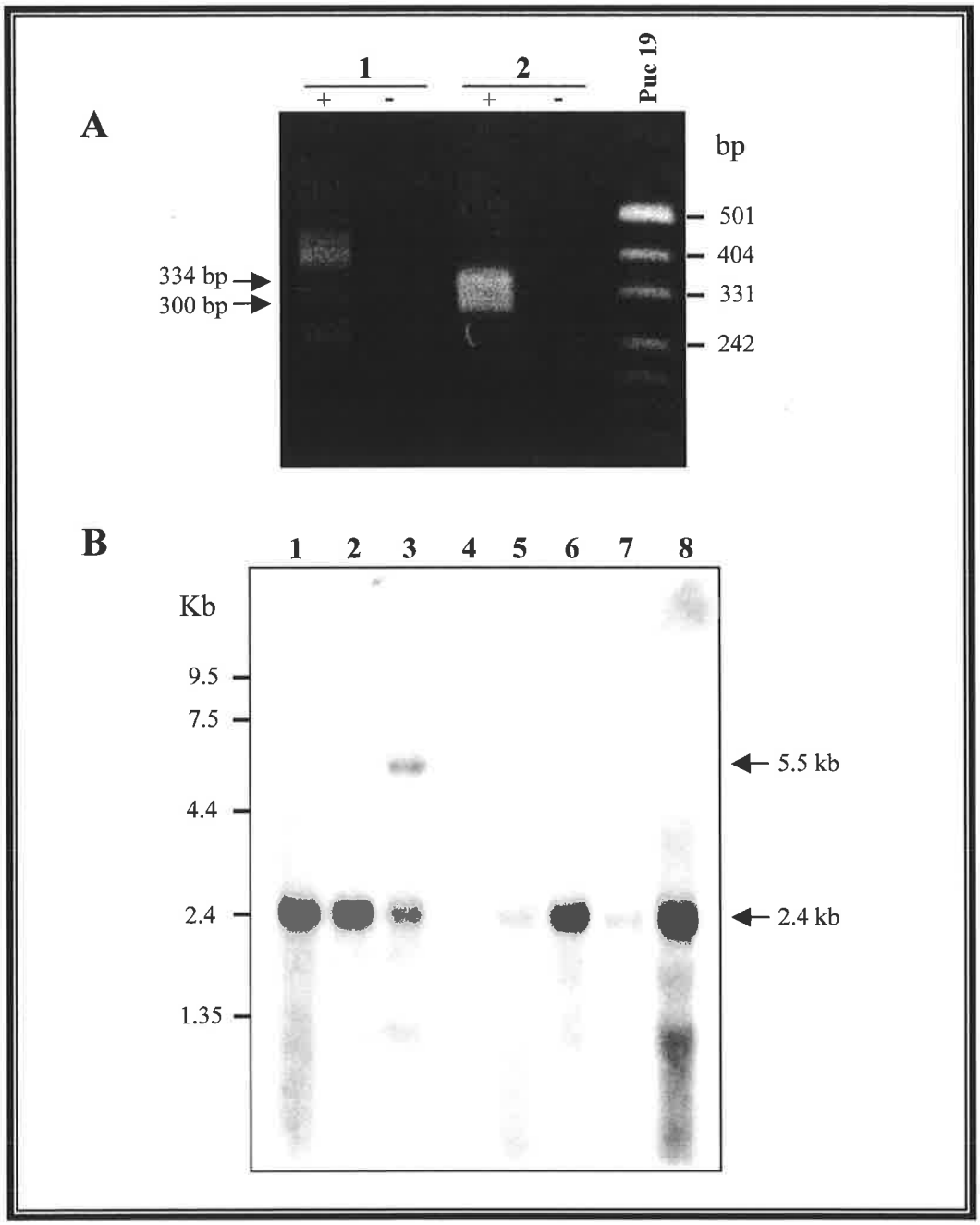


Figure 6.1: (**A**) 5' RACE of fetal brain RNA to obtain the 5' end of the T6 gene. Lane **1**: products generated after the first round of PCR with (+) or without (-) reverse transcriptase. Lane **2**: products generated from the second round of PCR with (+) or without (-) reverse transcriptase. The two fragments indicated by arrows were cloned and sequenced with the largest product extending the 5' sequence by 27 bp. (**B**) Northern blot analysis of the T6B transcript using the insert to the cDNA clone ze25b01. A single band of approximately 2.4 kb was highly expressed in adult **1**: Heart; **2**: Brain; **6**: Skeletal muscle; **8**: Pancreas. Fainter expression was observed in adult **3**: Placenta; **4**; Lung; **5**: Liver; **7**: Kidney. An additional band of ~5.5 kb was seen in the placenta alone. Northern analysis using a cDNA clone representing the entire T6 transcript identified an identical sized 2.4 kb mRNA.

Figure 6.2

Nucleotide and deduced amino acid sequence of the T6 transcript. Also shown is a portion of the 5' UTR of the gene. The proposed transcription start site is indicated by a red asterisk. The gene sequence extends over two pages. Numbers in the right hand column indicate the nucleotide and amino acid positions. The translation start site at position 76 and the stop codon at position 2,083 are indicated by green type. The polyadenylation signal is underlined. Horizontal arrowheads are located above the first and last nucleotide of a given exon. Each exon is represented by a red number with the corresponding trapped exon shown in brackets.

```
TGAAATGCTG CAGGCACGGT GCGGGGACAG GCCCCGGCGG GCAGAACGCG GCTGCGGGTT CCGGGCCACA      -300

TGGATGGGCA AGGCCGGGCA GGGCCGACTG TCGCCGGCTG CGCGCCCGTT ACCCCGAATC GAAGCCCGTC      -230

ATCCCCGCGG GGCACCTCTC CATGCCCGGT GCGCACGGCG CGCAGACACC CAGCTCTCCA GAGACCTGGC      -160

CCGGACCCCC CGGAGCCCGC CCAGACGCCG CCGCTGAAGC GGGAAAGTTC CGGAGAGCCG CGACGCGCAG      -90

GCGCCGCAGC CAGGTCCTGC CCCCACCCCT CGCGCCAAGA GTGCGCAGGC GCGCAGCCAG CTCCCCCGCC      -20

                                  *
CCCGACCCCG CGCGAAGAGT GCGCAGGCGC GCCGACAGCC GAGTTTTCTG CGCTTCCTTC TCCCTCTCTC       51

CAGACGTCGT GGTCGTTCGG TCCT ATGTCG CGCCGGGCCC TCCGGAGGCT GAGGGGGGAA CAGCGCGGCC      121
                             M  S  R  R  A  L  R  R  L  R  G  E  Q  R  G          15

AGGAGCCCCT CGGGCCCGGC GCCTTGCATT TCGATCTCCG TGATGACGAT GACGCGGAAG AAGAAGGGCC      191
Q  E  P  L  G  P  G  A  L  H  F  D  L  R  D  D  D  D  A  E  E  E  G  P           39

CAAGCGGGAG CTTGGTGTCC GGCGTCCCGG GGGCGCAGGG AAGGAGGGCG TCCGAGTCAA CAACCGCTTC      261
K  R  E  L  G  V  R  R  P  G  G  A  G  K  E  G  V  R  V  N  N  R  F              62

   1 ◀▶ 2 (ET17.14)
GAGCTGATAA ACATTGACGA TCTTGAGGAT GACCCTGTGG TGAACGGGGA GAGGTCTGGC TGTGCGCTCA      331
E  L  I  N  I  D  D  L  E  D  D  P  V  V  N  G  E  R  S  G  C  A  L              85

CAGACGCTGT GGCACCAGGG AACAAAGGAA GGGGTCAGCG TGGAAACACA GAGAGCAAGA CGGATGGAGA      401
T  D  A  V  A  P  G  N  K  G  R  G  Q  R  G  N  T  E  S  K  T  D  G  D           109

                  2 ◀▶ 3
TGACACCGAG ACAGTGCCCT CAGAGCAGTC TCATGCAAGT GGCAAACTCC GGAAGAAGAA AAAAAAACAG      471
D  T  E  T  V  P  S  E  Q  S  H  A  S  G  K  L  R  K  K  K  K  K  Q              132

                     3 ◀▶ 4
AAAAACAAGA AAAGCAGCAC GGGAGAAGCA TCGGAAAACG GACTAGAAGA TATCGATCGC ATCCTAGAGA      541
K  N  K  K  S  S  T  G  E  A  S  E  N  G  L  E  D  I  D  R  I  L  E              155

GGATTGAGGA CAGCACTGGG TTGAACCGTC CCGGCCCAGC TCCCCTGAGC TCCAGGAAGC ACGTTCTCTA      611
R  I  E  D  S  T  G  L  N  R  P  G  P  A  P  L  S  S  R  K  H  V  L  Y          179

      4 ◀▶ 5 (ET17.2)
CGTGGAGCAC AGACACTTGA ATCCAGACAC AGAACTGAAA AGGTATTTTG GTGCCCGGGC AATCCTGGGG      681
V  E  H  R  H  L  N  P  D  T  E  L  K  R  Y  F  G  A  R  A  I  L  G              202

   5 ◀▶ 6
GAGCAAAGGC CACGGCAGAG ACAACGTGTG TACCCCAAGT GCACATGGCT GACCACCCCT AAAAGCACCT      751
E  Q  R  P  R  Q  R  Q  R  V  Y  P  K  C  T  W  L  T  T  P  K  S  T              225

                  6 ◀▶ 7
GGCCCCGCTA CAGCAAACCA GGTCTGTCCA TGCGGCTGCT GGAATCAAAA AAAGGCCTCT CCTTCTTTGC      821
W  P  R  Y  S  K  P  G  L  S  M  R  L  L  E  S  K  K  G  L  S  F  F  A          249

GTTTGAGCAC AGTGAGGAGT ACCAGCAGGC TCAGCACAAG TTCCTGGTGG CCGTGGAGTC TATGGAGCCG      891
F  E  H  S  E  E  Y  Q  Q  A  Q  H  K  F  L  V  A  V  E  S  M  E  P              272

         7 ◀▶ 8
AACAACATCG TGGTTCTGCT CCAGACGAGC CCTTACCACG TTGACTCACT CCTGCAGCTC AGCGATGCCT      961
N  N  I  V  V  L  L  Q  T  S  P  Y  H  V  D  S  L  L  Q  L  S  D  A              295

                                 8 ◀▶ 9 (ET17.9)
GCCGCTTTCA AGAGGATCAG GAGATGGCTC GAGACCTCGT AGAGAGAGCG CTGTACAGCA TGGAATGTGC     1031
C  R  F  Q  E  D  Q  E  M  A  R  D  L  V  E  R  A  L  Y  S  M  E  C  A          319

                                                          9 ◀▶ 10
GTTCCACCCC CTGTTCAGTC TCACCAGTGG GGCCTGCCGG CTGGATTACC GCAGACCCGA GAACAGGAGC     1101
F  H  P  L  F  S  L  T  S  G  A  C  R  L  D  Y  R  R  P  E  N  R  S              342

TTCTACCTGG CCCTCTACAA GCAGATGAGC TTCCTGGAGA AGCGAGGCTG CCCGCGCACG GCGCTGGAGT     1171
F  Y  L  A  L  Y  K  Q  M  S  F  L  E  K  R  G  C  P  R  T  A  L  E              365
```

10 (ET6.14) ◀▶ 11

```
ACTGCAAGCT CATCCTGAGT CTCGAGCCGG ATGAGGACCC CCTCTGCATG CTGCTGCTCA TCGACCACCT   1241
Y  C  K  L   I  L  S   L  E  P   D  E  D  P   L  C  M   L  L  L   I  D  H  L   389

                                                    11 ◀▶ 12(ET17.23)
GGCCTTGCGG GCCCGGAACT ACGAGTACCT GATCCGCCTC TTCCAGGAGT GGGAGGCTCA TCGGAACCTG   1311
  A  L  R   A  R  N   Y  E  Y  L   I  R  L   F  Q  E   W  E  A  H   R  N  L   412

TCCCAGCTCC CTAATTTTGC CTTCTCTGTT CCACTGGCGT ATTTCCTGCT GAGCCAGCAG ACAGACCTCC   1381
 S  Q  L   P  N  F  A   F  S  V   P  L  A   Y  F  L  L   S  Q  Q   T  D  L    435

CTGAGTGTGA GCAGAGCTCT GCCAGGCAGA AGGCCTCTCT CCTGATACAG CAGGCGCTCA CCATGTTCCC   1451
P  E  C  E   Q  S  S   A  R  Q   K  A  S  L   L  I  Q   Q  A  L   T  M  F  P   459

12 ◀▶ 13
TGGAGTCCTC CTGCCCCTGC TCGAGTCTTG CAGTGTGCGG CCCGACGCCA GCGTTTCCAG TCACCGCTTC   1521
  G  V  L   L  P  L   L  E  S  C   S  V  R   P  D  A   S  V  S  S   H  R  F   482

                13 ◀▶ 14
TTTGGACCCA ATGCTGAAAT AAGCCAGCCC CCTGCCCTGA GCCAGCTGGT GAACCTGTAC CTTGGGAGGT   1591
 F  G  P   N  A  E  I   S  Q  P   P  A  L   S  Q  L  V   N  L  Y   L  G  R    505

CACACTTTCT CTGGAAAGAG CCCGCCACCA TGAGCTGGCT GGAGGAGAAC GTCCACGAGG TTCTGCAAGC   1661
S  H  F  L   W  K  E   P  A  T   M  S  W  L   E  E  N   V  H  E   V  L  Q  A   529

                                                14 ◀▶ 15
AGTGGACGCC GGGGACCCAG CCGTGGAAGC CTGTGAGAAC CGGCGGAAGG TGCTCTACCA GCGTGCACCC   1731
  V  D  A   G  D  P   A  V  E  A   C  E  N   R  R  K   V  L  Y  Q   R  A  P   552

                                                                15 ◀▶ 16
AGGAATATCC ACCGCCATGT GATCCTCTCT GAGATCAAGG AAGCCGTCGC TGCCCTGCCC CCGGACGTGA   1801
 R  N  I   H  R  H  V   I  L  S   E  I  K   E  A  V  A   A  L  P   P  D  V    575

CCACGCAGTC TGTGATGGGG TTTGATCCTC TGCCTCCTTC GGACACAATC TACTCCTACG TCAGGCCAGA   1871
T  T  Q  S   V  M  G   F  D  P   L  P  P  S   D  T  I   Y  S  Y   V  R  P  E   599

16 ◀▶ 17
GAGGCTAAGT CCTATCAGCC ATGGAAACAC CATTGCTCTC TTCTTCCGGT CACTGTTGCC AAACTATACC   1941
  R  L  S   P  I  S   H  G  N  T   I  A  L   F  F  R   S  L  L  P   N  Y  T   622

   17 ◀▶ 18
ATGGAGGGGG AGAGGCCCGA GGAAGGAGTG GCTGGGGGTT CTGAACCGCA ACCAGGGCCT GAACAGGCTG   2011
 M  E  G   E  R  P  E   E  G  V   A  G  G   S  E  P  Q   P  G  P   E  Q  A    645

ATGCTGGCTG TGCGCGACAT GATGGCCAAC TTCCACCTCA ACGACCTGGA GGCGCCGCAC GAGGACGACG   2081
D  A  G  C   A  R  H   D  G  Q   L  P  P  Q   R  P  G   G  A  A   R  G  R  R   669

CTGAGGGGGA GGGGGAGTGG GACTGAGCGT CCGCAGAGGT GACCGAAAAG CCGTATGATG ATGTTCCCGA   2151

TTTCTCTGTT GGTCGGAGTC GGCCAGTTGC CTGAAGTAGG GAAGCTGAGT GTGTCGCTCC CTGGTCCACT   2221

GTTTCTCCTA TAAATGTAAA TGGGTC                                                    2247
```

protein of *S. cerevisiae* (44% similarity over 366 amino acids). In addition, the high nucleotide sequence homology (85%) to mouse ESTs suggests this gene is conserved. Unfortunately, no significant homology was observed to known human proteins or proteins with known functions. The PSORT program (5.3.6) suggested the protein product may be targeted to the nucleus, however no functional domains were identified in the protein.

6.3.4 T6 Genomic Structure

The genomic structure of T6 was characterised by Ms. Joanna Crawford. The gene consists of 18 exons separated by 17 introns that collectively cover approximately 30 kb of genomic DNA. Table 6.3 summarises the sequences of the splice site junctions and relative splice site consensus strengths based on Shapiro and Senapathy (1987). All splice sites obeyed the ag/gt rule. Exon 1 contained the start codon while exon 18 contained the stop codon as well as the complete 3' UTR. The trapped exon ET17.14 corresponded to exon 2, ET17.2 was exon 5, ET17.9 was exon 9, ET6.14 was exon 10, and ET17.23 corresponded to exon 12.

6.3.5 Mutation Analysis

6.3.5.1 SSCP Results

To determine if this gene was a potential tumour suppressor gene involved in breast cancer, primers were designed from the intron sequences flanking each exon. These were used to test for the presence of mutations in breast tumour DNA. Primer sequences are listed in Table 6.2 and SSCP results for the analysis of all 27 tumour and matching normal DNA samples are shown in Table 6.4. Analysis of the first 10 exons by Ms. Joanna Crawford identified SSCP changes in exons 1.1, 9, and 10. These were considered to be polymorphisms as these changes were also seen in normal DNA from the same individuals. SSCP analysis of the remaining 8 exons by the candidate again identified changes seen in both normal and tumour DNA from the

TABLE 6.3

Splice Sites of the T6 Gene

Exon	Size (bp)	3' Splice site (intron/exon)	Consensus strength (%)	5' Splice site (exon/intron)	Consensus strength (%)
1	267	5'UTR		CTTCGAGCTG/**gt**gaggagcg	84.30
2	162	gtttttcc**ag**/ATAAACATTG	90.38	CTCAGAGCAG/**gt**gggggct	81.38
3	75	ctatcatt**ag**/TCTCATGCAA	76.81	AGAAGCATCG/**gt**acgtgagt	80.29
4	119	gggcttttt**ag**/GAAAACGGAC	78.66	TGGAGCACAG/**gt**gtggcccc	80.83
5	66	aattcccc**ag**/ACACTTGAAT	90.68	GGGAGCAAAG/**gt**aaggtcca	95.43
6	73	ttctccgc**ag**/GCCACGGCAG	90.90	AGCAAACCAG/**gt**gagggtct	92.15
7	131	cccctccc**ag**/GTCTGTCCAT	93.90	CAACATCGTG/**gt**gcgtggtc	77.37
8	100	cttctggt**ag**/GTTCTGCTCC	82.91	GACCTCGTAG/**gt**aaggcaga	95.43
9	94	gtgcccac**ag**/AGAGAGCGCT	86.13	CCGAGAACAG/**gt**gagtgcag	96.71
10	93	ttggatct**ag**/GAGCTTCTAC	79.79	TCATCCTGAG/**gt**gagtgtct	96.71
11	106	cctccgga**ag**/TCTCGAGCCG	68.93	GGAGTGGGAG/**gt**gggtgcga	85.94
12	160	tccctcgt**ag**/GCTCATCGGA	86.50	TTCCCTGGAG/**gt**gagtgagc	96.71
13	88	tcccctgc**ag**/TCCTCCTGCC	86.87	CTGAAATAAG/**gt**aaagagtg	81.02
14	159	tctcccccc**ag**/CCAGCCCCCT	87.51	GTGAGAACCG/**gt**gagctagg	82.66
15	91	ccggccct**ag**/GCGGAAGGTG	78.23	CCTGCCCCCG/**gt**aagggaga	87.22
16	80	ctttctga**ag**/GACGTGACCA	80.96	GGCCAGAGAG/**gt**acctccct	74.27
17	73	tttctttc**ag**/GCTAAGTCCT	95.46	TACCATGGAG/**gt**aggttgag	89.23
18	300	cttgctgc**ag**/GGGGAGAGGC	90.93	3'UTR	

Note. Consensus strengths are based on Shapiro and Senapathy, 1987. Exon sequences are shown in uppercase whereas intronic sequence is represented by lowercase letters. All splice site boundaries observed the gt/ag rule and are shown in bold.

TABLE 6.4

Results of SSCP Analysis of T6

Tumour	Exon									
loss 16q24.3	1.1	1.2	2	3	4	5	6	7	8	9
204	P	+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	P
309	+	+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
358	P	+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	P
367	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
413	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
549	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
559	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
589	P	+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	P
645	+	+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
666	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
757	+	+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
819	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
919	+	+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
Loss 16q22 to 16qter										
152	P	+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	P
380	+	+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
670	+	+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
683	+	+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
768	+	+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
594	+	+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
Loss whole 16q										
424	+	+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
438	+	+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
439	+	+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
448	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND
Complex loss										
355	+	+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
377	+	+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
555	+	+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
581	P	+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	P

Note. +/+ indicates the same SSCP pattern was observed in both the tumour DNA and normal DNA respectively from the same affected individual. A single + indicates that the tumour DNA only was examined for this individual to conserve samples. If an altered pattern was observed when compared to other tumour samples, the corresponding normal DNA was then examined. P indicates a polymorphism detected as a bandshift in tumour DNA and later confirmed to be present in the matching normal DNA. ND: not done.

TABLE 6.4 (Cont.)

Results of SSCP Analysis of T6

Tumour	Exon									
Loss 16q24.3	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18.1**	**18.2**
204	P	+/+	P	+	+	+	+/+	+	+	+
309	+/+	+/+	+/+	+	+	+	+/+	+	+	+
358	P	+/+	P	+	+	+	+/+	+	+	+
367	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
413	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
549	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
559	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
589	P	+/+	P	+	+	+	+/+	+	+	+
645	+/+	+/+	+/+	+	+	+	+/+	+	+	P
666	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
757	+/+	+/+	+/+	+	+	+	+/+	+	+	+
819	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+	+/+
919	+/+	+/+	+/+	+	+	+	+/+	+	+	+
Loss 16q22 to 16qter										
152	P	+/+	P	+	+	+	+/+	+	+	+
380	+/+	+/+	+/+	+	+	+	+/+	+	+	+
670	+/+	+/+	+/+	+	+	+	+/+	+	+	+
683	+/+	+/+	+/+	+	+	+	+/+	+	+	+
768	+/+	+/+	+/+	+	+	+	+/+	+	+	+
594	+/+	+/+	+/+	+	+	+	+/+	+	+	+
Loss whole 16q										
424	+/+	+/+	+/+	+	+	+	+/+	+	+	+
438	+/+	+/+	+/+	+	+	+	+/+	+	+	+
439	+/+	+/+	+/+	+	+	+	+/+	+	+	+
448	ND	+/+	P	+	+	+	+/+	+	+	P
Complex loss										
355	+/+	+/+	+/+	+	+	+	+/+	+	+	+
377	+/+	ND	ND	ND	ND	ND	ND	ND	ND	ND
555	+/+	+/+	+/+	+	+	+	+/+	+	+	P
581	P	+/+	P	+	+	+	+/+	+	+	+

Note. +/+ indicates the same SSCP pattern was observed in both the tumour DNA and normal DNA respectively from the same affected individual. A single + indicates that the tumour DNA only was examined for this individual to conserve samples. P indicates a polymorhism detected as a bandshift in tumour DNA and later confirmed to be present in the matching normal DNA. ND: not done.

same individual in exons 12, and 18.2. Figure 6.3 shows the altered conformation pattern and associated nucleotide change seen for exons 12 and 18.2. The frequency of the polymorphisms in the general population is shown in Table 6.5 with the relevant base changes also indicated.

TABLE 6.5

T6 Polymorphisms Detected During SSCP Analysis

Exon	Polymorphism	Frequency
1.1	ND	0.11 (12/110)
9	IVS8-4A/G	0.117 (14/120)
10	G1188A	0.102 (12/118)
12	IVS12+15T/C	0.105 (12/114)
18.2	3'UTR+124C/T	0.042 (5/118)

Note. Figures in brackets represent the number of chromosomes with the observed SSCP change as a fraction of the total number of chromosomes analysed. ND: not determined.

All five of the polymorphisms were found to be relatively rare with four occurring outside the coding region. The polymorphism detected in exon 10 identified a base change within the coding region at base 1,188. This presumably neutral polymorphism resulted in an acidic glutamine amino acid being replaced by a basic lysine residue. The absence of tumour specific SSCP changes suggests T6 is not involved in breast carcinogenesis.

6.3.5.2 Exon Skipping

Of all the breast cancer cell lines examined, each one gave rise to the expected sized RT-PCR product with all primer sets tested. This excluded exon skipping as a mutational mechanism.

6.3.6 Alternative T6 Transcripts

As shown in Figure 6.4, three alternate T6 transcripts (T6A, T6B, and T6C) were identified

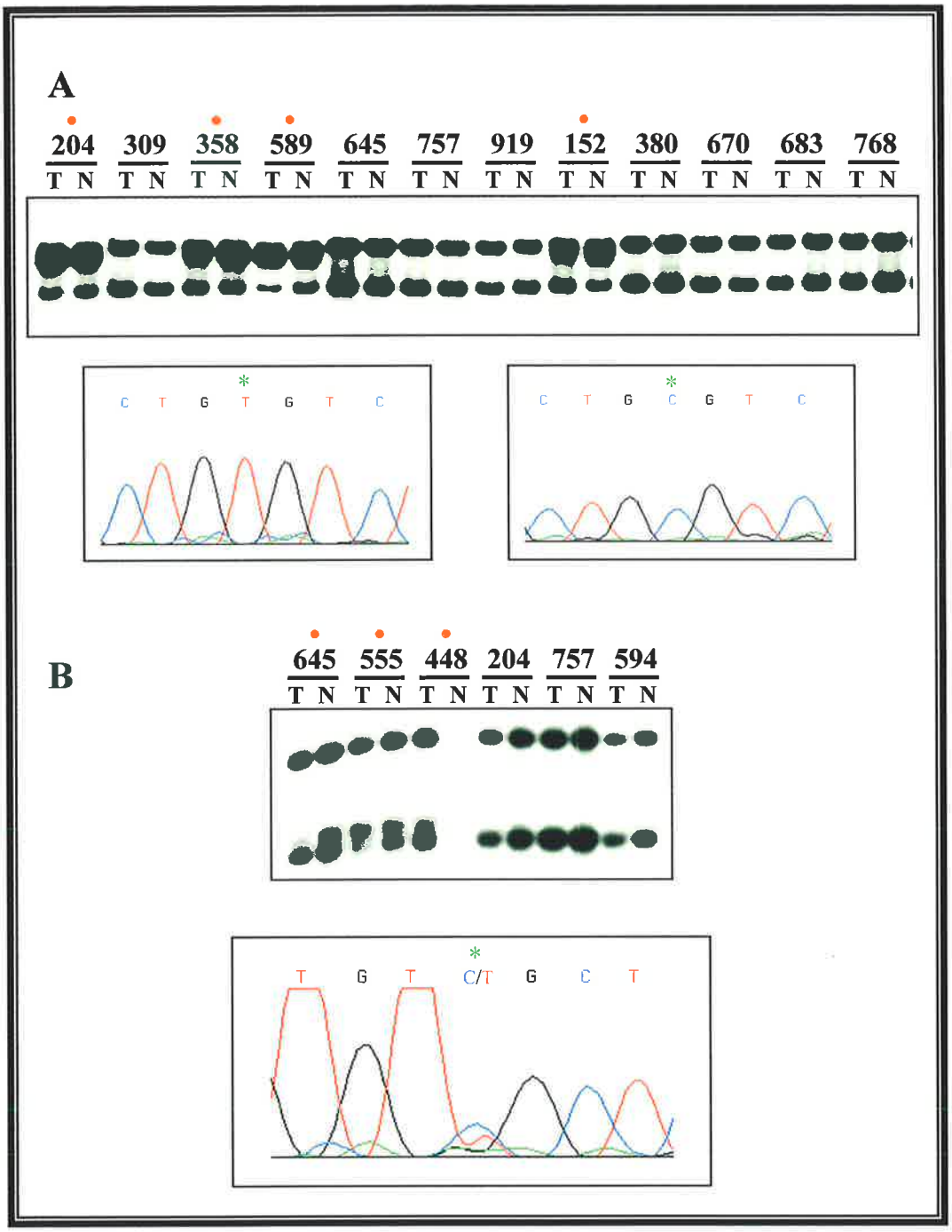


Figure 6.3: SSCP analysis of T6 exon 12 (**A**) and exon 18.2 (**B**). In each case, a section of the autoradiograph associated with the exon is shown on top, with each tumour sample indicated. **N**: normal DNA, **T**: tumour DNA. The sequence of the relevant base change is shown below each autoradiograph. The green asterisk indicates the nucleotide that is polymorphic, and the red dots represent the individual tumour samples that have the observed polymorphism. (**A**) The C nucleotide present in intron 12 was only present in approximately 10% of normal individuals tested (Table 6.5). (**B**) An individual homozygous for the rare T nucleotide (4% of the general population) was not identified such that the sequence of a heterozygote is shown.

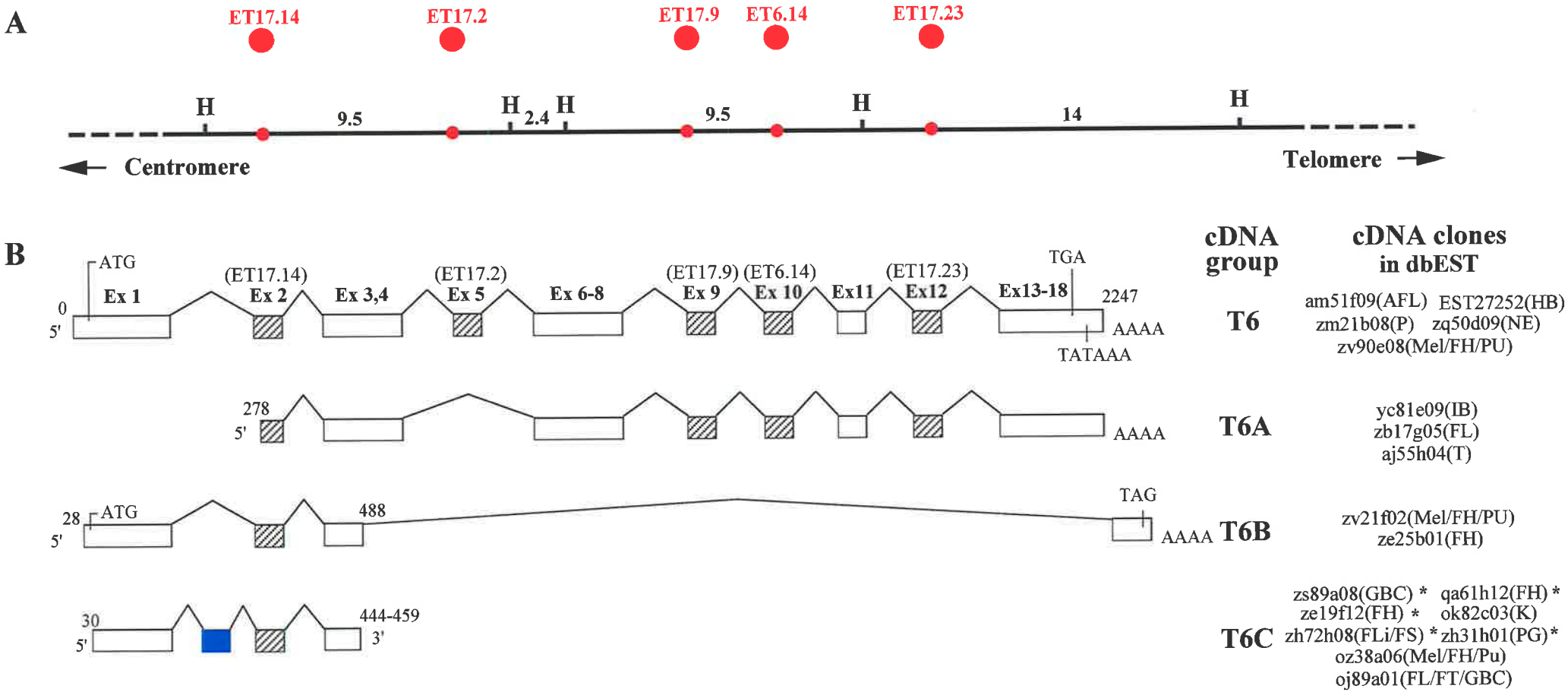


Figure 6.4: (**A**) A partial restriction map of the 16q24.3 region encompassing the T6 transcript. H: *Hind*III. The locations of the trapped exons corresponding to T6 are indicated on the map. The size of restriction fragments are in kilobases. (**B**) cDNA groups constituting T6. Sequence comparisons of cDNA clones identified four distinct transcripts (T6, T6A, T6B, and T6C). For transcripts T6A, T6B, and T6C the clone with the largest insert is represented. The number at the 5' end indicates how far into the complete T6 sequence the clone extends, while for T6B and T6C clones, numbers at the 3' ends indicate where in the T6 sequence this group of clones diverges (T6B) or originates (T6C). Of the 92 cDNA clones homologous to T6 in dbEST, only those clones which could be placed in one of the four groups are indicated. The cDNA libraries from which they were derived are indicated in brackets. AFL: adult frontal lobe; HB: human brain; NE: neuroepithelium; P: pancreas; Mel: melanocyte; FH: fetal heart; PU: pregnant uterus; IB: infant brain; FL: fetal lung; T: testis; GBC: germinal B cell; K: kidney; FLi: fetal liver; FS: fetal spleen; PG: pineal gland; FT: fetal testis. T6A clones were missing the trapped exon ET17.2. T6B clones were missing a large proportion of the T6 sequence and contained a different 3' origin. The final group of clones (T6C) appeared to originate from within the middle of T6 as a result of the binding of oligo-dT and subsequent initiation of cDNA synthesis at a region of high (65%) A residue content. Five of these clones (indicated by an asterisk) contained an additional 167 bp of sequence (represented by the blue box) 5' to exon 2 (ET17.14).

from BLASTN analysis of human ESTs, a search of the TIGR human gene index database (http://www.tigr.org/tdb/hgi/searching/hgi_seq_search.html), and trapped exon content.

6.3.6.1 T6A

T6A cDNA clones (THC230405), which included yc81e09, did not contain exon 5 (ET17.2). Of the 92 human ESTs homologous to T6 in dbEST, only eight had 5' sequence which extended beyond exon 6, and three of them did not contain exon 5. Therefore, the relative frequency of the T6A transcript within the 92 cDNA clones could not be determined due to the majority of clones not extending beyond exon 6 or 5' DNA sequences not being available. During exon skipping mutation analysis of T6 (6.3.5.2), RT-PCR on breast cancer cell lines and fetal heart RNA confirmed that the alternative transcript T6A was present in RNA from these sources. The absence of ET17.2 does not alter the ORF of the transcript since this exon is 66 bp in length. This indicates T6A may represent a functional alternatively spliced form of T6.

6.3.6.2 T6B

The T6B transcript (THC230404) was represented by only two cDNA clones (ze25b01 and zv21f02) from the 92 ESTs homologous to T6 in dbEST. This transcript contains only the first 489 nucleotides of T6 (including exon 1, exon 2, and 59 bp of exon 3) followed by 124 bp of unique sequence that results in an alternative stop codon and 3' end. The T6B transcript is 612 bp in length (Figure 6.5) and contains a single polyadenylation signal (AATAAA) 23 bp from the 3' end and an in-frame stop codon 10 bp after the point of divergence from T6 exon 3, such that it codes for a potential 141 amino acid protein.

RT-PCR experiments were performed to verify the presence of T6B transcripts (Figure 6.6).

```
TGCGCAGGCG CGCCGACAGC CGAGTTTTCT GCGCTTCCTT CTCCCTCTCT CCAGACGTCG TGGTCGTTCG    70

GTCCTATGTC GCGCCGGGCC CTCCGGAGGC TGAGGGGGGA ACAGCGCGGC CAGGAGCCCC TCGGGCCCGG   140
      M  S   R  R  A   L  R  R   L  R  G  E   Q  R  G   Q  E  P   L  G  P  G    22

CGCCTTGCAT TTCGATCTCC GTGATGACGA TGACGCGGAA GAAGAAGGGC CCAAGCGGGA GCTTGGTGTC   210
  A  L  H   F  D  L   R  D  D  D   D  A  E   E  E  G   P  K  R  E   L  G  V    45

                                                          1 ◀▶ 2
CGGCGTCCCG GGGGCGCAGG GAAGGGGGGC GTCCGAGTCA ACAACCGCTT CGAGCTGATA AACATTGACG   280
 R  R  P   G  G  A  G   K  G  G   V  R  V   N  N  R  F   E  L  I   N  I  D     68

ATCTTGAGGA TGACCCTGTG GTGAACGGGG AGAGGTCTGG CTGTGCGCTC ACAGACGCTG TGGCACCAGG   350
D  L  E  D   D  P  V   V  N  G   E  R  S  G   C  A  L   T  D  A   V  A  P  G    92

GAACAAAGGA AGGGGTCAGC GTGGAAACAC AGAGAGCAAG ACGGATGGAG ATGACACCGA GACAGTGCCC   420
  N  K  G   R  G  Q   R  G  N  T   E  S  K   T  D  G   D  D  T  E   T  V  P   115

      2 ◀▶ 3
TCAGAGCAGT CTCATGCAAG TGGCAAACTC CGGAAGAAGA AAAAAAAACA GAAAAACAAG AAAAGCAG TT  490
 S  E  Q   S  H  A  S   G  K  L   R  K  K   K  K  K  Q   K  N  K   K  S  S     138

TTTCCCCCTA GGGGTGGGAG GAAGCAAAAG ACTCTGTACC TATTTTGTAT GTGTATAATA ATTTGAGATG   560
F  S  P                                                                       141

TTTTTAATTA TTTTGATTGC TGGAATAAAG CATGTGGAAA TGACCCAAAC AT                      612
```

***Figure 6.5*:** Nucleotide sequence of the T6B transcript. The corresponding amino acids are indicated below the nucleotide sequence. Residue numbers are indicated in the right hand column. The gene contains sequence identical to exon 1, 2, and the first 59 bp of exon 3 from T6, and an additional 124 bp of sequence (indicated in red type) unique to T6B. Exon/Exon junctions are indicated by black triangles with the corresponding exon indicated in red adjacent to the triangle. The translation start site and open reading frame stop codon are indicated by green type. The polyadenylation signal is underlined.

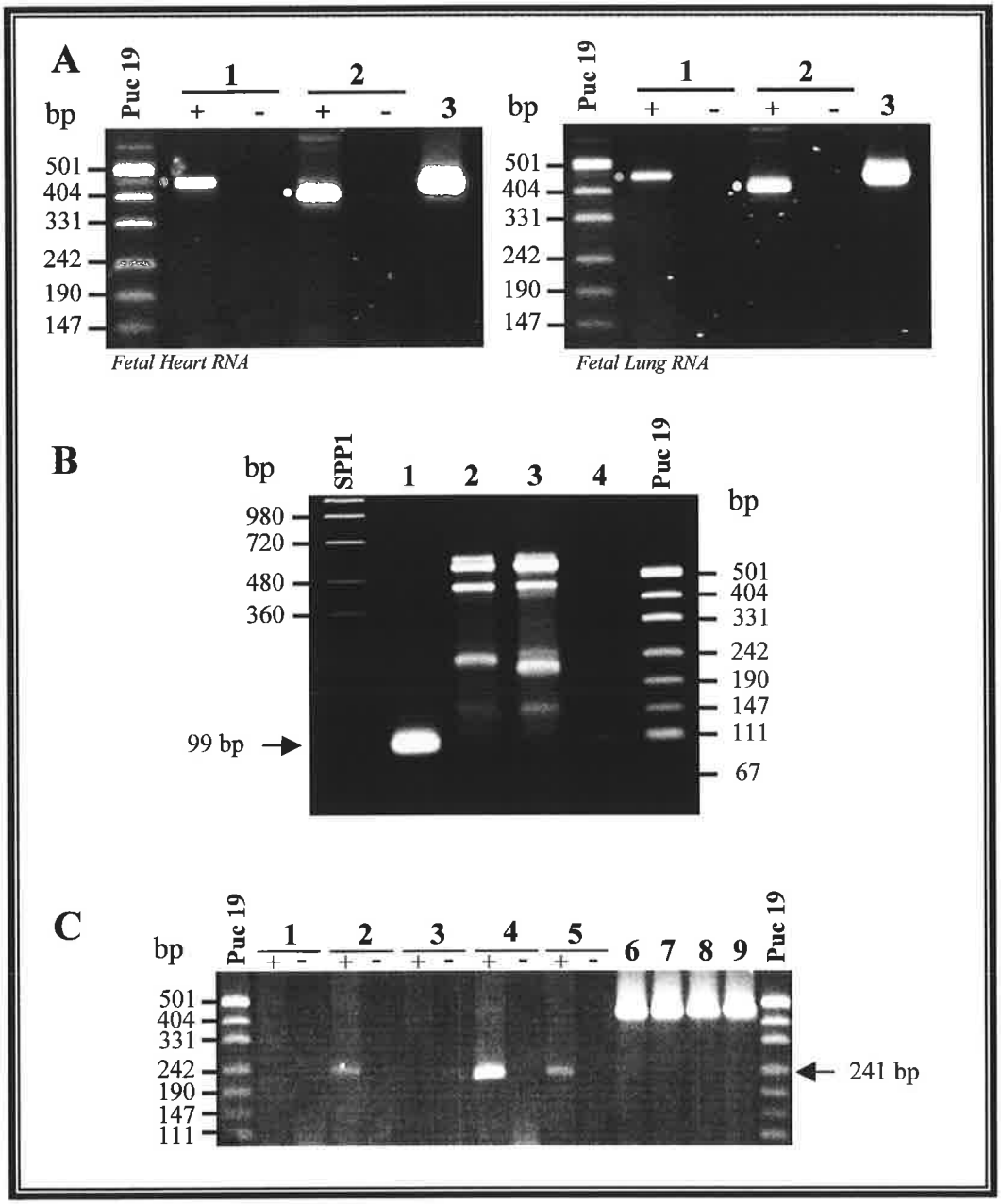


Figure 6.6: (**A**) Confirmation of the existence of T6B. RT-PCR on fetal heart and lung RNA with the AA3 primer and the AA2 primer (specific for the unique 3' end of T6B) identified the expected sized 455 bp fragment represented by the green dots (lane **1**). The T6 transcript 3' end was verified (lane **2**) with the use of the AA3 primer and the zq3 primer (specific for the 3' end of T6). The expected 417 bp fragment (represented by the yellow dots) was amplified in both tissues. (**B**) Mapping the T6B transcript by PCR using primers specific for the unique 3' end. Templates are **1**: Human genomic DNA; **2**: A9 DNA; **3**: CY18 DNA; **4**: no DNA. The expected 99 bp fragment was only amplified in the human DNA control and not in CY18. This suggests T6B does not map to chromosome 16. (**C**) Verification of the existence of the extra 167 bp "exon" seen in T6C transcripts. Primers AA3 and AA5 (situated either side of this sequence) were used in RT-PCR reactions on fetal **1**: Liver; **2**: Spleen; **3**: Heart; **4**: Lung; **5**: Brain RNA. A single band of 241 bp was detected in the spleen, lung, and brain only, which is the expected sized product if the 167 bp is not included. Control lanes with the esterase D primers are shown in **6-9** (liver, spleen, heart, and lung respectively). The integrity of the fetal brain cDNA had previously being verified. In (**A**) and (**C**) +: reverse transcription with enzyme; -: reverse transcription without enzyme.

When a primer located in a region shared by T6 and T6B (AA3) was used in combination with a primer unique to T6B (AA2), products of the expected size corresponding to T6B were identified in all tissues examined (Figure 6.6A).

6.3.6.2.1 Mapping T6B

Since the sequence of the 3' end of T6B did not correspond to T6 it was possible that this transcript represented an independent gene. Primers were designed specifically to the 3' end of T6B (AA9 and AA2) and used in a PCR reaction on human DNA and on DNA from the mouse/human somatic cell hybrid CY18 that contains chromosome 16 as its only human content. A band of the correct size was detected with human DNA as a template, however no product was seen when CY18 DNA was the template (Figure 6.6B). This suggested that the transcript was derived from another chromosome.

FISH analysis of metaphase chromosomes using the cDNA clone zq50d09 (representing the majority of the T6 transcript) detected signal on chromosome 16 at q24.3 (data of Elizabeth Baker, WCH). Although the homology between T6 and T6B extends for 489 bp, the zq50d09 clone contained only 211 bp of sequence in common with T6B. Therefore, when zq50d09 was used as a FISH probe, the amount of homology to T6B was likely to be too small to generate an observable FISH signal elsewhere in the genome corresponding to T6B. The cDNA clones available for the T6B transcript (zv21f02 and ze25b01) were less than 1 kb in size and therefore were also considered too small to be used as FISH probes for mapping purposes.

6.3.6.2.2 T6B Northern Analysis and Sequence Homologies

The insert of the cDNA clone corresponding to T6B (ze25b01) was used as a hybridisation probe on Northern blot filters containing normal adult tissues. A single ubiquitously expressed mRNA of approximately 2.4 kb was detected which was identical in size to that observed with a T6 probe. The strongest signal was observed in the heart, brain, liver, and pancreas (Figure 6.1B). An additional band of ~5.5 kb was also seen in the placenta alone. However, a mRNA corresponding to the size of the T6B transcript was not identified.

The 124 bp of unique nucleotide sequence in T6B displays 98% homology over its entire length to a portion of the 3' UTR of the mouse and human alpha-1 collagen type I (*COL1A1*) gene. PSORT analysis of this protein suggests it may be targetted to the nucleus, however BLASTN and P analysis of this transcript revealed no significant homology to known proteins or to protein domains.

6.3.6.3 T6C

T6C cDNA clones (THC213907) appeared to originate from within the middle of T6 due to the binding of oligo-dT, and subsequent initiation of cDNA synthesis, at a region of high (65%) A residue content (bases 445 to 484). Of the 92 human ESTs homologous to T6, 8 corresponded to T6C transcripts. Of these, 5 contained an additional 167 bp of sequence immediately 5' to exon 2 (ET17.14) which was not seen in all other overlapping cDNA clones (Figure 6.4B). Of the remaining three, 2 clones did not contain this extra sequence while the final clone did not extend far enough 5' to determine if the 167 bp was present or absent. Collectively, these 8 cDNA clones corresponding to T6C were derived from 10 different tissue sources, which suggested they may not be artifacts generated during the construction of the cDNA libraries. The presence of this extra sequence however results in a disruption of the

open reading frame of T6. Therefore, to confirm the existence of transcripts containing this sequence, RT-PCR using primers located either side of this sequence (AA3 and AA5) was performed on a number of fetal tissues from which these clones were originally derived. Of 5 tissues used (liver, spleen, heart, lung, brain), none confirmed the presence of transcripts containing this extra sequence (Figure 6.6C). Subsequently, the characterisation of the genomic structure of T6 (see 6.3.4) established that the extra 167 nucleotide "exon" present in T6 did not correspond to the intron sequence of the flanking T6 exons. However, the entire intron was not sequenced and it is suggested that this extra sequence may have originated from further within this intron.

6.4 Discussion

6.4.1 T6

The T6 gene has an ORF of 2,007 bp that codes for a protein of 669 amino acids. The gene spans approximately 30 kb of genomic DNA, is transcribed from the centomere toward the telomere, and consists of 18 exons, five of which were identified from exon trapping. A single polyadenylation signal at base 2,230 was observed and a single transcription start site based on RT-PCR results was detected.

Based on Northern analysis the gene is expressed as a single 2.4 kb mRNA in all tissues examined. The gene is also expressed in the breast, as detected by RT-PCR on breast cancer cell lines and the normal breast epithelial cell line, HBL-100. The T6 gene shares 85% nucleotide identity to its mouse homologue, and significant amino acid homology with proteins of unknown function from *S. pombe*, *C. elegans*, and *S. cerevisiae* (6.3.3) suggesting that the gene may be conserved. Although the gene is predicted to code for a protein that is targetted

to the nucleus, the lack of homology with any other proteins or protein motifs meant it was not possible to define a role for the protein product in the cell cycle. However, given that T6 is ubiquitously expressed and its 5' UTR is G-C rich (75% in 370 bp), it is most likely a housekeeping gene.

The results of SSCP mutation analysis on breast tumour DNA failed to detect nucleotide substitutions specific for tumour material and exon skipping was unable to identify breast cancer cell lines deleted for T6 exons. This suggests T6 is not a tumour suppressor gene mutated in breast cancer. Of the 5 polymorphisms detected, only one occurred in the coding region of T6 within exon 10 resulting in the substitution of a glutamic acid residue for a lysine. The significance of this change is not known but would be expected to have some effect on the protein product due to an alteration of its overall charge. One polymorphism in intron 8 occurred only 4 bases from the start of exon 9 however this change did not affect the consensus splice site score and would not be expected to have any effect on exon splicing. It is still possible that T6 may be involved in breast tumourigenesis if mutations occur outside the coding region or transcriptional unit. However, as explained with the *GAS11* gene, the sporadic nature of the tumours examined would dictate that not all disease causing mutations would occur outside the coding region.

6.4.2 T6 Isoforms

Based on sequence homology searches of dbEST, a number of possible T6 isoforms were identified although only a single ubiquitous band was detected on Northern blots. T6A transcripts did not contain exon 5 (ET17.2), but Northern analysis would be unable to resolve such a small size difference between T6 and T6A. However, the presence of T6A transcripts in breast cancer cell lines and the fetal heart was confirmed by RT-PCR. Coupled with the fact

that exon 5 is 66 bp in length and its absence would not disrupt the ORF, suggests T6A is a functional alternatively spliced form of T6.

The T6B transcript is 612 bp in length and codes for a putative protein of 141 amino acids. It contains sequence identical to exon 1, exon 2, and the first 59 bp of exon 3 from T6, and an additional 124 bp of sequence unique to T6B. This transcript does not appear to originate from chromosome 16. The exact location of T6B in the genome was not determined due to time constraints. The existence of RNA transcripts corresponding to T6B was confirmed by RT-PCR, however only two cDNA clones present in dbEST (from the 92 that showed homology to T6) represented T6B transcripts. This suggests T6B is not abundantly expressed.

Northern analysis using a T6B cDNA clone identified the 2.4 kb mRNA corresponding to T6 as well as a 5.5 kb mRNA seen in the placenta alone. Although the sequence of this larger transcript was not determined, it may have been derived from the use of an alternative polyadenylation signal 3' to the one used by either T6 or T6B. This extra sequence may contain signals specific for expression in the placenta. A mRNA signal corresponding to the 612 bp T6B gene was not seen from Northern hybridisations. It is possible that because of the small size of the associated mRNA, it may have been run off the gel during preparation of the Northern filter and hence not seen during hybridisations. Another explanation is that the gene may be poorly expressed such that it can only be detected by RT-PCR. Alternatively, the RNA blot probed consisted of adult tissues whereas all RT-PCRs and cDNA clones were derived from fetal tissues suggesting T6B may be only expressed during development stages.

BLASTN analysis of the unique 3' UTR of T6B revealed almost 100% homology to a section of the 3' UTR of the *COL1A1* gene. This corresponds to a stretch of nucleotides present in the

vicinity of the second polyadenylation signal (pA2) of this gene (Maatta *et al.*, 1991), which is also seen in the equivalent pA5 site of the *COL1A2* gene (Myers *et al.*, 1983). This region includes a GCATGT motif that immediately follows the polyadenylation site together with a conserved TGTACCTATTTTGTAT element present about 30 bp upstream. In protein gel shift assays, a 334 bp region of the *COL1A1* gene incorporating this conserved polyadenylation site was shown to exhibit cell-specific binding of nuclear proteins from a HeLa nuclear extract but not an NS-1 extract (Maatta *et al.*, 1991). In addition, nuclear proteins of human fetal and embryonic chicken tendon fibroblasts also recognised this region, with this specific binding inhibited by an excess of nonspecific competitors. This strongly suggested the conserved domain contained regulatory elements important for cell specific expression of the *COL1A1* gene. The identification of the same region in the T6B gene suggests it too may bind the same proteins as *COL1A1* to regulate its expression possibly by affecting the half-life of the T6B mRNA. T6B may therefore be an independent gene on another chromosome, which shares homology to both the T6 transcript and to 3' regulatory elements present in the *COL1A1* gene.

The T6C transcripts originated by incorrect binding of the RNA with the oligo-dT primer used for the construction of the cDNA libraries from where they derived. The majority of these clones contained an extra 167 bp of sequence between exons 1 and 2 that did not correspond to intronic DNA immediately adjacent to these exons. Inclusion of this sequence disrupts the ORF of T6 such that the next available methionine is present in exon 7. It is most likely that this 167 bp "exon" was derived from splicing occurring within intron 1 of T6. However, given that the entire intron was not sequenced, this cannot be confirmed at this stage. As all T6C cDNA clones were sequenced from normalised libraries and RT-PCR on total RNA failed to confirm the existence of this "exon" in the tissues from which T6C cDNA clones were derived, it appears that if these transcripts are functional, they are expressed at a relatively low

abundance. It was concluded that this transcript is most likely an artifact of cDNA library production.

Chapter 7

Cloning of the Gene for Cystinosis

Table of Contents

Page

7.1 Introduction **227**

7.1.1 Nephropathic Cystinosis 227

7.1.2 Clinical Course of Cystinosis 227

7.1.3 Basic Defect in Cystinosis 228

7.1.4 Mapping the Cystinosis Gene 228

7.2 Methods **230**

7.2.1 Exon Trapping 230

7.2.2 Exon Analysis 231

7.2.2.1 Colony PCR of Subcloned Trapped Exons 231

7.2.2.2 Trapped Exon Insert Amplification 231

7.2.2.3 Colony Master Plate Screening of Second Round Exon Trapped Products 232

7.2.2.4 Sequence Analysis and Physical Mapping 232

7.3 Results **233**

7.3.1 First Round Exon Trapping 233

7.3.2 Further Refinement of the Genetic Interval for the Cystinosis Gene 236

7.3.3 Second Round Exon Trapping 238

7.3.4 Cloning of the Cystinosis (*CTNS*) Gene 243

7.3.4.1 Screening CTNS for Small Mutations 245

7.4 Discussion **245**

7.1 Introduction

7.1.1 Nephropathic Cystinosis

Nephropathic cystinosis is an autosomal recessive disease with an incidence of approximately 1 in 200,000 live births. The disease is characterised clinically by generalised proximal renal tubular dysfunction (the renal Fanconi syndrome) and biochemically, by intracellular accumulation of cystine. This is caused by a defect in the transport of cystine out of the lysosome, a process mediated by an unidentified carrier (reviewed in McDowell *et al.*, 1997). Although rare, it is an important disorder as it serves as the prototype for lysosomal storage diseases due to deficiency of a membrane transporter as opposed to an enzyme deficiency (Gahl *et al.*, 1995). The diagnosis for cystinosis is based upon elevated leukocyte or fibroblast cystine levels as well as the presence of corneal crystals. Prenatal diagnosis is also possible based upon measurement of cystine levels in cultured amniotic fluid cells or chorionic villi (Gahl *et al.*, 1995).

Cystinosis has been subdivided into three types. The classical disorder, referred to as nephropathic cystinosis, occurs in 95% of cases and results in renal failure within the first decade of life. Late onset, or adolescent cystinosis, has only been diagnosed in a few patients whose renal disease occurs later compared with the infantile form. Finally, in adult cystinosis, the few patients that have been identified to date have corneal cystine crystals but do not have renal disease. Based on complementation studies, it has been suggested that the different forms of cystinosis are allelic (Pellet *et al.*, 1988).

7.1.2 Clinical Course of Cystinosis

Newborns with nephropathic cystinosis are clinically normal, however by 6 to 18 months of

age some of the effects of the renal tubular Fanconi syndrome are usually present, such as severe fluid and electrolyte disturbance (vomiting, dehydration, rickets), polyuria, and a failure to grow and gain weight. At the age of about 10, most patients present with end-stage renal failure due to deterioration of glomerular function. While the renal failure can be treated by renal transplantation, the donor kidney does not correct the metabolic disorder resulting in the accumulation of cystine in other organs, leading to additional complications. These include continued retinal damage and eventual blindness in many patients, hand muscle weakness and wasting, difficulty in swallowing, pancreatic exocrine and endocrine insufficiency, and delayed puberty (reviewed in McDowell *et al.*, 1997).

7.1.3 Basic Defect in Cystinosis

The intracellular cystine accumulation was demonstrated in the late 1960s, however further research failed to reveal a defect in the degradation of cystine (Gahl *et al.*, 1995). Although a lysosomal membrane transport defect was suspected, evidence for such a molecule was not available until 1982, when cystinotic leukocyte lysosomes were shown to exhibit virtually no cystine removal following cystine loading, as compared to normal cells. In addition, cystinotic cells had a prolonged half-life for cystine retention, but normal half-life for the removal of other amino acids (reviewed in McDowell *et al.*, 1997). Further experiments established that cystine transport was bi-directional and stimulated by ATP, and heterozygotes for cystinosis had half the velocity of cystine transport in their cellular lysosomes as compared to normal cells. This gene dose effect therefore verified that a lysosomal transport defect was the primary cause of cystinosis.

7.1.4 Mapping the Cystinosis Gene

The direct purification of the lysosomal cystine carrier protein using biochemical methods was

technically difficult such that a genomic approach to identify the protein was initiated. This began with genetic linkage mapping of 23 families segregating the disease, using 328 polymorphic marker loci covering the entire genome. Results from these studies indicated that the gene was most likely to be located on the short arm of chromosome 17 between D17S796 and D17S1583, a distance of approximately 7.6 cM (The Cystinosis Collaborative Research Group, 1995). Subsequent haplotype analysis with additional markers mapping to this region was successful in narrowing this interval to 3.6 cM between D17S1584 and D17S1583 (McDowell *et al.*, 1996) while another study reduced this to 3.1 cM between D17S1828 and D17S1798 (Jean *et al.*, 1996).

Following the initial linkage mapping studies, a physical map was constructed based on YACs covering most of the region between D17S1584 and D17S1583 (McDowell *et al.*, 1996). A selection of these YACs were then used to identify additional CA repeats located within the region, and in total, 10 were subsequently isolated and sequenced. Two of these repeats, D17S2167 and D17S2169, were sufficiently polymorphic for additional linkage mapping. Two cystinosis families were then found to have recombination events between D17S1828 and the new marker D17S2167 (McDowell *et al.*, 1997). The affected individuals in both these families shared only the interval between these markers indicating that this region must contain the cystinosis gene. Radiation hybrid mapping determined that the length of this interval was 10.2 cR_{8000}, or approximately 500 kb, a region small enough to allow positional cloning of the cystinosis gene.

A collaboration was established between the Department of Cytogenetics and Molecular Genetics at the Women's and Children's Hospital in Adelaide and those groups focussed on the cloning of the gene for cystinosis. A physical map based on cosmids has been partially

constructed across the critical region by a collaborating group in London. The aim of the project in this chapter therefore is to assist in the identification of the gene through the isolation of novel transcribed sequences present within cosmids mapping to this region. This will primarily involve the use of exon trapping from clones representing a minimum tiling path. Further characterisation of trapped exons will be performed by consortium members in London, followed by mutation analysis in affected individuals.

7.2 Methods

The cloning of the cystinosis gene involved collaborative efforts from a number of laboratories and only the methods involved in the contribution from the candidate towards this project will be presented. This primarily centred on the trapping of exons from cosmid clones mapping in the cystinosis critical region. All other methods used can be obtained from the publication arising from this work (Town *et al.*, 1998-Appendix A5).

7.2.1 Exon Trapping

The same exon trapping procedure as described in chapter 4 was used. Slight modifications to the method are listed below. Cosmid bacterial stabs were streaked for single colonies on L-Ampicillin plates and DNA was isolated from 200 ml overnight cultures as described in 2.2.1.1. Two rounds of trapping were performed. The first involved 5 cosmid clones indicated in Figure 7.1C, which were subcloned separately into the pSPL3B-CAM vector using *Bgl*II/*Bam*HI double digests and *Pst*I digests. A single flask of COS-7 cells was subsequently transfected with 2 µg of each subcloned cosmid (1 µg from the *Bgl*II/*Bam*HI subclones and 1 µg from the *Pst*I subclones). In the second round of trapping, 3 overlapping cosmid clones were used as indicated in Figure 7.3A. Each was separately subcloned into the trapping vector

with *Pst*I digests and *Bgl*II/*Bam*HI double digests as with the initial exon trapping. Each subcloned cosmid was then transfected into a separate flask of COS-7 cells. Subsequent procedures were followed as described in 4.2.7 to 4.2.11.

7.2.2 Exon Analysis

7.2.2.1 Colony PCR of Subcloned Trapped Exons

From the first round of exon trapping, a total of 176 exon trapped products were analysed by colony PCR (2.2.14.2). Products were grouped according to size and representative clones from each size group were analysed further (see 7.2.2.4). In the second round of trapping colony PCR of 24 randomly selected trapped products for each cosmid was performed. An additional 21 colonies were streaked onto the same master plate, such that in total, 45 colonies were kept from each cosmid trap which were subsequently screened by hybridisation (see 7.2.2.2 and 7.2.2.3).

7.2.2.2 Trapped Exon Insert Amplification

From the initial 24 clones analysed by colony PCR for each cosmid from the second round of trapping, clones representing each size group were chosen for insert amplification by colony PCR (2.2.14.2) and subsequent hybridisations to check for clone redundancy. Each trapped product contained exon sequence as well as a small amount of pSPL3B-CAM sequence adjacent to its splice donor and splice acceptor site. Therefore PCR amplification of exon sequences alone was achieved using the SA5 and SD5 primers (see Table 2.1) which lie immediately adjacent to the splice donor and splice acceptor site of the trapping vector. The entire PCR reaction was run on a 2% (w/v) agarose gel and the amplified insert excised with a sterile scalpel blade.

7.2.2.3 Colony Master Plate Screening of Second Round Exon Trapped Products

Each master plate containing 45 colonies from the second round of trapping (7.2.2.1), was replica plated onto fresh L-Ampicillin plates using a sterile pipette tip for each clone transferred. After growth overnight at 37°C, the plates were colony blotted as described in 2.2.17.5. PCR amplified inserts of clones from each size group (7.2.2.2) were then used individually as hybridisation probes to detect identical clones on the master plate (redundancy screening). Such positive clones were eliminated from subsequent analysis. Those clones that were negative but were in the same size group as the clone used for hybridisation were likely to represent unique trapped exons and so were analysed further (see 7.2.2.4). Screening of master plates continued until every clone was accounted for. Labelling of DNA fragments in agarose, hybridisation and washing of membranes, and stripping of membranes for re-use is described in chapter 2.

7.2.2.4 Sequence Analysis and Physical Mapping

Clones selected for further analysis were grown in 20 ml of L-Broth plus ampicillin (100 µg /ml) and DNA was isolated using Qiagen Tip-20 columns (2.2.1.3). The trapped products were sequenced using DyePrimer F and R sequencing kits (2.2.18.2). Trapped exon sequence homology searches of both the non-redundant and dbEST databases at NCBI were performed using the BLASTN algorithm (Althschul *et al.*, 1997). Unique trapped products were then sent to the collaborating group in London for physical mapping and identification of transcripts associated with the exons. Specific details of the procedures can be found in Town *et al.*, (1998).

7.3 Results

7.3.1 First Round Exon Trapping

Refined linkage analysis established that the cystinosis gene was located between the genetic markers D17S1828 and D17S2167, an interval of about 500 kb (7.1.4). Exon trapping therefore initially concentrated on cosmids mapping between these markers (Figure 7.1). Cosmids MT37 and MT41 both contained D17S1828 and overlapped by about 10 kb while MT148 and MT145 contained D17S2167 and overlapped by about 5 kb. An additional cosmid (MT71) was positive for D17S829, a microsatellite that had been genetically mapped between D17S1828 and D17S2167. However, the three groups of cosmids did not overlap with each other, therefore physical mapping across the region was continued by the London consortium members in order to link these clones while exon trapping proceeded. Table 7.1 shows the results of the first round of exon trapping from the initial five cosmids. Of the 176 trapped products analysed by colony PCR, 77 (43%) contained an insert size equal to that of a "vector-only" control product, and were omitted from further analysis. This left 99 clones, of which 30 appeared to have different sizes when compared to each other. Sequencing of these 30 clones established that 11 were the result of aberrant splicing involving the trapping vector, while one clone contained an *Alu* repeat. Of the remaining 18 clones, 10 showed homology to previously characterised genes mapping to chromosome 17. Five of these were trapped from the human *P2X1* receptor gene, 1 was trapped from the human aspartoacylase (*ASPA*) gene, 2 corresponded to exons 9, 10, and 17 of the human sarco/endoplasmic Ca^{++}-ATPase 3 (*SERCA3*) gene with the remaining 2 showing aberrant splicing of this gene. Two clones (ET31.124 and ET31.164) displayed significant homology to an EST in dbEST, and the remaining 6 clones did not show any homology to database entries (Table 7.2).

TABLE 7.1

Results of First Round Exon Trapping Experiments

Number of cosmids examined	5
Clones analysed by colony PCR	176
Vector only, no insert	77 (43%)
Number of trapped clones with inserts	99
Number of redundant clones	69
Unique trapped clones	30
Repetitive clones	1
Aberrant splicing of vector	11
Total trapped exons	18
No homology to existing sequences[a]	6
Significant homology to ESTs	2
Homology to known chromosome 16 genes[b]	10

Note. Cosmids examined are shown in Figure 7.1.
[a] 1 of these clones was generated from mis-priming of the dUSA4 oligo.
[b] 2 of these clones were derived from cryptic splicing events.

TABLE 7.2

Sequence Homology Searches with First Round Trapped Exons (January 1997)

Clone	Size	ORF	BLASTN
ET31.156	99bp	no	None
ET31.64	129bp	yes	None
ET31.60	97bp	yes	None
ET31.61	29bp	yes	None
ET31.96	35bp	yes	None
ET31.140	56bp	yes	None
ET31.124	88bp	yes	dbEST: R19743 (yg40c06.ri) 72% nt 27-114 3.4e-11 nr: L42810 (Rat Ca++/Calm. dep. prot. kin.) 86% nt 1231-1318 6.9e-20
ET31.164	104bp	yes	dbEST: R19743 (yg40c06.ri) 82% nt 243-293 5.7e-6 nr: L42810 (Rat Ca++/Calm. dep. prot. kin.) 88% nt 1448-1551 1.5e-26
ET31.17[a]	91bp	yes	nr: U45448 (Human P2X1 receptor) 100% nt 1072-1162 1.3e-25
ET31.106[a]	157bp	yes	nr: U45448 (Human P2X1 receptor) 100% nt 1072-1228 8.9e-32
ET31.3	70bp	yes	nr: U45448 (Human P2X1 receptor) 100% nt 554-623 2.0e-21
ET31.39	97bp	yes	nr: U45448 (Human P2X1 receptor) 100% nt 624-720 6.9e-32
ET31.74	142bp	yes	nr: U45448 (Human P2X1 receptor) 100% nt 802-943 1.3e-50
ET31.32	84bp	yes	nr: Z69881 (Human Serca3 gene) 100% nt 2532-2575 2.4e-16
ET31.27	192bp	yes	nr: Z69881 (Human Serca3 gene) 100% nt 1102-1293 6.2e-64
ET31.127[b]	171bp	yes	nr: Z69881 (Human Serca3 gene) 100% nt 2554-2616 2.2e-22
ET31.22[b]	89bp	yes	nr: Z69881 (Human Serca3 gene) 96% nt 1348-1376 1.4e-3
ET31.136	108bp	yes	dbEST: AA112139 (zn60c06) 100% nt 316-370 1.0e-12 nr: S67156 (Human aspartoacylase gene) 100% nt 685-792 1.1e-35

Note. In the BLAST homology column the following information is included: database, GenBank accession number (clone description), percent identity, region of nucleotide (nt) homology, and P value. ORF, open reading frames. ET: trapped exon followed by clone number.

[a] Overlapping clones.

[b] Aberrantly spliced clones.

As can be seen from Table 7.2, ET31.124 and ET31.164 also showed significant homology to a rat Ca^{++}/calmodulin dependant protein kinase gene (GenBank accession number L42810). These exons therefore most likely represent the human homologue to this gene. Of the 6 trapped products that did not show BLAST nucleotide homology, only 1 failed to have an open reading frame suggesting that the majority of these may belong to transcribed sequences. Each exon was successfully mapped back to its cosmid of origin (Figure 7.1). Interestingly, no products were trapped from the cosmid MT71. Further analysis of this cosmid with PCR indicated that it did not contain the D17S829 marker as originally stated, which was fortuitous given the trapping results. Detailed characterisation of each of the trapped exons was abandoned following additional linkage analysis of cystinosis families (see 7.3.2).

7.3.2 Further Refinement of the Genetic Interval for the Cystinosis Gene

During the exon trapping experiments, consortium members typed DNA from a collection of cystinosis families with novel microsatellites mapping between the region containing D17S2167 and D17S1828. In 23 out of 70 unrelated patients with nephropathic cystinosis, no amplification product for D17S829 was obtained. In addition, 14 patients initially thought to be homozygous for this marker were found to be missing one parental allele, suggesting they carried heterozygous deletions for this locus. Further studies showed that in every family, the deletion segregated with the disease. DNA from 100 unrelated and unaffected individuals was also typed for this marker. No homozygous deletions were detected indicating that the deletion found in the cytinosis patients was not a polymorphism. These results suggested that a deletion of the D17S829 marker in these patients may also be disrupting the gene responsible for cystinosis. As none of the previous cosmids analysed by exon trapping contained D17S289, a total of 6 cosmids containing this marker were subsequently isolated from a chromosome-17-specific cosmid library (ICRFc105) with an additional 8 identified from subsequent cosmid

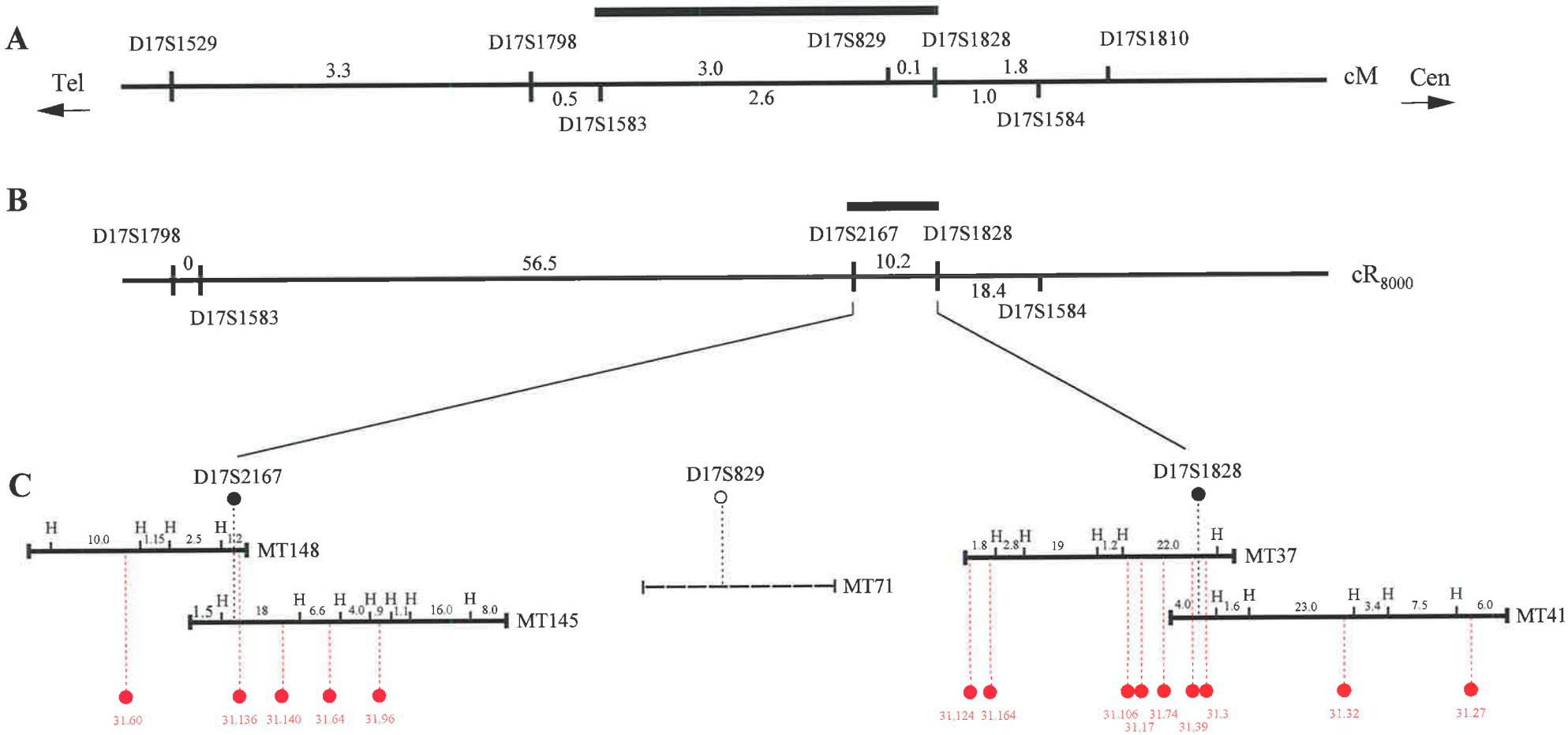


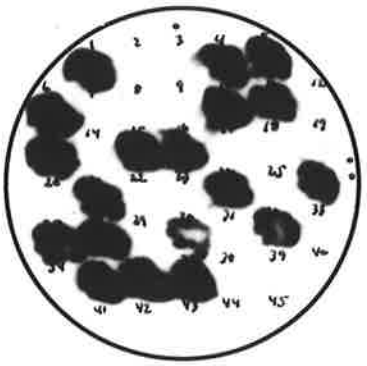
Figure 7.1: Integrated genetic and physical map of the cystinosis critical region at the start of the project. (**A**) Genetic map of 17p13.3 containing markers used in linkage mapping studies (see 7.1.4). Genethon markers are shown above the line, while distances between markers is in centimorgans (cM). The thick horizontal line indicates the region most likely to contain the cystinosis gene based on the results of Jean *et al* (1996) and The Cystinosis Collaborative Research Group (1995). (**B**) Results of radiation hybrid mapping studies in cystinosis families (McDowell *et al*., 1997). The distance between markers is measured in centirays (cR) using 8000 rad radiation hybrids. The thick horizontal line indicates the smallest region most likely to contain the cystinosis gene. The size of this critical region was estimated to be ~500 kb. (**C**) Physical map showing the cosmid clones that were used for the first round of exon trapping. Cosmids are not drawn to scale. H: *Hind*III. All restriction fragment sizes are indicated in kilobases. The location of the markers defining the critical cystinosis region are indicated by black circles. Red circles indicate the positions of the trapped exons. Cosmid MT71 was later found to be negative for the D17S829 marker and so does not map to the critical cystinosis region.

walking experiments. Data for all of the above results is presented in Town *et al.*, 1998. To identify transcribed sequences contained within these clones, further exon trapping was initiated.

7.3.3 Second Round Exon Trapping

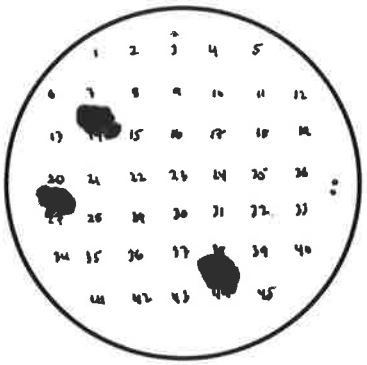
Three overlapping cosmids (AO956, DO2108, and EO4153) were chosen for additional exon trapping experiments (Figure 7.3A). Restriction enzyme analysis indicated that they collectively covered approximately 85 kb. As these cosmids most likely contained the cystinosis gene, in order to optimise the number of exons obtained, each individual cosmid was transfected separately into the COS-7 cells. As before, trapped exons were size selected based on colony PCR results. In addition, to eliminate the possibility of different exons being placed in the same size group, inserts representing each clone from within a defined size group, were used to screen the remaining clones (Figure 7.2). Colony PCR with exons trapped from cAO956 indicated that ETAO-1, -7, and –8 were the same size and therefore possibly the same exon. However, as seen in Figure 7.2A, the insert to the trapped exon ETAO-1 hybridised to itself as well as many other clones trapped from this cosmid, whereas ETAO-7 and –8 were negative for this probe. In addition, the insert to ETAO-7 did not hybridise to ETAO-8, indicating that they also were different, both to each other, and to ETAO-1 (Figure 7.2B and 7.2C). Tables 7.3 and 7.4 provide a summary of the final exon trapping results following extensive hybridisation screening.

A total of 11 trapped products were obtained with all but two showing no homology to database sequences in GenBank. In each case, the trapped product was successfully mapped back to the cosmid of origin by Southern analysis (Figure 7.3C). ETEO-8 and –34 overlapped with each other, suggesting ETEO-34 contains 2 exons, one of which is ETEO-8. Sequencing



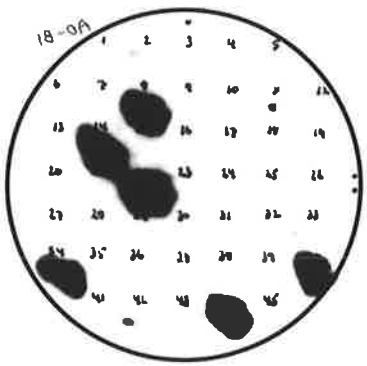
Probe: ETAO-1

Positive clones:
ETAO-1, -4, -5, -6, -10
ETAO-11, -13, -15, -16
ETAO-21, -24, -26, -27
ETAO-28, -30, -32, -35
ETAO-36, -37



Probe: ETAO-7

Positive clones:
ETAO-7, -20, -38



Probe: ETAO-8

Positive clones:
ETAO-8, -14, -22
ETAO-34, -40, -44

Figure 7.2: Screening exon trapped clones isolated from the cosmid AO956 by colony hybridisation with inserts from individual exons. (**A**) The ETAO-1 hybridises to itself as well as a number of other identical clones. ETAO-7 and -8 are negative for this probe even though they had the same sized insert by colony PCR. (**B**) ETAO-7 hybridises to itself and two other clones. It does not hybridise to ETAO-8 and is therefore different. (**C**) ETAO-8 identifies itself and 5 other clones. These results indicate that hybridisation screening of clones is important to identify different trapped exons possessing a similar size as determined by colony PCR .

TABLE 7.3

Results of Second Round Exon Trapping Experiments

Number of cosmids examined	3
Clones analysed by colony PCR	72
Vector only, no insert	27 (37%)
Number of trapped clones with inserts	65
Number of redundant clones	52
Unique trapped clones	13
Aberrant splicing of vector	2
Total trapped exons	11
No homology to existing sequences	9
Significant homology to ESTs	2
Homology to known chromosome 16 genes	0

Note. Cosmids examined are shown in Figure 7.3.

TABLE 7.4

Sequence Homology Searches with Second Round Trapped Exons (July 1997)

Clone	Size	ORF	BLASTN
ETAO-1	80bp	yes	None
ETAO-7	85bp	yes	None
ETAO-8	79bp	yes	None
ETAO-17	100bp	yes	None
ETEO-7	239bp	yes	None
ETEO-8[a]	83bp	yes	None
ETEO-34[a]	126bp	yes	None
ETEO-10	330bp	yes	None
ETEO-18	441bp	no	None
ETEO-11[b]	180bp	no	dbEST: N85789 (Human cDNA clone) 87% nt 63-202 4.4e-40 nr: AC001555 (BAC H75 subclone) 95% nt 855-1022 4.7e-55
ETDO-42	120bp	yes	dbEST: N36593 (yx86g08.r1) 99% nt 66-185 9.1e-44 nr: U90913 (Human cDNA clone 23665) 99% nt 74-193 1.9e-41

Note. In the BLAST homology column the following information is included: database, GenBank accession number (clone description), percent identity, region of nucleotide (nt) homology, and P value. ORF, open reading frames. ET: trapped exon followed by clone number.

[a] Overlapping clones.

[b] Homology is on the opposite strand.

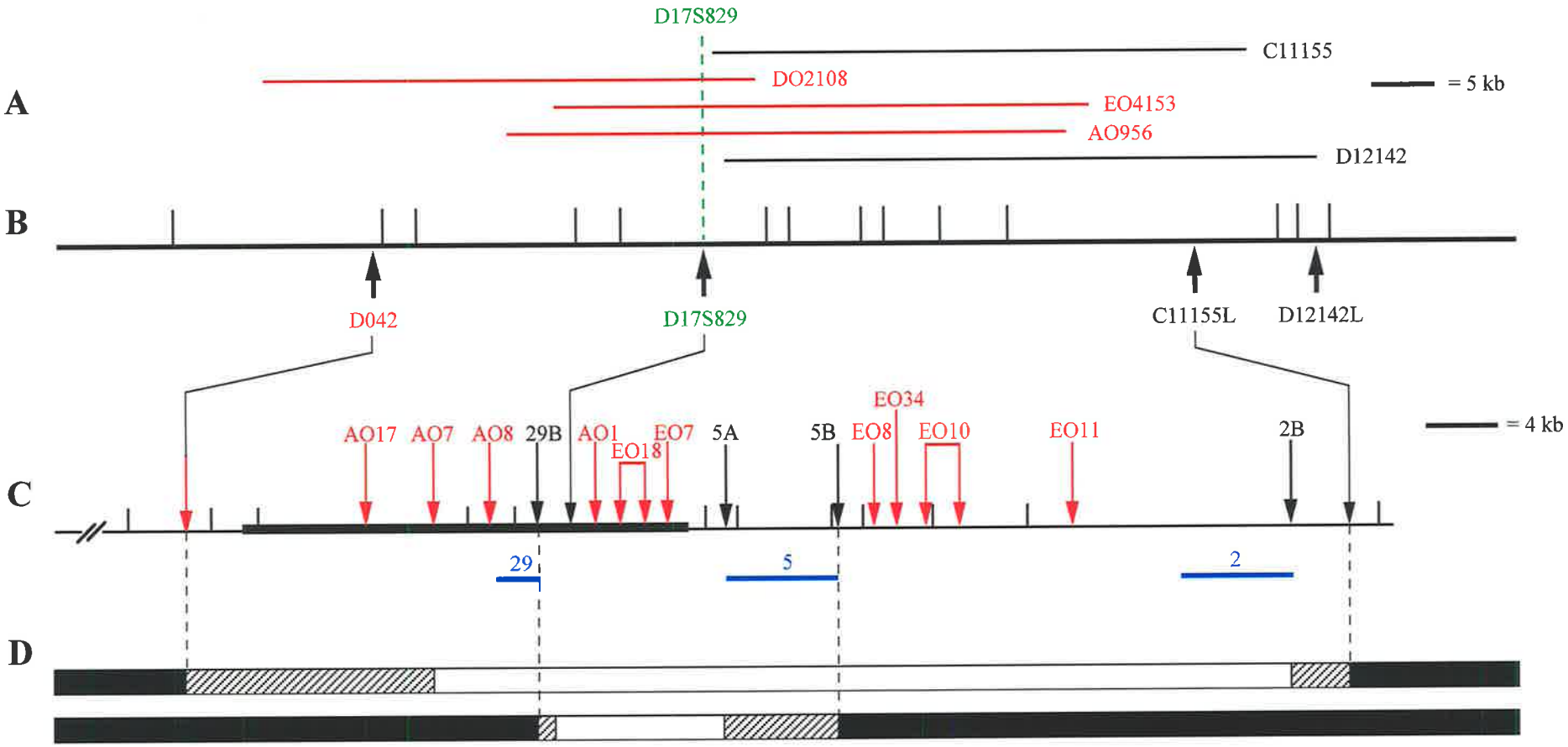


Figure 7.3 : Schematic representation of the cystinosis critical region as determined in Town *et al.* 1998. (**A**) Physical map of overlapping cosmids spanning the region containing the D17S829 locus. Cosmids marked in red were those used for the second round of exon trapping. (**B**) The solid bar represents genomic DNA with *Hind*III restriction enzyme sites marked by vertical lines. STSs used in mapping the contig and defining the extent of overlap between clones are indicated by vertical arrows below the restriction map. (**C**) Expanded view of the restriction map surrounding the D17S829 marker. Small vertical lines represent *Hind*III sites. Random subclones derived from this region and used as a source of genomic sequence are indicated by blue horizontal bars and numbers. The position of trapped exons (red arrows) and clone end sequences (black arrows) used to map the deletions are indicated. The position of the cystinosis gene is shown by a thick horizontal bar. (**D**) Extent of deletions found in cystinosis families. Horizontal (unfilled) boxes denote the minimum extent of deletions found in cystinosis families as defined by STS analysis. The shaded boxes indicate the regions which must contain the deletion breakpoints. The upper bar shows the deletion found in 22 cystinosis families, whereas the lower bar shows the smaller deletion found in family P11 only.

of ETAO-1, -7, and –8, indicated their sizes were almost identical (80 bp, 85 bp, and 79 bp respectively), as indicated by colony PCR. However their sequences were different which also agreed with the colony hybridisation experiments. For each exon, splice site boundary sequences were obtained from the direct sequencing of subcloned DNA from the region using primers derived from the "exonic" sequence. In each case, exon/intron junctions matched the consensus splice-site ag/gt sequence.

7.3.4 Cloning of the Cystinosis (*CTNS*) Gene

ESTs were generated from each of the trapped exons identified from the second round of experiments by London consortium members. In addition, STSs were obtained from the sequencing of cosmid ends from the contig, and random sequencing of *Bam*HI subclones of an overlapping PAC (K03130). DNA from cystinosis patients was subsequently screened with PCR primers designed to STSs from each of the trapped exons, for each end of the PAC subclones, and for the cosmid ends (C11155L and D01214L) to test for the presence or absence of these new markers. Of those patients found to carry homozygous deletions of D17S828 (7.3.2), 22 were missing all STSs between the trapped exon ETDO-42 and C11155L (Figure 7.3C and D). This region contained all eleven trapped exons identified from the second round of experiments and was estimated to be at least 65 kb in length. However, in one family (P11) the deletion only involved the markers between 29B and 5B, defining a minimum deletion interval of 9.5 to 16 kb (Figure 7.3C and D). This region contained only four trapped exons, ETAO-1, ETAO-7, ETAO-8, and ETEO-18.

As ETEO-7 had the largest insert of those exons mapping to the smallest deletion interval, it was used to screen a human adult kidney cDNA library (Clontech). Sequence analysis of a 1.8 kb insert of a positive clone revealed 100% homology to the trapped exons ETAO-1, -7, -8,

and part of ETEO-7. The screening of additional libraries coupled with 5' and 3' RACE allowed the identification of a 2.6 kb cDNA clone containing an open reading frame of 1,101 bp. The size of this cDNA was subsequently shown to be in agreement with the size of the transcript detected by Northern analysis.

The gene, designated *CTNS*, consists of twelve exons, with the ORF beginning in exon 3. ETAO-1, -7, and –8 corresponded to exons 3, 4, and 5 respectively, while ETEO-7 appeared to arise from aberrant splicing events as it was trapped from the opposite strand of the gene and contained most of exon 2 and a part of intron 1. ETEO-18 was also most likely derived from the use of cryptic splice sites as it too was trapped from the opposite strand of *CTNS* within intron 2. In addition, although ETAO-17 was mapped within the gene by Southern hybridisation, it did not contain any *CTNS* sequence, and may also be an aberrantly generated product located within an intron. In the 22 patients with the large homozygous deletion, PCR amplification of exons 2 to 12 revealed that the deletion spans all of the 5' end of the gene, with a 3' breakpoint in intron 10. In contrast, the deletion in family P11 affected only exons 1 to 3.

The *CTNS* gene spans approximately 23 kb of genomic DNA and codes for a protein of 367 amino acids. Homology searches with database sequences showed 48% identity and 66% similarity over 244 amino acids with a predicted 55.5 kD protein (C41C4.7) of *C. elegans*, as well as 43% identity and 64% similarity over 102 amino acids with the yeast transmembrane protein ERS1. *CTNS* appears to encode an integral membrane protein with six or seven membrane-spanning domains, which appear to be conserved between the three species.

7.3.4.1 Screening CTNS for Small Mutations

Single stranded conformation polymorphism (SSCP) analysis was performed for seven of the coding exons on 47 unrelated patients that did not have detectable homozygous deletions. In total, eleven different mutations were detected which included 6 small (less than 10 bp) deletions, and 5 missense mutations. When DNA from other family members was available, SSCP analysis confirmed that each mutation segregated with the disease, while SSCP analysis of 36 to 72 control chromosomes for each exon failed to identify the abnormal SSCP pattern seen in these affected families.

7.4 Discussion

Exon trapping of cosmid DNA from within the 500 kb cystinosis critical region was successful in the identification of 29 individual unique products. A total of five exons were trapped from the *P2X1* receptor gene, which had been previously mapped to chromosome 17 (Longhurst *et al.*, 1996). The identification of exons corresponding to the previously characterised *SERCA3* gene, also mapped to chromosome 17p at the D17S1828 marker (Dode *et al.*, 1998), provided a refined localisation. In addition, a product was trapped that corresponded to exon 4 of the Canavan disease causing gene, aspartoacylase (*ASPA*), which also had previously been mapped to this chromosomal region (Kaul *et al.*, 1994). In addition, 2 exons showed high homology to a rat protein kinase gene, suggesting the human homologue had been identifed.

A number of trapped products were then isolated from a group of three cosmid clones that spanned the D17S829 microsatellite repeat. This marker had been shown during the course of the exon trapping to be deleted in a number of affected individuals. Due to this fact, the method used to screen exons trapped from these cosmids was modified so as to increase the

number of potential exons identified. This involved screening all exons using a hybridisation based approach rather than solely relying on size determination from agarose gel electrophoresis. The result of this method confirmed that a colony hybridisation screening step is important to achieve maximum yields of trapped exons.

Although two of the exons identified from the second round of exon trapping displayed nucleotide homology to ESTs, both were excluded as candidates for the cystinosis gene based on physical mapping data. Three of the remaining nine trapped products mapped to the smallest region of deletion established for the interval, and were subsequently found to belong to a novel gene, *CTNS*. The D17S829 locus was shown to map to the third intron of *CTNS*, which localised this gene within the cystinosis region previously established by linkage analysis. A total of six trapped products showed homology to *CTNS*, three corresponded to exons 3, 4, and 5, while three others were aberrantly spliced clones from within the *CTNS* genomic interval.

Deletions involving the loss of the 5' end of the gene were characterised and eleven different small mutations were identified which segregated with the disease. All mutations identified to date have been shown to severely disrupt the open reading frame of the gene, suggesting possibly complete loss of function of the protein. Affected individuals were found to be either homozygous for a specific mutation or compound heterozygotes, consistent with an autosomal recessive mode of inheritance. These findings, coupled with the absence of these mutations in normal individuals, demonstrates that *CTNS* is the cystinosis gene.

CTNS was shown to be expressed at varying levels in all fetal and adult tissues tested, a finding which is consistent with the high levels of intracellular cystine identified in different cell types

and the multisystem involvement of the disorder (Gahl *et al.*, 1995). The encoded protein has been predicted to contain six or seven transmembrane domains, typical of an integral membrane protein. The absence of specific targetting sequences in the gene product suggests that the protein is not localised to the Golgi apparatus, ER lumen, peroxisome, mitochondria, or the nucleus. However, the presence of a GYXX-hydrophobic amino acid motif found in several lysosomal membrane proteins, suggests it may be a lysosomal membrane protein (Hunziker and Geuze, 1996). In addition, predicted heavy glycosylation of the N-terminal end of cystinosin, as seen in regions of lysosome-associated membrane glycoproteins, further supports the localisation of the protein to the membranes of lysosomes (Hunzicker and Geuze, 1996). This is consistent with the predicted role of the protein in lysosomal cystine removal, however further work is required to determine the exact role of cystinosin in the transport process.

Chapter 8

General Discussion

and

Future Directions

The work presented in this thesis enabled the positional cloning of the Fanconi anaemia group A (*FAA*) gene and the gene responsible for nephropathic cystinosis (*CTNS*). An additional aim of the study, which has not been achieved to date, is the identification of a breast cancer tumour suppressor gene, which has been localised to chromosome 16q24.3 between the microsatellite markers D16S3026 and D16S303. This ~750 kb interval has been shown to consist of 7 previously characterised genes (one of which was *FAA*), and 20 individual EST clones identified by a combination of exon trapping, EST mapping, and dbEST screening of partial genomic sequence.

Three of these ESTs (*GAS11*, *C16orf3*, and T6) were analysed in detail to determine if they were involved in breast carcinogenesis. All three were shown to be transcribed and expressed in breast tissue. However, mutation analysis failed to identify nucleotide changes specific for breast tumour DNA, which ruled them out as candidate tumour suppressor genes. Further characterisation of the remaining transcripts not reported in this thesis has shown that at least 14 individual transcription units now exist within the 750 kb candidate region. This does not include the 7 previously characterised genes, *GAS11*, *C16orf3*, T6, or the *SPG7* gene. Of these 14 transcripts, 3 have been examined for mutations in breast tumour DNA by SSCP analysis, and subsequently eliminated as candidate genes. Four transcription units have been proposed based on the presence of a single trapped exon and have not been analysed further. The remaining 7 transcription units are being examined presently. The development of the transcription map at 16q24.3 provides an important source of candidate genes for other diseases that may be mapped to this region in the future. An example is the gene for dehydrated hereditary stomatocytosis, which has recently been mapped by genetic linkage to 16q23-qter (Carella *et al.*, 1998).

The gene density within this region is quite high with an average of 1 gene every 30 kb. The gene identification methods used in this thesis involved exon trapping from cosmid DNA, analysis of ESTs mapped to 16q24.3 as part of the Human Gene Map construction at NCBI, and dbEST screening with partial cosmid sequence. These methods were also used by Cooper *et al.* (1998) in the analysis of the 1.65 Mb Best macular dystrophy disease region. Combined with previous mapping studies (Cooper *et al.*, 1997) they were able to identify at least 31 transcripts in the candidate region, or 1 gene per ~50 kb. In another study of a 900 kb region at Xp22.1-p22.2 (Warneke-Wittstock *et al.*, 1998), the same procedures identified at least 12 novel transcripts in addition to the two genes already localised to this region. Therefore the use of these three methods has proven successful in the identification of novel genes. It is highly probable that not all genes were identified in each of these studies, including the transcription map presented in this thesis. In a comparison of methods used in the positional cloning of the *BRCA1* gene (Harshman *et al.*, 1995), none of the four independent studies analysed identified all genes within the candidate region. This illustrates the enormous effort required to complete a positional cloning project successfully. Unfortunately, in most laboratories, a subset of transcript identification methods will have to be chosen. The success of the exon trapping procedure in the cloning of *FAA* and *CTNS* in this study as well as the success of this technique in other positional cloning projects suggests this technique should be one of those employed.

At the onset of this project, the rate-limiting step in positional cloning was the identification and characterisation of genes in the region of interest. During the course of this thesis, significant progress in the physical mapping of human ESTs developed to the point where a positional candidate approach to disease gene isolation can now be utilised. As a first step in this approach, ESTs known to map to a candidate region can be analysed for their possible role in the associated disease. If this is unsuccessful then more general techniques of transcript

identification, such as exon trapping, may then be employed to identify additional candidate genes. Although the release of the GeneMap '98 (Deloukas *et al.*, 1998) reports 30,000 mapped transcripts (~ one-third of all genes) the coverage of these transcripts in 16q24.3 is less than what is expected. Only three unique UniGene clusters are on the new map of the 14 that were identified by exon trapping. Therefore until such time as the majority of genes have been end-sequenced and mapped as ESTs, gene hunting methods such as exon trapping and direct selection still need to be employed for positional cloning projects.

The final phase of the Human Genome Project is in progress, the identification of the complete nucleotide sequence of man. Many studies have benefited greatly from the availability of only partial genomic sequence of the region under investigation (Osborne *et al.*, 1996; Bernot *et al.*, 1998; Hisama *et al.*, 1998; Warneke-Wittstock *et al.*, 1998). This thesis also used partial cosmid sequence in the 16q24.3 region to screen dbEST for homologous cDNA clones with the successful identification of three novel ESTs and 1 gene (beta III tubulin). An independent aim of our Department is to completely sequence the 16q24.3 region. The dense physical map produced during the course of this thesis will serve as an ideal sequencing template with analysis of this sequence further enhancing the transcription map produced to date.

An emerging tool for the identification of genes within large stretches of genomic DNA is the use of computational methods of gene prediction (Fickett, 1996). Although research into these methods has been going on for the past 15 years (Claverie, 1997), the field has evolved from the design of programs to identify protein coding regions of bacterial genomes to the challenge of predicting detailed multi-exon vertebrate genes. A number of programs have been developed and their performances compared (reviewed in Fickett, 1996; Claverie, 1997). Some of these include GENSCAN (Kulp *et al.*, 1996) which perfectly locates more than 80% of internal

coding exons, MZEF (Zhang, 1997), and GRAIL (Milanesi *et al.*, 1993). The latter 2 are simple exon finders and are best suited to low coverage genomic sequence data while GENSCAN is best suited to large contigs of complete genomic sequence similar to those being generated at present. Unfortunately, with most of the programs, a large percentage (60%) of predicted proteins are wrong, and nearly 100% of the predicted gene structures have incorrect 5' and 3' boundaries (Claverie, 1997). Therefore, computational methods of gene identification have to be combined and compared with the results of experimental data. Given that most 3' gene extremities will eventually be identified as ESTs, the key to the next generation of gene identification programs will be the efficient detection of vertebrate promoters, and thus the 5' end of genes. Undoubtedly, the availability of the complete nucleotide sequence of a disease gene region will have an enormous impact on positional cloning and positional candidate cloning. As well as computer generated predictions of possible gene content, a direct sequence homology scan of established EST and gene databases will identify homologous ESTs and provide an immediate gene structure for any transcripts identified. This signifies a new era in disease gene cloning, where transcript identification by exon trapping and direct selection may join zoo blots in being "classical approaches" for the identification of transcribed sequences.

Within the breast cancer tumour suppressor candidate region at 16q24.3, many transcripts have been characterised and found not to be involved in breast carcinogenesis. While there are still many other transcripts to be analysed, and possibly more genes to be found, the underlying basis for this approach is the definition of the localisation of the tumour suppressor gene by loss of heterozygosity (LOH) studies. The LOH data generated from the panel of 27 breast tumours used during the course of this thesis are shown in Figure 1.4. The chosen candidate region between D16S3026 and D16S303 was based on the restricted LOH seen in two tumours, BT559 and BT410. The availability and typing of additional microsatellite markers in

this panel of tumours has recently highlighted problems with the data generated from detailed LOH studies. Tumour BT410 has since been shown to have LOH for D16S2624 and D16S752 (data of Dr. Anne-Marie Cleton-Jansen), two markers that map proximal to D16S289 (see Figure 1.4) such that the candidate region is now defined by a single tumour (BT559). Interpretation of LOH data using microsatellite markers can often be difficult due to contaminating normal DNA and preferential PCR amplification of alleles. This is demonstrated by reports of extensive regions of intermittent loss (Tsuda *et al.*, 1994; Dorion-Bonnet *et al.*, 1995; Skirnisdottir *et al.*, 1995; Chen *et al.*, 1996*b*; Driouch *et al.*, 1997) which may be biologically unrealistic and more likely represent experimental artifacts. However given these complications, a number of tumours (seven) in the Leiden data still support the loss of 16q24.3 in breast tumourigenesis along with independent LOH data from the analysis of an additional 124 primary breast tumours (unpublished data of FAB consortium). Unfortunately a broader region to the one analysed during the course of this thesis may need to be considered in future studies.

Alternative methods to those presented in this thesis may need to be employed to examine the region proximal to D16S3026. A number have been proposed with the majority focussing on a more functional based approach. Microcell-mediated chromosome transfer has provided functional evidence for the presence of a tumour suppressor gene within a genomic region (Koi *et al.*, 1989). Many studies of this type have demonstrated that the introduction of normal chromosomes into malignant cell lines can restore gene function and reverse the malignant phenotype (Weissman *et al.*, 1987; Negrini *et al.*, 1992; Phillips *et al.*, 1996). In addition, the introduction of defined subchromosomal transferable fragments has led to the identification of specific regions of the genome most likely to contain tumour suppressor genes (Koi *et al.*, 1993; Tanaka *et al.*, 1993; Hu *et al.*, 1997; Steck *et al.*, 1997). Following the identification of

a defined region at 16q24.3 which suppresses tumourigenicity, the availability of a detailed physical map of this region (1.1 Mb presented in this thesis combined with ~ 1 Mb under construction proximal to the CY18A(D2)-qter region) could provide a source of cloned DNA with which to conduct additional functional complementation experiments. Whole cosmid, BAC or PAC genomic inserts can be cloned into mammalian expression vectors and tested for their ability to rescue the tumourigenic potential of selected breast cancer cell lines as detected by colony forming assays in soft agar (Epstein *et al.*, 1980) and reduced tumourigenicity in immunocompromised mice. A genomic insert shown to reduce the tumourigenicity of a cell line can then be examined to determine the genes present within the clone, which will be tumour suppressor candidates.

Another approach is the use of cDNA microarrays. This emerging technology allows a rapid survey of the tissue expression of a large number of genes and provides the ability to compare the expression pattern of genes between different tissues (Schena *et al.*, 1995; Iyer *et al.*, 1999). Recently, the temporal pattern of gene expression following serum treatment of growth-arrested human fibroblasts was examined using cDNA microarrays (Iyer *et al.*, 1999). In this study, cDNA made from purified mRNA from a serum stimulated fibroblast, was labelled with the fluorescent dye Cy5. This was mixed with a common reference probe (consisting of cDNA made from purified mRNA from quiescent fibroblast cells) labelled with a second fluorescent dye, Cy3, and hybridised to a DNA microarray containing 8,613 human ESTs and genes. The colour images from the hybridisation results were made by representing the Cy3 fluorescent image as green and the Cy5 image as red and merging the two colours. Yellow spots on the microarray filters therefore represented genes whose expression did not vary between the resting and stimulated fibroblast cell. Genes whose mRNAs were more abundant in the serum-deprived fibroblasts appeared green, whereas those genes whose

mRNAs were more abundant in the serum-treated cells appeared as red spots. As a result, the expression profile of individual genes in the human fibroblast in response to serum addition could be determined. This technique could also be applied to allow a comparison between the expression patterns of genes in a breast tumour and normal cell. Genes that are not expressed in the tumour sample (as a result of mutational inactivation) but are expressed in the corresponding normal tissue would be of interest as potential candidate tumour suppressor genes. Unfortunately this approach is costly, can only be performed in specialised laboratories at the present time, and will detect up-regulated or down-regulated genes from the whole genome. Given that a breast cancer tumour suppressor candidate region at 16q24.3 has already been identified, this approach therefore may not be the best option unless inhouse arrays of cDNA clones mapping to this region can be produced and analysed.

The exploration of any one of these techniques may assist in the identification of the breast cancer tumour suppressor gene mapping to chromosome 16q24.3. This will aid to improve our understanding of the development and progression of breast cancer and ultimately have an impact on the diagnosis and hopefully treatment of this disease.

References

Abbott, D.W., Freeman, M.L., and Holt, J.T. (1998). Double-strand break repair deficiency and radiation sensitivity in BRCA2 mutant cancer cells. *J. Natl. Cancer Inst.* **90**: 978-985.

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., Kerlavage, A.R., McCombie, W.R., and Venter, J.C. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**: 1651-1656.

Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., *et al.* (1995). Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377** Suppl: 3-17.

Adams, S.M., Helps, N.R., Sharp, M.G.F., Brammar, W.J., Walker, R.A., and Varley, J.M. (1992). Isolation and characterisation of a novel gene with differential expression in benign and malignant human breast tumours. *Hum. Mol. Genet.* **1**: 91-96.

Albertsen, H.M., Abderrahim, H., Cann, H.M., Dausset, J., Le Paslier, D., and Cohen, D. (1990). Construction and characterisation of a yeast artificial chromosome library containing seven haploid human genome equivalents. *Proc. Natl. Acad. Sci. USA* **87**: 4256-4260.

Aldaz, C.M., Chen, T., Sahin, A., Cunningham, J., and Bondy, M. (1995). Comparative allelotype of *in situ* and invasive human breast cancer: high frequency of microsatellite instability in lobular breast carcinomas. *Cancer Res.* **55**: 3976-3981.

Allikmets, R.L., Kashuba, V.I., Pettersson, B., Gizatullin, R., Lebedeva, T., Kholodnyuk, I.D., Bannikov, V.M., Petrov, N., Zakharyev, V.M., Winberg, G., Modi, W., Dean, M., Uhlen, M., Kisselev, L.L., Klein, G., and Zabarovsky, E.R. (1994). *Not*I linking clones as a tool for joining physical and genetic maps of the human genome. *Genomics* **19**: 303-309.

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

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389-3402.

Antequera, F., and Bird, A. (1993). Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. USA*. **90**: 11995-11999.

Apostolou, S. (1997). Physical mapping of human chromosome 16. *Thesis submission* (Department of Pediatrics, University of Adelaide, Australia).

Austruy, E., Cohen-Salmon, M., Antignac, C., Beroud, C., Henry, I., Nguyen, V.C., Brugieres, L., Junien, C., and Cornelisse, C. (1993). Isolation of a kidney cDNA down-expressed in Wilms' tumour, by a subtractive hybridisation approach. *Cancer Res.* **53**: 2888-2894.

Baens, M., Aerssens, J., Van Zand, K., Van den berghe, H., and Marynen, P. (1995). Isolation and regional assignment of human chromosome 12p cDNAs. *Genomics* **29**: 44-52.

Balaban, G., Gilbert, F., Nichols, W., Meadows, A.T., and Shields, J. (1982). Abnormalities of chromosome 13 in retinoblastomas from individuals with normal constitutional karyotypes. *Cancer Genet. Cytogenet.* **6**: 213-221.

Ballabio, A. (1993). The rise and fall of positional cloning? *Nature Genet.* **3**: 277-279.

Banfi, S., Borsani, G., Rossi, E., Bernard, L., Guffanti, A., Rubboli, F., Marchitiello, A., Giglio, S., Coluccia, E., Zollo, M., Zuffardi, O., and Ballabio, A. (1996). Identification and mapping of human cDNAs homologous to *Drosophila* mutant genes through EST database searching. *Nature Genet.* **13**: 167-174.

Bartkova, J., Lukas, J., Muller, H., Lutzhft, D., Strauss, M., and Bartek, J. (1994). Cyclin D1 protein expression and function in human breast cancer. *Intl. J. Cancer* **57**: 353-361.

Bassett, D.E. Jr., Boguski, M.S., Spencer, F., Reeves, R., Kim, S., Weaver, T., and Hieter, P. (1997). Genome cross-referencing and XREFdb: implications for the identification and analysis of genes mutated in human disease. *Nature Genet.* **15**: 339-344.

Baumann, P., Benson, F.E., and West, S.C. (1996). Human Rad51 protein promotes ATP-dependent homologous pairing and strand transfer reactions *in vitro*. *Cell* **87**: 757-766.

Bellanne-Chantelot, C., Lacroix, B., Ougen, P., Billault, A., Beaufils, S., Bertrand, S., Georges, I., Glibert, F., Gros, I., Lucotte, G., Susini, L., Codani, J-J., Gesnouin, P., Pook, S., Vaysseix, G., Lu-Kuo, J., Ried, T., Ward, D., Chumakov, I., Le Paslier, D., Barillot, E., and Cohen, D. (1992). Mapping the whole human genome by fingerprinting yeast artificial chromosomes. *Cell* **70**: 1059-1068.

Bergerheim, U., Nordenskjold, M., and Collins, V.P. (1989). Deletion mapping in human renal cell carcinoma. *Cancer Res.* **49**: 1390-1396.

Bernot, A., Heilig, R., Clepet, C., Smaoui, N., Da Silva, C., Petit, J-L., Devaud, C., Chiannilkulchai, N., Fizames, C., Samson, D., Cruaud, C., Caloustian, C., Gyapy, G., Delpech, M., and Weissenbach, J. (1998). A transcription map of the FMF region. *Genomics* **50**: 147-160.

Berry, R., Stevens, T.J., Walter, N.A.R., Wilcox, A.S., Rubano, T., Hopkins, J.A., Weber, J., Goold, R., Bento Soares, M., and 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., and 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., and 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.

Bird, A.P. (1987). CpG islands as gene markers in the vertebrate nucleus. *Trends in Genet.* **3**: 342-347.

Bodmer, W.R., Bailey, C.J., Bodmer, J., Bussey, H.J.R., Ellis, A., Gorman, P., Lucibello, F.C., Murday, V.A., Rider, S.H., Scambler, P., Sheer, D., Solomon, E., and Spurr, N.K. (1987). Localisation of the gene for familial adenomatous polyposis on chromosome 5. *Nature* **328**: 614-616.

Boguski, M.S., Lowe, T.M.J., and Tolstoshev, C.M. (1993). dbEST-database for "expressed sequence tags". *Nature Genet.* **4**: 332-333.

Boguski, M.S., and Schuler, G.D. (1995). ESTablishing a human transcript map. *Nature Genet.* **10**: 369-371.

Bonaldo, M.F., Lennon, G., and Soares, M.B. (1996). Normalisation and subtraction: two approaches to facilitate gene discovery. *Genome Res.* **6**: 791-806.

Botstein, D., White, R.L., Skolnick, M., and Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J Hum. Genet.* **32**: 314-331.

Bouffard, G.G., Idol, J.R., Braden, V.V., Iyer, L.M., Cunningham, A.F., Weintraub, L.A., Touchman, J.W., Mohr-Tidwell, R.M., Peluso, D.C., Fulton, R.S., Ueltzen, M.S., Weissenbach, J., Magness, C.L., and Green, E.D. (1997). A physical map of human chromosome 7: an integrated YAC contig map with average STS spacing of 79 kb. *Genome Res.* **7**: 673-692.

Brancolini, C., Benedetti, M., and Schneider, C. (1995). Microfilament reorganisation during apoptosis: the role of Gas2, a possible substrate for ICE-like proteases. *EMBO J.* **14**: 5179-5190.

Brancolini, C., and Schneider, C. (1994). Phosphorylation of the growth arrest-specific protein Gas2 is coupled to actin rearrangements during G_0-G_1 transition in NIH 3T3 cells. *J. Cell Biol.* **124**: 743-756.

Brenner, A.J., and Aldaz, C.M. (1995). Chromosome 9p allelic loss and p16/CDKN2 in breast cancer and evidence of p16 inactivation in immortal breast epithelial cells. *Cancer Res.* **55**: 2892-2895.

Brenner, A.J., and Aldaz, C.M. (1997). The genetics of sporadic breast cancer. *Prog. Clin. Biol. Res.* **396**: 63-82.

Brenner, D.G., Lin-Chao, S., and Cohen, S. (1989). Analysis of mammalian cell genetic regulation *in situ* by using retrovirus-derived "portable exons" carrying the *Escherichia coli* lacZ gene.. *Proc. Natl. Acad. Sci. USA* **86**: 5517-5521.

Brenner, S. (1990). The human genome: the nature of the enterprise. *CIBA Found. Symp.* **149**: 6-17.

Broca, P.O. *Traite des Tumeurs,* Vol 1. Paris: P Asselin, 1886-1889.

Brody, L.C., Abel, K.J., Castilla, L.H., Couch, F.J., Mckinley, D.R., Yin, G.Y., Ho, P.P., Merajver, S., Chandrasekharappa, S.C., Xu, J.Z., Cole, J.L., Struewing, J.P., Valdes,

J.M., Collins, F.S., and Weber, B.L. (1995). Construction of a transcription map surrounding the *BRCA1* locus of human chromosome 17. *Genomics* **25**: 238-247.

Brown, T. (1993). Analysis of DNA sequences by blotting and hybridisation. Current protocols in molecular biology, F.M. Ausbel, R. Brent, R.E. Kingston, D.D. Moore, J.G. seidman, J.A. Smith, and K. Struhl, eds. (New York: John Wiley & Sons, Inc.), pp. 2.9.1-2.9.20.

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., and 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., and Murray, J.C. (1994). Integrated human genome-wide maps constructed using the CEPH reference panel. *Nature Genet.* **6**: 391-393.

Bullrich, F., MacLachlan, T.K., Sang, N., Druck, T., Veronese, M.L., Allen, S.L., Chiorazzi, N., Koff, A., Heubner, K., Croce, C.M., and 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., and Olson, M.V. (1987). Cloning of large segments of exogeneous DNA into yeast by means of artificial chromosome vectors. *Science* **236**: 806-812.

Burn, T.C., Connors, T.D., Klinger, K.W., and Landes, G.M. (1995). Increased exon-trapping efficiency through modifications to the pSPL3 splicing vector. *Gene* **161**: 183-187.

Burn, T.C., Connors, T.D., Van Raay, T.J., Dackowski, W.R., Millholland, J.M., Klinger, K.W., and Landes, G.M. (1996). Generation of a transcriptional map for a 700-kb region surrounding the polycystic kidney disease type 1 (*PKD1*) and tuberous sclerosis type 2 (*TSC2*) disease genes on human chromosome 16p13.3. *Genome Res.* **6**: 525-537.

Cairo, G., Ferrero, M., Biondi, G., and Colombo, M.P. (1992). Expression of a growth arrest specific gene (gas-1) in transformed cells. *Br. J. Cancer* **66**: 27-31.

Callen, D.F. (1986). A mouse-human hybrid cell panel for mapping human chromosome 16. *Ann. Genet.* **29**: 235-239.

Callen, D.F., Baker, E., Eyre, H.J., Chernos, J.E., Bell, J.A., and Sutherland, G.R. (1990*a*). Reassessment of two apparent deletions of chromosome 16p to an ins(11;16) and a t(1;16) by chromosome painting. *Ann. Genet.* **33**: 219-221.

Callen, D.F., Baker, E., Eyre, H.J., and Lane, S.A. (1990*b*). An expanded mouse-human hybrid cell panel for mapping human chromosome 16. *Ann. Genet.* **33**:190-195.

Callen, D.F., Hyland, V.J., Baker, E., Fratini, A., Gedeon, A.K., Mulley, J.C., Fernandez, K.E.W., Breuning, M.H., and Sutherland, G.R. (1989). Mapping the short arm of human

chromosome 16. *Genomics* **4**: 348-354.

Callen, D.F., Hyland, V.J., Baker, E., Fratini, A., Simmers, R.N., Mulley, J.C., and 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., 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., and 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.

Carella, M., Stewart, G., Ajetunmobi, J.F., Perrotta, S., Grootenboer, S., Tchernia, G., Delaunay, J., Totaro, A., Zelante, L., Gasparini, P., and Iolascon, A. (1998). Genomewide search for dehydrated hereditary stomatocytosis (hereditary xerocytosis): mapping of locus to chromosome 16 (16q23-qter). *Am. J. Hum. Genet.* **63** (In Press).

Carter, B.S., Ewing, C.M., Ward, W.S., Treiger, B.F., Aalders, T.W., Schalken, J.A., Epstein, J.I., and Isaacs, W.B. (1990). Allelic loss of chromosomes 16q and 10q in human prostate cancer. *Proc. Natl. Acad. Sci. USA* **87**: 8751-8755.

Casari, G., De Fusco, M., Ciarmatori, S., Zeviani, M., Mora, M., Fernandez, P., De Michele, G., Filla, A., Cocozza, S., Marconi, R., Durr, A., Fontaine, B., and Ballabio, A. (1998). Spastic paraplegia and OXPHOS impairment caused by mutations in paraplegin, a nuclear-encoded mitochondrial metalloprotease. *Cell* **93**: 973-983.

Caspersson, T., Zech, L., and Johanson, C. (1970). Differential banding of alkylating fluorochromes in human chromosomes. *Exp. Cell Res.* **60**: 315-319.

Castilla, L.H., Couch, F.J., Erdos, M.R., Hoskins, K.F., Calzone, K., and Garber, J.E. (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., Phillips, R.A., Benedict, W.F., Godbout, R., Gallie, B.L., Murphree, A.L., Strong, L.C., and White, R.L. (1983). Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. *Nature* **305**: 779-784.

Chaconas, G., and van de Sande, J.H. (1980). 5'-^{32}P labelling of RNA and DNA restriction fragments. *Methods Enzymol.* **65**: 75-88.

Chen, H., Chrast, R., Rossier, C., Morris, M.A., Lalioti, M.D., and Antonarakis, S.E. (1996*a*). Cloning of 559 potential exons of genes of human chromosome 21 by exon trapping. *Genome Res.* **6**: 747-760.

Chen, T., Sahin, A., and Aldaz, C.M. (1996*b*). Deletion map of chromosome 16q in ductal carcinoma *in situ* of the breast: refining a putative tumour suppressor gene region. *Cancer Res.* **56**: 5605-5609.

Chomczynski, P., and Sacchi, N. (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* **162**: 156-159.

Chumakov, I., Rigault, P., Guillou, S., Ougen, P., Billaut, A., *et al.* (1992). Continuum of overlapping clones spanning the entire human chromosome 21q. *Nature* **359**: 380-387.

Chumakov, I., Rigault, P., Le Gall, I., Bellanne-Chantelot, C., Billault, A., *et al.* (1995). A YAC contig map of the human genome. *Nature* **377** Suppl: 175-297.

Chung, C.T., Niemela, S.L., and Miller, R.H. (1989). One-step preparation of competent *Escherichia coli*: transformation and storage of bacterial cells in the same solution. *Proc. Natl. Acad. Sci. USA* **86**: 2172-2175.

Church, D.M., Stotler, C.J., Rutter, J.L., Murrell, J.R., Trofatter, J.A., and Buckler, A.J. (1994). Isolation of genes from complex sources of mammalian genomic DNA using exon amplification. *Nature Genet.* **6**: 98-105.

Church, D.M., Yang, J., Bocian, M., Shiang, R., and Wasmuth, J.J. (1997). A high-resolution physical and transcript map of the Cri du Chat region of human chromosome 5p. *Genome Res.* **7**: 787-801.

Claverie, J-M. (1997). Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* **6**: 1735-1744.

Cleton-Jansen, A-M., Collins, N., Lakhani, S.R., Weissenbach, J., Devilee, P., Cornelisse, C.J., and Stratton, M.R. (1995). Loss of heterozygosity in sporadic breast tumours at the BRCA2 locus on chromosome 13q12-q13. *Br. J. Cancer* **72**: 1241-1244.

Cleton-Jansen, A-M., Moerland, E.W., Kuipers-Dijkshoorn, N.J., Callen, D.F., Sutherland, G.R., Hansen, B., Devilee, P., and Cornelisse, C.J. (1994). At least two different regions are involved in allelic imbalance on chromosome arm 16q in breast cancer. *Genes Chrom. Cancer* **9**: 101-107.

Cleton-Jansen, A-M., Moerland, E.W., Pronk, J.C., van Berkel, C.G.M., Apostolou, S., Crawford, J., Savoia, A., Auerbach, A.D., Mathew, C.G., Callen, D.F., and Cornelisse, C.J. (1999). Mutation analysis of the Fanconi anaemia A gene in breast tumours with loss of heterozygosity at 16q24.3. *Br. J. Cancer.* In Press.

Cohen, A.J., Li, F.P., Berg, S., Marchetto, D.J., Tsai, S., Jacobs, S.C., and Brown, R.S. (1979). Hereditary renal-cell carcinoma associated with a chromosomal translocation. *New Engl. J. Med.* **301**: 592-595.

Cohen, D., Chumakov, I., and Weissenbach, J. (1993). A first-generation physical map of the human genome. *Nature* **366**: 698-701.

Coles, C., Condie, A., Chetty, U., Steel, C.M., Evans, H.J., and Prosser, J. (1992). p53 mutations in breast cancer. *Cancer Res.* **52**: 5291-5298.

Collavin, L., Buzzai, M., Saccone, S., Bernard, L., Federico, C., DellaValle, G., Brancolini, C., and Schneider, C. (1998). cDNA characterisation and chromosome mapping of the human *GAS2* gene. *Genomics* **48**: 265-269.

Collins, F.S. (1992). Positional cloning: let's not call it reverse anymore. *Nature Genet.* **1**: 3-6.

Collins, F.S. (1995*a*). Positional cloning moves from perditional to traditional. *Nature Genet.* **9**: 347-350.

Collins, F., and Galas, D. (1993). A new five-year plan for the U.S. human genome project. *Science* **262**: 43-46.

Collins, F.S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., Walters, L., and the members of the DOE and NIH planning groups. (1998). New goals for the U.S. Human Genome Project: 1998-2003. *Science* **282**: 682-689.

Collins, F.S., and Weissman, S.M. (1984). Directional cloning of DNA fragments at a large distance from an initial probe: a circularization method. *Proc. Natl. Acad. Sci. USA.* **81**: 6812-6816.

Collins, J.E., Cole, C.G., Smink, L.J., Garrett, C.L., Leversha, M.A., *et al.* (1995*b*). A high-density YAC contig map of human chromosome 22. *Nature* **377** Suppl: 367-379.

Collins, N., McManus, R., Wooster, R., Mangion, J., Seal, S., Lakhani, S.R., Ormiston, W., Daly, P.A., Ford, D., Easton, D.F., and Stratton, M.R. (1995*c*). Consistent loss of the wild type allele in breast cancers from a family linked to the BRCA2 gene on chromosome 13q12-13. *Oncogene* **10**: 1673-1675.

Cooper, D.N., Smith, B.A., Cooke, H.J., Niemann, S., and Schmidtke, J. (1985). An estimate of unique DNA sequence heterozygosity in the human genome. *Hum. Genet.* **69**: 201-205.

Cooper, P.R., Nowak, N.J., Higgins, M.J., Church, D.M., and Shows, T.B. (1998). Transcript mapping of the human chromosome 11q12-q13.1 gene-rich region identifies several newly described conserved genes. *Genomics* **49**: 419-429.

Cooper, P.R., Nowak, N.J., Higgins, M.J., Simpson, S.A., Stoehr, H., Weber, B.H., Gerhard, D.S., de Jong, P., and Shows, T.B. (1997). A sequence ready high resolution physical map of the Best macular dystrophy gene region in 11q12-q13. *Genomics* **41**: 185-192.

Corbo, L., Maley, J.A., Nelson, D.L., and Caskey, C.T. (1990). Direct cloning of human transcripts with hnRNA from hybrid cell lines. *Science* **249**: 652-655.

Couch, F.J., Farid, L.M., Deshano, M.L., Tavtigian, S.V., Calzone, K., *et al.* (1996*a*). BRCA2 germline mutations in male breast cancer cases and breast cancer families. *Nature Genet.* **13**: 123-125.

Couch, F.J., Rommens, J.M., Neuhausen, S.L., Belanger, C., Dumont, M., *et al.* (1996*b*). Generation of an integrated transcription map of the *BRCA2* region on chromosome 13q12-q13. *Genomics* **36**: 86-99.

Coulson, A., Sulston, J., Brenner, S., and Karn, J. (1986). Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. Natl. Acad. Sci. USA.* **83**: 7821-7825.

Cox, D.R., Burmeister, M., Price, E.R., Kim, S., and Myers, R.M. (1990). Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of

mammalian chromosomes. *Science* **250**: 245-250.

Davies, K.E., Pearson, P.L., Harper, P.S., Murray, J.M., O'Brien, T., Sarfarazi, M., and Williamson, R. (1983). Linkage analysis of two cloned sequences flanking the Duchenne muscular dystrophy locus on the short arm of the human X chromosome. *Nucleic Acids Res.* **11**: 2303-2312.

Dear, P.H., Bankier, A.T., and Piper, M.B. (1998). A high-resolution metric HAPPY map of human chromosome 14. *Genomics* **48**: 232-241.

Deloukas, P., Schuler, G.D., Gyapy, G., Beasley, E.M., Soderlund, C., *et al.* (1998). A physical map of 30,000 human genes. *Science* **282**: 744-746.

Del Sal, G., Collavin, L., Ruaro, M.E., Edomi, P., Saccone, S., Valle, G.D., and Schneider, C. (1994). Structure, function, and chromosome mapping of the growth-suppressing human homologue of the murine gas1 gene. *Proc. Natl. Acad. Sci. USA* **91**: 1848-1852.

Del Sal, G., Ruaro, M.E., Philipson, L., and Schneider, C. (1992). The growth arrest-specific gene, gas1, is involved in growth suppression. *Cell* **70**: 595-607.

De Sario, A., Geigl, E-M., Palmieri, G., D'Urso, M., and Bernardi, G. (1996). A compositional map of human chromosome band Xq28. *Proc. Natl. Acad. Sci. USA*. **93**: 1298-1302.

Devilee, P., and Cornelisse, C.J. (1994). Somatic genetic changes in human breast cancer. *Biochimica et Biophysica Acta* **1198**: 113-130.

de Winter, J.P., Waisfisz, Q., Rooimans, M.A., van Berkel, C.G.M., Bosnoyan-Collins, L., Alon, N., Carreau, M., Bender, O., Demuth, I., Schindler, D., Pronk, J.C., Arwert, F., Hoehn, H., Digweed, M., Buchwald, M., and Joenje, H. (1998). The Fanconi anaemia group G gene *FANCG* is identical with *XRCC9*. *Nature Genet.* **20**: 281-283.

Dib, C., Faure, S., Fizames, C., Samson, D., Drouot, N., Vignal, A., Millasseau, P., Marc, S., Hazan, J., Seboun, E., Lathrop, M., Gyapy, G., Morissette, J., and Weissenbach, J. (1996). A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* **380**: 152-154.

Digweed, M. (1993). Human genetic instability syndromes: single gene defects with increased risk of cancer. *Toxicol. Letters* **67**: 259-281.

Dizikes, G.J. (1995). Update on the human genome project. *Genetic Testing* **15**: 973-988.

Dode, L., De Greef, C., Mountian, I., Attard, M., Town, M.M., Casteels, R., and Wuytack, F. (1998). Structure of the human sarco/endoplasmic reticulum Ca^{2+}-ATPase 3 gene. *J. Biol. Chem.* **273**: 13982-13994.

Doggett, N.A., Goodwin, L.A., Tesmer, J.G., Meincke, L.J., Bruce, D.C., *et al.* (1995). An integrated physical map of human chromosome 16. *Nature* **377** Suppl: 335-365.

Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., *et al.* (1987). A

genetic linkage map of the human genome. *Cell* **51**: 319-337.

Dorion-Bonnet, F., Mautalen, S., Hostein, I., and Longy, M. (1995). Allelic imbalance study of 16q in human primary breast carcinomas using microsatellite markers. *Genes Chrom. Cancer* **14**: 171-181.

Dorssers, L.C., Van Agthoven, T., Dekker, A., Van Agthoven, T.L., and Kok, E.M. (1993). Induction of antiestrogen resistance in human breast cancer cells by random insertional mutagenesis using defective retroviruses: identification of Bcar-1, a common integration site. *Mol. Endocrinol.* 7: 870-878.

Draberova, E., Lukas, Z., Ivanyi, D., Viklicky, V., and Draber, P. (1998). Expression of class III beta-tubulin in normal and neoplastic human tissues. *Histochem Cell Biol.* **109**: 231-239.

Driouch, K., Dorion-Bonnet, F., Briffod, M., Champeme, M-H., Longy, M., and Lidereau, R. (1997). Loss of heterozygosity on chromosome arm 16q in breast cancer metastases. *Genes Chrom. Cancer* **19**: 185-191.

Dubovsky, J., Sheffield, V.C., Duyk, G.M., and Weber, J.L. (1995). Sets of short tandem repeat polymorphisms for efficient linkage screening of the human genome. *Hum. Mol. Genet.* **4**: 449-452.

Dutrillaux, B., Gerbault-Seureau, M., and Zafrani, B. (1990). Characterisation of chromosomal anomalies in human breast cancer. A comparison of 30 paradiploid cases with few chromosome changes. *Cancer Genet. Cytogenet.* **49**: 203-217.

Duyk, G.M., Kim, S., Myers, R.M., and 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.

Easton, D.F. (1994). Cancer risks in A-T heterozygotes. *Int. J. Radiat. Biol.* **66**: S177-S182.

Easton, D.F., Bishop, D.T., Ford, D., Crockford, G.P., and Breast Cancer Linkage Consortium. (1993). Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. *Am. J. Hum. Genet.* **52**: 678-701.

Edwards, A., Civitello, A., Hammond, H.A., and Caskey, C.T. (1991). DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *Am. J. Hum. Genet.* **49**: 746-756.

El-Ashry, D., and Lippman, M.E. (1994). Molecular biology of breast carcinoma. *World J. Surg.* **18**: 12-20.

Elo, J.P., Harkonen, P., Kyllonen, A.P., Lukkarinen, O., Poutanen, M., Vihko, R., and Vihko, P. (1997). Loss of heterozygosity at 16q24.1-q24.2 is significantly associated with metastatic and aggressive behavior of prostate cancer. *Cancer Res.* **57**: 3356-3359.

Elvin, P., Slynn, G., Black, D., Graham, A., Butler, R., Riley, J., Anand, R., and Markham, A.F. (1990). Isolation of cDNA clones using yeast artificial chromosome probes. *Nucleic*

Acids Res. **18**: 3913-3917.

Engh, G., Van den Sachs, R., and Trask, B.J. (1992). Estimating genomic distance from DNA sequence location in cell nuclei by a random walk model. *Science* **257**: 1410-1412.

Epstein, L.B., Shen, J., Abele, J.S., and Reese, C.C. (1980). Sensitivity of human ovarian carcinoma cells to interferon and other antitumour agents assessed by an *in vitro* semi-solid agar technique. *Ann. NY Acad. Sci.* **350**: 228-244.

Evans, C., Bouzyk, M., Cox, S., Warne, D., Bryant, S.P., and Spurr, N.K. (1996). Chromosomal assignment of 79 cDNAs from a range of human tissues. *Genomics* **31**: 130-134.

Evdokiou, A., Webb, G.C., Peters, G.B., Dobrovich, A., O'Keefe, D.S., Forbes, I.J., and Cowled, P.A. (1993). Localisation of the human growth arrest-specific gene (*GAS1*) to chromosome bands 9q21.3-q22, a region frequently deleted in myeloid malignancies. *Genomics* **18**: 731-733.

Fan, J-B., DeYoung, J., Lagace, R., Lina, R.A., Xu, Z., Murray, J.C., Buetow, K.H., Weissenbach, J., Goold, R.D., Cox, D.R., and Myers, R.M. (1994). Isolation of yeast artificial chromosome clones from 54 polymorphic loci mapped with high odds on human chromosome 4. *Hum. Mol. Genet.* **3**: 243-246.

Fanconi, G. (1928). Familiare infantile perniziosaartige Anamie (pernizioses Blutbild und Konstitution). *Jahrb Kinderheil* **117**: 257-280.

Farshid, M., Hsia, C.C., and Tabor, E. (1994). Alterations of the RB tumour suppressor gene in hepatocellular carcinoma and hepatoblastoma cell lines in association with abnormal p53 expression. *J. Viral Hep.* **1**: 45-53.

Fearon, E.R., Cho, K.R., Nigro, J.M., Kern, S.E., Simons, J.W., Ruppert, J.M., Hamilton, S.R., Preisinger, A.C., Thomas, G., Kinzler, K.W., and Vogelstein, B. (1990). Identification of a chromosome 18q gene that is altered in colorectal cancers. *Science* **247**: 49-56.

Fearon, E.R., and Vogelstein, B. (1990). A genetic model for colorectal tumourigenesis. *Cell* **61**: 759-767.

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

Ferrero, G.B., Franco, B., Roth, E.J., Firulli, B.A., Borsani, G., Delmas-Mata, J., Weissenbach, J., Halley, G., Schlessinger, D., Chinault, A.C., Zoghbi, H., Nelson, D.L., and Ballabio, A. (1995). An integrated physical and genetic map of a 35 Mb region on chromosome Xq22.3-Xp21.3. *Hum. Mol. Genet.* **4**: 1821-1827.

Fickett, J.W. (1996). Finding genes by computer: the state of the art. *Trends in Genet.* **12**: 316-320.

Filippova, G.N., Fagerlie, S., Klenova, E., Myers, C., Dehner, Y., Goodwin, G.H., Neiman,

P.E., Collins, S., and Lobanenkov, V.V. (1996). An exceptionally conserved transcriptional repressor, *CTCF*, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian *c-myc* oncogenes. *Mol. Cell Biol.* **16**: 2802-2813.

Filippova, G.N., Lindblom, A., Meincke, L.J., Klenova, E.M., Neiman, P.E., Collins, S.J., Doggett, N.A., and Lobanenkov, V.V. (1998). A widely expressed transcription factor with multiple DNA sequence specificity, *CTCF*, is localised at chromosome segment 16q22.1 within one of the smallest regions of overlap for common deletions in breast and prostate cancers. *Genes Chrom. Cancer* **22**: 26-36.

Flint, J., Thomas, K., Micklem, G., Raynham, H., Clark, K., Doggett, N.A., King, A., and Higgs, D.R. (1997). The relationship between chromosome structure and function at a human telomeric region. *Nature Genet.* **15**: 252-257.

Florijn, R.J., Blonden, L.A.J., Vrolijk, J., Wiegant, J., Vaandrager, J.W., Baas, F., Den Dunnen, J.T., Tanke, H.J., Van Ommen, G.J.B., and Raap, A.K. (1995). High resolution DNA FiberFISH genomic DNA mapping and colour barcoding of large genes. *Hum. Mol. Genet.* **4**: 831-836.

Foote, S., Vollrath, D., Hilton, A., and Page, D.C. (1992). The human Y chromosome: overlapping DNA clones spanning the euchromatic region. *Science* **258**: 60-66.

Friedman, L.S., Ostermeyer, E.A., Szabo, C.I., Dowd, P., Lynch, E.D., Rowell, S.E., and 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.

Friend, S.H., Bernards, R., Rogelj, S., Weinberg, R.A., Rapaport, J.M., Albert, D.M., and Dryja, T.P. (1986). A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* **323**: 643-646.

Fukushima, A., Okuba, K., Sugino, H., Hori, N., Matoba, R., Niiyama, T., Murakawa, K., Yoshii, J., Yokoyama, M., and Matsubara, K. (1994). Chromosomal assignment of HepG2 3'-directed partial cDNA sequences by Southern blot hybridisation using monochromosomal hybrid cell panels. *Genomics* **22**: 127-136.

Futreal, P.A., Liu, Q., Shattuck-Eidens, D., Cochran, C., Harshman, K., *et al.* (1994). BRCA1 mutations in primary breast and ovarian carcinomas. *Science* **266**: 120-122.

Gahl, W.A., Schneider, J.A., and Aula, P.P. (1995). Lysosomal transport disorders: cystinosis and sialic acid storage disorders. In *The Metabolic and Molecular Basis of Inherited Disease.* (eds Scriver, C.R. Beaudet, A.L., Sly, W.S., and Valle, D.) 3763-3797 (McGraw-Hill, New York, 1995).

Gecz, J., Bielby, S., Sutherland, G.R., and Mulley. J.C. (1997). Gene structure and subcellular localisation of FMR2, a member of a new gene family of putative transcription activators. *Genomics* **44**: 201-213.

Gecz, J., Villard, L., Lossi, A.M., Millasseau, P., Djabali, M., and Fontes, M. (1993). Physical and transcriptional mapping of DXS56-PGK1 1 Mb region: identification of three new

transcripts. *Hum. Mol. Genet.* **2**: 1389-1396.

Gemmill, R.M., Chumakov, I., Scott, P., Waggoner, B., Rigault, P., Cypser, J., Chen, Q., Weissenbach, J., Gardiner, K., Wang, H., Pekarsky, Y., Le Gall, I., Le Paslier, D., Guillou, S., Li, E., Robinson, L., Hahner, L., Todd, S., Cohen, D., and Drabkin, H.A. (1995). A second-generation YAC contig map of human chromosome 3. *Nature* **377** Suppl: 299-319.

Geraghty, M.T., Brody, L.C., Martin, L.S., Marble, M., Kearns, W., Pearson, P., Monaco, A.P., Lehrach, H., and Valle, D. (1993). The isolation of cDNAs from OATL1 at Xq11.2 using a 480 kb YAC. *Genomics* **16**: 440-446.

Giles, R.H., Petrij, F., Dauwerse, H.G., den Hollander, A.I., Lushnikova, T., van Ommen, G-J. B., Goodman, R.H., Deaven, L.L., Doggett, N.A., Peters, D.J.M., and Breuning, M.H. (1997). Construction of a 1.2-Mb contig surrounding, and molecular analysis of, the human CREB-binding protein (CBP/CREBBP) gene on chromosome 16p13.3. *Genomics* **42**: 96-114.

Glavac, D., and Dean, M. (1993). Optimisation of the single-strand conformation polymorphism (SSCP) technique for detection of point mutations. *Hum. Mut.* **2**: 404-414.

Godfrey, T.E., Cher, M.L., Chhabra, V., and Jensen, R.H. (1997). Allelic imbalance mapping of chromosome 16 shows two regions of common deletion in prostate adenocarcinoma. *Cancer Genet. Cytogenet.* **98**: 36-42.

Goldstein, L.S.B. (1993). With apologies to scheherazade: tails of 1001 kinesin motors. *Annu. Rev. Genet.* **27**: 319-351.

Goss, S.J., and Harris, H. (1975). New method for mapping genes in human chromosomes. *Nature* **255**: 680-684.

Groden, J., Thliveris, A., Samowitz, W., Carlson, M., Gelbert, L., *et al.* (1991). Identification and characterisation of the familial adenomatous polyposis coli gene. *Cell* **66**: 589-600.

Grodzicker, T., Williams, J., Sharp, P., and Sambrook, J. (1974). Physical mapping of temperature-sensitive mutations of adenoviruses. *Cold Spring Harbor Symp Quant. Biol.* **39**: 439-446.

Grunstein, M., and Hogness, D.S. (1975). Colony hybridisation: a method for the isolation of cloned DNAs that contain a specific gene. *Proc. Natl. Acad. Sci. USA* **72**: 3961-3965.

Gullick, W.J., Love, S.B., Wright, C., Barnes, D.M., Gusterson, B., Harris, A.L., and Altman, D.G. (1991). C-erbB-2 protein overexpression in breast cancer is a risk factor in patients with involved and uninvolved lymph nodes. *Br. J. Cancer* **63**: 434-438.

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., Bonilla, E., and Martin, J.B. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**: 234-238.

Gyapy, G., Morisette, J., Dib, C., Fizames, C., Millasseau, P., Marc, S., Bernardi, G., Lathrop, M., and Weissenbach, J. (1994). The 1993-4 Genethon human genetic linkage map. *Nature Genet.* **7**: 246-339.

Gyapy, G., Schmitt, K., Fizames, C., Jones, H., Vega-Czarny, N., Spillett, D., Muselet, D., Prud'Homme, J-F., Dib, C., Auffrey, C., Morissette, J., Weissenbach, J., and Goodfellow, P.N. (1996). A radiation hybrid map of the human genome. *Hum. Mol. Genet.* **5**: 339-346.

Hall, J.M., Lee, M.K., Newman, B., Morrow, J.E., Anderson, L.A., Huey, B., and King, M-C. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250**: 1684-1689.

Hansen, R., and Oren, M. (1997). p53; from inductive signal to cellular effect. *Curr. Opin. Genet. Devel.* **7**: 46-51.

Harshman, K., Bell, R., Rosenthal, J., Katcher, H., Miki, Y., *et al.* (1995). Comparison of the positional cloning methods used to isolate the *BRCA1* gene. *Hum. Mol. Genet.* **4**: 1259-1266.

Hawkins, J.D. (1988). A survey on intron and exon lengths. *Nucleic Acids Res.* **16**: 9893-9908.

Hayashi, K. (1992). PCR-SSCP: a method for detection of mutations. *GATA* **9**: 73-79.

Heng, H.H.Q., Squire, J., and Tsui, L.C. (1992). High resolution mapping of mammalian genes by *in situ* hybridisation to free chromatin. *Proc. Natl. Acad. Sci. USA* **89**: 9509-9513.

Herman, J.G., Latif, F., Weng, Y., Lerman, M.I., Zbar, B., Liu, S., Samid, D., Duan, D-S.R., Gnarra, J.R., Linehan, W.M., and Baylin, S.B. (1994). Silencing of the VHL tumour-suppressor gene by DNA methylation in renal carcinoma. *Proc. Natl. Acad. Sci. USA* **91**: 9700-9704.

Hillier, L., Lennon, G., Becker, M., Fatima Bonaldo, M., Chiapelli, B., *et al.* (1996). Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6**: 807-828.

Hiraguri, S., Godfrey, T., Nakamura, H., Graff, J., Collins, C., Shayesteh, L., Doggett, N., Johnson, K., Wheelock, M., Herman, J., Baylin, S., Pinkel, D., and Gray, J. (1998). Mechanisms of inactivation of E-cadherin in breast cancer cell lines. *Cancer Res.* **58**: 1972-1977.

Hisama, F.M., Oshima, J., Yu, C-E., Fu, Y-H., Mulligan, J., Weissman, S.M., and Schellenberg, G.D. (1998). Comparison of methods for identifying transcription units and transcription map of the Werner syndrome gene region. *Genomics* **52**: 352-357.

Horowitz, J.M., Yandell, D.W., Park, S.H., Canning, S., Whyte, P., Buchkovich, K., Harlow, E., Weinberg, R.A., and Dryja, T.P. (1989). Point mutational inactivation of the retinoblastoma antioncogene. *Science* **243**: 937-940.

Houlgatte, R., Mariage-Samson, R., Dupart, S., Tessier, A., Bentolila, S., Lamy, B., and

Auffray, C. (1995). The Genexpress Index: a resource for gene discovery and the genic map of the human genome. *Genome Res.* **5**: 272-304.

Hu, R.J., Lee, M.P., Connors, T.D., Johnson, L.A., Burn, T.C., Su, K., Landes, G.M., and Feinberg, A.P. (1997). A 2.5-Mb transcript map of a tumor-suppressing subchromosomal transferable fragment from 11p15.5, and isolation and sequence analysis of three novel genes. *Genomics* **46**: 9-17.

Hudson, T.J., Stein, L.D., Gerety, S.S., Ma, J., Castle, A.B., *et al.* (1995). An STS-based map of the human genome. *Science* **270**: 1945-1954.

Hunziker, W., and Geuze, H.J. (1996). Intracellular trafficking of lysosomal membrane proteins. *BioEssays* **18**: 379-389.

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., and Savoia, A. (1997). The genomic organisation of the Fanconi anemia group A (*FAA*) gene. *Genomics* **41**: 309-314.

Ichikawa, H., Hosoda, F., Arai, Y., Shimizu, K., Ohira, M., and Ohki, M. (1993). A *Not*I restriction map of the entire long arm of human chromosome 21. *Nature Genet.* **4**: 361-366.

Iida, A., Isobe, R., Yoshimoto, M., Kasumi, F., Nakamura, Y., and Emi, M. (1997). Localisation of a breast cancer tumour-suppressor gene to a 3-cM interval within chromosomal region 16q22. *Br. J. Cancer* **75**: 264-267.

Ilyas, M., and Tomlinson, I.P. (1997). The interactions of APC, E-cadherin, and β-catenin in tumour development and progression. *J. Pathol.* **182**: 128-137.

International Batten Disease Consortium. (1995). Isolation of a novel gene underlying Batten disease, CLN3. *Cell* **82**: 949-957.

Ioannou, P.A., Amemiya, C.T., Garnes, J., Kroisel, P.M., Shizuya, H., Chen, C., Batzer, M.A., and de Jong, P.J. (1994). A new bacteriophage P1-derived vector for the propogation of large human DNA fragments. *Nature Genet* **6**: 84-89.

Isola, J.J., Kallioniemi, O-P., Chu, L.W., Fuqua, S.A.W., Hilsenbeck, S.G., Osborne, C.K., and Waldman, F.M. (1995). Genetic aberrations detected by comparative genomic hybridisation predict outcome in node-negative breast cancer. *Am. J. Pathol.* **147**: 905-911.

Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Hudson Jr, J., Boguski, M.S., Lashkari, D., Shalon, D., Botstein, D., and Brown, P.O. (1999). The transcriptional program in the response of human fibroblasts to serum. *Science* **283**: 83-87.

James, M.R., Richard III, C.W., Schott, J-J., Yousry, C., Clark, K., Bell, J., Terwilliger, J.D., Hazan, J., Dubay, C., Vignal, A., Agrapart, M., Imai, T., Nakamura, Y., Polymeropoulos, M., Weissenbach, J., Cox, D.R., and Lathrop, G.M. (1994). A radiation hybrid map of 506

STS markers spanning human chromosome 11. *Nature Genet.* **8**: 70-76.

Jean, G., Fuchshuber, A., Town, M.M., Gribouval, O., Schneider, J.A., Broyer, M., van't Hoff, W., Niaudet, P., and Antignac, C. (1996). High-resolution mapping of the gene for cystinosis, using combined biochemical and linkage analysis. *Am. J. Hum. Genet.* **58**: 535-543.

Jeffreys, J.J., Wilson, V., and Thein, S.L. (1985). Hypervariable "minisatellite" regions in human DNA. *Nature* **314**: 67-73.

Joenje, H., Oostra, A.B., Wijker, M., Di Summa, F., Van Berker, C., Ebell, W., Van Weel, M., Pronk, J.C., Buchwald, M., and Arwert, F. (1997). Evidence for at least eight Fanconi anemia genes. *Am. J. Hum. Genet.* **61**: 940-944.

Jones, K.W., Chevrette, M., Shapero, M.H., and Fournier, R.E.K. (1992). Generation of region- and species-specific expressed gene probes from somatic cell hybrids. *Nature Genet.* **1**: 278-283.

Jones, P.A. (1996). DNA methylation errors and cancer. *Cancer Res.* **56**: 2463-2467.

Jordanova, A., Kalaydjieva, L., Savov, A., Claustres, M., Schwarz, M., Estivill, X., Angelicheva, D., Haworth, A., Casals, T., and Kremensky, I. (1997). SSCP analysis: a blind sensitivity trial. *Hum. Mut.* **10**: 65-70.

Kahloun, A.E., Chauvel, B., Mauvieux, V., Dorval, I., Jouanolle, A-M., Gicquel, I., le Gall, J-Y., and David, V. (1993). Localisation of seven new genes around the HLA-A locus. *Hum. Mol. Genet.* **2**: 55-60.

Kallioniemi, A., Kallioniemi, O-P., Piper, J., Tanner, M., Stokke, T., Chen, L., Smith, H.S., Pinkel, D., Gray, J.W., and Waldman, F.M. (1994). Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridisation technique. *Proc. Natl. Acad. Sci. USA* **91**: 2156-2160.

Kallioniemi, A., Kallioniemi, O-P., Sudar, D., Rutovitz, D., Gray, J.W., Waldman, F., and Pinkel, D. (1992). Comparative genomic hybridisation for molecular cytogenetic analysis of solid tumours. *Science* **258**: 818-821.

Kan, Y., and Dozy, A. (1978). Antenatal diagnosis of sickle-cell anaemia by DNA analysis of amniotic-fluid cells. *Lancet* **2**: 910-912.

Kashiwaba, M., Tamura, G., Suzuki, Y., Maesawa, C., Ogasawara, S., Sakata, K., and Satodate, R. (1995). Epithelial-cadherin gene is not mutated in ductal carcinomas of the breast. *Jpn. J. Cancer Res.* **86**: 1054-1059.

Kaul, R., Balamurugan, K., Gao, G.P., and Matalon, R. (1994). Canavan disease: genomic organisation and localisation of human *ASPA* to 17p13-ter and conservation of the *ASPA* gene during evolution. *Genomics* **21**: 364-370.

Kerangueven, F., Essioux, L., Dib, A., Noguchi, T., Allione, F., Geneix, J., Longy, M., Lidereau, R., Eisinger, F., Pebusque, M.J., Jacquemier, J., Bonaiti-Pellie, C., Sobol, H.,

and Birnbaum, D. (1995). Loss of heterozygosity and linkage analysis in breast carcinoma: indication for a putative third susceptibility gene on the short arm of chromosome 8. *Oncogene* **10**: 1023-1026.

Khan, A.S., Wilcox, A.S., Polymeropoulos, M.H., Hopkins, J.A., Stevens, T.J., Robinson, M., Orpana, A.K., and Sikela, J.M. (1992). Single pass sequencing and physical and genetic mapping of human brain cDNAs. *Nature Genet.* **2**: 180-185.

Kim, U-J., Shizuya, H., Kang, H-L., Choi, S-S., Garrett, C.L., Smink, L.J., Birren, B.W., Korenberg, J.R., Dunham, I., and Simon, M.I. (1996). A bacterial artificial chromosome-based framework contig map of human chromosome 22q. *Proc. Natl. Acad. Sci. USA.* **93**: 6297-6301.

King, C.R., Schimke, R.N., Arthur, T., Davoren, B., and Collins, D. (1986). Proximal 3p deletion in renal cell carcinoma cells from a patient with von Hippel-Lindau disease. *Cancer Genet. Cytogenet.* **27**: 345-348.

Kinzler, K.W., and Vogelstein, B. (1997). Gatekeepers and caretakers. *Nature* **386**: 761-763.

Klenova, E.M., Nicolas, R.H., Paterson, H.F., Carne, A.F., Heath, C.M., Goodwin, G.H., Neiman, P.E., and Lobanenkov, V.V. (1993). *CTCF*, a conserved nuclear factor required for optimal transcriptional activity of the chicken *c-myc* gene, is an 11-Zn-finger protein differentially expressed in multiple forms. *Mol. Cell Biol.* **13**: 7612-7624.

Knudson, A.G. (1971). Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl. Acad. Sci.* **68**: 820-823.

Kogan, S.C., Doherty, B.S., and Gitsher, J. (1987). An improved method for prenatal diagnosis of genetic diseases by analysis of amplified DNA sequences. *N. Eng. J. Med.* **317**: 985-990.

Kohara, Y., Akiyama, K., and Isono, K. (1987). The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* **50**: 495-508.

Koi, M., Johnson, L.A., Kalikin, L.M., Little, P.F.R., Nakamura, Y., and Feinberg, A.P. (1993). Tumour cell growth arrest caused by subchromosomal transferable DNA fragments from human chromosome 11. *Science* **260**: 361-364.

Koi, M., Shimizu, M., Morita, H., Yamada, H., and Oshimura, M. (1989). Construction of mouse A9 clones containing a single human chromosome tagged with neomycin-resistance gene via microcell fusion. *Jpn. J. Cancer Res.* **80**: 413-418.

Korn, B., Sedlacek, Z., Manca, A., Kioschis, P., Konecki, D., Lehrach, H., and Poustka, A. (1992). A strategy for the selection of transcribed sequences in the Xq28 region. *Hum. Mol. Genet.* **1**: 235-242.

Kozman, H.M., Phillips, H.A., Callen, D.F., Sutherland, G.R., and 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.

Krauter, K., Montgomery, K., Yoon, S-J., LeBlanc-Straceski, J., Renault, B., *et al.* (1995). A second-generation YAC contig map of human chromosome 12. *Nature* **377** Suppl: 321-333.

Kremmidiotis, G., Baker, E., Crawford, J., Eyre, H.J., Nahmias, J., and Callen, D.F. (1998). Localisation of human cadherin genes to chromosome regions exhibiting cancer-related loss of heterozygosity. *Genomics* **49**: 467-471.

Krizman, D.B., and Berget, S.M. (1993). Efficient selection of 3'-terminal exons from vertebrate DNA. *Nucleic Acids Res.* **21**: 5198-5202.

Kruglyak, L. (1997). The use of a genetic map of biallelic markers in linkage studies. *Nature Genet.* **17**: 21-24.

Kull, F.J., Sablin, E.P., Lau, R., Fletterick, R.J., and Vale, R.D. (1996). Crystal structure of the kinesin motor domain reveals a structural similarity to myosin. *Nature* **380**: 550-555.

Kulp, D., Haussler, D., Reese, M.G., and Eeckman, F.H. (1996). A generalised hidden Markov model for the recognition of human genes in DNA. In *Proc. Fourth International Conference on Intelligent Systems for Molecular Biology*, pp. 134-142.

Kumar, S. (1995). ICE-like proteases in apoptosis. *Trends Biochem. Sci.* **20**: 198-202.

Lagoda, P.J.L., Trent, J.M., and Meese, E.U. (1994). Chromosome specific cDNA libraries: reduction of unspecific priming events by purification of heteronuclear RNA. *Mol. Biol. Reports* **19**: 89-92.

Lammie, G.A., and Peters, G. (1991). Chromosome 11q13 abnormalities in human cancer. *Cancer Cells* **3**: 413-420.

Landegent, J.E., Jansen in de Wal, N., Ommen, G.J.B., Baas, F., De Vijlder, J.J.M., Van Duijn, P., and Van der Ploeg, M. (1985). Chromosomal localisation of a unique gene by non-autoradiographic *in situ* hybridisation. *Nature* **317**: 175-177.

Lanfrancone, L., Pelicci, G., and Pelicci, P.G. (1994). Cancer genetics. *Curr. Opin. Genet. Devel.* **4**: 109-119.

Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. (1992). CpG islands as gene markers in the human genome. *Genomics* **13**: 1095-1107.

Lasko, D., Cavenee, W., and Nordenskjold, M. (1991). Loss of constitutional heterozygosity in human cancer. *Annu. Rev. Genet.* **25**: 281-314.

Latil, A., Cussenot, O., Fournier, G., Driouch, K., and Lidereau, R. (1997). Loss of heterozygosity at chromosome 16q in prostate adenocarcinoma: identification of three independent regions. *Cancer Res.* **57**: 1058-1062.

Lennon, G., Auffray, C., Polyeropoulos, M., and Soares, M.B. (1996). The I.M.A.G.E. consortium: an integrated molecular analysis of genomes and their expression. *Genomics* **33**: 151-152.

Leppert, M., Dobbs, M., Scambler, P., O'Connell, P., Nakamura, Y., Stauffer, D., Woodward, S., Burt, R., Hughes, J., Gardner, E., Lathrop, M., Wasmuth, J., Lalouel, J-M., and White, R. (1987). The gene for familial polyposis coli maps to the long arm of chromosome 5. *Science* **238**: 1411-1415.

Levran, O., Erlich, T., Magdalena, N., Gregory, J.J., Batish, S.D., Verlander, P.C., and Auerbach, A.D. (1997). Sequence variation in the Fanconi anaemia gene *FAA*. *Proc. Natl. Acad. Sci. USA* **94**: 13051-13056.

Li, S., MacLachlan, T.K., De luca, A., Claudio, P.P., Condorelli, G., and Giordano, A. (1995). The cdc2-related kinase, PISSLRE, is essential for cell growth and acts in G_2 phase of the cell cycle. *Cancer Res.* **55**: 3992-3995.

Liaw, D., Marsh, D.J., Li, J., Dahia, P.L., Wang, S.I., Zheng, Z., Bose, S., Call, K.M., Tsou, H.C., Peacocke, M., Eng, C., and Parsons, R. (1997). Germline mutations of the *PTEN* gene in Cowden disease, an inherited breast and thyroid cancer syndrome. *Nature Genet.* **16**: 64-66.

Lih, C-J., Cohen, S.N., Wang, C., and Lin-Chao, S. (1996). The platelet-derived growth factor α-receptor is encoded by a growth-arrest-specific (*gas*) gene. *Proc. Natl. Acad. Sci. USA* **93**: 4617-4622.

Litt, M., and Luty, J.A. (1989). A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am. J. Hum. Genet.* **44**: 397-401.

Liu, N., Lamerdin, J.E., Tucker, J.D., Zhou, Z.Q., Walter, C.A., Albala, J.S., Busch, D.B., and Thompson, L.H. (1997). The human *XRCC9* gene corrects chromosomal instability and mutagen sensitivities in CHO UV40 cells. *Proc. Natl. Acad. Sci. USA* **94**: 9232-9237.

Liu, P., Legerski, R., and Siciliano, M.J. (1989). Isolation of human transcribed sequences from human-rodent somatic cell hybrids. *Science* **246**: 813-815.

Livak, K.J., Marmaro, J., and Todd, J.A. (1995). Towards fully automated genome-wide polymorphism screening. *Nature Genet.* **9**: 341-342.

Lo Ten Foe, J., Rooimans, A.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., and Joenje, H. (1996). Expression cloning of a cDNA for the major Fanconi anemia gene, *FAA*. *Nature Genet.* **14**: 320-323.

Longhurst, P.A., Schwegel, T., Folander, K., and Swanson, R. (1996). The human P2x1 receptor: molecular cloning, tissue distribution, and localisation to chromosome 17. *Biochim. Biophys. Acta.* **1308**: 185-188.

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., and Deaven, L.L. (1993). Construction and characterisation of partial digest DNA libraries made from flow-sorted human chromosome 16. *GATA* **10**: 69-76.

Lovett, M., Kere, J., and Hinton, L.M. (1991). Direct selection: a method for the isolation of cDNAs encoded by large genomic regions. *Proc. Natl. Acad. Sci. USA* **88**: 9628-9632.

Maatta, A., Bornstein, P., and Penttinen, R.P.K. (1991). Highly conserved sequences in the 3'-untranslated region of the COL1A1 gene bind cell-specific nuclear proteins. *FEBS* **279**: 9-13.

Malkin, D., Li, F.P., Strong, L.C., Fraumeni Jr, J.F., Nelson, C.E., Kim, D.H., Kassel, J., Gryka, M.A., Bischoff, F.Z., Tainsky, M.A., and Friend, S.H. (1990). Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* **250**: 1233-1238.

Manfioletti, G., Brancolini, C., Avanzi, G., and Schneider, C. (1993). The protein encoded by a growth arrest-specific gene (gas6) is a new member of the vitamin K-dependant proteins related to protein S, a negative coregulator in the blood coagulation cascade. *Mol. Cell. Biol.* **13**: 4976-4985.

Manfioletti, G., Ruaro, M.E., Del Sal, G., Philipson, L., and Schneider, C. (1990). A growth arrest-specific (*gas*) gene codes for a membrane protein. *Mol. Cell. Biol.* **10**: 2924-2930.

Maniatis, T., Hardison, R., Lacy E., Lauer, J., O'Connell, C., Quon, D., Sim, G.K., and Efstratiadis, A. (1978). The isolation of structural genes from libraries of eucaryotic DNA. *Cell* **15**: 687-701.

Mansouri, M., Spurr, N., Goodfellow, P.N., and Kemler, R. (1988). Characterisation and chromosomal localisation of the gene encoding the human cell adhesion molecule uvomorulin. *Differentiation* **38**: 67-71.

Marra, M.A., Kucaba, T.A., Dietrich, N.L., Green, E.D., Brownstein, B., Wilson, R.K., McDonald, K.M., Hillier, L.W., McPherson, J.D., and Waterston, R.H. (1997). High throughput fingerprint analysis of large-insert clones. *Genome Res.* **7**: 1072-1084.

Matise, T.C., Perlin, M., and Chakravarti, A. (1994). Automated construction of genetic linkage maps using an expert system (MultiMap): a human genome linkage map. *Nature Genet.* **6**: 384-390.

Maw, M.A., Grundy, P.E., Millow, L.J., Eccles, M.R., Dunn, R.S., Smith, P.J., Feinberg, A.P., Law, D.J., Paterson, M.C., Telzerow, P.E., Callen, D.F., Thompson, A.D., Richards, R.I., and Reeve, A.E. (1992). A third Wilms' tumour locus on chromosome 16q. *Cancer Res.* **52**: 3094-3098.

McCormick, M.K., Campbell, E., Deaven, L., and Moyzis, R.K. (1993). Low-frequency chimaeric yeast artificial chromosome libraries from flow-sorted human chromosome 16 and 21. *Proc. Natl. Acad. Sci. USA* **90**: 1063-1067.

McDowell, G., Isogai, T., Tanigemi, A., Hazelwood, S., Ledbetter, D., Polymeropoulos, M.H., Lichter-Konecki, U., Konecki, D., Town, M.M., van't hoff, W., Weissenbach, J., and Gahl, W.A. (1996). Fine mapping of the cystinosis gene using an integrated genetic and physical map of a region within human chromosome band 17p13. *Biochem. Mol. Med.* **58**: 135-141.

McDowell, G., Town, M.M., van't Hoff, W., and Gahl, W.A. (1997). Clinical and molecular aspects of nephropathic cystinosis. *J. Mol. Med.* **76**: 295-302.

Miki, Y., Katagiri, T., Kasumi, F., Yoshimoto, T., and Nakamura, Y. (1996). Mutation analysis in the BRCA2 gene in primary breast cancers. *Nature Genet.* **13**: 245-247.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., *et al.* (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266**: 66-71.

Milanesi, L., Kolchanov, N., Rogozin, I., Kel, A., and Titov, I. (1993). Sequence functional inference. In Bishop, M.J. (ed.) *Guide to Human Genome Computing*. Academic Press, Cambridge, UK, pp. 249-312.

Mitelman, F., Mertens, F., and Johansson, B. (1997). A breakpoint map of recurrent chromosomal rearrangements in human neoplasia. *Nature Genet.* **15**: 417-474.

Moerland, E., Breuning, M.H., Cornelisse, C.J., and Cleton-Jansen, A-M. (1997). Exclusion of BBC1 and CMAR as candidate breast tumour-suppressor genes. *Br. J. Cancer* **76**: 1550-1553.

Mollenhauer, J., Wiemann, S., Scheurlen, W., Korn, B., Hayashi, Y., Wilgenbus, K.K., von Deimling, A., and Poustka, A. (1997). *DMBT1*, a new member of the SRCR superfamily, on chromosome 10q25.3-26.1 is deleted in malignant brain tumours. *Nature Genet.* **17**: 32-39.

Monaco, A.P., Neve, R.L., Colletti-Feener, C., Bertelson, C.J., Kurnit, D.M., and Kunkel, L.M. (1986). Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. *Nature* **323**: 646-650.

Moyzis, R.K., Torney, D.C., Meyne, J., Buckingham, J.M., Wu, J.R., Burks, C., Sirotkin, K.M., and Goad, W.B. (1989). The distribution of interspersed repetitive DNA sequences in the human genome. *Genomics* **4**:273-289.

Muleris, M., Almeida, A., Gerbault-Seureau, M., Malfoy, B., and Dutrillaux, B. (1994). Detection of DNA amplification in 17 primary breast carcinomas with homogeneously staining regions by a modified comparative genomic hybridisation technique. *Genes Chrom. Cancer* **10**: 160-170.

Mulley, J.C., and Sutherland, G.R. (1993). Integrating maps of chromosome 16. *Curr. Opin. Genet. Dev.* **3**: 425-431.

Murakawa, K., Matsubara, K., Fukushima, A., Yoshii, J., and Okubo, K. (1994). Chromosomal assignment of 3'-directed partial cDNA sequences representing novel genes expressed in granulocytoid cells. *Genomics* **23**: 379-389.

Murray, J.C., Buetow, K.H., Weber, J.L., Ludwigsen, S., Scherpbier-Heddema, T., *et al.* (1994). A comprehensive human linkage map with centimorgan density. *Science* **265**: 2049-2054.

Mushegian, A.R., Bassett, D.E. Jr., Boguski, M.S., Bork, P., and Koonin, E.V. (1997). Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs. *Proc. Natl. Acad. Sci. USA* **94**: 5831-5836.

Mushegian, A.R., Garey, J.R., Martin, J., and Liu, L.X. (1998). Large-scale taxonomic profiling of eukaryotic model organisms: a comparison of orthologous proteins encoded by the human, fly, nematode, and yeast genomes. *Genome Res.* **8**: 590-598.

Myers, J.C., Dickson, L.A., De Wet, W.J., Bernard, M.P., Chu, M-L., Di Liberto, M., Pepe, G., Sangiorgi, F.O., and Ramirez, F. (1983). Analysis of the 3' end of the human pro-alpha 2 (I) collagen gene. Utilisation of multiple polyadenylation sites in cultured fibroblasts. *J. Biol. Chem.* **258**: 10128-10135.

NIH/CEPH Collaborative Mapping Group. (1992). A comprehensive genetic linkage map of the human genome. *Science* **258**: 67-86.

Negrini, M., Castagnoli, A., Sabbioni, S., Recanatini, E., Giovannini, G., Possati, L., Stanbridge, E.J., Nenci, I., and Barbanti-Brodano, G. (1992). Suppression of tumourigenesis by the breast cancer cell line MCF-7 following transfer of normal human chromosome 11. *Oncogene* **7**: 2013-2018.

Nelson, D.L., Ledbetter, S.A., Corbo, L., Victoria, M.F., Ramirez-Solis, R., Webster, T.D., Ledbetter, D.H., and Caskey, C.T. (1989). *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.

Neuhausen, S., Gilewski, T., Norton, L., Tran, T., McGuire, P., Swensen, J., Hampel, H., Borgen, P., Brown, K., Skolnick, M., Shattuck-Eidens, D., Jhanwar, S., Goldgar, D., and Offit, K. (1996). Recurrent BRCA2 6174delT mutations in Ashkenazi Jewish women affected by breast cancer. *Nature Genet.* **13**: 126-128.

Nickerson, D.A., Kaiser, R., Lappin, S., Stewart, J., Hood, L., and Landegren, U. (1990). Automated DNA diagnostics using an ELISA-based oligonucleotide ligation assay. *Proc. Natl. Acad. Sci. USA*. **87**: 8923-8927.

Nielsen, K.V., Blichert-Toft, M., and Andersen, J. (1989). Chromosome analysis of *in situ* breast cancer. *Acta Oncol.* **28**: 919-922.

Ohashi, K., Nagata, K., Toshima, J., Nakano, T., Arita, H., Tsuda, H., Suzuki, K., and Mizuno, K. (1995). Stimulation of sky receptor tyrosine kinase by the product of growth arrest-specific gene 6. *J. Biol. Chem.* **270**: 22681-22684.

Olson, M., Dutchik, J.E., Graham, M.Y., Brodeur, G.M., Helms, C., Frank, M., MacCollin, M., Scheinman, R., and Frank, T. (1986). Random-clone strategy for genomic restriction mapping in yeast. *Proc. Natl. Acad. Sci. USA* **83**: 7826-7830.

Olson, M., Hood, L., Cantor, C., and Botstein, D. (1989). A common language for physical mapping of the human genome. *Science* **245**: 1434-1435.

Onyango, P., Lubyova, B., Gardellin, P., Kurzbauer, R., and Weith, A. (1998). Molecular

cloning and expression analysis of five novel genes in chromosome 1p36. *Genomics* **50**: 187-198.

Orita, M., Suzuki, Y., Sekiya, T., and Hayashi, K. (1989). Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction. *Genomics* **5**: 874-879.

Orkin, S.H. (1986). Reverse genetics and human disease. *Cell* **47**: 845-850.

Osborne, L.R., Martindale, D., Scherer, S.W., Shi, X-M., Huizenga, J., Heng, H.H.Q., Costa, T., Pober, B., Lew, L., Brinkman, J., Rommens, J., Koop, B., and Tsui, L-C. (1996). Identification of genes from a 500-kb region at 7q11.23 that is commonly deleted in Williams syndrome patients. *Genomics* **36**: 328-336.

Pandis, N., Heim, S., Bardi, G., Idvall, I., Mandahl, N., and Mitelman, F. (1992). Whole-arm t(1;16) and i(1q) as sole anomalies identify gain of 1q as a primary chromosomal abnormality in breast cancer. *Genes Chrom. Cancer* **5**: 235-238.

Parimoo, S., Patanjali, S.R., Shukla, H., Chaplin, D.D., and Weissman, S.M. (1991). cDNA selection: efficient PCR approach for the selection of cDNAs encoded in large chromosomal DNA fragments. *Proc. Natl. Acad. Sci. USA* **88**: 9623-9627.

Pavletich, N.P., Chambers, K.A., and Pabo, C.O. (1993). The DNA-binding domain of p53 contains the 4 conserved regions and the major mutation hotspots. *Genes Dev.* **7**: 2556-2564.

Pellet, O.L., Smith, M.L., Greene, A.A., and Schneider, J.A. (1988). Lack of complementation in somatic cell hybrids between fibroblasts from patients with different forms of cystinosis. *Proc. Natl. Acad. Sci. USA* **85**: 3531-3534.

Peterson, A., Patli, N., Robbins, C., Wang, L., Cox, D.R., and Myers, R.M. (1994). A transcript map of the Down Syndrome critical region on chromosome 21. *Hum. Mol. Genet.* **3**: 1735-1742.

Phelan, C.M., Lancaster, J.M., Tonin, P., Gumbs, C., Cochran, C., *et al.* (1996). Mutation analysis of the BRCA2 gene in 49 site-specific breast cancer families. *Nature Genet.* **13**: 120-122.

Phillips, K.K., Welch, D.R., Miele, M.E., Lee, J-H., Wei, L.L., and Weissman, B.E. (1996). Suppression of MDA-MB-435 breast carcinoma cell metastasis following the introduction of human chromosome 11. *Cancer Res.* **56**: 1222-1227.

Pierceall, W., Woodard, A., Morrow, J., Rimm, D., and Fearon, E. (1995). Frequent alterations in E-cadherin and α- and β-catenin expression in human breast cancer cell lines. *Oncogene* **11**: 1319-1326.

Polymeropoulos, M.H., Xiao, H., Glodek, A., Gorski, M., Adams, M.D., Moreno, R.F., Fitzgerald, M.G., Venter, J.C., and Merril, C.R. (1992). Chromosomal assignment of 46 brain cDNAs. *Genomics* **12**: 492-496.

Polymeropoulos, M.H., Xiao, H., Sikela, J.M., Adams, M.D., Venter, J.C., and Merril, C.R. (1993). Chromosomal distribution of 320 genes from a brain cDNA library. *Nature Genet.* **4**: 381-386.

Poustka, A., Pohl, T.M., Barlow, D.P., Frischauf, A-M., and Lehrach, H. (1987). Construction and use of human chromosome jumping libraries from *Not*I-digested DNA. *Nature* **325**: 353-355.

Pronk, J.C., Gibson, R.A., Savoia, A., Wijker, M., Morgan, N.V., Melchionda, S., Ford, D., Temtamy, S., Ortega, J.J. Jansen, S., Havenga, C., Cohn, R.J., de Ravel, T.J., Roberts, I., Westerveld, A., Easton, D.F., Joenje, H., Mathew, C.G., and Arwert, F. (1995). Localisation of the Fanconi anemia complementation group A gene to chromosome 16q24.3. *Nature Genet.* **11**: 338-340.

Pullman, W.E., and Bodmer, W.F. (1992). Cloning and characterisation of a gene that regulates cell adhesion. *Nature* **356**: 529-532.

Quackenbush, J., Davies, C., Bailis, J.M., Khristich, J.V., Diggle, K., Marchuck, Y., Tobin, J., Clark, S.P., Rodkins, A., Marcano, S., Churukian, A.C., Hutchinson, J.S., Probst, S., Romberg, L., Wei, Y.H., Nowak, N.J., Garner, H.R., Smith, M.W., Selleri, L., and Evans, G.A. (1995). An STS content map of human chromosome 11: localisation of 910 YAC clones and 109 islands. *Genomics* **29**: 512-525.

Qui, M., and Byers, P.H. (1998). Constitutive skipping of alternatively spliced exon 10 in the ATP7A gene abolishes Golgi localisation of the menkes protein and produces the occipital horn syndrome. *Hum. Mol. Genet.* **7**: 465-469.

Radford, D.M., Fair, K.L., Phillips, N.J., Ritter, J.H., Steinbrueck, T., Holt, M.S., and Donis-Keller, H. (1995). Allelotyping of ductal carcinoma *in situ* of the breast: deletion of loci on 8p, 13q, 16q, 17p and 17q. *Cancer Res.* **55**: 3399-3405.

Rajan, J.V., Wang, M., Marquis, S.T., and Chodosh, L.A. (1996). *Brca2* is coordinately regulated with *Brca1* during proliferation and differentiation in mammary epithelial cells. *Proc. Natl. Acad. Sci. USA* **93**: 13078-13083.

Reed, K.C., and Mann, D.A. (1985). Rapid transfer of DNA from agarose to nylon membranes. *Nucl. Acids. Res.* **13**: 7207-7221.

Reed, P.W., Davies, J.L., Copeman, J.B., Bennett, S.T., Palmer, S.M., Pritchard, L.E., Gough, S.C., Kawaguchi, Y., Cordell, H.J., Balfour, K.M., Jenkins, S.C., Powell, E.E., Vignal, A., and Todd, J.A. (1994). Chromosome-specific microsatellite sets for fluorescence-based semi-automated genome mapping. *Nature Genet.* **7**: 390-395.

Reeders, S.T., Breuning, M.H., Davies, K.E., Nicholls, R.D., Jarman, A.P., Higgs, D.R., Pearson, P.L., and Weatherall, D.J. (1985). A highly polymorphic DNA marker linked to adult polycystic kidney disease on chromosome 16. *Nature* **317**: 542-544.

Renwick, J.H., and Schulze, J. (1965). Male and female recombination fraction of the nail-patella: ABO linkage in man. *Ann. Hum. Genet.* **28**: 379-392.

Richards, R.I., Holman, K., Lane, S., Sutherland, G.R., and Callen, D.F. (1991). Human chromosome 16 physical map: mapping of somatic cell hybrids using multiplex PCR deletion analysis of sequence tagged sites. *Genomics* **10**: 1047-1052.

Riethman, H.C., Moyzis, R.K., Meyne, J., Burke, D.T., and Olson, M.V. (1989). Cloning human telomeric DNA fragments into *Saccharomyces cerevisiae* using a yeast-artificial-chromosome vector. *Proc. Natl. Acad. Sci. USA* **86**: 6240-6244.

Rockmill, B., Sym, M., Schertham, H., and Roeder, G.S. (1995). Role of two RecA homologs in promoting meiotic chromosome synapsis. *Genes Dev.* **9**: 2684-2695.

Rommens, J.M., Iannuzzi, M.C., Kerem, B-S., Drumm, M.L., Melmer, G., Dean, M., Rozmahel, R., Cole, J.L., Kennedy, D., Hidaka, N., Zsiga, M., Buchwald, M., Riordan, J.R., Tsui, L-P., and Collins, F.S. (1989). Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* **245**: 1059-1065.

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., and Hayden, M.R. (1993). A transcription map of the region containing the Huntington disease gene. *Hum. Mol. Genet.* **2**: 901-907.

Royle, N.J., Clarkson, R.E., Wong, Z., and Jeffreys, A.J. (1988). Clustering of hypervariable minisatellites in the proterminal region of human autosomes. *Genomics* **3**: 352-360.

Ruddle, F.H. (1984). The William Allan memorial award address: reverse genetics and beyond. *Am. J. Hum. Genet.* **36**: 944-953.

Russo, J., and Russo, I.H. (1991). Mammary tumorigenesis. *Prog. Exp. Tumor Res.* **33**: 175-191.

Ryan, K.M., and Birnie, G.D. (1996). Myc oncogenes: the enigmatic family. *Biochem. J.* **314**: 713-721.

Saar, K., Schindler, D., Wegner, R.D., Reis, A., Wienker, T.F., Hoehn, H., Joenje, H., Sperling, K., and Digweed, M. (1998). Localisation of a Fanconi anaemia gene to chromosome 9p. *Eur. J. Hum. Genet.* **6**: 501-508.

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

Saito, H., Inazawa, J., Saito, S., Kasumi, F., Koi, S., Sagae, S., Kudo, R., Saito, J., Noda, K., and Nakamura, Y. (1993). Detailed deletion mapping of chromosome 17q in ovarian and breast cancers: 2-cM region on 17q21.3 often and commonly deleted in tumours. *Cancer Res.* **53**: 3382-3385.

Sakai, K., Nagahara, H., Abe, K., and Obata, H. (1992). Loss of heterozygosity on chromosome 16 in hepatocellular carcinoma. *J. Gastr. Hepat.* **7**: 288-292.

Sakai, T., Toguchida, J., Ohtani, N., Yandell, D.W., Rapaport, J.M., and Dryja, T.P. (1991). Allele specific hypermethylation of the retinoblastoma tumour-suppressor gene. *Am. J.*

Hum. Genet. **48**: 880-888.

Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). Molecular cloning: a laboratory manual. Second Edition. Cold Spring Harbour Laboratory Press, New York.

Sasaki, M., Okamoto, M., Sato, C., Sugio, K., Soejima, J., Iwama, T., Ikeuchi, T., Tonomura, A., Miyaki, M., and Sasazuki, T. (1989). Loss of constitutional heterozygosity in colorectal tumours from patients with familial polyposis coli and those with nonpolyposis colorectal carcinoma. *Cancer Res.* **49**: 4402-4406.

Sato, T., Akiyama, F., Sakamoto, G., Kasumi, F., and Nakamura, Y. (1991*a*). Accumulation of genetic alterations and progression of primary breast cancer. *Cancer Res.* **51**: 5794-5799.

Sato, T., Saito, H., Morita, R., Koi, S., Lee, J.H., and Nakamura, Y. (1991*b*). Allelotype of human ovarian cancer. *Cancer Res.* **51**: 5118-5122.

Saurin, A.J., Borden, K.L.B., Boddy, M.N., and Freemont, P.S. (1996). Does this have a familiar RING? *Trends Biochem.* **21**: 208-214.

Savov, A., Angelicheva, D., Jordanova, A., Eigel, A., and Kalaydjieva, L. (1992). High percentage acrylamide gels improve resolution in SSCP analysis. *Nucleic Acids Res.* **20**: 6741-6742.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**: 467-470.

Schmitt, H., Kim, U.J., Slepak, T., Blin, N., Simon, M.I., and Shizuya, H. (1996). Framework for a physical map of the human 22q13 region using bacterial artificial chromosomes (BACs). *Genomics* **33**: 9-20.

Schneider, C., King, R.M., and Philipson, L. (1988). Genes specifically expressed at growth arrest of mammalian cells. *Cell* **54**: 787-793.

Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapy, G., *et al.* (1996). A gene map of the human genome. *Science* **274**: 540-546.

Schwartz, D.C., and Cantor, C.R. (1984). Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. *Cell* **37**: 67-75.

Scott, I.C., Halila, R., Jenkins, J.M., Mehan, S., Apostolou, S., Winqvist, R., Callen, D.F., Prockop, D.J., Peltonen, L., and Kadler, K.E. (1996). Molecular cloning, expression and chromosomal localisation of a human gene encoding a 33 kDa putative metallopeptidase (*PRSM1*). *Gene* **174**: 135-143.

Scully, R., Chen, J., Plug, A., Xiao, Y., Weaver, D., Feunteun, J., Ashley, T., and Livingston, D.M. (1997). Association of BRCA1 with Rad51 in mitotic cells. *Cell* **88**: 265-275.

Sealy, P.G., Whittaker, P.A., and Southern, E.M. (1985). Removal of repeated sequences from hybridisation probes. *Nucl. Acids Res.* **13**: 1905-1922.

Seizinger, B.R., Rouleau, G.A., Ozelius, L.J., Lane, A.H., Farmer, G.E., *et al.* (1988). Von Hippel-Lindau disease maps to the region of chromosome 3 associated with renal cell carcinoma. *Nature* **332**: 268-269.

Serova, O.M., Mazoyer, S., Puget, N., Dubois, V., Tonin, P., Shugart, Y.Y., Goldgar, D., Narod, S.A., Lynch, H.T., and Lenoir, G.M. (1997). Mutations in BRCA1 and BRCA2 in breast cancer families: are there more breast cancer-susceptibility genes? *Am. J. Hum. Genet.* **60**: 486-495.

Shapiro, M.B., and Senapathy, P. (1987). RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.* **15**: 7155-7174.

Sharan, S.K., Morimatsu, M., Albrecht, U., Lim, D-S., Regel, E., Dinh, C., Sands, A., Eichele, G., Hasty, P., and Bradley, A. (1997). Embryonic lethality and radiation hypersensitivity mediated by Rad51 in mice lacking *Brca2*. *Nature* **386**: 804-810.

Shattuck-Eidens, D., McClure, M., Simard, J., Labrie, F., Narod, S., *et al.* (1995). A collaborative survey of 80 mutations in the BRCA1 breast and ovarian cancer susceptibility gene: implications for presymptomatic testing and screening. *JAMA* **273**: 535-541.

Shaw, S.H., Farr, J.E., Thiel, B.A., Matise, T.C., Weissenbach, J., Chakaravarti, A., and Richard III, C.W. (1995). A radiation hybrid map of 95 STSs spanning human chromosome 13q. *Genomics* **27**: 502-510.

Shinohara, A., Ogawa, H., and Ogawa, T. (1992). Rad51 protein involved in repair and recombination in *S. cerevisiae* is a RecA-like protein. *Cell* **69**: 457-470.

Shizuya, H., Birren, B., Kim, U-J., Mancino, V., Slepak, T., Tachiiri, Y., and 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**: 8794-9797.

Simard, J., Tonin, P., Durocher, F., Morgan, K., Rommens, J., and Gingras, S. (1994). Common origins of BRCA1 mutations in Canadian breast and ovarian cancer families. *Nature Genet.* **8**: 392-398.

Skirnisdottir, S., Eiriksdottir, G., Baldursson, T., Barkardottir, R.B., Egilsson, V., and Ingvarsson, 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 (Pred. Oncol.)* **64**: 112-116.

Slamon, D.J., Godolphin, W., Jones, L.A., Holt, J.A., Wong, S.G., Keith, D.E., Levin, W.J., Stuart, S.G., Udove, J., Ullrich, A., and Press, M.F. (1989). Studies of the HER-2/neu proto-oncogene in human breast and ovarian cancer. *Science* **244**: 707-712.

Smith, C.L., and Cantor, C.R. (1989). Evolving strategies for making physical maps of mammalian chromosomes. *Genome* **31**: 1055-1058.

Smith, S., Easton, D., Evans, D., and Bonder, B. (1992). Allele losses in the region 17q12-q21 in familial breast and ovarian cancer involve the wild-type chromosome. *Nature Genet.* **2**: 128-131.

Soares, M.B., Bonaldo, M.F., Jelene, P., Su, L., Lawton, L., and Efstratiadis, A. (1994). Construction and characterisation of a normalized cDNA library. *Proc. Natl. Acad. Sci. USA* **91**: 9228-9232.

Sood, R., Blake, T., Aksentijevich, I., Wood, G., Chen, X., *et al.* (1997). Construction of a 1-Mb restriction-mapped cosmid contig containing the candidate region for the familial mediterranean fever locus (*MEFV*) on chromosome 16p13.3. *Genomics* **42**: 83-95.

Southern, E.M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **98**: 503-507.

Sparkes, R.S., Murphree, A.L., Lingua, R.W., Sparkes, M.D., Field, L.L., Funderburk, S.J., and Benedict, W.F. (1983). Gene for hereditary retinoblastoma assigned to chromosome 13 by linkage to esterase D. *Science* **219**: 971-973.

Sreekantaiah, C., Baer, M.R., Preisler, H.D., and Sandberg, A.A. (1989). Involvement of bands 9q21-q22 in five cases of acute nonlymphocytic leukemia. *Cancer Genet. Cytogenet.* **39**: 55-64.

Stallings, R.L., Torney, D.C., Hildebrand, C.E., Longmire, J.L., Deaven, L.L., Jett, J.H., Doggett, N.A., and Moyzis, R.K. (1990). Physical mapping of human chromosomes by repetitive sequence fingerprinting. *Proc. Natl. Acad. Sci. USA* **87**: 6218-6222.

Steck, P.A., Pershouse, M.A., Jasser, S., Yung, W.K.A., Lin, H., Ligon, A.H., Langford, L.A., Baumgard, M.L., Hattier, T., Davis, T., Frye, C., Hu, R., Swedlund, B., Teng, D.H.F., and Tavtigian, S.V. (1997). Identification of a candidate tumour suppressor gene, MMAC1, at chromosome 10q23.3 that is mutated in multiple advanced cancers. *Nature Genet.* **15**: 356-362.

Steinmetz, M., Stephan, D., and Lindahl, K. (1986). Gene organisation and recombinational hotspots in the murine major histocompatability complex. *Cell* **44**: 895-904.

Stewart, E.A., McKusick, K.B., Aggarwal, A., Bajorek, E., Brady, S., *et al.* (1997). An STS-based radiation hybrid map of the human genome. *Genome Research* **7**: 422-433.

Stitt, T.N., Conn, G., Gore, M., Lai, C., Bruno, J., Radziejewski, C., Mattsson, K., Fisher, J., Gies, D.R., Jones, P.F., Masiakowski, P., Ryan, T.E., Tobkes, N.J., Chen, D.H., DiStefano, P.S., Long, G.L., Basilico, C., Goldfarb, M.P., Lemke, G., Glass, D.J., and Yancopoulos, G.D. (1995). The anticoagulation factor protein S and its relative, Gas6, are ligands for the Tyro 3/Axl family of receptor tyrosine kinases. *Cell* **80**: 661-670.

Strathdee, C.A., Gavish, H., Shannon, W.R., and Buchwald, M. (1992). Cloning of cDNAs for Fanconi's anemia by functional complementation. *Nature* **356**: 763-767.

Stratton, M.R., Ford, D., Neuhausen, S., Seal, S., Wooster, R., Friedman, L.S., King, M.C., Egilsson, V., Devilee, P., McManus, R., Daly, P.A., Smyth, E., Ponder, B.A.J., Peto, J.,

Cannon-Albright, L., Easton, D.F., and Goldgar, D.E. (1994). Familial male breast cancer is not linked to the BRCA1 locus on chromosome 17q. *Nature Genet.* **7**: 103-107.

Strong, L.C., Riccardi, V.M., Ferrell, R.E., and Sparkes, R.S. (1981). Familial retinoblastoma and chromosome 13 deletion transmitted via an insertional translocation. *Science* **213**: 1501-1503.

Struewing, J.P., Brody, L.C., Erdos, M.R., Kase, R.G., Giambarresi, T.R., Smith, S.A., Collins, F.S., and Tucker, M.A. (1995). Detection of eight BRCA1 mutations in 10 breast/ovarian cancer families, including 1 family with male breast cancer. *Am. J. Hum. Genet.* **57**: 1-7.

Suggs, S.V., Wallace, R.B., Hirose, T., Kawashima, E.H., and Itakura, K. (1981). Use of synthetic oligonucleotides as hybridisation probes: isolation of cloned cDNA sequences for human beta-2-microglobin. *Proc. Natl. Acad. Sci. USA* **78**: 6613-6617.

Suthers, G.K., Hyland, V.J., Callen, D.F., Oberle, I., Rocchi, M., Thomas, N.S., Morris, C.P., Schwartz, C.E., Schmidt, M., Ropers, H.H., Baker, E., Oostra, B.A., Dahl, N., Wilson, P.J., Hopwood, J.J., and Sutherland, G.R. (1990). Physical mapping of new DNA probes near the Fragile X mutation (*FRAXA*) by using a panel of cell lines. *Am. J. Hum. Genet.* **47**: 187-195.

Suzuki, H., Komiya, A., Emi, M., Kuramochi, H., Shiraishi, T., Yatani, R., and Shimazaki, J. (1996). Three distinct commonly deleted regions of chromosome arm 16q in human primary and metastatic prostate cancers. *Genes Chrom. Cancer* **17**: 225-233.

Szabo, C.I., and King, M-C. (1995). Inherited breast and ovarian cancer. *Hum. Mol. Genet.* **4**: 1811-1817.

Tanaka, K., Yanoshita, R., Konishi, M., Oshimura, M., Maeda, Y., Mori, T., and Miyaki, M. (1993). Suppression of tumourigenicity in human colon carcinoma cells by introduction of normal chromosome 1p36 region. *Oncogene* **8**: 2253-2258.

Tanner, M.M., Tirkkonen, M., Kallioniemi, A., Collins, C., Stokke, T., Karhu, R., Kowbel, D., Shadravan, F., Hintz, M., Kuo, W.L., Waldman, F.M., Isola, J.J., Gray, J.W., and Kallioniemi, O-P. (1994). Increased copy number at 20q13 in breast cancer: defining the critical region and exclusion of candidate genes. *Cancer Res.* **54**: 4257-4260.

Tavtigian, S.V., Simard, J., Rommens, J., Couch, F., Shattuck-Eidens, D., *et al.* (1996). The complete BRCA2 gene and mutations in chromosome 13q-linked kindreds. *Nature Genet.* **12**: 333-337.

The Cystinosis Collaborative Research Group. (1995). Linkage of the gene for cystinosis to markers on the short arm of chromosome 17. *Nature Genet.* **10**: 246-248.

The FAB Consortium. (1996). Positional cloning of the Fanconi anaemia group A gene. *Nature Genet.* **14**: 324-328.

The Huntington's Disease Collaborative Research Group. (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes.

Cell **72**: 971-983.

Thomas, G.A., and Raffel, C. (1991). Loss of heterozygosity on 6q, 16q, and 17p in human central nervous system primitive neuroectodermal tumours. *Cancer Res.* **51**: 639-643.

Thompson, F., Emerson, J., Dalton, W., Yang, J-M., McGee, D., Villar, H., Knox, S., Massey, K., Weinstein, R., Bhattacharyya, A., and Trent, J. (1993). Clonal chromosome abnormalities in human breast carcinomas I. Twenty-eight cases with primary disease. *Genes Chrom. Cancer* **7**: 185-193.

Tijo, J.H., and Levin, A. (1956). The chromosome number of man. *Hereditas* **42**: 1-6.

Touchman, J.W., Bouffard, G.G., Weintraub, L.A., Idol, J.R., Wang, L., Robbins, C.M., Nussbaum, J.C., Lovett, M., and Green, E.D. (1997). 2006 expressed-sequence tags derived from human chromosome 7-enriched cDNA libraries. *Genome Res.* **7**: 281-292.

Town, M., Jean, G., Cherqui, S., Attard, M., Forestier, L., Whitmore, S.A., Callen, D.F., Gribouval, O., Broyer, M., Bates, G.P., van't Hoff, W., and Antignac, C. (1998). A novel gene encoding an integral membrane protein is mutated in nephropathic cystinosis. *Nature Genet.* **18**: 319-324.

Tribioli, C., Mancini, M., Plassart, E., Bione, S., Rivella, S., Sala, C., Torri, G., and Toniolo, D. (1994). Isolation of new genes in distal Xq28: transcriptional map and identification of a human homologue of the ARD1 N-acetyl transferase of *Saccharomyces cerevisiae*. *Hum. Mol. Genet.* **3**: 1061-1067.

Trofatter, J.A., MacCollin, M.M., Rutter, J.L., Murrell, J.R., Duyao, M.P., Parry, D.M., Eldridge, R, Kley, N., Menon, A.G., Pulaski, K., Haase, V.H., Ambrose, C.M., Munroe, D., Bove, C., Haines, J.L., Martuza, R.L., MacDonald, M.E., Seizinger, B.R., Short, M.P., Buckler, A.J., and Gusella, J.F. (1993). A novel moesin-, ezrin-, radixin-like gene is a candidate for the neurofibromatosis 2 tumour suppressor. *Cell* **72**: 791-800.

Tsuda, H., Callen, D.F., Fukutomi, T., Nakamura, Y., and Hirohashi, S. (1994). Allele loss on chromosome 16q24.2-qter occurs frequently in breast cancers irrespectively of differences in phenotype and extent of spread. *Cancer Res.* **54**: 513-517.

Tsuda, H., Hirohashi, S., Shimosato, Y., Hirota, T., Tsugane, S., Yamamoto, H., Miyajima, N., Toyoshima, K., Yamamoto, T., Yokota, J., Yoshida, T., Sakamoto, H., Terada, M., and Sugimura, T. (1989). Correlation between long-term survival in breast cancer patients and amplification of two putative oncogene-coamplification units: hst-1/int-2 and c-erbB-2/ear-1. *Cancer Res.* **49**: 3104-3108.

Tsurugi, K., and 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.

Valverde, P., Healy, E., Jackson, I., Rees, J.L., and Thody, A.J. (1995). Variants of the melanocyte stimulating hormone receptor gene are associated with red hair and fair skin in humans. *Nature Genet.* **11**: 328-330.

van Deutekom, J.C.T., Lemmers, R.J., Grewal, P.K., van Geel, M., Romberg, S., Dauwerse, H.G., Wright, T.J., Padberg, G.W., Hofker, M.H., Hewitt, J.E., and Frants, R.R. (1996). Identification of the first gene (FRG1) from the FSHD region on human chromosome 4q35. *Hum. Mol. Genet.* **5**: 581-590.

van de Vijver, M.J. (1993). Molecular genetic changes in human breast cancer. *Adv. Cancer Res.* **61**: 25-56.

Varley, J.M., Swallow, J.E., Brammar, W.J., Whittaker, J.L., and Walker, R.A. (1987). Alterations to either c-erbB-2(neu) or c-myc proto-oncogenes in breast carcinomas correlate with poor short-term prognosis. *Oncogene* **1**: 423-430.

Varnum, B.C., Young, C., Elliot, G., Garcia, A., Bartley, T.D., Fridell, Y.W., Hunt, R.W., Trail, G., Clogston, C., Toso, R.J., Yanagihara, D., Bennett, L., Sylber, M., Merewether, L.A., Tseng, A., Escobar, E., Liu, E.T., and Yamane, H.K. (1995). Axl receptor tyrosine kinase stimulated by the vitamin K-dependent protein encoded by growth-arrest-specific gene 6. *Nature* **373**: 623-626.

Venter, J.C., Smith, H.O., and Hood, L. (1996). A new strategy for genome sequencing. *Nature* **381**: 364-366.

Voglino, G., Castello, S., Silengo, L., Stefanuto, G., Friard, O., Ferrara, G., and Fessia, L. (1997). An intronic deletion in TP53 gene causes exon 6 skipping in breast cancer. *Eur. J. Cancer* **33**: 1479-1483.

Vulpe, C., Levinson, B., Whitney, S., Packman, S., and Gitschier, J. (1993). Isolation of a candidate gene for Menkes disease and evidence that it encodes a copper-transporting ATPase. *Nature Genet.* **3**: 7-13.

Wagner, M.J., Ge, Y., Siciliano, M., and Wells, D.E. (1991). A hybrid cell mapping panel for regional localisation of probes to human chromosome 8. *Genomics* **10**: 114-125.

Walter, M.A., Spillett, D.J., Thomas, P., Weissenbach, J., and Goodfellow, P.N. (1994). A method for constructing radiation hybrid maps of whole genomes. *Nature Genet.* **7**: 22-28.

Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., *et al.* (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**: 1077-1082.

Warneke-Wittstock, R., Marquardt, A., Gehrig, A., Sauer, C.G., Gessler, M., and Weber, B.H.F. (1998). Transcript map of a 900-kb genomic region in Xp22.1-p22.2: identification of 12 novel genes. Genomics **51**: 59-67.

Washington, S.S., Bowcock, A.M., Gerken, S., Matsunami, N., Lesh, D., Osborne-Lawrence, S.L., Cowell, J., Ledbetter, D.H., White, R.L., and Chakravarti, A. (1993). A somatic cell hybrid panel of human chromosome 13. Genomics **18**: 486-495.

Weber, J.L., and May, P.E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.* **44**: 388-396.

Weinberg, R.A. (1995). The retinoblastoma protein and cell cycle control. *Cell* **81**: 323-330.

Weissenbach, J. (1993). A second generation linkage map of the human genome based on highly informative microsatellite loci. *Gene* **135**: 275-278.

Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G., and Lathrop, M. (1992). A second-generation linkage map of the human genome. *Nature* **359**: 794-801.

Weissman, B.E., Saxon, P.J., Pasquale, S.R., Jones, G.R., Geiser, A.G., and Stanbridge, E.J. (1987). Introduction of a normal human chromosome 11 into a Wilms' tumour cell line controls its tumourigenic expression. *Science* **236**: 175-180.

Welcher, A.A., Suter, U., De Leon, M., Snipes, G.J., and Shooter, E.M. (1991). A myelin protein is encoded by the homologue of a growth arrest-specific gene. *Proc. Natl. Acad. Sci. USA* **88**: 7195-7199.

Whitmore, S.A., Apostolou, S., Lane, S., Nancarrow, J.K., Phillips, H.A., Richards, R.I., Sutherland, G.R., and Callen, D.F. (1994). Isolation and characterisation 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.

Whitney, M., Thayer, M., Reifsteck, C., Olson, S., Smith, L., Jakobs, P.M., Leach, R., Naylor, S., Joenje, H., and Grompe, M. (1995). Microcell mediated chromosome transfer maps the Fanconi anaemia group D gene to chromosome 3p. *Nature Genet.* **11**: 341-343.

Wijker, M., Morgan, N.V., Herterich, S., van Berkel, C.G.M., Tipping, A.J., *et al.* (1998). Heterogeneous spectrum of mutations in the Fanconi anaemia group A gene. *Eur. J. Hum. Genet.* (In Press).

Wilcox, A.S., Khan, A.S., Hopkins, J.A., and 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. *Nucleic Acids Res.* **19**: 1837-1843.

Wilkie, A.O.M., and Higgs, D.R. (1992). An unusually large $(CA)_n$ repeat in the region of divergence between subtelomeric alleles of human chromosome 16p. *Genomics* **13**: 81-88.

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., *et al.* (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378**: 789-791.

Wooster, R., Neuhausen, S.L., Mangion, J., Quirk, Y., Ford, D., *et al.* (1994). Localisation of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science* **265**: 2088-2090.

Wyman, A., and White, R. (1980). A highly polymorphic locus in human DNA. *Proc. Natl. Acad. Sci. USA* **77**: 6754-6758.

Yaspo, M-L., Gellen, L., Mott, R., Korn, B., Nizetic, D., Poustka, A., and 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.

Yoshiura, K., Kanai, Y., Ochiai, A., Shimoyama, Y., Sugimura, T., and Hirohashi, S. (1995). Silencing of the E-cadherin invasion-suppressor gene by CpG methylation in human carcinomas. *Proc. Natl. Acad. Sci. USA* **92**: 7416-7419.

Yunis, J., and Ramsay, N. (1978). Retinoblastoma and subband deletion of chromosome 13. *Am. J. Dis. Child.* **132**: 161-163.

Zhang, M.Q. (1997). Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl. Acad. Sci. USA* **94**: 565-568.

Zhang, W., Hirohashi, S., Tsuda, H., Shimosato, Y., Yokota, J., Terada, M., and Sugimura, T. (1990). Frequent loss of heterozygosity on chromosomes 16 and 4 in human hepatocellular carcinoma. *Jpn. J. Cancer Res.* **81**: 108-111.

Zucman-Rossi, J., Legoix, P., and Thomas, G. (1996). Identification of new members of the Gas2 and Ras families in the 22q12 chromosome region. *Genomics* **38**: 247-254.

Appendix

A1

Apostolou, S., Whitmore, S.A., Crawford, J., and Lennon, G., et al., (1996) Positional cloning of the Fanconi anaemia group A gene.
Nature Genetics, v. 14 (3), pp. 320-323.

A2

Ianzano, L., d'Apolito, M., Centra, M., and Savino, M., et al., (1997) The genomic organization of the Fanconi anemia group A (*FAA*) gene.
Genomics, v. 41 (3), pp. 309-314.

A3

Whitmore, S.A., Crawford, J., Apostolou, S., and Eyre, H., et al., (1998) Construction of a high-resolution physical and transcription map of chromosome 16q24.3: a region of frequent loss of heterozygosity in sporadic breast cancer.
Genomics, v. 50 (1), pp. 1-8.

A4

Whitmore, S.A., Settasatian, C., Crawford, J., and Lower, K.M., et al., (1998) Characterization and screening for mutations of the growth arrest-specific 11 (*GAS11*) and*C16orf3*genes at 16q24.3 in breast cancer.
Genomics, v. 52 (3), pp. 325-331.

A5

Town, M., Jean, G., Cherqui, S., Attard, M., and Forestier, L., et al., (1998) A novel gene encoding an integral membrane protein is mutated in nephropathic cystinosis. *Nature Genetics, v. 18 (4), pp. 319-324.*

Amendments in response to examiner I:

On page 178, it states that the *GAS11* gene was found to be expressed in mammary cells based on Northern analysis with RNA from selected breast cancer cell lines. While RNA from the normal breast epithelial cell line HBL-100 was not included on the Northern membrane, RT-PCR results on RNA from this cell line (5.2.10.3 on page 169 and 171; 5.3.9.3 on page 191) did confirm that *GAS11* is expressed in HBL-100 and therefore normal breast epithelium. For the *C16orf3* gene, RT-PCR amplification from fetal RNA was unsuccessful in all tissues examined. Given that this gene is intronless and has a small coding region, mutation analysis by SSCP was performed without prior confirmation of the genes expression in normal breast tissue.

On page 33, paragraph 1, it was stated that about 50% of late adenomas of the colon show a mutation in the *DCC* tumour suppressor gene on 18q. This should be changed to state that about 50% of late adenomas have loss of heterozygosity (LOH) at 18q, however mutations in the *DCC* gene appear quite rare. Therefore the target of 18q LOH is most likely directed at another locus.

Page 106, line 4: criterion, not criteria.

Page 156, second paragraph: *H-ras* and *K-ras*, not *h-ras* and *k-ras*.