

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

University of Alberta

**Transcriptional Mapping in a Terminal Microdeletion of Human
Chromosome 22q**

by

Chi Cheung Andrew Wong



A thesis submitted to the Faculty of Graduate Studies and Research in partial
fulfillment of the requirements for the degree of Doctor of Philosophy

in

Molecular Biology and Genetics

Department of Biological Sciences

Edmonton, Alberta

Fall 1998



National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-34859-8

Canada

University of Alberta

Library release form

Name of Author: Chi Cheung Andrew Wong

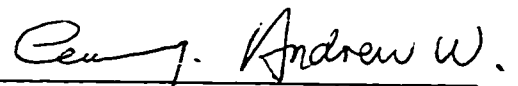
Title of Thesis: Transcriptional mapping in a terminal microdeletion of human chromosome 22q.

Degree: Doctor of Philosophy

Year of this Degree Granted: 1998

Permission is hereby granted to the University of Alberta to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly, or scientific research purpose only.

The author reserves all other publication and other rights in association with the copyright in the thesis, except as hereinbefore provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.



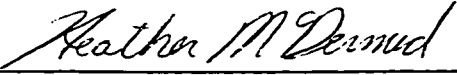
25 University Manor East
Hershey, PA 17033 USA

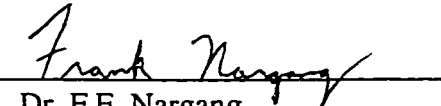
Date: 2nd October 1998

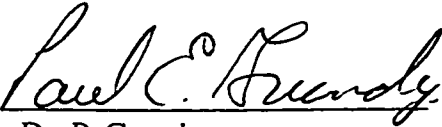
University of Alberta

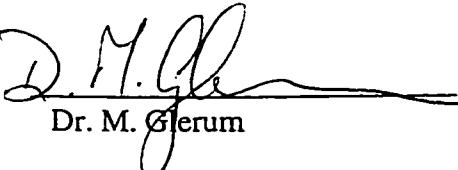
Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled "Transcriptional mapping in a terminal microdeletion of human chromosome 22q" submitted by Chi Cheung Andrew Wong in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Molecular Biology and Genetics.


Dr. H. E. McDermid (supervisor)


Dr. F.E. Nargang


Dr. P. Grundy


Dr. M. Glerum


Dr. D. Mager

Date: Oct 1 / 98

Abstract

A microdeletion at the terminal region of chromosome 22q was identified in a 12-year-old boy with mild mental retardation and delay of expressive speech. These features of patient NT show overlap with the clinical features of the cytologically visible 22q13.3 deletion syndrome. This microdeletion can therefore be used to narrow the focus of gene searches for involvement in the deletion syndrome. To clone the microdeletion region, I constructed a cosmid/P1 contig, which covers the terminal 150 kb of 22q, and encompasses the 140 kb microdeletion. The microdeletion breakpoint was mapped within a VNTR locus D22S163, and the deletion was confirmed to be terminal. Four cosmids spanning the contig were sequenced by a collaborator. I then used sequence analysis to identify potentially expressed sequences within the microdeletion region. Those expressed sequences were further confirmed by RT-PCR, cDNA library screening, and Northern blot analysis. One previously cloned gene, acrosin (ACR), was mapped between two newly identified genes. Ankyrin-like proline rich (ALPR) was mapped proximal to ACR. It has a complex structure which includes three alternatively spliced 3' ends. ALPR is similar to a predicted *C. elegans* protein C33B4.3. Both proteins contain an ankyrin repeat domain and stretches of polyproline sequences, both of which imply protein-protein interaction. Reporter gene fusion analysis shows that C33B4.3 is widely expressed in various organs, with a sexually dimorphic expression pattern in *C. elegans*. ALPR also shares extensive homology with rat cortactin binding protein 1, which has an implied function in cytoskeletal remodeling. The most distally mapped gene on 22q, RAB-like (RABL), is homologous to proteins in the RAB family whose members are involved in vesicular trafficking. RABL is duplicated on 2q13. Both loci show expression in all tissues tested. This work sets the stage for future work on studying the relationship between the deletion of the candidate genes and the phenotype of the 22q13.3 deletion/microdeletion syndrome patients.

**This thesis is dedicated to my mother, Cheung Kam Lin.
For her love, care, and belief that her son can make his dreams come true.**

Acknowledgements

First I would like to thank my supervisor, Dr. Heather McDermid, for giving me an opportunity to do my research, for her constructive criticism to keep me in a logical train of thought, and for her generous support whenever difficulties were encountered. I would like to thank my supervisory committee: Drs. Paul Grundy and Frank Nargang for their encouragement and for helping to solve some mysteries in the project. Thanks to my collaborators Dr. Jonathan Flint at Oxford University; Drs. Yi Ning and David Ledbetter at National Institute of Health; Angela Dorman, Diana Willingham and Dr. Bruce Roe at the University of Oklahoma Advanced Center for Genome Technology for providing with me useful information to make this research possible. Thanks to Dr. Dave Pilgrim for teaching me how to work with *Caenorhabditis elegans*. Also I thank all the people in Dave's lab, especially Dave Hansen who taught me various techniques in reporter gene fusion studies, and Wanyuan Ao who provided me with *C. elegans* total RNA. My sincere thanks are extended to all the people in Heather's lab. I especially thank Dr. Alan Mears for teaching me how to do densitometric analysis, Dr. Ali Riazi for teaching me various techniques such as FISH and MISH (mouse embryo whole mount *in situ* hybridization), and Dr. Valérie Trichet for collaborating in the sectioned embryo *in situ* hybridization experiments. Also I would like to express my great appreciation to Dana Shkolny, who helped in cell culturing, setting up sequencing reactions, and continuing the experiments when I left my lab bench to concentrate on my thesis writing. Angela Johnson also made direct contributions to the project by doing microinjections and maintaining the transgenic lines of *C. elegans*. Thanks to Nancy Nesslinger for helping me to get started on the project, and the following people for being great colleagues (no special order): Karen Romanyk, Kerry McTaggart, Tim Footz, Graham Banting, Polly Brinkman-Mills, Dean Mah, Jill Frizzley, Jennifer Brown, Liza de Castro, Gloria Hsi, Austin Chen, Stephanine Maier, Heather Wilson and Dr. Song Hu. Thanks to Dr. Paul Wong and his lab for allowing me to use a corner of his lab to set up my PCR reactions. I thank the great technicians in this department, especially Randy Mandryk, Barb Hepperle, Bill Clark, and Gary Ritzel. Finally, I would like to thank my Mom, my brother Frank, and my sister Shirley for their love and support.

TABLE OF CONTENT

INTRODUCTION	1
Chromosomal abnormalities	1
Telomeres: stabilizing the chromosome by capping the end	2
The Mechanisms Causing Deletions	7
<i>Unequal crossover between repetitive sequences</i>	7
<i>Terminal deletion due to breakage at a fragile site</i>	10
<i>Unbalanced translocations</i>	16
Molecular Aetiology of Deletions	17
<i>Haploinsufficiency</i>	17
<i>Segmental aneusomy syndromes (SAS)</i>	22
Phenotypes Associated with Deletions at the Ends of Chromosomes	25
Delineating the Chromosomal Region Involved in Terminal Deletions	26
Microdeletions as a Tool to Study Deletion Syndromes	28
Unknown Cases of Multiple Congenital Anomalies/Mental Retardation	29
Interstitial/Terminal Deletions of Chromosome 22q	30
<i>DiGeorge syndrome (DGS)</i>	30
<i>22q13.3 deletion syndrome</i>	33
Use of a Microdeletion to Study a Subset of Features in 22q13.3 Deletion Syndrome	33
Research Summary	34
MATERIALS AND METHODS	36
Cell Lines	36
Probes	36
DNA Studies	38
<i>Genomic DNA extraction</i>	38
<i>Plasmid and cosmid DNA preparations</i>	38
<i>Southern blot hybridization</i>	39
Dosage Analysis	40
BAL31 Analysis	40
<i>Embedding genomic DNA in agarose</i>	40
<i>BAL 31 exonuclease digestion</i>	41
Sequence annotations on the cosmid contig	41
Sequence Analysis	43
<i>Multiple Sequences alignment</i>	43
<i>Determining the G-C content of exons predicted by genscan program</i>	43
PCR and RT-PCR	43
<i>Total RNA isolation for reverse transcription</i>	45
<i>Reverse transcription</i>	45
<i>PCR</i>	47
<i>Subcloning PCR products</i>	48
cDNA Library Screening	48
<i>Probes and cDNA libraries</i>	48
<i>In vivo excision of insert into pBluescript SK+ phagemid</i>	50
Confirmation of the Authenticity of cDNA Clones	50
Sequencing of cDNA clones	51
Northern Analysis	51
Comparison of RABL2 and RABL22 Expression Level	53
3' Rapid Amplification of cDNA Ends (RACE)	53
Gene Expression Studies in <i>C. elegans</i>	54
<i>Preparation of reporter gene fusion protein constructs</i>	54
<i>Examination of fusion protein expression pattern</i>	55

<i>Integration of Extrachromosomal Arrays using UV irradiation</i>	56
Inhibition of Specific Gene Transcription by Double-stranded RNA Interference.....	56
RESULTS	58
Production of a Cosmid Contig from D22S163 to the 22q Telomere.....	58
Localization of the Breakpoint of the NT Microdeletion	62
BAL31 Analysis.....	64
Localization of ACR to the NT Microdeletion	64
Sequence Analysis of the NT microdeletion region	68
<i>Distribution of interspersed repeat sequences in the microdeletion region</i>	68
<i>Blast search for the microdeletion region sequences that match the entries in public databases</i>	69
<i>Putative Expressed sequences determined by exon prediction programs</i>	73
Ankyrin-Like Proline Rich (ALPR) gene	74
<i>Cloning of ALPR</i>	74
<i>Northern analysis of ALPR</i>	88
<i>Construction of an ALPR cDNA contig</i>	89
<i>Determining the G-C content of putative ALPR exons</i>	90
<i>Homology between ALPR and other proteins in the public databases</i>	90
Characterization of C33B4.3, the <i>C. elegans</i> homolog of ALPR.....	105
<i>Life cycle and basic anatomy of C. elegans</i>	105
<i>Confirmation of the polyproline sequences</i>	106
<i>Reporter gene fusion studies of C33B4.3</i>	106
<i>Double stranded RNA interference study</i>	116
<i>Identification of the RABL locus</i>	119
<i>Duplication of the RABL locus on chromosome 2q and 22q</i>	119
<i>Sequence analysis of RABL</i>	121
<i>Comparison of the relative expression level between RABL2 and RABL22</i>	125
<i>Northern analysis of the RABL gene</i>	127
<i>Cloning the 1.4 kb alternative transcript of RABL by 3' RACE</i>	127
DISCUSSION	130
Cloning of the NT Microdeletion region.....	130
Genomic Organization of the NT Microdeletion Region	134
Transcription Mapping within the NT Microdeletion	139
1. <i>Localization of the acrosin (ACR) gene</i>	139
2. a) <i>Structure of ALPR</i>	140
2 b) <i>Possible function(s) of the ALPR protein</i>	150
2 c) <i>Functional analysis of the C33B4.3 in C. elegans</i>	156
3 a) <i>Structure of RABL</i>	164
3 b) <i>Possible function of RABL</i>	164
3c) <i>Gene duplication in RABL</i>	169
4. <i>Relevance to NT microdeletion</i>	170
Future Research.....	171
<i>Construction of a full ALPR cDNA contig</i>	171
<i>Functional analysis of the mammalian ALPR gene</i>	172
<i>Determine the basic function of ALPR using C. elegans</i>	173
REFERENCES.....	174
APPENDIX : SEQUENCE ANNOTATIONS ON AWCONTIG	202

LIST OF TABLES

Number	Title	Page
Table 1	Terminal deletions in different chromosomes arms.....	11
Table 2	Genetic disease involved in the haploinsufficiency of genes.....	19
Table 3	PCR and RT-PCR products amplified in this study.....	37
Table 4	Primer sequences and their positions on AWcontig.....	44
Table 5	Primer sequences and their positions on <i>C. elegans</i> cosmid C33B4.3.....	46
Table 6	cDNA libraries and probes used for screening the libraries.....	49
Table 7	Subclones of cDNA clones and internal primers used for manual sequencing.....	52

LIST OF FIGURES

Number	Title	Page
Figure 1	Chromosomal Bouquet Formation.....	6
Figure 2	Deletion as a result of unequal crossover.....	9
Figure 3	Cosmid/P1 contig spanning the NT microdeletion.....	59
Figure 4	RFLP analysis showing Xh1.2 is deleted in NT.....	61
Figure 5	Rearrangement bands detected by the D22S163 probe.....	63
Figure 6	Sau 3AI-digested NT family genomic DNA hybridized to D22S163 probe under two different electrophoretic condition.....	65
Figure 7	BAL 31 sensitivity of rearrangement band in NT.....	66
Figure 8	Densitometric analysis of ACR.....	67
Figure 9	Genomic organization within the NT microdeletion region.....	70
Figure 10	Genomic organization of the ALPR gene.....	75
Figure 11	Nucleotide sequence and putative open reading frame of cDNA clone sc24.....	78
Figure 12	Nucleotide sequence and putative open reading frame of cDNA clone I511.....	79
Figure 13	Nucleotide sequence of cDNA clone FLS.....	80
Figure 14	Northern analysis of ALPR.....	81
Figure 15	Nucleotide sequence and putative open reading frame of cDNA clone fli.....	84
Figure 16	Nucleotide sequence and putative open reading frame of cDNA clone FL2.....	85
Figure 17	Multiple nucleotide sequence alignment of FL2 against its rat and mouse homolog.....	86
Figure 18	The illegitimate RT-PCR product R1F-M1314R.....	91
Figure 19	Determination of the G-C content of putative ALPR exons.....	92
Figure 20	Multiple alignment of proteins that share homology with ALPR..	95
Figure 21	Protein identities between ALPR and its homologous proteins....	104
Figure 22	The amino acid sequence of the <i>C. elegans</i> protein C33B4.3.....	107
Figure 23	C33B4.3 expression in the two fold embryo stage of <i>C. elegans</i> ..	111
Figure 24	C33B4.3 expression in the three fold embryo stage of <i>C.</i> <i>elegans</i>	112
Figure 25	C33B4.3 expression in the L1 stage of <i>C. elegans</i>	113
Figure 26	C33B4.3 expression in the whole adult hermaphrodite of <i>C.</i> <i>elegans</i>	114

Figure 27	C33B4.3 expression in the hermaphrodite tail of <i>C. elegans</i>	115
Figure 28	C33B4.3 expression in the male of <i>C. elegans</i>	117
Figure 29	C33B4.3 expression in the ventral nerve cord of <i>C. elegans</i>	118
Figure 30	Partial monochromosomal hybrid panel probed with RABL cDNA.....	120
Figure 31	Genomic organization of the RABL gene on chromosome 22.....	122
Figure 32	The cDNA sequence of RABL2 and RABL22.....	123
Figure 33	Comparison of the relative expression level between RABL2 and RABL22.....	126
Figure 34	Northern blot analysis of RABL.....	128
Figure 35	Comparison between sequence at the NT microdeletion breakpoint and a normal chromosome.....	132
Figure 36	The genomic organization of the subtelomeric region in yeast, human, and human chromosome 22.....	136
Figure 37	Alternative splicing to produce premature translation termination and in frame translation of the H55337 exon.....	146
Figure 38	Two hypotheses for reverse transcription with an RNA template that contains secondary structure.....	149
Figure 39	lin-3/let-23 signal transduction pathway in vulva development in the hermaphrodite and spicule formation in the male <i>C. elegans</i> ..	160
Figure 40	The cycling of a RAB protein in vesicle transportation.....	166

Abbreviations

ACR	Acrosin
ALPR	Ankyrin-like, proline rich
ARSA	Arylsulfatase A
AS	Angelman syndrome
BAC	Bacterial artificial chromosome
BCM	Baylor College of Medicine
Beauty	Blast enhanced alignment utility
Blast	Basic local alignment search tool
CATCH22	Cardiac Abnormality, Abnormal faces, T cell deficit due to Thymic hypoplasia. Cleft palate. Hypocalcemia due to hypoparathyroidism resulting from 22q11 deletion
CBP	CREB binding protein
CLTCL	Clathrin heavy chain-like
CMT1A	Charot-Marie Tooth disease type 1A
CTAB	Mixed alkyltrimethylammonium bromide
CTP	Citrate transport protein
DAPI	4,6-diamidino-2-phenylindole
DDBJ	DNA Data Bank of Japan
DGS	DiGeorge syndrome
DRD4	dopamine D4 receptor
DTAB	Dodecyl-trimethyl-ammonium-bromide
EDTA	Disodium ethylene diamine tetraacetate
EGF	Epidermal growth factor
EGTA	Ethylene glycol-bis[β -Aminoethyl ether] N, N, N', N'-tetraacetic acid
ELN	Elastin
EMBL	European Molecular Biology Laboratory
EPH	A cell receptor protein – tyrosine kinase that overexpresses in erythropoietin – producing human hepatocellular carcinoma cell line (ETL-1)
EST	Expressed Sequence Tag
FB	Designation of a deletion syndrome patient
FBS	Fetal bovine serum
FISH	Fluorescence <i>in situ</i> hybridization
FSHD	Facioscapulohumeral muscular dystrophy
FZD3	The human homologue to the Drosophila fizzled gene
GABP	GA-binding protein
GAP	GTPase-activating protein
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
GCG	Genetics computer group
GDI	GDP-dissociation inhibitor
GDP	Guanine diphosphate
GEF	Guanine nucleotide exchange factor
GSCL	Gooseoid-like

GTP	Guanine triphosphate
HBSS	Hanks' balanced salt solution
HNPP	Hereditary neuropathy with liability to pressure palsies
IPW	Imprinted in Prader-Willi
kb	Kilobase
L1MK1	L1M kinase-1
LB	Luria-Bertani medium
LINE	Long interspersed nuclear element
LTR	Long terminal repeat
MaLR	Mammalian LTR-retrotransposon
Mb	Megabase
MBP	Myelin basic protein
MIR	Mammalian-wide interspersed repeat
MITE	Mariner insect transposon-like element
MRI	Magnetic resonance imaging
MRX	Non-specific X-linked mental retardation
Myr	Million year
NLS	Nuclear localization signal
NT	Designation of the patient who has the 22q terminal microdeletion
ORF	Open reading frame
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PDGF	Platelet-derived growth factor
PFGE	Pulsed field gel electrophoresis
PMSF	Phnazine methosulfate
PPS	Peripheral pulmonic stenosis
PWS	Prader-Willi syndrome
RABL	RAB-like
RACE	Rapid amplification of cDNA ends
REP-1	RAB escort protein 1
RFC2	Replication factor C subunit 2
RFLP	Restriction fragment length polymorphism
RIEG1	Rieger syndrome type I
RT-PCR	Reverse transcription- polymerase chain reaction
SAS	Segmental aneusomy syndrome
scRNA	Small cytoplasmic RNA
SDS	Sodium dodecyl sulfate
SINE	Short interspersed nuclear element
SNRPN	Small nuclear ribonucleoprotein polypeptide N
SSC	0.15 M NaCl, 0.015 M Na citrate
SSPE	0.15 M NaCl, 0.01 M NaH ₂ PO ₄ , 0.01 M EDTA-Na ₂
STEP	Striatum-enriched phosphatase
STX1A	Syntaxin 1A gene
SVAS	Supravalvular aortic stenosis
TAE	40 mM Tris pH7.2, 20 mM sodium acetate, 1 mM EDTA-Na ₂
TBE	0.089 M Tris, 0.089M Boric acid, 0.002M EDTA, pH 8.4

TDF	Testis determining factor
TE	10 mM Tris pH8.0, 1 mM EDTA
TNE	10 mM Tris pH8.0, 10mM NaCl, 2 mM EDTA
TOP3	Topoisomerase III
UTR	Untranslated region
UV	Ultraviolet
VCFS	Velocardiofacial syndrome
VNTR	Variable number of tandem repeats
WSCP	Williams syndrome cognitive profile

Introduction

Chromosomal abnormalities

Human and other primates are unusual among mammals – they are thought to lose a high percentage of conceptions, usually before these pregnancies are clinically recognizable (McFadden and Friedmann 1997). This is often due to chromosome abnormalities. Therefore, chromosomal abnormalities are an important cause of infertility and miscarriage in humans. Studies have shown various forms of chromosomal abnormalities in 1.3% of couples seen at infertility clinics (Gueneri et al. 1987), 20-50% of human conceptions (McFadden and Friedmann 1997), and at least 6% of spontaneous abortions of clinically recognized pregnancies (Stern et al. 1996). Chromosomal abnormalities are also a significant source of congenital anomalies and mental retardation in newborns. One study showed that approximately 0.9% of newborn infants have some forms of chromosomal abnormalities which often lead to congenital anomalies (Jacobs et al. 1992).

There are three major types of chromosomal abnormalities: polyploidy, aneuploidy and structural abnormalities. In polyploidy, the nucleus has a multiple of the basic haploid chromosome number other than two. Polyploid embryos are usually lost early in pregnancy. Aneuploidy refers to the gain or loss of one or more chromosomes, which in most cases also leads to spontaneous abortion. However, viable offspring are often produced through structural chromosomal abnormalities, which involve various chromosomal rearrangements such as duplications, deletions, or translocations. Most chromosomal rearrangements involve at least two breaks in one or more chromosomes, and the two breakpoints usually fuse together to form an aberrant chromosome. However, this rule does not apply to terminal deletions, which only require one break in the chromosome. A specific terminal deletion is the topic of this thesis.

Telomeres: stabilizing the chromosome by capping the end

Terminal deletions remove the end of a chromosome arm. The broken end is then highly unstable. This phenomenon has been studied for many years in many organisms. For example, McClintock (1938) showed that a broken chromosome in maize fused with its sister chromatid after replication to form a dicentric chromosome, which broke again at mitotic anaphase. A “break and fuse cycle” was then repeated during the next meiosis and mitosis.

In most organisms (except some insect species in the order *Diptera* and plants in the genus *Allium*), the normal ends of linear chromosomes in the nucleus are stabilized by a G-rich repeat sequence called a telomere (Biessmann and Mason 1997), which may also play a role in stabilizing broken chromosome ends. In humans (Moyzis et al. 1988) and mice (Kipling and Cook 1990; Starling et al. 1990) the telomeric sequence is the canonical repeat (TTAGGG)_n. This repeat sequence is maintained by a mechanism other than conventional semi-conservative DNA replication. There is a commonly described “end replication” problem for linear DNA replication. During conventional DNA replication, DNA polymerase initiates elongation of a new strand from an RNA primer at the lagging strand. The new strand extends from 5' to 3'. As a result, the sequence at the start of the elongation to the 5' end of the lagging strand is fully replicated. However, the sequence where the RNA primer is removed from the lagging strand is not replicated, leaving a 3' overhang of the lagging strand. If there is no other DNA replication mechanism to add sequences at the 3' end, the linear DNA will shorten after each round of DNA replication (Linger et al. 1995). The solution to this problem was explained by the discovery of telomerase activity. Telomerase is a ribonucleoprotein. It contains an RNA template as well as two protein subunits, p80 and p95 (Collins et al. 1995a). Recently the RNA template (Feng et al. 1996) and the p80 subunit (Kilian et al. 1997; Meyerson et al. 1997; Nakayama et al. 1997) of human telomerase have been cloned. The RNA template is a complementary sequence of 1.5 telomeric repeats. It provides a template for the addition of the telomeric repeat at the end of the chromosomes in the

presence of the p80 and p95 subunits, which presumably are the catalytic unit of a reverse transcriptase and telomeric sequence binding protein, respectively (Collins et al. 1995a). Telomerase can also add telomeric repeats directly to a broken end, a mechanism which is known as telomere healing. Broken chromosomes can also be stabilized by telomere capture through non-homologous recombination. These two mechanisms involve two distinct properties of telomeres; namely, the maintenance of telomere length by telomerase and the dynamic structure of the repeats that are close to the telomere.

Telomere healing has been well characterized in the ciliated protozoa. In *Paramecium* or *Tetrahymena*, telomere healing is found in their macronucleus during normal development (Barion et al. 1987, Forney and Blackburn 1988, Spangler et al. 1988). For example, the unicellular organism *Tetrahymena thermophila* contains of two nuclei, the macronucleus and micronucleus. The diploid micronucleus is transcriptionally silent, and undergoes meiosis in the sexual process of conjugation. The macronucleus, on the other hand, is transcriptionally active and undergoes programmed chromosome fragmentation, resulting in 200 different linear chromosomes from the fragmentation of 5 pairs of chromosomes (Forney and Blackburn 1988). The broken ends of the small chromosomes are healed by *de novo* addition of telomere sequence. Telomere healing is also involved in stabilizing broken chromosomes in other organisms. In humans, a total of seven terminal deletion breakpoints have been cloned. In all cases the tip of chromosome 16p was deleted, which caused hemizyosity of the α -globin genes and resulted in α -thalassaemia (Wilkie et al. 1992a, Lamb et al. 1993, Flint et al. 1994). In addition to the human telomeric repeats (TTAGGG)_n, the only common feature between these breakpoints was the presence of a few nucleotides that were similar to the telomeric sequence prior to the added telomeric repeats. These results suggested that the telomerase recognized telomere-like sequence at the breakpoints, and added telomeric repeats to them. The sequence requirement for the telomerase to recognize a broken end is unclear. When studying telomere healing in *Tetrahymena*, Harrington and Greider (1991) showed that sequence that is complementary to the telomerase RNA template is not required, if the breakpoint is close to two internal telomeric repeats. Other *in vitro* studies of humans (Morin 1991; Murnane and Yu 1993) showed that two nucleotides are the minimal complementary sequence required for the recognition of the breakpoint by telomerase.

Lamb et al. (1993) hypothesized that when a chromosome is broken, exonucleases degrade the end of the broken chromosome until a sequence that can be recognized by telomerase is reached. Although exonuclease activity has been found, it is specific for the C rich strand of the telomere (Wellinger et al. 1996; Makarov et al. 1997). There are five possible combinations for a two-nucleotide sequence to match the telomeric repeat sequence (TT, TA, AG, GG, and GT in the TTAGGG sequence), and sixteen different combinations of bases for a two-nucleotide sequence ($2^4 = 16$). If two nucleotides are the minimal complementary sequence for telomerase recognition, there is at least a 31.25% chance that the telomerase can recognize the broken end ($5/16 = 0.3125$). Other than telomere healing, there are at least two possible alternative fates for broken chromosomes. The broken chromosome can either cause checkpoint-mediated cell cycle arrest, as it is shown in yeast (Bennett et al. 1993; Sandell and Zakian 1993) and mammalian cell lines (Lock and Ross 1990), or it can be stabilized by telomere capture that involves non-homologous recombination.

The mechanism of telomere capture relates to the genomic organization of the telomeric region. The sequence that is adjacent to the telomeric repeats, (TTAGGG)_n, is called the telomere associated sequence. In yeast (Pryde and Louis 1997), *Pneumocystis carinii* (Underwood et al. 1996), *Plasmodium falciparum* (Pace et al. 1995), *Trypanosoma brucei* (Eid and Sollnerwebb. 1995), plants (Fajkus et al. 1995), or even the protoeukaryotic cell *Giardia duodenalis* (Upcroft et al. 1997), similar telomere associated sequences are found associated with different telomeres. For example, 0 – 4 tandem repeats of Y' elements, which are either 5.2 kb or 6.7 kb in size and contain 2 overlapping ORFs, found associated in some yeast telomeres (Pryde and Louis 1997). In humans, three different types of telomere associated sequences called TelBam3.4, TelBam8, and TelBam11 have been mapped to different telomeres (Brown et al. 1990). These TelBam sequences are variably associated with three different length polymorphisms at the subtelomeric region of 16p. For example, TelBam3.4 is associated with "A" and "C" length alleles, whereas TelBam11 is associated with the "B" allele. B and C alleles are 180 kb and 260 kb longer than A, respectively (Wilkie et al. 1991). Furthermore, these polymorphic alleles are found at telomeres of other chromosomes. The A allele maps to Xq and Yq, whereas the B and C alleles are on 9q, 10p, and 18p.

Sequences of the same length allele in different chromosomes are similar, yet they could be distinguished by restriction mapping analysis (Wilkie et al. 1991). This suggests that the subtelomeric region is dynamic in nature, where chromosome ends are exchanged between non-homologous chromosomes. The subnuclear organization of chromosomes during meiosis provides the possible structure for such non-homologous chromosome exchanges. During interphase, human telomeres are associated with the nuclear matrix (Luderus et al. 1996). Starting at late preleptotene, the telomeres are moved to the nuclear envelope and clustered to form a structure known as chromosomal bouquet, which is not resolved until pachytene (Scherthan et al. 1996, Fig.1). A similar structure to the chromosomal bouquet has been extensively studied in the fission yeast *Schizosaccharomyces pombe* (Nimmo et al. 1998; Cooper et al. 1998). During meiotic prophase I, the nucleus of *S. pombe* elongates and forms a structure known as a horsetail (de Lange 1998). Like the chromosomal bouquet, the telomeres in *S. pombe* are clustered in the horsetail and associated with the spindle pole body. Centromeres then move back and forth through the cell. This movement is believed to “straighten up” the chromosomes and facilitate the search for homology between homologous chromosomes. The clustering of the telomeres at the spindle polar body is mediated by a telomere sequence binding protein Taz1 (Cooper et al. 1998), which also regulates the telomere length (Shore 1997). A defect in telomere clustering caused by a Taz1 mutation results in chromosome missegregation, reduction in recombination and low spore viability (Nimmo et al. 1998). The chromosomal bouquet in humans presumably serves the same purpose as the horsetail cluster in yeast. Because of the telomere clustering, exchanges between the telomeres of non-homologous chromosomes may also be facilitated. Such non-homologous chromosome exchanges have been found in humans. Other than the polymorphic alleles of the telomere associated sequences found in 16p and other chromosomes (Wilkie et al. 1991), sequence exchange has also been found between 4qter and 10qter in 20% of the population. The non-homologous chromosomal exchange region found in 4q is associated with the autosomal dominant myopathy disorder,

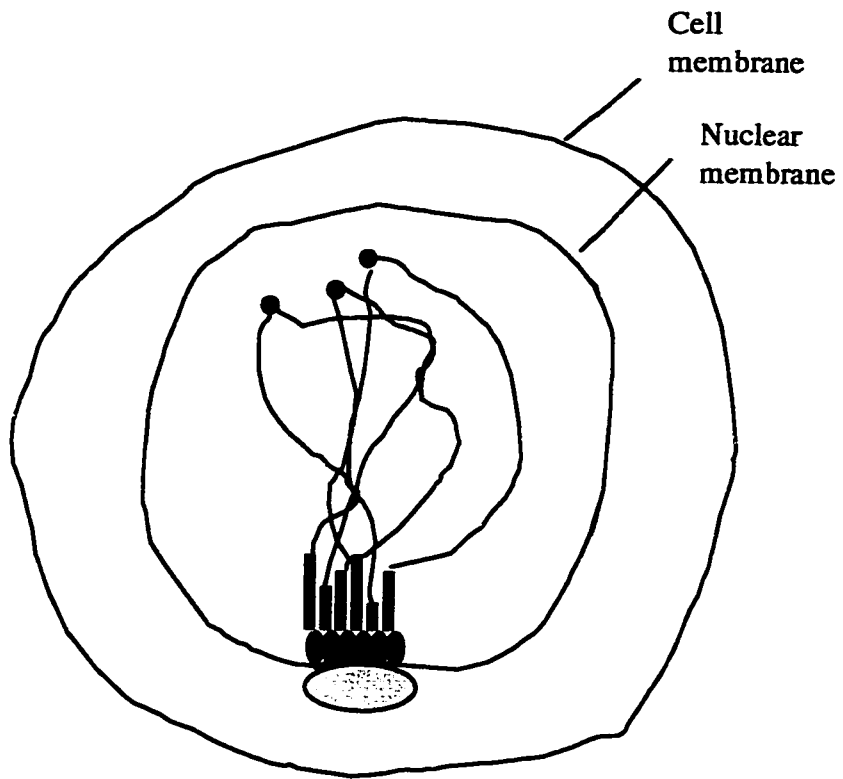


Figure 1 Chromosomal Bouquet formation. During late preleptotene, the telomeres are moved to the nuclear envelope and clustered at the centrosome (blue oval). Telomeric repeats (TTAGGG)_n (black ovals) attach to the nuclear envelope. The subtelomeric repeat region (blue boxes) from non-homologous chromosomes are close to each other, facilitating the exchange of non-homologous chromosomes. The chromosomes (thin lines with green circles as centromeres) start searching for their homologous pair (Redrawn after Scherthan et al. 1996; de Lange 1998).

facioscapulohumeral muscular dystrophy (FSHD), and genotyping of families for this disease lead to the discovery of the exchange (van Deutekom et al. 1996). Non-homologous chromosome exchanges between chromosome ends have also been found in the subtelomeric repeat region of yeast. This region lacks DNase I hypersensitive sites (Louis et al. 1994), where double strand breaks are initiated, resulting in reciprocal sister chromatid exchange during meiosis (Wu and Lichten 1995). The lack of these sites suggests that the non-homologous chromosome exchanges are not mediated by conventional homologous recombination. Therefore, the recombination between non-homologous chromosomes at the subtelomeric region may not produce reciprocal products. When a chromosome is truncated, the broken end will be brought close to the subtelomeric regions of other chromosomes in the chromosomal bouquet. Since full homology between the sequences is not required for this recombination, any repetitive sequences such as an Alu element will initiate the recombination with another non-homologous chromosome end. The deleted chromosome will then acquire subtelomeric sequences from a non-homologous chromosome. Such chromosome stabilization has been found at 16p (Flint et al. 1996). It is also implicated in one 16p- case where the breakpoint is within an Alu element and no complementary sequence to the telomerase RNA template is found adjacent to the telomeric repeat (Flint et al. 1994).

The Mechanisms Causing Deletions

Unequal crossover between repetitive sequences

The mechanism involved in chromosome deletion is not completely understood. However, there are well documented cases of interstitial deletions caused by unequal crossover between repetitive sequences. One well known example is the Charcot-Marie Tooth disease type 1A (CMT1A)/hereditary neuropathy with liability to pressure palsies (HNPP) region, which maps to a 1.5 Mb region on chromosome 17p11.2-p12. Both duplications and deletions have been found in this region, resulting in CMT1A and HNPP respectively. The duplication/deletion region is flanked by two large CMT1A-REP repeats. The duplication/deletion is believed to be a pair of reciprocal unequal cross-over products caused by the misalignment of the proximal and distal CMT1A-REP repeats

during meiosis (Fig. 2A) (Reiter et al. 1996). Within the CMT1A-REP there is a hot spot for recombination where the recombination rate in patients is 53 times higher than that of the normal individuals (Reiter et al. 1996). Close to the recombination hot spot there is a *Mariner* insect transposon-like element (MITE). *Mariner* is a class II transposable element originally isolated from *Drosophila mauritiana* (Robertson 1993). This type of DNA transposon moves from one location in a genome to another by excision and reintegration (Morgan 1995). Although the MITE element within CMT1A-REP does not contain any functional transposase gene, the double strand break at the recombination hot spot could be mediated by an endogenous transposase. The double strand break then promotes the recombination that results in an unequal crossover event (Reiter et al. 1996).

A similar example that involves repetitive elements flanking an interstitial deletion region is found in Williams syndrome. In a study to delineate the deletion region of this syndrome at 7q11.2, Robinson et al. (1996) identified three loci of the microsatellite marker D7S489 in the region. Two loci (D7S489L and D7S489M) flank the deletion on either side, whereas one locus (D7S489U) is mapped between D7S489L and D7S489M and is deleted occasionally (Fig. 2B). These loci are expressed, which is evident by their matches to partially sequenced cDNAs ("EST"s) in the Genbank sequence database (Robinson et al. 1996). The deletion in Williams syndrome is likely caused by misalignment between D7S489L and D7S489M, or D7S489U and D7S489L during meiosis. Like the CMT1A/HNPP region, the recombination rate within the critical region of Williams syndrome patients is elevated compared to normal individuals (Dutly et al. 1998), further supporting the hypothesis that the deletion is caused by the unequal crossover events. However, unlike the CMT1A/HNPP syndromes, there is no disorder associated with the duplication of the Williams syndrome region. Robinson et al. (1996) suggested that the increased dosage of genes within the Williams syndrome region may be lethal. Alternatively, the increased dosage may not produce any phenotypic effect and therefore has not been detected.

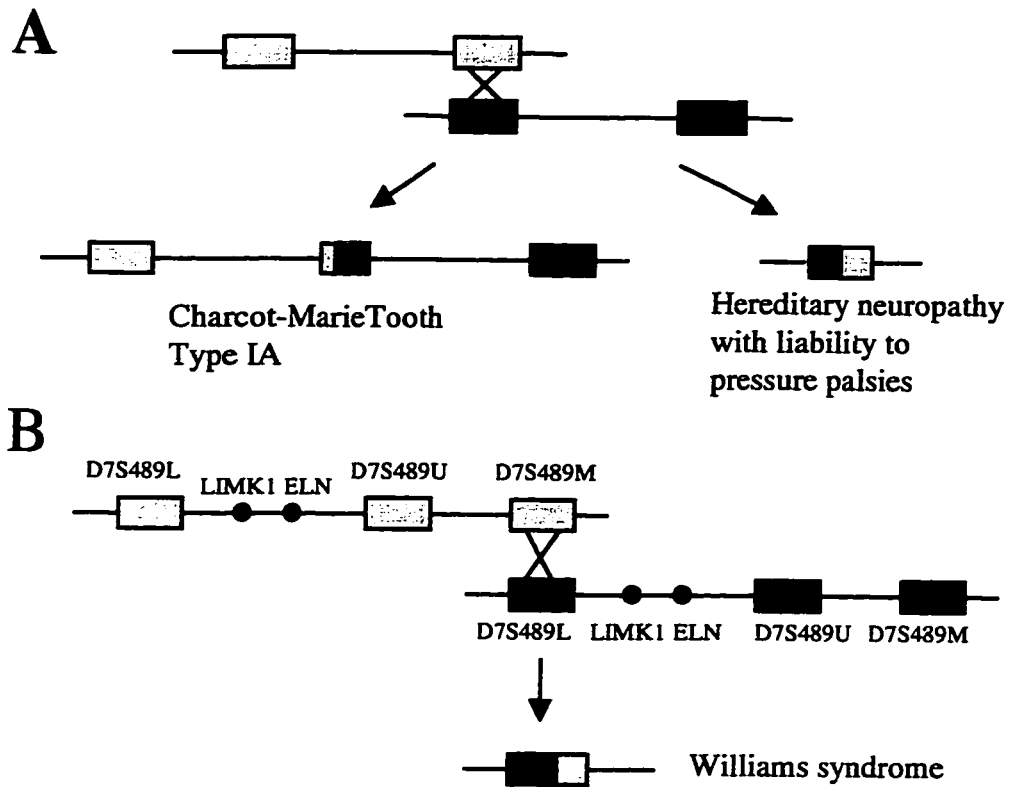


Figure 2 Deletion as a result of unequal cross-over in (A) hereditary neuropathy with liability to pressure palsies. The reciprocal product results in the duplication disease Charcot-Marie Tooth Type 1A. Each box represents a CMT1A-REP repeat. (B) Williams syndrome. The three copies of the D7S489 locus flank the breakpoint of the microdeletion. The reciprocal product has not been found.

While it is likely that unequal crossover causes other interstitial deletions, it is unlikely to be the mechanism of terminal deletions, which only involve one break in the chromosomes.

Terminal deletion due to breakage at a fragile site

Very little is known about how one break occurs in a chromosome causing a terminal deletion, however many such deletions are known (Table 1). Although there are cases of inherited terminal deletions (Wilkie et al. 1990b, Descartes et al. 1996), most cases arise *de novo*. However, some reports show that deletion breakpoints in the Jacobsen syndrome region have been mapped close to a fragile site (Voullaire et al. 1987; Jones et al. 1994). Jacobsen syndrome is associated with the terminal deletion of 11q (Michaelis et al. 1998). Voullaire et al. (1987) identified a Jacobsen syndrome patient who also inherited a fragile site on 11q (FRA11B) from his mother. Fragile sites, detected cytogenetically when cells are grown under special chemical treatments, appear as a constriction in the chromosome *in vitro* and presumably render it susceptible to breakage *in vivo* (Jones et al. 1994). By positional cloning, Jones et al. (1994) mapped an ~100 kb region that contains the fragile site. They also mapped the breakpoint of the Jacobsen syndrome patient from Voullaire et al. (1987) within the same 100 kb region. Although they showed that the tip of the truncated chromosome contained telomeric repeats by FISH, they did not show whether the chromosome also contained the subtelomeric region. Therefore, it was not confirmed whether the deletion is terminal or interstitial. Although both the fragile site and the breakpoint were within a 100 kb region, the distance between the fragile site and the breakpoint was not determined. As a result, it is not yet possible to establish a causal relationship between the presence of the 11q fragile site and the Jacobsen syndrome breakpoint.

Jones et al. (1995) further examined three different families who showed the transmission of the FRA11B fragile site as well as Jacobsen syndrome. They demonstrated that the fragile site was located at the 5' end of the CBL2 proto-oncogene.

Table 1 Terminal deletions in different chromosome arms

Chromosome arm	Cytogenetic position involved	Phenotype	References
1p	1p36.22->pter	Large anterior fontanelles, low set ears, hypotonia, developmental delay, cardiomyopathy, hydrocephalus.	Keppler-Noreuil et al. (1995)
1q	1q43-> qter	Severe psychomotor developmental delay, generalized hypotonia, seizures, autistic-like behavior.	Murayama et al. (1991)
2p		No case reported ¹	
2q	2q37->qter	Low birth weight, growth retardation, developmental delay, macrocephaly, large forehead, upslanting palpebral fissures, lower-set ears with small canals, low-set nipples, cardiac defect, genital abnormalities, syndactyly/clinodactyly, simian creases and seizures, Wilm's tumor.	Conard et al. (1995) Wenger et al. (1997) Viot-Szoboszlai et al. (1998)
3p	3p25->pter	Mental retardation, growth retardation, microcephaly, ptosis, micrognathia.	Phipps et al. (1994)
3q	3q37->qter	Non-specific developmental delay, growth retardation, hypotonia, ear abnormalities, bilateral microphthalmia or anophthalmia (2 cases)	Chitayat et al. (1996)
4p	4p16.3->pter	Wolf-Hirschhorn syndrome: severe growth deficiency, mental retardation with onset of convulsions by the 2nd year of life, microcephaly, sacral dimples, facial features, midline defects, flexion/contracture deformities of hands and feet, heart defects, hemangiomas, hypoplastic nipples, eye defects.	Altherr et al. (1997)
4q	4q33->qter 4q34.2->qter	Micrognathia, abnormalities of the fingers, mental retardation, flat nasal bridge. Upper lip abnormality, narrow high arched palate, short upturned nose, high nasal bridge, low set ear.	Menko et al. (1992) Descartes et al. (1996)
5p	5p15->pter	cri-du-chat syndrome: Microcephaly, round face, hypertension, micrognathia, prominent nasal bridge, epicanthal folds, hypotonia, severe psychomotor and mental retardation, high-pitched cry.	Overhauser et al. (1994)

Table 1 (continued.....1)

Chromosome arm	Cytogenetic position involved	Phenotype	References
5q	5q35.1->qter	oral, facial, digital anomalies.	Kleckowska (1993)
6p	6p23->pter	Psychomotor and growth retardation, ventricular septal defect, patent ductus arteriosus, facial features.	Plaja et al. (1994)
6q	Proximal breakpoint from 6q21 to 6q25.3	Microcephaly; broad, prominent nasal bridge, long philtrum, micrognathia or retrognathia, short neck, hypotonia, congenital heart defect, seizures, mental retardation, growth deficiency	Meng et al. (1992)
7p	7p15 (some interstitial cases)	Craniosynostosis.	McPherson et al. (1976)
7q	7q32->qter	Low birth weight, pre/postnatal growth and developmental retardation, microcephaly, eye anomalies, flat/broad nasal bridge with bulbous nasal tip, male genital abnormalities, abnormal palm/sole creases, holoprosencephaly.	Frints et al. (1998)
8p	8p22.1->pter	Developmental delay, microcephaly, mild mental retardation, congenital heart defect, behavioural problems.	Claeys et al. (1997)
8q	8q24->qter	Langer-Giedion syndrome/trichorhino-phalangeal syndrome type II: sparse scalp hair, large protruding ears, bushy eyebrows, a broad nasal bridge, bulbous nose, elongated upper lip with a thin upper vermilion border, cone-shaped epiphyses, malocclusion, mandibular retrognathia, dental abnormalities, multiple cartilaginous exostoses, mental retardation.	Langer et al. (1984)
9p	qter->p2304 9p22->pter	1 case report showed the deletion associated with Gilles de la Tourettes syndrome (Pauls et al. [1986]) Trigonocephaly, long philtrum, psychomotor retardation, upward slant of palpebral fissures, dolichomesophalany.	Taylor et al. (1991)
9q		No case reported ¹	

Table 1 (continued.....2)

Chromosome arm	Cytogenetic position involved	Phenotype	References
10p	10p13->pter	DiGeorge Syndrome-like phenotypes: Hypoparathyroidism, heart defects, retarded growth, dysplastic or low set ears, Renal/renal tract anomalies.	Daw et al. (1996)
10q	10q25.3->qter	Triangular-shaped face with broad nasal bridge, downslanting palpebral fissures, strabismus, bow-shaped upper lips, malformed, posteriorly angulated ear, various degree of psychomotor retardation, mental retardation.	Petersen et al. (1998)
11p		No case reported ¹	
11q	11q23.3->qter 11q24.1->qter	Intrauterine growth retardation, short stature, severe to moderate mental and motor retardation, trigonocephaly, hypertelorism, ptosis of upper lids, short nose, anteverted nares, carp-shaped mouth, short neck, cardiac anomalies, delayed myelination. Jacobsen syndrome: Growth and mental retardation, trigonocephaly, facial and digit anomalies, cardiac defects thrombocytopenia, pancytopenia.	Ono et al. (1994) Michaelis et al. (1998)
12p		No case reported ¹	
12q		No case reported ¹	
13p		Acrocentric ²	
13q	13q32.3->qter	Opitz GBBB syndrome-like phenotype: Midline defects include craniofacial abnormalities, microcephaly, hypertelorism, brachycephaly, facial asymmetry, ear anomalies, male external genital anomalies, mental retardation/developmental delay.	Urioste et al. (1995)
14p		Acrocentric ²	
14q	14q32.3->qter	Postnatal onset of relatively small head size in comparison to body size, high forehead with lateral hypertrichosis, broad nasal bridge, high arched palate, residual or apparent epicanthic folds, single palmar crease, mild to moderate developmental delay	Ortigas et al. (1997)

Table 1 (continued.....3)

Chromosome arm	Cytogenetic position involved	Phenotype	References
15p		Acrocentric ²	
15q		No case reported ¹	
16p	16p13.3->pter	Alpha Thalassemia/Mental Retardation Syndrome: α -thalassemia and mental retardation	Wilike et al. (1990b)
16q		No case reported ¹	
17p	17p13.3->pter	Miller-Dieker syndrome : lissencephaly, mental retardation, facial features, heart defects, growth retardation, seizures	Dobyns et al. (1991)
17q		No case reported ¹	
18p		No case reported ¹	
18q	18q21.3->qter	Developmental delay, mental retardation, incomplete myelination, microcephaly, facial and limb abnormalities, genitourinary malformations, neurological abnormalities, hearing abnormalities, growth failure	Brkanac et al. (1998)
19p		No case reported ¹	
19q	19q13.2->qter	Gliomas	Deimling et al. (1994)
20p		No case reported ¹	
20q		No case reported ¹	
21p		Acrocentric ²	
21q	21q22.3->qter	Holoprosencephaly	Estabrooks et al.(1990)
22p		Acrocentric ²	
22q	22q13.3-qter	Generalized developmental delay, normal or accelerated growth, hypotonia, severe delays in expressive speech, mild dysmorphic facial features.	Nesslinger et al. (1994)
Xp	Male: Xp22.32->pter	X-linked ichthyosis, chondrodysplasia punctata, Kallmann syndrome, mental retardation.	Naritomi et al. (1992)
	Female Xp22.31->pter	Goltz and Aicardi syndrome: Microphthalmia, cloudy cornea, chorioretinal coloboma, mild skin defects, agenesis of corpus callosum	Ballabio et al. (1989)
Xq	Xq24->qter	Variable phenotypes, the common one is low birthweight	Maraschio et al. (1996)
Yp	Undefined	Sex reversal of XY females	Disteche et al. (1986)
Yq	Undefined	Variable phenotypes include infertility, short stature, normal intelligence or mental retardation, dysmorphic features	Podruch et al. (1982)

Legend for Table 1

1. A medline search gave no indication that these chromosome ends have been found with *de novo* terminal deletions.
2. Acrocentric chromosomes are composed of repetitive DNA and tandem arrays of the rRNA gene locus. Since rRNA genes are present in multiple copies on all five acrocentric pairs, the deletion of one acrocentric p arm does not apparently have any deleterious phenotypic effects.

The expression of the fragile site is the result of the expansion of $(CCG)_n$ repeats. This expansion was shown in two patients, indicative of the inheritance of the fragile site. Furthermore, the deletion breakpoint of one patient was mapped within 20 kb of the FRA11B fragile site. Therefore, Jones et al. (1995) showed a physical link between the expression of FRA11B and 11q deletion. It is not clear, however, whether the fragile site caused the chromosome truncation. Other human disorders that involve fragile sites such as fragile-X syndrome or a mild mental retardation associated with FRAXE are not associated with terminal deletions. However, further studies are necessary to determine whether the expansion of $(CCG)_n$ is one of the mechanisms that causes chromosome breakage *in vivo*.

Unbalanced translocations

Due to the dynamic nature of the subtelomeric region, non-homologous chromosome exchanges are found at chromosome ends (see above). The type of telomere associated sequence of one chromosome may be more similar to the type on a non-homologous chromosome than its homologue. For example, a pair of 16p chromosomes could be heterozygous for the A and B allele which have different length and sequence (Wilkie et al. 1991 and see above). When the 16p chromosome which carries the A allele searches for its homologue at the chromosomal bouquet stage, it may have a higher chance to pair with the Xq, which also carries an A allele, than with its homologue which carries the non-similar B allele. Wilkie et al. (1991) suggested that this may be the mechanism for the high frequency of non-homologous pairing observed in oocytes, most of them at telomeres (Speed 1998). If the mispairing occurs at the telomere, it may lead to reciprocal translocation.

Reciprocal translocation can lead to the equivalent of terminal deletions in the offspring of carriers. Various chromosome combinations are found in the gametes of an individual who carries a balanced reciprocal translocation. Some gametes will be normal. Alternatively, the gamete may contain both derivative chromosomes so that the resulting offspring carries the balanced translocation. However, if only one of the two derivative chromosomes is found in a gamete, it will produce an unbalanced offspring with a

deletion of part of one chromosome and duplication of part of the other chromosome. If the duplication segment is small, the derivative will appear to be a terminal deletion (reviewed by Ledbetter et al. 1992). This is important clinically, since the cryptic duplicated material may cause additional phenotypic features. In addition, if a parent carries a translocation, an unbalanced offspring may recur, unlike with *de novo* deletions.

Molecular Aetiology of Deletions

Haploinsufficiency

Gene deletions result either in the gene's function being abolished or in the reduced dosage of the gene. If the gene is located on either the X or Y chromosome in males, its deletion will cause a null mutation. For example, the deletion of the testis determining factor gene (TDF) on the Y chromosome causes the sex reversal of XY females (Disteche et al. 1986). Also, when a gene is located in an imprinted region, where expression is restricted to the allele of a particular parentally derived chromosome, the deletion of the actively expressing allele will cause a null mutation. The best known examples of deletions in an imprinted region are Prader-Willi syndrome (PWS) and Angelman syndrome (AS). Both syndromes involve the interstitial deletion of 15q11-q13. There are many genes within the PWS region, including SNRPN (Glenn et al., 1993), IPW (Wevrick et al. 1994), PAR1, PAR5, and PAR7 (Sutcliffe et al. 1994), ZNF127 (Jong et al. 1993), and necdin (MacDonald and Wevrick 1997). Only the paternal alleles of these genes are expressed. On the other hand, AS involves the deletion of the maternal chromosome. To date only the UBE3A/E6-AP gene appears to be associated with AS. Several patients without a deletion have been shown to have a mutation of UBE3A (Tatsuya et al. 1997). Although biallelic expression of UBE3A has been found in many tissues (Vu and Hoffman 1997), it appears to be specifically imprinted in the brain (Rougeulle et al. 1997; Vu and Hoffman 1997).

Most deletions do not occur within imprinted regions. These deletions therefore cause reduction of gene dosage. Many genes are not sensitive to dosage and the deletion of those genes would not result in any clinical abnormalities. For example, almost all enzymes are present in excess and are therefore insensitive to dosage. Topoisomerase III

(TOP3) has been mapped within the critical region of Smith-Magenis syndrome (Elsea et al. 1998). Elsea et al. (1998) showed that the deletion of TOP3 did not cause increased radiation sensitivity and cell cycle checkpoint defects, which would be the expected phenotype for a TOP3 mutation (Fritz et al. 1997). Also, Smith-Magenis patients do not have hyper-recombination and cell cycle checkpoint defects (Greenberg et al. 1996), indicating the deletion of TOP3 does not have any phenotypic effect in Smith-Magenis syndrome patients. Genes for which two copies are needed for a normal phenotype are said to show haploinsufficiency when deleted. Table 2 lists candidate genes for haploinsufficiency states. All mutations that show haploinsufficiency transmit in a dominant fashion. They may be lost due to either deletion or other mutations such as premature termination of translation, where the mRNA becomes unstable and susceptible to degradation (Maquat 1995). Haploinsufficiency has been implicated in many genes with different functions. The genes listed in Table 2 can be divided into two major groups. The first group consists of structural proteins, which includes type X collagen, α -tecorin, myelin basic protein, and the α -1 subunit of collagen VI (Table 2). The second group consists of molecules that play a role in signal transduction pathways. Structural proteins provide the building blocks of the organs, and reduced amounts of such building blocks will cause malformation of the organ. For example, α -tecorin is one of the major noncollagenous components of the mammalian tectorial membrane in the inner ear. Reduced expression of α -tecorin has a deleterious effect on inner ear development, resulting in a form of non-syndromic deafness (Table 2, Hughes et al. 1998).

The second group of haploinsufficient genes includes ligands (for example, jagged 1), transcription modulators (for example, SOX10), and transcription factors, all of which are components in signal transduction pathways. Gene dosage is critical in some such pathways. One example is lateral inhibition during development. In *Drosophila*, neural precursors are formed in a spaced pattern separated by intervening epidermal cells. This pattern formation is controlled by *Notch*. In the neurogenic region, the default state for the cells is the neural fate. However, some spaced cells express *Notch*, a receptor molecule that receives the inhibition signal from the adjacent cells, so that they acquire

Table 2 Genetic diseases involved in the haploinsufficiency of genes

Chromosomal Location	Syndrome Involved	Gene Involved	Function of the gene	References
1p35-p31.3	GLUT-1 deficiency syndrome	GLUT-1	A major glucose transporter	Seidner et al. (1998)
3p14.1-p12.3	Waardenburg syndrome type 2A	MITF	Transcription factor	Nobukuni et al. (1996)
4q25-q27	Rieger syndrome	RIEG	A bicoid-related homeobox	Flomen et al. (1998)
6p21	Cleidocranial dysplasia	OSF2/CBFA1	Transcription factor	Lee et al. (1997)
6p21-q22.3	Schmid metaphyseal chondrodysplasia	Type X collagen	Structural protein for calcifying cartilage	Chan et al. (1998)
7p13	Greig syndrome	GLI3	Kruppel-related zinc finger family gene	Wild et al. (1997)
9q22.3-q31	Nevoid basal cell carcinoma syndrome	PATCHED	Transcription repressor	Wicking et al. (1997)
9q34.1	Hemorrhagic telangiectasia	Endoglin	A heavily glycosylated disulfide-linked dimer that regulates TGF- β 1 signalling pathway	Shovlin et al. (1997)
9q34.1	Nail-patella syndrome	LMX1B	Transcription factor	Vollrath et al. (1998)
11q24	Non-syndromic deafness	TECTA	Major noncollagenous components of mammalian tectorial membrane in the inner ear.	Hughes et al. (1998)

Table 2 (continued)

Chromosomal Location	Syndrome Involved	Gene Involved	Function of the gene	References
12q23-24.1	Ulnar-mammary syndrome	TBX3	T-Box transcriptional factor	Bamshad et al. (1997)
12q24.1	Holt-Oram syndrome	TBX5	T-Box transcriptional factor	Basson et al. (1997)
18q22-qter	18q- syndrome	myelin basic protein	Major protein for myelination	Gay et al. (1998)
20p12	Alagille syndrome	jagged 1	Ligand in the Notch signal pathway	Li et al. (1997)
21q22.3	Bethen myopathy	COL6A1	collagen VI subunit α 1 major protein for muscle	Lamande et al. (1998)
22q13	Waardenburg syndrome type IV	SOX10	Transcriptional modulator	Pingault et al. (1998)
Xpter-p22.32	Turner syndrome	PHOG	Homeodomain-containing transcription factor	Ellison et al. (1997)

the secondary fate as epidermal cells (Artavanis-Tsakonas and Simpson, 1991). In mosaic experiments, wild-type cells adopted an epidermal fate if adjacent cells expressed a lower level of *Notch* activity than their own, but these cells produced neural precursors if adjacent cells expressed a higher level of *Notch* protein (Heitzler and Simpson 1991). These results showed that the relative level of inhibition signal between two cells determines their cell fates. *Jagged 1*, a ligand of *Notch*, is deleted in Alagille syndrome patients (Table 2). The haploinsufficiency of *Jagged 1* may be analogous to the reduced inhibition signal received from the adjacent cells in *Drosophila*. It is possible that the change of cell fate during development causes the malformations in Alagille syndrome (liver, heart, skeletal and facial abnormalities).

Transcription factors may also be sensitive to gene dosage. For example, the gap genes encode transcription factors, whose patterns of expression define the boundaries of body segments during embryogenesis in *Drosophila*. The gap gene *Hunchback* is expressed in the anterior region of the embryo, whereas *Krüppel* is expressed posterior to *Hunchback*. *Krüppel* forms a protein gradient along the anterior-posterior axis. It activates the expression of a third gap gene *knirps* in the posterior region, where the concentration of *Krüppel* is low (Pankratz et al. 1989). At the anterior end, *Hunchback* activates the expression of the pair-rule gene *even-skipped (eve)*, which induces the formation of a stripe pattern in embryos. Stripe 2 is at the boundary of *Hunchback/Krüppel* expression. While *hunchback* activates the expression of *eve*, *Krüppel* represses *eve* so that it forms the border of stripe 2 (Small et al. 1991). Therefore, *Krüppel* acts as both transcription activator and transcription repressor. Sauer and Jäckle (1993) showed that the activator/repressor function of *Krüppel* depends on the conformation of the protein. When the concentration of *Krüppel* is low, it exists as a monomer and acts as an activator, whereas a high concentration of *Krüppel* forms homodimers and represses transcription. Haploinsufficiency of *Krüppel* lowers the concentration of the protein, resulting in a change of the *Krüppel* function and subsequently disrupts pattern formation. A similar phenomenon may cause Greig syndrome (fusion of fingers, wide set eyes, broad forehead), in which GLI3, a member of the *Krüppel*-family of genes, is haploinsufficient (Table 2, Wild et al. 1997)

Segmental aneusomy syndromes (SAS)

There are a large number of genes located within a visible chromosomal deletion region. Schmickel et al. (1986) first introduced the term “contiguous gene syndrome” for disorders resulting from the deletion or duplication of adjacent genes in a specific chromosomal region. Since not all the genes within a deletion region are responsible for the clinical features of the syndrome, the term “contiguous gene” cannot correctly describe the aetiology of the syndrome. Therefore, “segmental aneusomy syndrome” (SAS) has been suggested as a replacement for the old term. An SAS is defined as a syndrome that is the result of dosage imbalance of certain critical genes within a duplication/deletion region (Budarf and Emanuel 1997). SAS implies that there is more than one gene responsible for manifesting the clinical features. Although deletions involving large chromosomal segments are probably SASs, there are cases in which mutations in a single gene within the deletion region gives the complete clinical features of the syndrome. In order to determine whether a deletion syndrome is a SAS or single gene disorder, one has to look for non-deletion patients who share the same phenotype as the deletion syndrome patient. Through the search for mutations within a candidate gene of the deletion syndrome in those non-deletion patients, some deletion syndromes have been redefined as single gene disorders. For example, Rubinstein-Taybi syndrome is characterized by facial abnormalities, broad thumbs, broad big toes and mental retardation (Rubinstein and Taybi 1963; Hennekam et al. 1990). The common cause of Rubinstein-Taybi syndrome is a small deletion (microdeletion) which encompasses 130 kb of chromosome 16p13.3 (Petrij et al. 1995). One candidate gene, the CREB binding protein (CBP), maps within the microdeletion (Chen and Korenberg 1995; Wydner et al. 1995). Other than translocation breakpoints that disrupted the gene, Petrij et al. (1995) found point mutations within CBP by protein truncation assay. Since patients that only have point mutations show the syndrome, Petrij et al. (1995) concluded that the loss of one functional copy of the CBP is the cause of the developmental abnormalities in Rubinstein-Taybi syndrome.

Another syndrome associated with a deletion which is now known to be a single gene disorder is the Rieger syndrome type I (RIEG1). RIEG1 shows a variable

phenotype. The major clinical features include tooth defects (Brooks et al. 1989), malformations of the anterior segment of the eye (Rieger 1935), and the failure of the periumbilical skin to involute (Jorgenson et al. 1978). Ligutic et al. (1981) first showed the association of RIEG1 with a 4q interstitial deletion. Since then some RIEG1 patients with interstitial deletions between 4q25 to 4q27 have been reported (For example, Shiang et al. 1987; Vaux et al. 1992). Semina et al. (1996) mapped a homeo box gene RIEG within the RIEG1 deletion region. When mutation analysis was performed in six RIEG1 families, three families showed missense mutations, two families had mutations within the introns that altered splicing sites, and one family had a nonsense mutation (Semina et al. 1996). Those mutations confirmed that RIEG is the only gene on 4q responsible for the clinical features of RIEG1.

Alternatively, there are well-known examples of SAS, where more than one gene in a deleted region contributes to the phenotype. For instance, Williams syndrome is a neurodevelopmental disorder with a variable phenotype. The common clinical features include mental retardation; distinctive facial features; infantile hypercalcaemia, a friendly outgoing personality; and cardiovascular anomalies, including narrowing of the aortic valve orifice (supravalvular aortic stenosis [SVAS]) and narrowing at the peripheral pulmonary artery (peripheral pulmonic stenosis [PPS]) (Williams et al. 1961; Beuren et al. 1962). In addition, some patients showed a specific cognitive profile (WSCP) characterized by relative strength in language and auditory rote memory and pronounced weakness in visuospatial constructive cognition (Udwin et al. 1987; Bellugi et al. 1994). Williams syndrome is associated with an ~2 Mb deletion at 7q11.2 (Wu et al. 1998). Many genes have been mapped to the deletion region, including the elastin gene (ELN)(Curran et al. 1993; Ewart et al. 1993), the replication factor C subunit 2 (RFC2), a novel RNA binding protein (WSCR1), a gene with similarity to restin (WSCR4)(Osborne et al. 1996), the human homologue to the Drosophila frizzled gene (FZD3)(Wang et al. 1997), Syntaxin 1A gene (STX1A)(Osborne et al. 1997), and LIM kinase-1 (LIMK1)(Frangiskakis et al. 1996; Tassabehji et al. 1996). Of these genes, ELN has been associated with SVAS, a hoarse voice and rapidly aged-appearing skin of the Williams syndrome, since elastin is required for the normal ultrastructure of the aorta, vocal ligaments, and the skin (Ewart et al. 1993). This is further supported by mutations

detected in SVAS patients who did not show other Williams syndrome features (Curran et al. 1993; Olson et al. 1995). However, some Williams syndrome patients with deletions of ELN did not show SVAS (Nickerson et al. 1995), which is presumably due to variable penetrance and a different genetic background. The LIMK1 gene, which is located 15.4 kb 3' of ELN, has been associated with the WSCP cognitive abnormality (Frangiskakis et al. 1996). Frangiskakis et al. (1996) identified a Williams syndrome family who have SVAS and WSCP. Other clinical features such as mental retardation, developmental delay, and facial features are absent in this family. Molecular studies revealed that this family carries an 83.6 kb microdeletion within the Williams syndrome critical region, which includes only ELN and LIMK1 (Frangiskakis et al. 1996). LIMK1 is a novel kinase with two zinc-binding motifs (LIM domains) at the N-terminal. It may play a role in an intracellular signaling pathway by LIM-mediated protein-protein interactions (Frangiskakis et al. 1996). The fact that LIMK1 is highly expressed in the central nervous system in rat and human (Bernard et al. 1994; Pröschel et al. 1995) makes it a likely candidate gene for involvement in cognitive development. However, when Wu et al. (1998) analyzed four non-deletion Williams syndrome patients, they did not detect any mutation in LIMK1. Those non-deletion patients did not have SVAS, but had Williams syndrome features including mental retardation, facial features, and infantile hypercalcaemia. Their cognitive profile, however, has not been described (Nickerson et al. 1995). Therefore, it is inconclusive whether LIMK1 is responsible for the WSCP in Williams syndrome, until specific mutations in patients with only WSCP can be identified. However, it seems clear that Williams syndrome is an SAS, with ELN associated with SVAS, possibly LIMK1 associated with WSCP, and other genes associated with other clinical features.

Another SAS involves the terminal deletion of 18q. 18q- syndrome is a rare disorder with clinical features listed in Table 1. One common feature for 18q- syndrome is white matter abnormalities of the brain, caused by delayed myelination (Gay et al. 1998). The myelin basic protein gene (MBP), which plays a role in myelinogenesis, maps within the 18q- deletion region (Kamholz et al. 1987). To determine whether the reduced dosage of MBP causes abnormal white matter development, Gay et al. (1998) studied the white matter development in twenty 18q- syndrome patients by magnetic resonance

imaging (MRI) and compared this to the extent of their deletions. The results showed that nineteen out of twenty patients showed abnormal white matter development. The only patient with normal white matter development has an interstitial deletion, which does not include MBP (Gay et al. 1998). This suggests that MBP is a critical gene in white matter development and is dosage sensitive. Since the clinical features other than neurologic manifestations are unlikely to be a result of MBP deletion, 18q- appears to be a SAS.

Phenotypes Associated with Deletions at the Ends of Chromosomes

Specific phenotypes have been associated with deletions of the terminal band for most chromosome arms (Table 1). An exception is the p arms of the acrocentric chromosomes, which are composed of repetitive DNA and tandem arrays of the rDNA gene locus. Since rDNA genes are present in multiple copies on all five acrocentric pairs, the deletion of one acrocentric p arm does not appear to have any deleterious effects. Common features for terminal deletion syndromes include mental and growth retardation, developmental delay, and distinguishable facial features. With some chromosome arms, the size of the deletion determines the phenotype. For example, deletions at the end of 4q give two different phenotypic spectrums. When the deletion includes the two most distal bands 4q34-35, only mild cranial-facial anomalies are seen, whereas if the deletion includes 4q33, mental retardation is evident (Table 1). The Jacobsen syndrome, which has the minimal deletion region of 11q24.1-qter, shares a similar phenotype with patients who have the larger deletion of 11q23.3-qter. However, myelination delay is only found in the larger 11q23 deletion (Table 1), suggesting that a gene that regulates myelination maps proximal to the Jacobsen syndrome deletion region. Xp terminal deletion has been found in both male and female patients (Table 1). However, no male patient has been found with a deletion of Xp22.31-pter; presumably there is a gene in this region which causes lethality in males when lost (Ballabio et al. 1989). Some terminal deletions, such as in chromosome Xq⁻ or Yq⁻, exhibit a wide range of phenotype so that it is difficult to define them as syndromes.

For some chromosome arms, no terminal deletion has been reported (Table 1). Such regions may not contain any genes that are dosage sensitive, and therefore a

deletion cannot be detected phenotypically. Alternatively, they could cause lethality in early embryonic stages resulting in spontaneous abortion. Another possibility is that those deletions are hard to detect by karyotyping, as occurs in some terminal regions of chromosomes (Pedersen and Kerndrup 1986; Ortigas et al. 1997). Therefore the related syndrome could exist but has not been associated with the deletion.

Delineating the Chromosomal Region Involved in Terminal Deletions

Giemsa banding (G-banding) is a standard method for analyzing chromosomal anomalies. G-banding involves the selective removal of chromosomal proteins by chemical treatment, allowing the binding of Giemsa stain to exposed DNA (Comings 1978). G-banding produces a series of light and dark bands along the length of the chromosomes. Deletions can be visualized as the absence of a particular band or part of a band. Skilled high resolution karyotyping of G-banded chromosomes can detect a deletion encompassing at least 2 to 3 Mb of genomic sequence (Ledbetter and Cavenee, 1989). However, it would be difficult to identify the genes responsible for the associated phenotype in such a large region. This is especially true of the deletions in the terminal bands, which are believed to be gene rich (Saccone et al. 1992). It has been estimated that such gene rich regions may have an average of one gene every 23.4 kb (Fields et al. 1993)

When studying a deletion associated with a SAS, a main objective is to find and characterize the genes which may be dosage sensitive. It is important to define a deleted region that is common to patients who show overlapping clinical features. Such a region is called the critical region, which is defined as the smallest deleted region that produces a common phenotype. However, extra caution has to be taken when determining the genotype: phenotype correlation because chromosomal abnormalities typically show wide phenotypic variability from patient to patient. Due to variable penetrance, patients with larger deletions may by chance show a less severe phenotype than those who have smaller ones. Two patients with the same sized deletion/duplication will usually differ somewhat in phenotype. For example, most Down syndrome patients are full trisomy 21, but can vary substantially in phenotype. Some clinical features, such as congenital heart

disease or narrowing of the duodenum (duodenal stenosis), are absent in some patients (Korenberg et al. 1992). Genetic background that affects the penetrance of genes may contribute to the variable phenotype.

Many molecular genetic and cytogenetic techniques have been used to delineate a common region of deletion. Since deletions cause hemizyosity for the region, the boundaries of the critical region can be determined by assaying the number of copies of loci within the deleted region. Those loci include known genes, anonymous DNA markers that contain variable number of tandem repeats (VNTR, minisatellites), or variable length di- or trinucleotide repeats known as microsatellites (Hudson et al. 1992). There are abundant resources of VNTR or microsatellite markers. For example, the Généthon (Dib et al. 1996) or Cooperative Human Linkage Centre (<http://www.chlc.org>) genetic linkage maps provide high density microsatellite markers that cover the whole human genome. The copy number of polymorphic loci can be determined by genotyping. However, there are limitations for this technique. Firstly, polymorphic markers are not always informative. Secondly, genotyping requires DNA samples from the parents of the patients in order to determine the origins of the alleles. Parental DNA samples are not always available. If DNA markers are uninformative, DNA fragments within the deletion region can be used for quantitative dosage analysis. Radioactive probes are hybridized to restriction endonuclease-digested, electrophoresed, and blotted genomic DNA from the deletion patient and a normal individual. The dosage of the probe in the patient genomic DNA can be determined by comparing the intensity of the signals from the deletion patient and normal individual on the autoradiograph. A reliable comparison can be obtained by applying a non-parametric statistical method to replicates of the experiment (Mears et al. 1994). Another commonly used technique is fluorescence *in situ* hybridization (FISH). Large DNA fragments are biotin labeled and hybridized to metaphase spreads from the patient. The deletion can be scored by the presence of fluorescence signals on one or both of the chromosome homologues.

Microdeletions as a Tool to Study Deletion Syndromes

Another useful resource for delineating a deletion syndrome critical region is the microdeletion. Microdeletions are small, cytogenetically-undetectable deletions. If they share with the larger typical deletion a subset or all of the syndrome features, they can be used to narrow the search for candidate genes. Since a microdeletion is not detectable by karyotyping, another molecular cytogenetic technique such as FISH is often used to detect it. For example, Altherr et al. (1997) studied the critical interval of Wolf-Hirschhorn syndrome which involves the terminal deletion of 4p16.3 (Table 1). They found a patient with features of the syndrome but no visible deletion. Using FISH to identify and define the proximal and distal boundaries of an interstitial microdeletion, Altherr et al. (1997) narrowed the Wolf-Hirschhorn syndrome critical region down to ~750 kb, which excludes four genes mapping to the larger region.

Other than redefining the critical region to narrow the focus of gene searches, microdeletions associated with a subset of phenotypic features can be used to study the genotype: phenotype correlation of a deletion syndrome. A good example is the 83.6 kb microdeletion within the Williams syndrome critical region, which was used to establish a relationship between the haploinsufficiency of LIMK1 and the WSCP cognitive abnormality (Frangiskakis et al. 1996 and see above). In another example, there are interstitial/terminal deletions of various sizes in the cri-du-chat syndrome critical region that define genes associated with different features of the syndrome (Overhauser et al. 1994). This 5p terminal deletion syndrome shows clinical features such as small head size (microcephaly), a round face, wide set eye (hypertelorism), small chin (micrognathia), prominent nasal bridge, epicanthal folds, low muscle tone (hypotonia), severe psychomotor and mental retardation, and a high-pitched cry (Table 1). By studying deletions that span different regions of 5p, Overhauser et al. (1994) found that patients with large deletions of the most proximal region (5p13-14) have mild to severe mental retardation and microcephaly. Individuals with the small interstitial deletion of 5p14 are normal. Deletions with the proximal breakpoint at 5p15.1 show mild mental retardation as well as other clinical features, and small deletions at the most distal band 5p15.3 only

show the characteristic cry. As a result, the phenotype map for cri-du-chat syndrome, from centromere to telomere, is mild to severe mental retardation/microcephaly - mild mental retardation - facial features/psychomotor retardation/micrognathia/hypotonia - high pitched cry.

Unknown Cases of Multiple Congenital Anomalies/Mental Retardation

Although terminal deletions are clearly involved in different phenotypic features, growth/mental retardation, developmental delay, and abnormal facial features are commonly found in all interstitial and terminal deletions (see above). These common congenital defects are found in liveborns with considerable frequency. For example, mental retardation (with intelligence quotient less than 70) occurs at a frequency of 3% of the population (Flint et al. 1995). Although there are many known causes for mental retardation, such as Down syndrome and Fragile-X syndrome, about half the cases identified still remain unexplained with no obvious chromosomal abnormality (Flint et al. 1995). This raises the question of whether some of those unexplained congenital defects are actually caused by microdeletions or cryptic translocations at chromosome ends, and thus represent a subset of the features of a larger deletion syndrome. Such a hypothesis is based on several observations. First, since the telomeric region is gene rich (Saccone et al. 1992), the rearrangements at chromosome ends seem more likely to have phenotypic effects. Second, chromosomal rearrangements at chromosomal ends appear to occur more frequently due to the dynamic nature of the subtelomeric region (see above). However, these cases of chromosomal rearrangements may be underrepresented, because if the chromosomal rearrangements involve small segments at the tips of the chromosomes (cryptic rearrangements), they might not be detected by karyotyping. There are cases of patients showing clinical features of various terminal deletion syndromes but apparently normal karyotypes. Subsequent FISH analysis showed that they carry unbalanced cryptic translocations. (Altherr et al. 1991; Goodship et al. 1992; Kuwano et al. 1991; Overhauser et al. 1989; and reviewed by Ledbetter 1992). If cryptic rearrangements produce terminal deletion syndromes, it is possible that some even smaller cryptic rearrangements could produce a subset of clinical features of a terminal deletion

syndrome. Those cryptic rearrangements can only be detected by molecular techniques such as genotyping the polymorphic markers at the subtelomeric regions (Wilkie 1993) or FISH. FISH probes from a complete set of human genomic clones that are unique and at the most distal region of each chromosome arm have been used for detecting such cryptic rearrangements (National Institutes of Health and Institute of Molecular Medicine Collaboration, 1996).

In order to determine whether there is an association of congenital defects to microdeletions at chromosome ends, Flint et al. (1995) screened 99 patients with mental retardation of unknown cause for subtelomeric cryptic rearrangements. They firstly genotyped VNTR markers from 28 different chromosome ends. When a putative cryptic rearrangement on a chromosome end was found, FISH was used to confirm the case and determine whether it was a microdeletion or cryptic translocation. Through this scheme of screening, Flint et al. (1995) found 3 out of 99 patients who showed cryptic chromosomal rearrangements. In one case a deletion of chromosome 13q was present due to an unbalanced cryptic translocation between 13q and the Xp/Yp pseudoautosomal region (three copies for Xp/Yp, and one copy for 13q), whereas two cases involved chromosome 22q. One has an unbalanced cryptic translocation between 22q and 9q (three copies for 9q and one copy for 22q), and the other has a chromosome 22q microdeletion. Although only 3% of the patients with unexplained mental retardation showed chromosomal rearrangement in this study, Flint et al. (1995) estimated that the actual frequency of subtelomeric rearrangement could be as high as 6%. This is because they only tested 28 chromosome ends, VNTR markers were informative in only 76% of the cases, and some markers could be one to two Mb from telomeres.

Interstitial/Terminal Deletions of Chromosome 22q

DiGeorge syndrome (DGS)

Chromosome 22 is the second smallest human chromosome, and is one of the five acrocentric chromosomes that carry the rDNA arrays on the p-arm (the other four acrocentric chromosomes are 13, 14, 15, and 21). Chromosome 22 is believed to be one

of the most gene rich chromosomes (Saccone et al. 1996). Therefore, chromosome rearrangements associated with chromosome 22q show many different congenital defects. For example, duplications at the centromeric region are found in cat eye syndrome patients (McDermid et al. 1986), and translocations in different segments of 22q give rise to various cancers (Nowell and Hungerford 1960; Rowley 1973; Berger et al. 1979; Aurias et al. 1983).

Among various congenital chromosomal anomalies found on 22q, an interstitial microdeletion at 22q11.2 occurs frequently in the population (1 in 4000; Burn et al. 1995). This microdeletion is associated with several overlapping but clinically distinguishable syndromes. DiGeorge syndrome (DGS) is a development field defect. The major aetiology involves the third and fourth pharyngeal pouch derivatives, including thymus and parathyroid gland aplasia/hypoplasia and conotruncal cardiac malformations (Lammer and Opitz 1986). Thymus and parathyroid gland malformations subsequently cause neonatal hypocalcemia and susceptibility to infection due to a deficit of T cells (DiGeorge 1968). In addition, DGS patients also show distinguishable facial features which include wide set eyes (hypertelorism), cleft lip and palate, bifid uvula, and small/low-set ears (Greenberg 1993). Both conotruncal-anomaly-face syndrome (Takao et al. 1980) and velocardiofacial syndrome (VCFS)(Goldberg et al. 1993) are also characterized by abnormal facial features and cardiac anomalies in association with the same microdeletion. Variable clinical features such as psychiatric disorders of adulthood (Karayiorgou et al. 1995), mild to moderate learning difficulties (Schprintzen et al. 1978), and renal and urological tract malformations (Novak and Robinson 1994; Devriendt et al. 1996) are often present in these patients. Wilson et al. (1993) named these syndromes collectively as CATCH22 (Cardiac Abnormality, Abnormal faces, and T cell deficit due to Thymic hypoplasia. Cleft palate, Hypocalcemia due to hypoparathyroidism resulting from 22q11 deletion). The microdeletion syndrome in general is also known as DGS/VCFS. All features of these syndromes are compatible with defects of neural crest cells.

Since the association between DGS and 22q11.2 was demonstrated in the early 80's (for example, De la Chapelle et al. 1981), numerous efforts have been made to localize the critical region of the syndrome. Recently, the critical region has been

narrowed down to a <160 kb region (reviewed by Budarf and Emanuel 1997). Three likely candidate genes within the critical region, from centromere to telomere, are Goosecoid-like (GSCL), citrate transport protein (CTP), and clathrin heavy chain-like (CLTCL). GSCL encodes a homeobox protein, with expression mainly in the pons region of the developing brain. GSCL is also expressed in the primordial germ cells when these cells migrate from the epithelium of the hindgut to the developing gonads. The major affected region in DGS (i.e. the colonization of cephalic neural crest cells at the third and fourth pharyngeal arches), show no GSCL activity (Galili et al. 1998). Furthermore, when 17 non-deleted VCFS patients were screened for mutations at GSCL, no mutation was detected (Funke et al. 1997). Therefore, it is not likely that the deletion of GSCL alone can produce the complete phenotype of DGS/VCFS. CTP transports citrate across the mitochondrial inner membrane. It is not clear whether the deletion of CTP is responsible for any phenotype of DGS/VCFS. CLTCL is homologous to the clathrin heavy chain gene (Sirotkin et al. 1996), which is the structural component of coated pits and coated vesicles that are formed during endocytosis and membrane receptor trafficking (Brodsky 1988; Robinson 1994). A breakpoint from a balanced (21;22)(p12;q11) translocation disrupts the CLTCL gene (Holmes et al. 1997). The patient who carries the balanced translocation shows some DGS/VCFS features, including typical VCFS facial features, cataracts, and mild mental retardation. It has therefore been suggested that haploinsufficiency of CLTCL is responsible for a part of the DGS/VCFS phenotype. Future research in DGS/VCFS will concentrate on characterizing gene(s) that are directly related to the developmental defect. Genes that fall outside of the distal and proximal boundaries of the critical region are not totally excluded from being DGS/VCFS candidates. Position effect may play a role in altering gene expression as well. For example, DGS1/ES2 located proximal to GSCL (Gong et al. 1997) is highly expressed in the pons during development, in the same region where GSCL expression is found (Lindsay et al. 1998). If there is a cis-regulatory element at the telomeric end of GSCL that controls both GSCL and DGS1/ES2 expression, a deletion distal to DGS1/ES2 may also affect its expression.

22q13.3 deletion syndrome

Terminal deletion of chromosome 22q is rare. It can be visualized by cytogenetic analysis, when the most distal subband, q13.3 is deleted. Nesslinger et al. (1994) studied seven patients with 22q13.3 deletion, and showed a common phenotype including generalized developmental delay, normal or accelerated growth, low muscle tone (hypotonia), severe delays in expressive speech, and mild facial dysmorphic features. At the time when Nesslinger et al. (1994) conducted their studies, the patients were too young (under 5 years old) for an assessment of mental ability. Follow-up studies show that all 22q13.3 deletion syndrome patients have various degrees of mental retardation (Katy Phelan, personal communication). By dosage analysis, Nesslinger et al. (1994) mapped the proximal deletion breakpoints, and found no obvious correlation between the size of the deletion and the severity of the phenotype. The proximal boundary of the critical region is D22S94, the most proximal marker that is deleted in the patient with the smallest deletion. Arylsulfatase A (ARSA), the most distal probe known on 22q at the time, was deleted in all patients. Since nothing distal to ARSA had been mapped (Dumanski et al. 1991; Collins et al. 1995b) it was not possible to assess the extent of the deletions below this locus. All deletions have since been mapped to include at least the VNTR locus D22S163, which maps distal to ARSA by pulse field gel electrophoresis (PFGE)(McDermid, personal communication).

Use of a Microdeletion to Study a Subset of Features in 22q13.3 Deletion Syndrome

The critical region of the 22q13.3 deletion syndrome covers a genetic distance of 25.5 cM (Nesslinger et al. 1994), which represents 5 Mb as measured by pulsed-field gel electrophoresis (PFGE)(H.E. McDermid, unpublished data). This large size therefore makes impractical a search for genes that may be involved in the production of the phenotype of this syndrome. However, Flint et al. (1995) identified a child with a microdeletion at the end of chromosome 22q through the screening of 99 mentally retarded individuals for subtelomeric cryptic chromosomal rearrangements (see above). Flint et al. (1995) described a 12-year-old boy (NT) with delayed expressive speech, mild

mental retardation (IQ = 64), normal facial features, and a negative family history for mental retardation. Although his high-resolution karyotype was normal, the paternal allele of the subtelomeric minisatellite probe D22S163 (MS607) was found to be deleted. ARSA and other 22q13.3 probes were present in two copies, indicating that NT carried a microdeletion. The patient NT shows overlapping features with those in the 22q13.3 deletion syndrome and the microdeletion falls within the critical region. This microdeletion may therefore be useful to narrow the focus of gene searches for the 22q13.3 deletion syndrome.

Research Summary

There were several objectives of my research: (1) To complete the molecular characterization of the microdeletion, to determine the location and size of the deletion. (2) To determine the origin of the microdeletion. (3) To identify and characterize genes within the NT microdeletion that may be responsible for a subset of clinical features found in 22q13.3 deletion syndrome.

To determine the location and size of the microdeletion, I constructed a cosmid contig that spans the microdeletion region. The cosmid contig was constructed by a cosmid walk starting from cosmids that contain the D22S163 locus, which was shown to be hemizygous in NT by RFLP analysis (Flint et al. 1995). The direction of the cosmid walk was determined by dosage analysis of the cosmid end fragments. When one end of a cosmid was found not deleted while the other end was, the cosmid walk was extended to the deleted end.

To understand the origin of the microdeletion, I mapped its breakpoint. I firstly narrowed down the location of the breakpoint within a cosmid by dosage analysis. I then used the DNA fragments from this cosmid as probes, and hybridized to Southern blots that contained restriction endonuclease digested and electrophoresed DNA from the NT family. If the breakpoint was in or near the probe, a novel rearrangement band would be detected by the probe. I further tested whether the breakpoint fragment was sensitive to the digestion of BAL31 exonuclease, which is indicative of terminal deletion.

To search for genes within the microdeletion, I sent the cosmid contig for sequencing to Dr Bruce Roe at the University of Oklahoma's Advanced Center for Genome Technology. I searched the resulting cosmid sequences against various public databases to search for known genes or expression sequence tags (ESTs), which are end sequences from random cDNAs. I also found putative expressed sequences by using exon predication programs. I confirmed the expression of those putative genes by RT-PCR and Northern analysis. One putative candidate gene is homologous to a *Caenorhabditis elegans* hypothetical protein. Since many homologous proteins between species also share functional homology, I studied the expression of the *C. elegans* protein by reporter gene fusion in order to understand its function. Another new gene identified within the region is probably the last gene on 22q and is duplicated on 2q13.

In this project I analyzed the genomic organization of the subtelomeric region in chromosome 22q. I mapped two new putative genes and one known gene within the microdeletion. Therefore, this research sets the stage for future work on studying the relationship between the deletion of the candidate genes and the phenotype of the 22q13.3 deletion/microdeletion syndrome patients.

Materials and Methods

Cell Lines

Epstein Barr virus (EBV)-transformed lymphoblastoid cell lines for NT and his parents were obtained from Dr Jonathan Flint (Flint et al. 1995). Cell lines from the human/rodent somatic cell hybrid mapping panel #2 and a normal human lymphoblastoid cell line (GM03657) were obtained from the NIGMS Human Genetic Mutant Cell Repository. Fibroblasts from FB, who has the 22q13.3 deletion syndrome has been described elsewhere (Nesslinger et al. 1994). RJK88 is a hamster fibroblast line (Fusco et al. 1983). Lymphoblast lines were cultured in T25 flasks (Corning) with RPMI 1640 media (GIBCO-BRL) and supplemented with 10% fetal bovine serum, 1% penicillin/streptomycin and 1% L-glutamine. FB fibroblasts were cultured in T75 flasks (Corning) with D-MEM media (GIBCO-BRL) and supplemented with 10% fetal bovine serum, 1% penicillin/streptomycin and 1% L-glutamine. For subculturing fibroblast lines, the cells were first washed in Hank's balanced salt solution (HBSS) and then trypsinized with 0.5% trypsin, 0.53 mM EDTA-4Na (GIBCO-BRL). Human/rodent somatic hybrid lines were cultured as described in NIGMS Human Genetics Mutant Cell Repository catalog and were cultured by Dana Shkolny.

Probes

The following probes were used in this study: pHU4A, a cDNA of ACR (Baba et al. 1989); D22S163 (probe MS607), a minisatellite probe (Armour et al. 1990); D21S15 and D21S110, two reference probes on chromosome 21 (Stewart et al. 1985; Spinner et al. 1989). Two probes were produced from cosmid N66C4: a 5.0-kb HindIII fragment (H5.0), a 1.2-kb XhoI fragment (Xh1.2). For the PCR products that were used as probes, they are listed in Table 3 "PCR and RT-PCR products" section. EST clones are listed in appendix.

Table 3: PCR and RT-PCR products amplified in this study

Name of PCR/RT-PCR product	DNA template	Tissue of total RNA	Reverse transcription primer	Forward primer	Reverse primer	Product size (bp)	Description
F3-R7		spleen	Ank1 R1	Ank F3	Ank R7	491	Partial cDNA of ALPR
F1-R1		fetal brain and spleen	Ank R3, OFLS4, OFH2	Ank F1	Ank R1	400	Partial cDNA of ALPR
F2-R1		fetal brain and spleen	Ank R3, OFLS4, OFH2	Ank F2	Ank R1	358	Partial cDNA of ALPR
55337F-R1		spleen	OFLS4 OFH2	55337F	Ank R1	510	Partial cDNA of ALPR
sc 24, sc31 sc 32		spleen	Ank R1	F3	553R	1268	Partial cDNA of ALPR
fli		fetal liver, fetal brain	OFLS4	M1314F	OFLS3	584	Partial cDNA of ALPR
R1F-M1314R		fetal brain	OFLS4	R1F	M1314	413	Illegitimate product of ALPR partial cDNA
F5-R3	N85A3			Ank F5	Ank R3	167	"Last exon" Candidate exon of ALPR
U2	N94H12			U2F	U2R	811	Partial U2 SnRNP polypeptide A pseudogene
3'AR	N94H12			3'AR F1	3'AR R1	1244	Minisatellite locus 3'AR
RABLF1-E0.91F, 2C, 22A		Various tissues	pul31	RABLF1	E0.91F	524	partial cDNA of RABL
Internal intron		Various tissues	pul31	IIF	E1.03R	398	Internal intron at the 3'UTR of RABL
RABLF1-R1	Chromosome 2 hybrid line			RABLF1	RABLR1	1081	Chromosome 2 genomic DNA
B4.3F-R	C. elegans genomic DNA			B4.3F	B4.3R	6129	Promoter + 4 exons of C33B4.3
B4.3HP		C. elegans total RNA	B4.33	B4.3H	B4.3P	1041	Partial DNA of C33B4.3
B431HR1-B4.332		C. elegans total RNA	B4.333	B431HR1	B4.322	1541	Partial DNA of C33B4.3
B4.3FP-B4.322		C. elegans total RNA	B4.333	B4.3FP	B4.322	898	Partial DNA of C33B4.3

DNA Studies

Genomic DNA extraction

Genomic DNA extraction was a modification of a previously described procedure (Gustincich et al. 1991; Mears 1995 PhD thesis). Briefly, cells from three to four T25 (lymphoblastoid cells) or T75 (fibroblasts) flasks were pooled in a 50 ml centrifuge tube (Corning or Fisher) and pelleted by spinning in a clinical centrifuge for 10 minutes at 800 rpm. Cells were then resuspended in 5 ml HBSS. DTAB solution (8% DTAB [Sigma Chemical], 1.5 M NaCl, 100 mM Tris-HCl pH 8.6, 50 mM EDTA) was preheated to 68 °C and 15 ml was added to the cell suspension. After inverting the centrifuge tube three to four times, the solution was incubated at 68 °C for 15 min. To the cell lysate 15 ml of chloroform was added and mixed by inversion (3-4 times). The chloroform-DTAB mixture was transferred to a 30 ml Oakridge centrifuge tube and centrifuged in a HB4 rotor at 10,000 rpm for 15 min. The aqueous layer was transferred to a fresh Oakridge centrifuge tube. 17 ml of dH₂O and 1.7 ml CTAB solution (5% CTAB [Sigma Chemical] in 0.4 M NaCl) were added. After gentle mixing by inversion, it was centrifuged at 10,000 rpm for 15 min. The DNA-CTAB pellet was resuspended in 5 ml 1.2 M NaCl overnight to exchange the detergent. Genomic DNA was then precipitated by adding 18 ml 95% ethanol. The DNA pellet was washed one time in 70% ethanol and resuspended in various volumes of TE pH 8.0 (10 mM Tris-HCl, pH 8.0; 1 mM EDTA).

Plasmid and cosmid DNA preparations

A cosmid contig was constructed to span the NT microdeletion region (See the results "Production of a Cosmid Contig from D22S163 to the 22q Telomere"). Probes used in this study were subcloned in different plasmid vectors. Small quantities of plasmid DNA were isolated by the conventional "mini-prep" method (Sambrook et al. 1989). To isolate large amounts of plasmid and cosmid DNA, bacterial clones were inoculated in 100 ml LB with 60 µg/ml ampicillin (for plasmids) or 500 ml LB with 50 µg/ml

kanamycin (for cosmids), incubated overnight at 37 °C with shaking, and isolated by using Qiagen-tip 500, following the manufacturer's instruction.

Southern blot hybridization

Southern blot hybridization was used for RFLP analysis, densitometric analysis, hybrid panel analysis, and hybridization of cDNA back to genomic clones to confirm the authenticity of the cDNA clones. Different sources of DNA (genomic, cosmid or plasmid) were digested with various enzymes, mixed with ~5 µl loading dye (Orange G in 20% Ficoll and 100 mM EDTA), and then loaded and run on 0.8% TBE (0.089 M Tris, 0.089 M Boric acid, 0.002 M EDTA, pH 8.4) agarose gel. DNAs were transferred to a nylon membrane (GeneScreen Plus, DuPont) by conventional methods (Sambrook et al. 1989). For transferring high molecular weight DNA, the gel was first soaked in depurination solution (0.25 M HCl) for 10 min, and washed thoroughly with distilled water before proceeding to the southern blotting step.

Probes were prepared by excising the cloned DNA from a low melt agarose gel. Approximately 50 ng of DNA was labeled using a conventional random primer method (Feinberg and Vogelstein, 1983). At the end of the labeling reaction, 50 µl of "stop buffer" (blue dextran in 0.2 M EDTA) was added. The labeled product was then purified by a G-50 Sephadex column. The radioactive probe was boiled for 10 min before adding to the membranes. For genomic DNA fragment probes that may contain repetitive sequences, the probes were pre-annealed for four hours at 65 °C with sonicated placental DNA. Hybridization was performed in a variation on the solution of Church and Gilbert (1984) (7% SDS, 0.263 M sodium phosphate pH 6.5, 1 mM EDTA, 5X Denhardt's solution) overnight at 65 °C in a hybridization oven (Tyler). The membrane was washed at a final stringency of 0.1X SSC, 0.2% SDS at 65 °C and exposed to Kodak X-OMAT film for various times.

Dosage Analysis

The copy number of probes in the NT microdeletion region was preferably determined by RFLP analysis of NT and his parents. For uninformative probes, quantitative analysis of Southern blots was performed. Five replicates of DNA from a normal individual, NT and a known 22q13.3 deletion patient (FB) were hybridized with a probe in the microdeletion region and a control probe simultaneously. The hybridization signal was quantified using a BioRad GS-670 Imaging Densitometer. The intensity of the bands were either measured by the peak value of a lane, or the signal density in the area of the band using Molecular Analyst version 1.4 program (Bio-Rad). The ratios between the intensity of the test and control probe were ranked from 1 (lowest) to 10 (largest). The ranks for the normal, NT, and FB DNA were summed and compared using a table of critical values for the Wilcoxon Rank-Sum test (Verdooren 1963). If the sum of ranks was lower than that of the critical value at the p-value 0.05 (95% confidence), the difference between the control and the two patients' DNA was considered significant. To test whether the patients' values were half that of the control, the ratios from NT and FB were doubled and the Wilcoxon Rank-Sum test was repeated. If the difference between control and deletion patients was now insignificant, it was concluded that the NT and FB DNA had one copy of the test probe and was therefore deleted in the microdeletion region.

BAL31 Analysis

Embedding genomic DNA in agarose

NT lymphoblastoid cells were suspended in HBSS to a final concentration 1.4 - 2.0 X 10⁷ cells/ml. An equal volume of melted 1% InCert agarose (FMC) was added to the cells, mixed and pipetted into a plug-former on ice. After the plug was solidified, it was incubated in ESP solution (0.5 M EDTA, 1% sodium laurosyl sarcosine, 1 mg/ml Proteinase K) for 2 days at 50 °C with gentle shaking. Proteinase K was then removed by

washing the plug two times in TE + 1 mM PMSF for 1 hour, followed by three times in TE for 1 hour. Agarose-embedded DNA was used to avoid DNA shearing.

BAL 31 exonuclease digestion

Plugs of 50 μ l were added to the BAL31 digest reaction mix and adjusted to a final volume of 1 ml. DNA was digested with 20 U BAL31 (New England Biolab) at 30 °C for 0, 0.5, 1, and 1.5 h. DNA plugs were then equilibrated in ice-cold 100 mM EGTA (ethylene glycol-bis[β -Aminoethyl ether] N, N, N', N'-tetraacetic acid) to inactivate BAL31, followed by an ice-cold TE pH 8.0 wash before digestion with EcoRV. DNA plugs were then loaded in a 0.6% agarose gel and separated by standard electrophoresis, Southern blotted, and probed with D22S163.

Sequence annotations on the cosmid contig

Four cosmid clones were sent to Dr Bruce Roe at University of Oklahoma for sequencing. Three clones (N85A3, N94H12, and N1G3) were completely sequenced and the sequences were submitted to GenBank (with accession number AC000036, AC002056, and AC002055 respectively). The sequencing of one clone (N66C4, AC000050) is in progress and five sequence contigs are now available in Genbank. The cosmid sequences were combined into one large sequence called "AWcontig" by the following procedures: The five sequence contigs of N66C4 were ordered and combined according to the information available. One fragment that was overlapping with N85A3 sequence was assigned as the distal end. The order of the rest of the sequence contigs was determined based on the position and order of exons in the cDNA clone sc24, which were located within N66C4. Cosmid sequences were entered into a local database by the program TOBLAST in the GCG package. The position of sequence overlap between cosmids was determined by the BLAST program. The overlapping sequence of the distal clone was removed, and the cosmid sequences were joined. Thus the resulting "AWcontig" contains at least four gaps in the N66C4 region, and a continuous region of 123.3 kb.

The putative expressed sequences were determined by exon prediction programs including GENSCAN (<http://CCR-081.mit.edu/GENSCAN.html>) and Grail 2 program through MacX in a UNIX environment. Grail 2 searches genomic sequence for opening reading frames (ORFs) that are bound by a pair of translation start/splicing donor, splicing acceptor/donor or splicing acceptor/translation stop sites. The quality of the exons are rated based on the presence of other genomic context information, such as splice junctions, or non-coding scores of 60-base regions on either side of a putative exon (Xu et al. 1994). The Genscan program was designed specifically for predicting exons in genomic sequence which has high G-C content. Therefore it is suitable for human or other mammalian genomic sequences. Instead of putting all putative exons into one large transcription unit, Genscan also determines multiple transcriptional units in one large genomic sequence (Burge and Karlin 1997). The predicted exons were translated into amino acid sequences (both programs have an installed translation function), and searched against a protein database through BEAUTY at the BCM search launcher ; (<http://kiwi.imgen.bcm.tmc.edu.8088/search-launcher/launcher.html>).

To search for the sequence in public DNA databases, human repetitive sequences were masked by the Repeat Masker program (Smit A.F.A. and Green P. Repeat Masker at <http://ftp.genome.washington.edu/RM/RepeatMasker.html>). The sequence was then sought in Blast nr (All non-redundant GenBank + EMBL + DDBJ + PDB sequences), Blast month (All new or revised GenBank + EMBL + DDBJ + PDB sequences), and Blast dest (Non-redundant database of GenBank + EMBL + DDBJ EST Division) through Gap-Blast at the NCBI web site (<http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-newblast?Jform=1>). The TIGR EST database was also searched through <http://www.ncbi.nlm.nih.gov/cgi-bin/THCBlast/nph-thcbblast>. The last update of the database search was on June 9, 1998.

The sequences of new cDNAs isolated by RT-PCR or cDNA library screening (see below) were aligned against the AWcontig sequence through BLAST in the GCG package.

Sequence Analysis

Multiple Sequences alignment

For the alignment of multiple nucleic acid sequences, Clustral W 1.7 program in the BCM search launcher (<http://kiwi.imgen.bcm.tmc.edu.8088/search-launcher/launcher.html>) was used. For alignment of multiple protein sequences, sequences were first aligned by the pairwise alignment program gap Blast (<http://www.ncbi.nlm.nih.gov/gorf/bl2.html>) using a BLOSUM62 matrix (Henikoff and Henikoff, 1992). The alignments were then combined into the multiple alignment matrix manually.

Determining the G-C content of exons predicted by genscan program

“Last exon” and “genscanex1” are the two exons predicted by the Genscan exon prediction program. The G-C content of these exons as well as two ALPR cDNA clones sc24 and I511 was evaluated with the GCG program WINDOW using a window of 100 bp and shifts of 10 bp.

PCR and RT-PCR

Primer Design

Primers for PCR or RT-PCR reactions were designed by using the Primer (v0.5) program through the UK HGMP Resource Centre web site (<http://www.hgmp.mrc.uk>). Criteria for choosing a primer were: low homology to human repetitive sequences, G-C content equal or higher than 50%, primer length between 20- 22 base pairs, and low number of complementary nucleotides at the 3' end. Primers that are on AWcontig are listed in Table 4. Two primers called 3'ARR and sc243 have sequences that are different from the genomic sequence (Table 4), due to a sequencing error of the clones. An internal primer is usually made from the sequence that is far away from the start of the sequencing, where the sequence quality is low. The primer AnkF6 is at position 98 to 117

Table 4 Primer sequences and their positions on AWcontig

Name of primer	Primer sequence (5' to 3')	Position on AWcontig
3'AR F	cct ctg cct ctg tga gtg tc	88090-88081
3'AR R	cat gtg agt gcc cat ttc cc*	86846-86865
55337F	cag atg agg cag cat gac ac	37647-37667
5533R	gtg tca tgc tgc ctc atc tg	37667-37647
Ank1 F1	gtc att gat gac aaa gtg gct g	43045-43066
Ank1 F2	agg gct ttg gtt ttg tgc tc	43088-43107
Ank F3	cta tgg gct ttt cca gcc gc	11708-11727
Ank F4	cgg caa gtt cct gga tga gg	11744-11763
Ank F5	ctg gag caa gtt cga cgt gg	69938-70007
Ank F51	cgc ttc gag gac cat gag at	70049-70068
Ank F6	aag gcg aga tcc cgc tgc ac#	
Ank1 R1	ctc agc tgt cat gga ctt gg	45010-44990
Ank1 R2	gca ggg tca gtg tgg tgc tg	44987-44968
Ank R3	cgc tcg atg ttc atg cgg tg	70155-70136
Ank R4	gaa gtc gtc ctt ggt aag cg	70108-70039
Ank R5	cgt cct ctt ctg gct tcc tt	43952-43933
Ank R6	tta gca ggt ccg tgg cgt tg	15387-15368
Ank R7	ctg tcc ttg tag tca ggt agg g	15627-15606
Ank R9	gag aca aga tgg aca gta gag g	55249-55228
Ank R10	gcc agc atc tca tcc agc tt	53809-53790
M1314F	gcc tga aga cga caa acc aa	61179-61158
M1314R	tgc tgc gat ggg cga ctt ct	61303-61284
M1314R2	aag tgt gtc cgt cag cga ag	60917-60898
OFH-1	tcc tca cac tca gcg atg acc	71522-71542
OFH-2	gac tgc tgc tgt gtg acc tgg	71812-71792
OFH-3	ctc ggg ggt ctt tga ctc tc	71496-71427
OFLS-1	ctg ctt ctc cct ccg ctc tt	62769-62788
OFLS-2	cct ttt cca act ctc agc ca	62590-62609
OFLS-3	ctc agg ggt ctg gtc ctg ta	62722-62703
OFLS-4	cag gcc agg cac tgt gct at	62855-62836
R1F	cca agt cca tga cag ctg ag	44990-45010
R7F	ctt cac ctg act aca agg aca g	14606-15627
sc 243	aat cgc cac cgc ttc gca ta@	34025-34006
sc 245	gcc atc atc gca ggg aac tt	23796-23815
U2F	ggc ttt cca cag cgc ggg gg	94338-94310
U2R	tgc ctc act gct cag gac cc	93527-93546
E0.91F	tcc tct gat ggg gtc tcg at	107668-107687
E1.03F	ttc ctt ggg ctt ctc ctg ag	107253-107234
E1.03R	ctc agg aga agc cca agg aa	107234-107253
pul 51	cca gag cat gca tgc ctc ct	1144690-114671
pul 31	cat act agc aag cca caa gta	106473-406494
pul 32	ggg agt tct gtt tgt aag aca c	107946-107966
IIF	ggg tgg gtg gtg ccc ttt ta	107632-107611
RABLF1	cca cag cag ctg tcc acg ta	115615-115596
RABLR1	cac cca gac ttg gtt aca ag	114544-114563

*Primer sequence is different from AWcontig, which is cat gag agt gcc cat ttc cc

@Primer sequence is different from AWcontig, which is agt cgc cgc cgc ttc gca ta

These two primers were designed before the genomic sequence AWcontig was available. The discrepancy between the genomic and cDNA sequences is due to sequencing error in the cDNA sequence.

This primer sequence is not present in AWcontig because the cDNA sequence where this primer is not located on AWcontig.

of I511, and does not match anywhere in the genomic sequence (see the result section “Cloning of ALPR”). For the *C. elegans* C33B4.3 gene study, since it was not necessary to search for the homology to human repetitive repeats, primers were designed by the Primer 3 program (<http://www-genome.wi.mit.edu/cgi-bin/primer/primer3.cgi>). Table 5 lists the primers on *C. elegans* cosmid DNA C33B4.

Total RNA isolation for reverse transcription

Various human tissues were provided by Dr Jeff Oulette in the Department of Pathology, University of Alberta Hospital. Human fetal tissues were provided by the Human Fetal Tissue Bank, through Dr Steve Bamforth in the Department of Medical Genetics. Total RNA from the tissues was extracted by using TRIzol reagent (GIBCO-BRL) following the manufacturer’s instructions. Briefly, 10 ml of TRIzol reagent were added to 1g of tissue, or 1 ml of TRIzol reagent to 10 cm² of cultured cells. Tissues were homogenized in TRIzol reagent using a homogenizer. The homogenized samples were incubated at room temperature for 5 min to permit the complete dissociation of nucleoprotein complexes. To the homogenized sample 0.2 ml of chloroform per 1 ml of TRIzol was added, followed by vigorously shaking and incubation at room temperature for 3 min. The samples were centrifuged at 9500 rpm for 15 min at 4 °C. The aqueous layer was drawn from the samples and mixed with 0.5 ml isopropanol per 1 ml of TRIzol reagent. The samples were spun again. The RNA pellets were washed with 75% ethanol and dissolved in various volumes of DEPC-treated dH₂O. The concentration and purity (A_{260}/A_{280} ratio) of total RNA were determined by a photospectrometer (Gene Quant, Pharmacia). Total RNA from human fetal tissues was prepared by Dr Valérie Trichet.

Reverse transcription

Approximately 1 µg of total RNA from various tissues was mixed with 10 pmoles of reverse transcription primer (gene specific primer), and DEPC- treated dH₂O to a final volume of 20 µl. The mixture was heated at 70 °C for 5 min followed by chilling on ice for 1 min. After the total RNA and primer were denatured, 4 µl of 5X “first strand buffer”

Table 5 Primer sequences and their positions on *C. elegans* cosmid C33B4.3

Primer name	Primer sequence (5' to 3')	Position on C33B4
B4.3F1	gga ctg cag ccg atc aat gtt ctt acc gga aga	30019-29994
B4.3FP	agc ttc aac acc tca acc aa	21781-21762
B4.3H	ttc cgc aag ctt tca act ac	25070-25051
B431HR1	gaa cta ttg tgc atc gac cg	23165-23146
B4.3P	gat gga gct gca gac att gg	23657-23676
B4.3R1	ggg gat cca cac cag gat ttc cga caa tat gc	23893-23918
B4.33	tgg aat cac cag aat ccg ag	23512-23531
B4.332	ccg acc aga cat cta cct tc	19639-19658
B4.333	aaa tcg aga ccg atc gca ct	19526- 19545

(GIBCO-BRL), 2 μ l of 0.1 M DTT, and 1 μ l of 10 mM dNTP mix was added. This was pre-incubated at 42 °C before 1 μ l (200 units) of reverse transcriptase SuperScript II (GIBCO-BRL) was added. The reaction mix was incubated at 42 °C for 30 min, followed by incubating at 55 °C for 5 min. The RNA strand was digested by adding 1 μ l (3.8 units) of RNase H (GIBCO-BRL). The reaction was further incubated at 55 °C for 10 min. The final volume of the reverse transcription reaction is 20 μ l, and 1 μ l was used for RT-PCR.

PCR

PCR was performed in either “1.1X” buffer (final concentration = 50 mM Tris-HCl pH 9.0, 1.5 mM MgCl₂, 0.4 mM β -mercaptoethanol, 0.1 mg/ml non-acetylated BSA [Boehringer Mannheim], and 0.2 mM dNTPs) or GIBCO BRL 10X buffer (final concentration = 20 mM Tris-HCl pH 8.4, 50 mM KCl, 1.5 mM MgCl₂, and 0.2 mM dNTPs) with 20 pmoles forward and reverse primers, 2.5 U of *Taq* polymerase, and either 100 ng of genomic DNA, 1 ng of cloned DNA, or 1 μ l of reverse transcription product as template. PCR was run in 30 cycles using a PTC-100 thermocycler (MJ Research) with various annealing temperatures and extension times depending on the T_m (melting temperature) of the primers and the size of the target products. If the reaction required “hot start”, the reaction was heated for 2 min at 95 °C before *Taq* polymerase was added. For PCR, one negative control (no template control) was run with the actual PCR reactions. For RT-PCR, one no template control, and one no RT control (an RT reaction was set up without the addition of reverse transcriptase) were used to monitor the contamination of PCR by genomic DNA and the PCR product which was amplified by the same pair of primers before. For amplifying cDNA clones which require high sequence fidelity, *TaqPlus* Precision polymerase (Stratagene) was used. The reaction was composed of 5 μ l of 10X *TaqPlus* Precision buffer, 0.4 μ l of 25 mM dNTPs, 20 pmol of forward and reverse primer, template DNA, 2.5 U *TaqPlus* Precision *Taq* polymerase, and dH₂O to a final volume 50 μ l. Table 3 listed different RT-PCR and PCR products.

Subcloning PCR products

Most PCR products were subcloned in pGEM-T or pGEM-T Easy vectors (Promega) following the manufacturer's instructions. Exceptions were B4.3HP, which was subcloned into pBluescript SK+ (Stratagene) with HindIII and PstI sites; and B4.3F-R, which was subcloned into various *C. elegans* reporter gene vectors (see the "Gene Expression Studies in *C. elegans*" section). Ligation reactions were incubated at room temperature for 2 h to overnight. Transformation was performed with competent cells XL1-Blue or XL10-Gold (Stratagene), by conventional methods (Sambrook et al. 1989).

cDNA Library Screening

Probes and cDNA libraries

Probes included cDNAs amplified by RT-PCR, EST clones, and cDNAs that were isolated from the cDNA library. Table 6 lists the probes and cDNA libraries used in this study.

The original titres of the phage cDNA libraries were determined by plating out different concentrations of the library with XL1-Blue plating cells. Twelve to fifteen 15 x 150 mm plates with 5×10^4 pfu/plate were used for plaque lifts. Plaques were transferred to Hybond-N nylon membranes (Amersham) following the manufacturer's instructions. Membranes were hybridized with radioactive cDNA probes in 5X SSPE, 5X Denhardt's solution, 0.5% SDS overnight at 65 °C. Membranes were washed at a final stringency of 0.1X SSPE, 0.1% SDS at 65 °C and exposed to Kodak X-OMAT film overnight. Primary positives were plated out at 100 pfu/plate and the screening was repeated. Usually, a third round of screening (tertiary screen) was required to confirm the plaque was from a single clone.

Table 6: cDNA libraries and probes used for screening the libraries

Probes	cDNA library	Source of the library
F3-R7, 55337F-R1,	Human fetal brain in λ ZapII	Stratagene Inc.
FLS, FL2, I511,		
R1F-M1314R, F5-R3, pul		
F3-R7, I511, FL2,	Human colon carcinoma	Dr. J Rommens
R1F-M1314R, F5-R3, pul	cell line Caco-2 in λ Zap	
FLS	Human fetal liver in λ gt11	Clontech Inc.
FL2	Human adult heart in λ gt11	Dr. R. Farahani
MB	Mouse brain in λ ZapII	Dr. J Rommens

In vivo excision of insert into pBluescript SK+ phagemid

In a 50 ml tube, 200 μ l of $OD_{600}= 1.0$ XL1-Blue cells, 200 μ l of phage stock ($> 5 \times 10^5$ pfu/ml), and 1 μ l of R408 helper phage ($> 1 \times 10^6$ pfu/ml) were mixed and incubated at 37 °C for 15 min. The co-infected culture was then added to 5 ml of 2X YT media (10 g NaCl, 10 g Yeast Extract, 10 g Bacto-Tryptone) and incubated 3 h at 37 °C with shaking. The culture was then heated at 70 °C for 20 minutes. The debris of ruptured *E. coli* cells was spun down at 4000 g for 5 min. The supernatant that contained the pBluescript phagemid was transferred into a new tube, and 10 μ l of phagemid stock was incubated with 200 μ l $OD_{600}= 1.0$ XL1-Blue cells at 37 °C for 15 min. Different amounts of cells were plated out on LB/ampicillin plates and incubated overnight at 37 °C.

Confirmation of the Authenticity of cDNA Clones

Cosmid DNAs were digested with different enzymes, electrophoresed and Southern blotted. cDNA isolated from different sources (RT-PCR, EST clones, cDNA library screening) were used as probes and hybridized to the cosmid blots. Clones that showed positive signals, were further hybridized to a Southern blot which contained total human DNA, DNA from the Human Genetics Mutant Cell Repository human/rodent somatic cell hybrid mapping panel #2, and hamster cell line RJK88 DNA digested with different enzymes. This procedure was to test the copy number of the gene in human genome. To further characterize the genomic regions where RABL2 and RABL22 are located, another partial hybrid panel southern blot was prepared by digesting DNA isolated from cell line GM/NA10826B (chromosome 2 hybrid cell line), GM/NA10888 (chromosome 22 hybrid cell line), normal human (GM03657), and hamster cell line RJK88 with HindIII, BamHI, and EcoRI. The membrane was then probed with pul.

Sequencing of cDNA clones

Authentic cDNA clones were sequenced either by an ABI 373A automated sequencing machine in Dr Ken Roy's laboratory in Department of Biological Sciences, or by using the Thermo Sequenase radiolabeled terminator cycle sequencing kit (Amersham). For manual cycle sequencing, the sequencing reactions were run on an 8% polyacrylamide gel with 7M urea and 40% formamide to eliminate the compressions at the GC rich region of the sequences. The gel was soaked in 5% acetic acid/ 15% methanol to remove the urea and formamide before drying in a gel vacuum dryer at 80 °C for 2 h. The sequencing gel was exposed overnight with a Kodak BioMax MR film.

One 10 h run of manual cycle sequencing could read sequence up to 400 bp, the high quality sequence range is ~350 bp. Sequencing reactions were usually started from the primers in the vectors that flank the multiple cloning site (T7 or T3), or the forward and reverse primers if it was a PCR product. If sequences from both ends were not overlapping, internal primers were used to fill in the gaps. For large clones, inserts were digested with various enzymes and the resulting fragments were subcloned into pBluescript SK+. This approach reduced the number of internal primers needed to fill up the gaps. The internal primers and subcloned fragments are listed in Table 7

Northern Analysis

Human multiple tissue Northern Blots were purchased from Clontech. The Northern Blots were first prehybridized at 42 °C with 5 ml hybridization solution (50% formamide, 5X SSPE, 2X Dehardt's reagent, 0.1% SDS and 100µg/ml sheared herring sperm DNA) for 1 hr. The hybridization solution was then replaced by probes in 5 ml fresh hybridization solution and hybridized overnight. The final washes of the blots were in 0.1X SSPE, 0.1% SDS at 50 °C. The blots were exposed to x- ray film at -70 °C for 1 to 1 1/2 weeks.

Table 7: Subclones of cDNA clones and internal primers used for manual sequencing

Name of clone	Size of clone	Subclones	Internal primers
sc24	1268 bp	None	R7F, sc245, sc243
I511	2008 bp	I511HXh0.9 I511H0.4	Ank F6, 5533R, 55337F, R1F, Ank R9, Ank R10 Ank F1, Ank R1
AR	~2 kb	AR 1.2, AR 0.7	OAR5, OAR4, OAR3, OAR2, OAR1*
FLS	524 bp	None	OFLS1
FL2	1160 bp	None	OFH1, OFH2
MB	991 bp	None	OMB-1
pul	1056 bp	None	pul51, pul32
hc 5	1712 bp	None	pul31, E1.03F, E0.91F
hc 7	757 bp	None	pul31, E0.91F
hc 8	757 bp	None	pul31, E0.91F
B431HR1- B4333	1541 bp	None	B4.3FP

* primer sequences (from 5' to 3') are as follows:

OAR-1: ctg ggt tac att cag att aca g
OAR-2: aat tcc agc agg gta aaa cgg g
OAR-3: cat ccc tcg agg tac gat gc
OAR-4: atc ctt ctt ctt ccg gac cg
OAR-5: gca ccg acc aca gcc ctc ac

Comparison of RABL2 and RABL22 Expression Level

Reverse transcription was done with pul31 as the gene specific primer and total RNA from cell line GM/NA10826B (chromosome 2), GM/NA10888 (chromosome 22) and various human tissues as templates. PCR was done in triplicate for each tissue. PCR conditions were 25 cycles of 92 °C for 30s, 63 °C-0.1 °C/cycle for 30s, 72 °C for 30s with hot start and final extension at 72 °C for 5 min. Reactions were then run on a 1% agarose gel with TAE (40 mM Tris pH 7.2, 20 mM Sodium Acetate, 1 mM EDTA-Na₂) electrophoresis buffer. PCR products were excised from the gel, purified by GENECLEAN (Bio101, La Jolla, CA), and digested with 10U BsrG1 (New England Biolab) at 60 °C for 4 hours to overnight. Digested products were then run on a 1% agarose gel. The band intensity of the BsrG1-uncleaved product (RABL2) and BsrG1-cleaved product (RABL22) were compared by eye. As a control experiment, partial cDNAs of RABL2 (use 1 ng of pul as template) and RABL22 (use chromosome 22 cell line GM/NA10888 reverse transcription product as template) were amplified and subcloned into pGEM-T Easy vector (Promega). cDNA clones were sequenced to confirm that no sequence errors were introduced during amplification. Approximately 0.001ng of RABL2 and RABL22 cDNA clones were mixed in 1:0, 2:1, 1:1, 1:2, and 0:1 ratio and amplified using the same conditions as described above. Band intensities between PCR products that were cleaved (RABL22) and were not cleaved (RABL2) by BsrG1 were compared by eye to determine whether the ratio of PCR products of the RABL2 and RABL22 clones matched approximately that in the ratio of the starting templates DNA.

3' Rapid Amplification of cDNA Ends (RACE)

Reverse transcription was performed by using ~1µg of heart total RNA and ~ 10 pmol. GIBCO BRL polyT adapter primer (5' GGC CAC GCG TCG ACT ACT TTT TTT TTT TTT TTT T 3'). The reverse transcription product was purified by a GlassMax column (GIBCO BRL) to remove polyT adapter primers. Purified reverse transcription product

was eluted in a final volume of 40 μ l, and 5 μ l was used as a template for PCR. The PCR reaction was done using pul51 and GIBCO BRL universal amplification primer (5' GGC CAC GCG TCG ACT AC 3'), and *TaqPlus* Precision *Taq* polymerase (Stratagene) in 30 cycles of 92°C for 30 s, 62°C- 0.2°C/cycle for 30 s, and 72°C for 2 min with hot start. PCR products were run on 1% agarose gel with TAE electrophoresis buffer. Two major bands (1.7 and 0.7 kb) were excised from the gel, purified by GENECLEAN and subcloned into pGEM-T Easy vector (Promega).

Gene Expression Studies in *C. elegans*

Preparation of reporter gene fusion protein constructs

A 6129 bp genomic DNA fragment of *C. elegans*, which consists of a ~4.7 kb putative promoter region and four exons of *C. elegans* gene C33B4.3, was amplified by PCR using primers B4.3F/ B4.3R, *C. elegans* genomic DNA, and *TaqPlus* Precision *Taq* polymerase (Stratagene). PCR conditions were 30 cycles of 68 °C - 0.3 °C/cycle for 8 min, 94 °C for 30s with hot start. The PCR product was digested with BamHI and PstI and subcloned into reporter gene vectors constructed by Dr Andrew Fire at the Department of Embryology, Carnegie Institution of Washington, Baltimore, MD ([ftp: ciw1.ciwemb.edu](http://ftp.ciw1.ciwemb.edu)). The reporter gene vectors included pPD95.75, an in-frame promoterless green fluorescence protein (*gfp*) vector; pPD95.77 and pPD95.79, two out-of-frame promoterless *gfp* vectors; pPD95.03, an in-frame promoterless *lacZ* vector with nuclear localization signal (NLS); and pPD95.57, an out-of-frame promoterless NLS-*lacZ* vector. Transformation was done with competent cell XL1-Blue, and mini-prep DNA was done without RNase treatment. DNA from four individual clones of the same construct was pooled to reduce the possibility of non-functional construct due to sequence error introduced during PCR amplification. The pooled DNA was purified through a GlassMax column (GIBCO BRL) to remove most of the RNA in the sample.

Examination of fusion protein expression pattern

The *gfp* constructs were co-injected with plasmid MMO16B (the plasmid clone that contains the wildtype *unc-119* gene) into the gonad of the hermaphrodites strain *unc-119 (ed3); him-8* by Dave Hanson. The *lacZ* constructs were also co-injected with MMO16B into the gonad of the hermaphrodites of the same strain by Angela Johnson. Worms were grown in NGM agar plates with a lawn of OP50 (Wood 1988). Wild type progeny indicated that the extrachromosomal constructs were transmitting to the next generations. To look for *gfp* expression, worms were examined directly using fluorescence microscopy with a Zeiss Axioskop Routine Microscope. To examine *lacZ* fusion protein expression, X-gal reaction staining was applied: worms were washed off the plates with M9 buffer (3 g KH_2PO_4 , 6 g Na_2HPO_4 , 5 g NaCl, 1 ml 1M MgSO_4 , dH_2O to 1 liter), pelleting by quickly centrifuging at maximum speed, and then washed one time in M9 to remove bacteria and agarose debris. The worms were finally resuspended in 25 μl of M9 buffer. Microscope slides with size 75 X 25 mm were coated with poly-L-lysine by spreading 0.1 mg/ml stock onto the slides and incubating them for 30 min at 37 °C. Worms were transferred onto the poly-L-lysine coated slides and covered with 40 X 22 mm glass cover slips. The slides were frozen on metal containers in the -80 °C freezer for 15 min. After the freezing, the cover slips were quickly cracked off with a single-sided razor blade, and the slides were immediately submerged for 5 min in 100% acetone prechilled to -20 °C. The slides were washed for one minute each in 75%, 50%, and 25% acetone and then allowed to dry briefly at room temperature. To the slides 100 μl of "Mix S" (620 μl dH_2O , 250 μl 0.8 M Na-phosphate buffer pH 7.5, 1 μl 1 M MgCl_2 , 4 μl 1% SDS, 100 μl 100 mM Redox buffer [100 mM Potassium Ferricyanide, 100 mM Potassium Ferrocyanide], 15 μl 5 mg/mL kanamycin, 2 μl 1 mg/ml DAPI, 8 μl 3% X-Gal in N, N -dimethylformamide)(Fire, 1992) was added, covered by cover slips, and the cover slips were sealed with nail polish. The slides were incubated at room temperature on moistened Whatman chromatography paper (3 MM) in petri dishes for 1 h. The

worms were examined for β -galactosidase staining using bright-field microscopy and using fluorescence microscopy for DAPI staining.

Integration of Extrachromosomal Arrays using UV irradiation

Transgenic worms that carried the *lacZ* fusion protein construct were selected for young adult hermaphrodites, and ten young worms were picked onto each of five plates (fifty in total). The plates were irradiated under UV at 350 μ J using a UV stratalinker 1800 machine (Stratagene). The irradiated P₀ worms were allowed to lay eggs for three to four days. Equal numbers (i.e. 40) of young F₁ worms were picked from each parental plate onto single plates to generate a total of 200 plates. The worms were allowed to produce an F₂ brood. Plates were scanned for the highest number of wildtype progeny. The best ten plates were selected and the rest of the 190 plates were discarded. For these pools of ten plates, ten single F₂ worms from each plate were transferred to new plates. The worms were allowed to produce F₃ broods. The F₃ plates were then screened for 100% wildtype progeny.

Inhibition of Specific Gene Transcription by Double-stranded RNA Interference

The procedure of double-stranded RNA interference was as described by Fire et al. (1998). Plasmid clone B4.3HP, which contained a partial cDNA of C33B4.3, was linearized by HindIII or SmaI endonuclease digestion. The insert of the plasmid that contained *apx-1*, a positive control gene, was amplified by PCR with T7 and T3 primer. *In vitro* transcription was done to produce the sense and antisense RNAs of both genes using a MEGAscript kit (Ambion) following the manufacturer's instruction. Double strand RNA was prepared by mixing equal amount (i.e. 5 μ l) of sense RNA, antisense RNA and microinjection buffer (20 mM KPO₄ pH 7.5, 3 mM Kcitrate, 2% PEG6000). The RNA mix was incubated at 68 °C for 10 min, followed by 37 °C for 30 min. Double-stranded RNA was microinjected into the gonad of wildtype (N2) hermaphrodites by Angela Johnson. The hermaphrodites were transferred to new plates each day after the

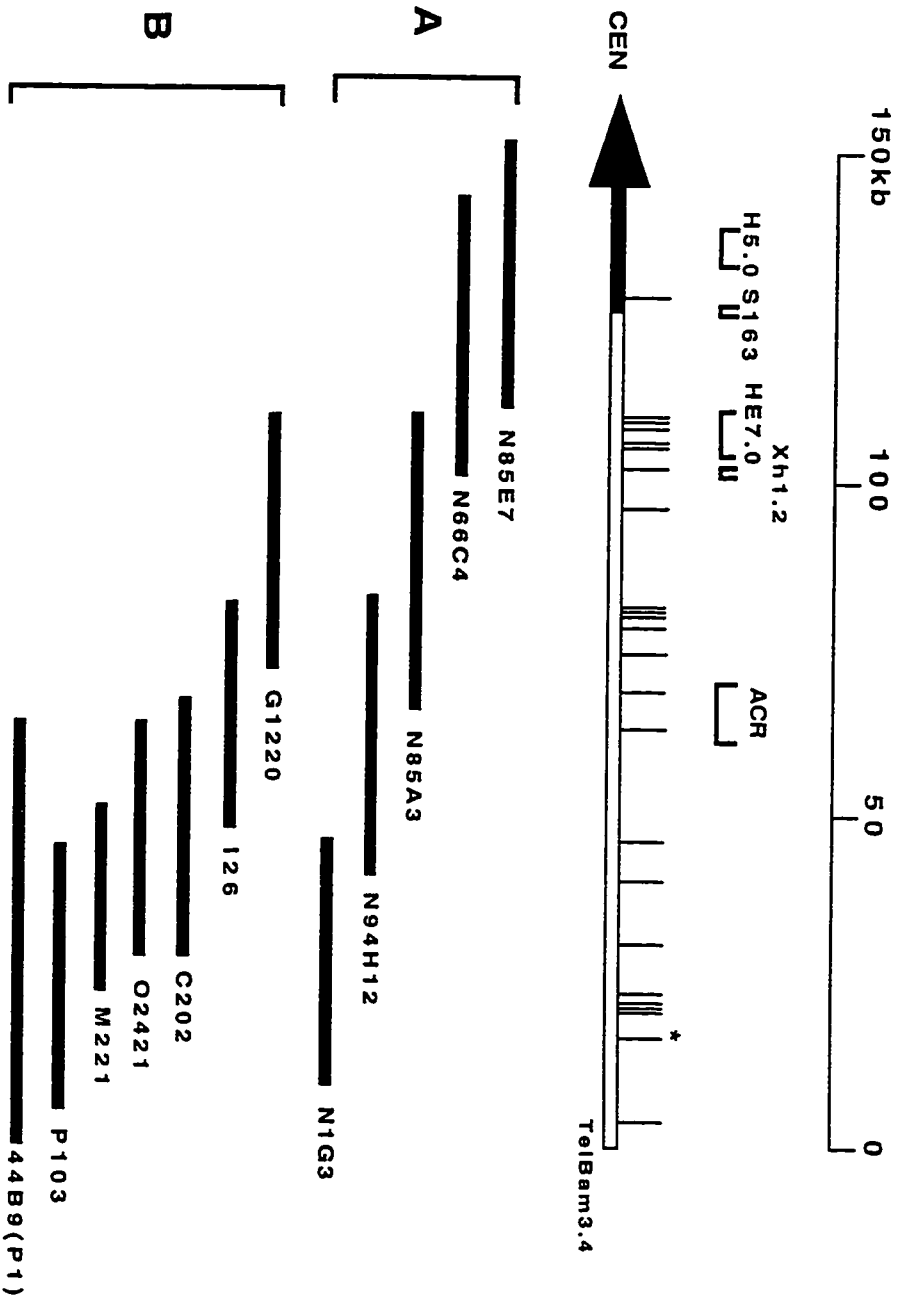
microinjection up to four days. Progeny were examined under the dissecting microscope for any observable change of phenotype.

Results

Production of a Cosmid Contig from D22S163 to the 22q Telomere

To characterize the NT microdeletion, a cosmid contig was constructed that spans the deleted region. A cosmid walk was started from the cosmid clones N85E7 and N66C4, which were isolated using a probe for D22S163, which was previously shown to be deleted in NT (Flint et al. 1995). Cosmid end fragments were isolated by detecting SmaI-digested vector-insert junction fragments with a vector probe on Southern blots. These vector-insert junction fragments were then used to identify XhoI and XhoI/HindIII fragments flanking but not including vector sequences. End fragments of these cosmids were tested by densitometric or RFLP analysis, to assess the copy number in NT. Densitometric analysis showed that the proximal end of N85E7 was not deleted in NT (data not shown). However, RFLP analysis using the distal end fragment of N66C4 (Xh1.2, Fig. 3) showed that the paternal allele was deleted in NT (Fig. 4). The cosmid walk was therefore extended from the distal end of N66C4 for three rounds of end-fragment walking (Fig. 3, contig A). At the same time a second cosmid walk was being done by Dr Yi Ning at NIH. It started from a P1 clone containing the 22q putative telomere associated sequence TelBam 3.4 and walked proximally (Fig. 3, contig B) (Ning et al. 1996). We agreed to combine the two contigs, which overlapped extensively (Wong et al. 1997). I constructed a BamHI restriction map of both contigs by using BamHI fragments as probes against HindIII and/or SmaI cosmid digests electrophoresed and Southern blotted. This allowed unambiguous assignment of contiguous BamHI fragments. Restriction mapping of both contigs revealed that the two contigs overlap by about three cosmid lengths (~100 kb) and span ~150 kb between the proximal end of N85E7 and the putative 22q telomere.

Figure 3 Cosmid/P1 contig, spanning the NT microdeletion. The upper bar (with the arrow) shows the chromosome, with the deleted region represented by an open bar. Vertical lines on the chromosome represent BamHI sites. *A*, cosmid contig, constructed by screening the chromosome 22-specific library LL22NC03. *B*, cosmid contig, constructed by screening the chromosome 22-specific library ICRFc106 (modified from Ning et al. 1996). In contig B, the full name of each cosmid is "ICRFc106". followed by the designations shown in the figure. Locus D22S163 is abbreviated as "S163". The BamHI site marked with an asterisk (*) was found only in 44B9. The clone 44B9 was isolated by screening a P1 genomic library with the telomere associated sequence TelBam3.4



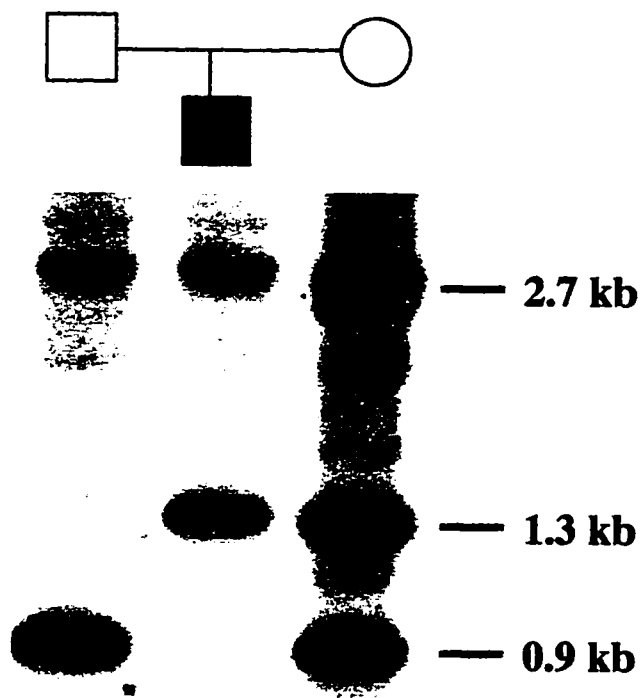


Figure 4 RFLP analysis, showing that Xh1.2 is deleted in NT. The autoradiograph shows the NT family genomic DNA digested with TaqI and probed with Xh1.2. Each individual shows a constant 2.7 kb band. The father is homozygous for the 0.9 kb allele, and the mother has both 1.3 kb and 0.9 kb allele. NT inherited the 1.3 kb allele from the mother but no allele from the father.

Localization of the Breakpoint of the NT Microdeletion

Since the proximal end of N85E7 was not deleted, while the distal end of N66C4 showed a deletion of paternal allele by RFLP analysis, these results indicated that the breakpoint of the NT microdeletion is within the N66C4 and N85E7 region. An additional probe (H5.0), which is 6.5 kb proximal to D22S163 (Fig. 3), showed no deletion by densitometric analysis (data not shown), further narrowing down the breakpoint. In order to look for DNA rearrangement fragments indicative of the breakpoint, genomic DNAs from a normal control and the NT family were digested with HindIII, SmaI, EcoRV, and NgoMI, electrophoresed and blotted. Since Flint et al. (1995) had shown that D22S163 was deleted in NT, the blot was hybridized to a 6.5 kb fragment that was adjacent to the H5.0 probe. This fragment also contains a part of the D22S163 locus. The fragment detected novel rearrangement bands in all different restriction endonuclease digestions of NT genomic DNA. In order to confirm that the D22S163 locus is deleted in NT, the same blot was probed with D22S163. Contrary to what I expected, the D22S163 probe showed the same hybridization pattern as the 6.5 kb fragment: novel bands for all four enzyme digests. D22S163 detected the presence of a large rearrangement band in NT with EcoRV and NgoMI digests (Fig. 5). For HindIII and SmaI the novel bands were smeared in the 9-11 kb (Fig. 5). This smear is characteristic of breakpoint -junction fragments fused with telomeric sequences that are heterogeneous in length (Wilkie et al. 1990b). D22S163 also detected the presence of a large rearrangement band in NT with EcoRV and NgoMI digests (Fig. 5). The presence of these wide rearrangement bands suggested that NT has a terminal deletion and that the breakpoint is close to D22S163. One discrepancy between this study and that of Flint et al. (1995) is that I detected HindIII-, SmaI-, EcoRV-, and NgoMI-rearrangement fragments with the D22S163 probe, while Flint et al. (1995) detected a paternal deletion with the same probe using Sau3AI. I hypothesized that under normal electrophoretic conditions, the smaller Sau3AI smeared rearrangement band could be diffused to the point of being not visible with a standard exposure (Fig. 6A). The Sau3AI digestions of

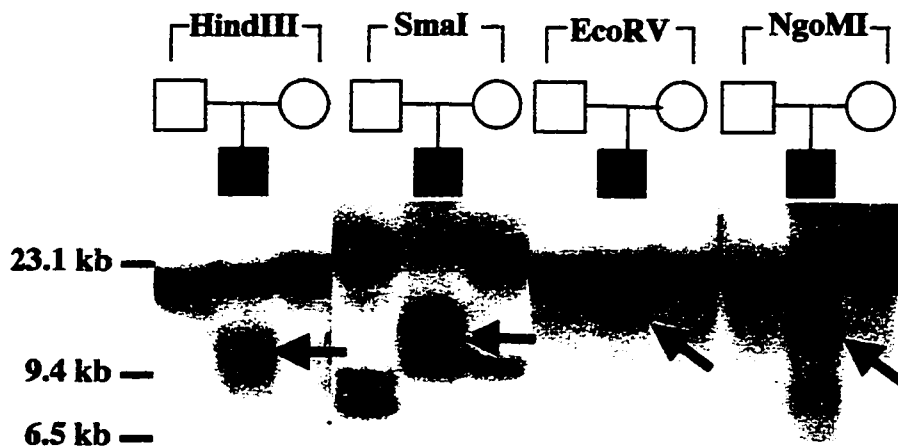


Figure 5 Rearrangement bands, detected by the D22S163 probe. The autoradiograph shows the NT family genomic DNA digested with HindIII, SmaI, EcoRV, and NgoMI, followed by hybridization to the D22S163 probe. Novel smeared bands in 9-11 kb range were present in HindIII and SmaI digest of NT DNA (arrows). SmaI also identifies a polymorphic locus with a 9.4-kb allele in NT and his mother and 8.0/9.0 kb alleles in the father, which confirms the paternal deletion of NT. D22S163 also detected the presence of large rearrangement bands in NT with EcoRV and NgoMI (arrows). Slight differences in the apparent size of the large bands common to all individuals are due to anelectrophoresis artifact.

the NT family DNAs were repeated and the fragments were electrophoresed over a short distance to maximize the chance of seeing a smeared rearrangement band. Under these conditions, a novel smeared band was seen in NT (Fig. 6B), indicating that D22S163 contains or is adjacent to the breakpoint rather than being deleted. This also confirms the explanation given by Flint et al, (1995) as to why cosmids containing D22S163 did not show a deletion for NT when FISH was used, since only part of the cosmid would have been deleted.

BAL31 Analysis

If NT is a terminal deletion, then the rearrangement fragment should be sensitive to BAL31 exonuclease digestion. When agarose-embedded NT DNA was subjected to BAL31 digestion from 0 to 1.5 h, followed by EcoRV digestion, the D22S163 probe detected a reduction in the size of the rearrangement band (Fig. 7). The larger, normal band detected by D22S163 served as a control to show that the genomic DNA was intact. Since D22S163 is normally located 130 kb from the telomere, it is not sensitive to BAL31 digestion on a normal chromosome.

Localization of ACR to the NT Microdeletion

Pulse field gel electrophoresis (PFGE) studies of the region indicated that ARSA, D22S163, and the acrosin locus (ACR) all mapped to a 190 kb Not I fragment (McDermid, unpublished data). When the ACR cDNA probe was hybridized to a Southern blot containing BamHI digests of all the cosmid DNAs in the contig, fragments of 4.6, 6.9 and 17 kb in cosmids N94H12, I26, and C202 were positive. The BamHI fragment pattern matched the ACR restriction map reported by Vazquez-Levin et al. (1992). Therefore, the ACR locus maps to the NT microdeletion, ~60 kb distal to D22S163 and ~70 kb from the telomere (Fig. 3). Sequence analysis results also confirmed this finding (see below). Densitometric analysis confirmed that ACR is deleted in NT (Fig. 8), as well as in the 22q13.3 deletion patient FB.

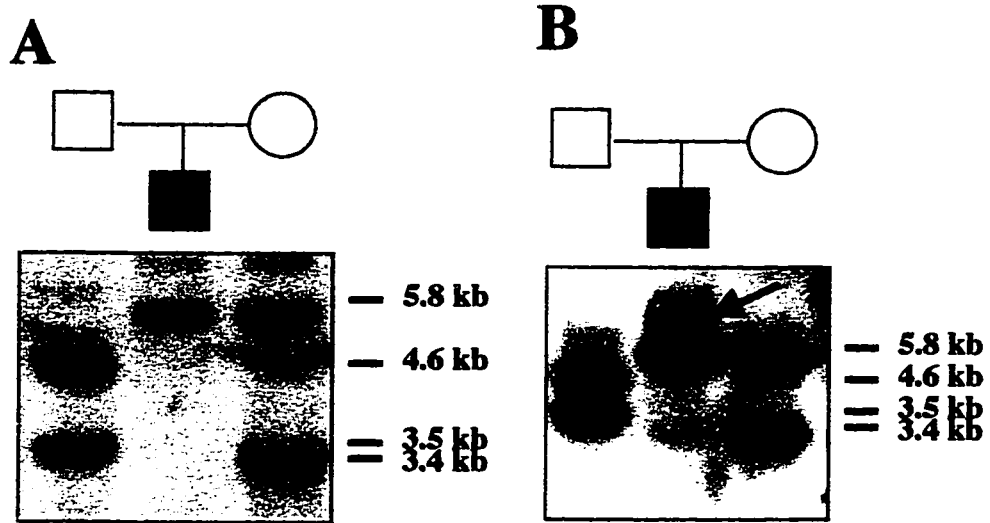


Figure 6 Sau3AI-digested NT family genomic DNA, hybridized to the D22S163 probe under two different electrophoretic conditions. In A, using conditions reported in Flint et al. (1995), NT appears to inherit one 5.8 kb allele from his mother but no allele from his father, suggesting a paternal deletion of D22S163 in NT. In B, electrophoresing the DNA over a much shorter distance and exposing the autoradiograph for a longer time, D22S163 detected a smeared rearrangement band in NT (arrow), indicating the microdeletion breakpoint is close to, or within D22S163.

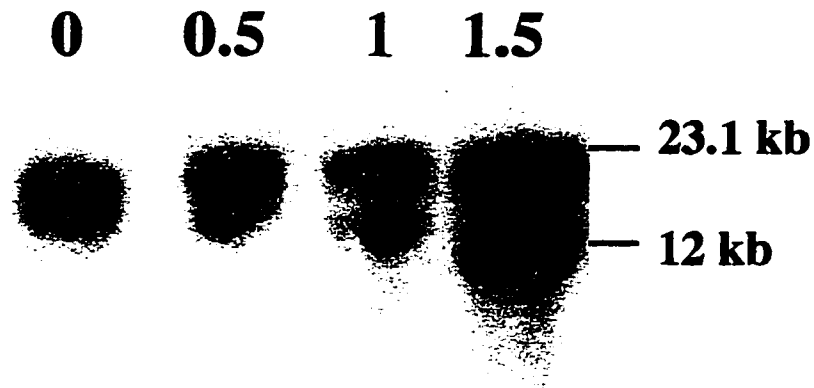


Figure 7 BAL31 sensitivity of the rearrangement band in NT. Genomic DNA from NT was digested with BAL31 before digestion with EcoRV. The number at the top of each lane indicates the incubation time of BAL31 (in hours). DNA was then separated by standard electrophoresis in a 0.6% agarose gel. The resulting blot probed with D22S163 shows a decrease in size of the rearrangement band after BAL31 digestion. The positions of DNA size markers from lambda phage digested with HindIII and the 1 kb ladder (GIBCO-BRL) are shown on the right.

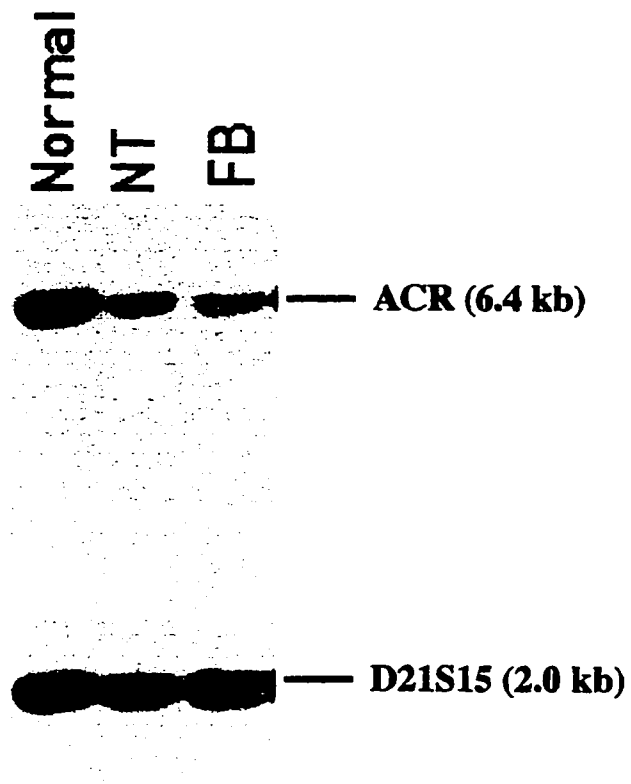


Figure 8 Densitometric analysis of ACR. The copy number of ACR was compared by densitometric analysis of TaqI digested DNA from a normal individual, NT, and FB. The ACR gene probe was compared to a reference probe from chromosome 21 (D21S15). Both NT and FB indicate the presence of only one copy of ACR.

Sequence Analysis of the NT microdeletion region

Four overlapping cosmid clones (N66C4, N85A3, N94H12 and N1G3) were sent to Dr Bruce Roe at the University of Oklahoma Advanced Center for Genome Technology for sequencing by the “shotgun” method. The cosmid sequences were assembled into one continuous sequence as described in the Materials and Methods. This continuous sequence was called AWcontig. Since N66C4 was not completely sequenced at the time of writing, the AWcontig contains gaps in this region. The total length of N66C4 in the database is 43.9 kb, which is the average size of a cosmid. Therefore, the size of the sequence in the gaps is likely to be small. Restriction mapping analysis showed that there is ~12 kb between the distal end of N1G3 and the putative 22q telomere (Fig. 3).

The AWcontig was put through different sequence analysis programs such as Repeat Masker, Blast search against various public databases, and exon prediction programs Grail v1.2 and Genscan. The results are summarized in the Appendix. Figure 9 shows the genomic organization in the NT microdeletion region, excluding the unsequenced 12 kb at the distal end of AWcontig. However, distance calculation for loci relative to the putative telomere include this 12 kb. ACR is ~70 kb proximal to the telomere, compared to the 65 kb estimated by restriction mapping (Fig. 3). D22S163 originally estimated to be ~130 kb proximal to the telomere, is now ~140 kb. However, the position of D22S163 relative to the telomere is still an estimate, because N66C4 was not completely sequenced and the most distal 12 kb was not sequenced.

Distribution of interspersed repeat sequences in the microdeletion region

In this study, the human interspersed repeats were divided into two groups. The first group consisted of Short Interspersed Nuclear Elements (SINEs) and Long Interspersed Nuclear Elements (LINEs), which are the biggest fraction of human interspersed repeats (28% of the total human genome; Smit [1996]). The second group consisted of retrovirus-like elements which are characterized by long terminal repeats

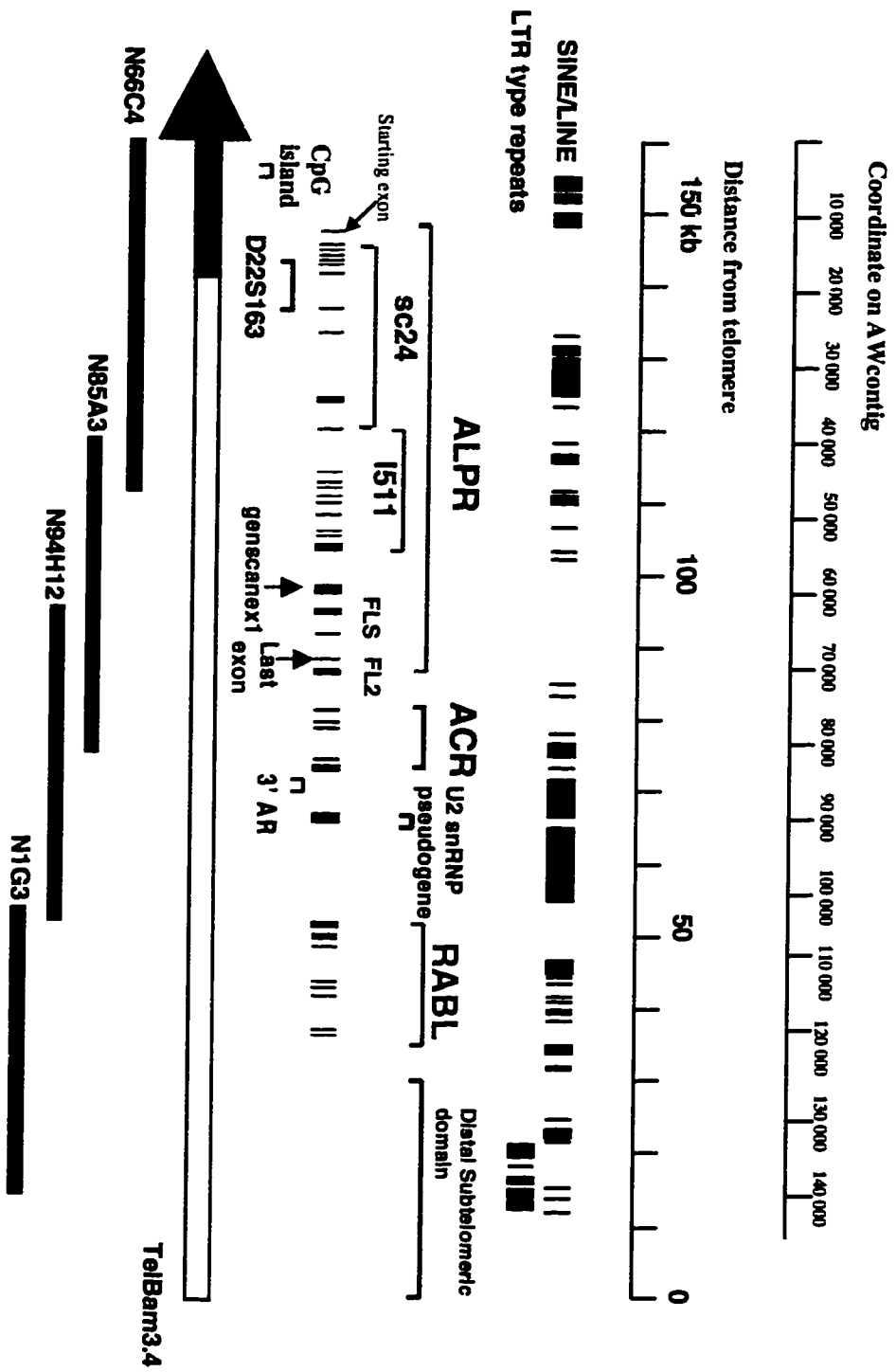
(LTR). This group constitutes ~4% of the total human genome (Smit 1996). SINE/LINE repeats are concentrated in several regions (Fig. 9). The first region is at the centromeric end of N66C4. These repeats may extend proximally into N85E7 (Fig. 3). The cosmid walk at the proximal end of the contig was unsuccessful, perhaps due to the high amount of repeats at the proximal end of N85E7, making the region difficult to clone. Other SINE/LINE repeats are within the introns of the three genes. There is a large repeat cluster between acrosin and RABL, which spans ~16 kb of the region. SINEs consist of Alu and MIR class repeats, which occupied 9.6% and 0.9% of AWcontig, respectively. LINEs consist of two classes, L1 and L2. There are 9 L2 elements in AWcontig, which occupied ~1% of the contig. All of the L2 elements are truncated. There are 20 L1 elements in AWcontig, most of them are truncated and are members of the old "M" family, which were inserted into the genome before the mammalian radiation (Smit et al., 1995). However, there is one full length young element L1PA2, which is interrupted by the processed pseudogene U2 SnRNP specific polypeptide A'(see below).

LTR type repeats are of two major classes: the mammalian LTR-retrotransposons (MaLRs) and a variety of endogenous retroviruses (ERVs). AWcontig has both classes, which are clustered at the subtelomeric region (Fig. 9.). The terminal 19.7 kb of AWcontig, which constitutes the subtelomeric repeats, contains approximately 10 LTR type repeat elements which occupy 35% of the region.

Blast search for the microdeletion region sequences that match the entries in public databases

After the removal of repetitive sequences by RepeatMasker, AWcontig sequence was put through Blast to search against various public databases. The Blast search against database nr (all non-redundant GenBank + EMBL + DDBJ + PDB sequences) and month (all new or revised GenBank + EMBL + DDBJ + PDB sequences) identified a CpG island clone at the beginning of AWcontig (Appendix); locus D22S163, the gene acrosin (ACR), and the minisatellite 3' AR which is 2.6 kb distal to ACR. D22S163 is a VNTR locus which contains two minisatellite repeats, MS607A and MS607B (Armour and

Figure 9 Genomic organization within the NT microdeletion region. The chromosome is shown by a solid bar with an arrow pointing towards to the centromere. The open bar represents the deleted region. One CpG island, and two minisatellite loci (D22S163 and 3'AR) are shown above the arrow. Three genes are labeled with different colours. ALPR (in blue) consists of 2 partial cDNAs (sc24 and I511), 2 EST contigs (FLS and FL2), and 3 putative exons (starting exon, genscanex1 and Last Exon). The number and position of exons are shown below the cDNA clones. Acrosin (ACR, in green) contains five exons and has previously been cloned. RABL is labeled with red and the positions of its nine exons are shown. Between ACR and RABL there is a processed pseudogene of U2 snRNP specific polypeptide A'. The distal subtelomeric domain is shown distal to RABL The proximal subtelomeric domain extends to the RABL gene region (see text). Two groups of repeat sequences, SINE/LINE and LTR type repeats, are represented by the solid boxes above the genes. Two scales are shown at the top of the figure. The uppermost scale shows the coordinates in the Awcontig, the lower one shows the distance of the genes from the putative telomere. The cosmid contig is shown below the chromosome.



Jefferys 1991). The sequences in Genbank are the flanking sequences surrounding the MS607A minisatellite array. One is 500 bp proximal (accession number X58043) and the other is 866 bp distal (accession number X58044) to the MS607A. Acrosin is a serine protease present in the acrosome of the sperm head (Klemm et al. 1991). It contains 5 exons that encompass 7080 bp in the contig (Fig. 9). 3'AR is the most distal VNTR locus on 22q. There is another minisatellite repeat that is close to the telomere (position 138528-139523 in AWcontig, Appendix), but that minisatellite locus is also on the subtelomeric repeat region of 21q, and thus it is not a unique locus. When the PCR-cloned 3'AR probe was hybridized with a blot contained TaqI-digested DNAs from NT family, the results confirmed that the paternal allele of 3'AR in NT was deleted (data not shown).

At position 93328-94328 of AWcontig, which is 6.5 kb distal to ACR, the Blast search identified a sequence related to the U2 snRNP specific polypeptide A'. It is within a 16 kb SINE/LINE repeat cluster and it is inserted within a full length young L1 element L1PA2, which arose after the divergence of the human genome from the New World monkey (Smit et al. 1995). This U2 snRNP specific polypeptide A' locus is a processed pseudogene because it does not contain introns, it lacks 20 bp at the 5' end when compared with the U2 snRNP specific polypeptide A' cDNA sequence, and many nucleotide substitutions causing nonsense mutations were found. When a PCR cloned U2 snRNP specific polypeptide A' partial cDNA probe (Table 3) was hybridized to a hybrid panel, bands in multiple lanes were seen, indicating that there are multiple copies of U2 snRNP specific polypeptide A' gene in the human genome. These copies localized to chromosomes 2, 15, 20, and 22 (data not shown). The fact that this locus is within the L1PA2 element suggests that the mRNA of U2 snRNP specific polypeptide A' transcript may have been trapped in the L1 element, and retrotransposed into different chromosomal positions while the L1 element was active. The locus of the authentic U2 snRNP specific polypeptide A' has not been mapped.

When the AWcontig was searched against the EST database, clusters of ESTs and single ESTs in various regions were identified (Appendix). EST clusters usually represent true transcripts, since the cDNAs are found in different libraries (see below). There are two single ESTs on the unique sequence of the contig. Neither of them have open reading

frames and polyadenylation tails; those clones probably represent genomic DNA contamination in the cDNA libraries. One such clone is AL (position 45612-46609, Appendix), which has an Alu element in the middle of the clone. Although it has a poly-A tail, the poly-A sequence is also found in the genomic sequence. Therefore AL is likely a cloning artifact. However, AL is also the last entry on chromosome 22 in the Human Transcript Map (<http://www.ncbi.nlm.nih.gov/cgi-bin/SCIENCE96>). Therefore it is the last “gene” on chromosome 22 in that database. ESTs were also found in the subtelomeric repeat region, where the sequence matches to many different chromosome ends. There is an EST contig that is within the subtelomeric region between position 124042 to 128223. This EST contig seems to have conserved splicing donor and receptor signals. Flint et al. (1997b) also observed that ESTs within subtelomeric regions could be split and contain splicing donor and acceptor signals. However, those EST contig sequences are not identical to that of the genomic sequence in the subtelomeric region. Also, no expression was found when the EST sequence specific primers were used for RT-PCR (Flint et al. 1997b). Therefore, those EST contigs likely represent non-processed pseudogenes. No further study was done on those ESTs.

Putative Expressed sequences determined by exon prediction programs

When exon prediction programs Grail v1.2 and Genscan were used to determine the putative expressed sequences in AWcontig, two groups of exons were predicted. The first group started at position 11210, and ended at 15754. The second group began at position 37562, and ended at 45023 (Fig. 10). When the ORFs of these two groups of exons were put through the Beauty program to search against protein databases, it showed that both groups matched a predicted *C. elegans* protein C33B4.3. Primers were designed within the exons to clone the putative gene by RT-PCR (see below). The efforts to clone putative expressed sequences by RT-PCR, together with EST database searches, resulted in the discovery of two new putative genes within the NT microdeletion region. These two genes were called ALPR (Ankyrin like, proline rich) and RABL (RAB like). The Grail and Genescan programs also identified exons of acrosin and RABL.

Ankyrin-Like Proline Rich (ALPR) gene

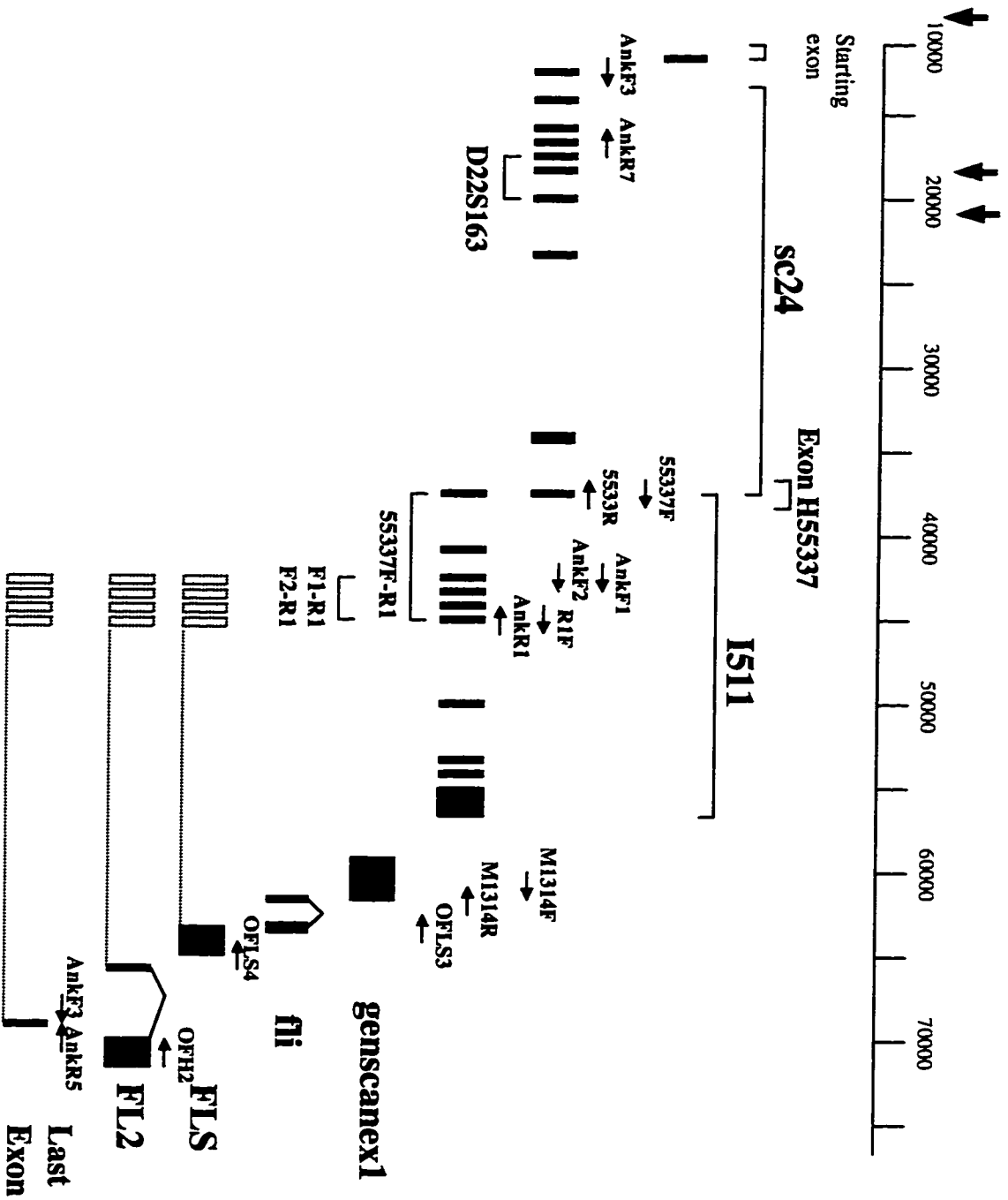
Cloning of ALPR

ALPR was firstly identified by two groups of exons predicated by Grail v1.2 (at position 11210-15754 and 37562-45023, Fig. 10). The ORFs of these two groups of exons match an unknown predicted *C. elegans* protein C33B4.3. Attempts were made to clone the putative gene by RT-PCR. Together with the characterization of EST contigs, partial cDNAs of ALPR were cloned (Fig. 10). They included 2 RT-PCR products (sc24, fli), 1 cDNA clone from a fetal brain cDNA library (I511) and 2 EST contigs (FLS and FL2). Protein similarity analysis identified three more putative exons that were part of ALPR (starting exon, genscanex1, Last Exon). Each partial ALPR clone is described below:

The starting exon of ALPR has the start codon at position 11210 and ends at the splice donor site at 11273 (Fig. 10), which was predicted by Grail v1.2. The ORF matches that of the starting exon of C33B4.3 protein. No primer was made to amplify this exon with others by RT-PCR because this exon is G-C rich. The exon has 79% G-C content over 62 nucleotides. Because the starting exon contains the start codon, it may represent the first exon of ALPR. However, the length of the 5' UTR is unknown.

Using RT-PCR from the next predicted exon, I made a partial clone of ALPR called sc24. Primers were made from this exon (primer Ank F3) and exon H55337 (primer 5533R, Fig. 10), which was the first exon from the next group of Grail exons. Total RNAs from various tissues were used as templates for reverse transcription. A 1.2 kb partial cDNA was amplified from the spleen reverse transcription product. It was subcloned into pGEM-T and called sc24 (spleen clone 24). Sequencing of sc24 showed that it contained 10 exons, which included the proximal 6 Grail exons, exon H55337 and 3 new exons which had not been identified before. There are three exons embedded in the D22S163 locus. D22S163 is a VNTR locus which contains two minisatellite repeats called MS607A and MS607B (Armour and Jeffery 1991). The first of the three exons is at the centromeric end of the 5' flanking sequence. The intron in between the latter two

Figure 10 The genomic organization of the ALPR gene. The scale above shows the coordinates in AWcontig. Arrows above the scale indicates gaps in the sequence. The exons are shown below in black bars, with grail-predicted exons highlighted in green. Primers used for RT-PCR are shown above the clones. The exons outlined by dotted lines and joined with a dotted line to different alternative 3' ends (FLS, FL2 and Last Exon) indicates that those exons and the 3' ends are found in the same transcripts (see text).



exons is mostly composed of MS607A minisatellite repeats. The ORF of sc24 matches that of the C33B4.3 protein. However, there is a frame shift in the H55337 exon which introduces a stop codon. Figure 11 shows the whole sequence and ORF of sc24.

The second group of Grail exons consists of 6 exons including exon H55337. RT-PCR using total spleen RNA as template and primer 55337F and AnkR1 amplified this group of exons and the product was called 55337F-R1 (Table 3, Fig. 10). 55337F-R1 was used as a probe to screen a fetal brain cDNA library (Table 6), and a cDNA clone I511 was isolated. I511 included the 6 Grail exons as well as 4 new exons (Fig. 10). The last exon was a terminal exon with poly-A tail at the position 54772-55801. The first 136 nucleotides of I511 did not match anywhere in AWcontig, which suggested that either I511 is chimeric, or the genomic sequence that corresponds to this 136 nucleotide has yet to be sequenced. Figure 12 shows the sequence and ORF of I511. Other than its homology to C33B4.3, the ORF of I511 also shows protein homology to the rat cortactin binding protein 1 (see below). When I511 was used as a probe to re-screen the fetal brain cDNA library, another cDNA clone DS17 was isolated. DS17 shares sequence and protein homology with ALPR (see below and Fig. 21), but is not a part of the ALPR locus.

Approximately 3.2 kb distal to I511, Genscan predicted a 2252 bp exon called genscanex1 (Fig. 10). It is a G-C rich exon (see below). Its ORF contains polyproline tracts, which are also found in the C33B.4 protein. The ORF also shows split matches to rat cortactin binding protein 1 (see below)

An EST contig was found at the position 62403 to 62902. One EST clone, called FLS, was chosen for further studies (Fig. 10). FLS is a cDNA clone from a fetal liver/spleen library. It is a 3'UTR sequence with a poly-A tail and no ORF. Figure 13 shows the sequence of FLS. Three pieces of evidence suggest that FLS is an alternative 3' end of ALPR: First, when a primer designed from the FLS sequence (primer OFLS4) was used to synthesize the first strand cDNA by reverse transcription, partial cDNAs which contained four I511 exons (F1-R1 and F2-R1) could be amplified. This result showed that I511 and the putative full transcript of FLS share the same exons. Second, when FLS was hybridized to Northern blots with different human tissues (Fig. 14 B, next

Figure 11 Nucleotide sequence and the putative open reading frame of cDNA clone sc24. The exon boundaries are marked by vertical lines in the sequences. The H55337 exon, starting from position 1155 to the end of the sequence, has a different ORF from that of I511 (Fig. 12). The stop codon in the H55337 exon is marked by an asterisk (*)

> SC24, 1268 bp

```

1   CTC AAC TAT GGG CTT TTC CAG CCG CCC TCC CGG GGC CGC GCC GGC AAG TTC CTG GAT GAG
   L  N  Y  G  L  F  Q  P  P  S  R  G  R  A  G  K  F  L  D  E
61  GAG CGG CTC CTG CAG GAG TAC CCG CCC AAC CTG GAC ACG CCC CTG CCC TAC CTG GAG|TTT
   E  R  L  L  Q  E  Y  P  P  N  L  D  T  P  L  P  Y  L  E  F
121 CGA TAC AAG CGG CGA GTT TAT GCC CAG AAC CTC ATC GAT GAT AAG CAG TTT GCA AAG CTT
   R  Y  K  R  R  V  Y  A  Q  N  L  -I  D  D  K  Q  F  A  K  L
181 CAC ACA AAG|GCG AAC CTG AAG AAG TTC ATG GAC TAC GTC CAG CTG CAT AGC ACG GAC AAG
   H  T  K  A  N  L  K  K  F  M  D  Y  V  Q  L  H  S  T  D  K
241 GTG GCA CGC CTG TTG GAC AAG GGG CTG GAC CCC AAC TTC CAT GAC CCT GAC TCA GGA G|AG
   V  A  R  L  L  D  K  G  L  D  P  N  F  H  D  P  D  S  G  E
301 TGC CCC CTG AGC CTC GCA GCC CAG CTG GAC AAC GCC ACG GAC CTG CTA AAG GTG CTG AAG
   C  P  L  S  L  A  A  Q  L  D  N  A  T  D  L  L  K  V  L  K
361 AAT GGT GGT GCC CAC CTG GAC TTC CGC ACT CGC GAT GGG CTC ACT GCC GTG CAC TGT GCC
   N  G  G  A  H  L  D  F  R  T  R  D  G  L  T  A  V  H  C  A
421 ACA CGC CAG CGG AAT GCG GCA GCA CTG ACG|ACC CTG CTG GAC CTG GGG GCT TCA CCT GAC
   T  R  Q  R  N  A  A  A  L  T  T  L  L  D  L  G  A  S  P  D
481 TAC AAG GAC AGC CGC GGC TTG ACA CCC CTC TAC CAC AGC GCC CTG GGG GGT GGG GAT GCC
   Y  K  D  S  R  G  L  T  P  L  Y  H  S  A  L  G  G  G  D  A
541 CTC TGC TGT GAG CTG CTT CTC CAC GAC CAC GCT CAG CTG GGG ATC ACC GAC GAG AAT GGC
   L  C  C  E  L  L  L  H  D  H  A  Q  L  G  I  T  D  E  N  G
601 TGG CAG GAG ATC CAC CAG|GCC TGC CGC TTT GGG CAC GTG CAG CAT CTG GAG CAC CTG CTG
   W  Q  E  I  H  Q  A  C  R  F  G  H  V  Q  H  L  E  H  L  L
661 TTC TAT GGG GCA GAC ATG GGG GCC CAG AAC GCC TCG GGG AAC ACA GCC CTG CAC ATC TGT
   F  Y  G  A  D  M  G  A  Q  N  A  S  G  N  T  A  L  H  I  C
721 GCC CTC TAC AAC CAG|GAG AGC TGT GCT CGT GTC CTG CTC TTC CGT GGA GCT AAC AGG GAT
   A  L  Y  N  Q  E  S  C  A  R  V  L  L  F  R  G  A  N  R  D
781 GTC CGC AAC TAC AAC AGC CAG ACA GCC TTC CAG|GTG GCC ATC ATC GCA GGG AAC TTT GAG
   V  R  N  Y  N  S  Q  T  A  F  Q  V  A  I  I  A  G  N  F  E
841 CTT GCA GAG GTT ATC AAG ACC CAC AAA GAC TCG GAT GTT G|TACCA TTC AGG GAA ACC CCC
   L  A  E  V  I  K  T  H  K  D  S  D  V  V  P  F  R  E  T  P
901 AGC TAT GCG AAG CGG CGG CGA CTG GCT GGC CCC AGT GGC TTG GCA TCC CCT CGG CCT CTG
   S  Y  A  K  R  R  R  L  A  G  P  S  G  L  A  S  P  R  P  L
961 CAG CGC TCA GCC AGC GAT ATC AAC CTG AAG GGG GAG GCA CAG CCA GCA GCT TCT CCT GGA
   Q  R  S  A  S  D  I  N  L  K  G  E  A  Q  P  A  A  S  P  G
1021 CCC TCG CTG AGA AGC CTC CCC CAC CAG CTG CTG CTC CAG CGG CTG CAA GAG GAG AAA GAT
   P  S  L  R  S  L  P  H  Q  L  L  L  Q  R  L  Q  E  E  K  D
1081 CGT GAC CGG GAT GCC GAC CAG GAG AGC AAC ATC AGT GGC CCT TTA GCA GGC AGG GCC GGC
   R  D  R  D  A  D  Q  E  S  N  I  S  G  P  L  A  G  R  A  G
1141 CAA AGC AAG ATC AG|TGCT CAG CAT TGG GGA GGG CGG TTT CTG GGA GGG AAC CGT GAA AGG
   Q  S  K  I  S  A  Q  H  W  G  G  R  F  L  G  G  N  R  E  R
1201 CCG CAC GGG CTG GTT CCC GGC CGA CTG CGT GGA GGA AGT GCA GAT GAG GCA GCA TGA CAC
   P  H  G  L  V  P  G  R  L  R  G  G  S  A  D  E  A  A  *  H
1261 ACG GCC TG
      T  A

```

Figure 12 Nucleotide sequence and the putative open reading frame of cDNA clone I511. The exon boundaries are marked by vertical lines in the sequence. The H55337 exon starts at position 137 and ends at 250. The open reading frame for this exon is different from that in Fig. 11. Although I511 has a poly-A tail, no polyadenylation signal is found in the 3' UTR.

```

> I511, 2008bp
1   CCC CGC CGC CCG CCG CCC CGG GGC CCG AAG CGG AAA CTT TAC AGC GCC GTC CCC GGC CGC
   P R R P P P R G P K R K L Y S A V P G R
61  AAG TTC ATC GCC GTG AAG GCG CAC AGC CCG CAG GGT GAA GGC GAG ATC CCG CTG CAC CGC
   K F I A V K A H S P Q G E G E I P L H R
121 GGC GAG GCC GTG AAG G|TGCTC AGC ATT GGG GAG GGC GGT TTC TGG GAG GGA ACC GTG AAA
   G E A V K V L S I G E G G F W E G T V K
181 GGC CGC ACG GGC TGG TTC CCG GCG GAC TGC ETG GAG GAA GTG CAG ATG AGC CAG CAT GAC
   G R T G W F P A D C V E E V Q M R Q H D
241 ACA CGG CCT G|AAACG CGG GAG GAC CGG ACG AAG CGG CTC TTT CGG CAC TAC ACA GTG GGC
   T R P E T R E D R T K R L F R H Y T V G
301 TCC TAC GAC AGC CTC ACC TCA CAC AG|CGAT TAT GTC ATT GAT GAC AAA GTG GCT GTC CTG
   S Y D S L T S H S D Y V I D D K V A V L
361 CAG AAA CGG GAC CAC GAG GGC TTT GGT TTT GTG CTC CGG GGA GCC AAA G|CAGAG ACC CCC
   Q K R D H E G F G F V L R G A K A E T P
421 ATC GAG GAG TTC ACG CCC ACG CCA GCC TTC CCG GCG CTG CAG TAT CTC GAG TCG GTG GAC
   I E E F T P T P A F P A L Q Y L E S V D
481 GTG GAG GGT GTG GCC TGG AGG GCC GGG CTG CGC ACG GGA GAC TTC CTC ATC GAG|GTG AAC
   V E G V A W R A G G L R T G D F L I E V N
541 GGG GTG AAC GTG GTG AAG GTC GGA CAC AAG CAG GTG GTG GCT CTG ATT CGC CAG GGT GGC
   G V N V V K V G H K Q V V A L I R Q G G
601 AAC CGC CTC GTC ATG AAG GTT GTG ACA AGG AAG CCA GAA GAG GAC GGG GCT CGG
   N R L V M K V V S V T R K P E E D G A R
661 CGC AGA G|CCCCA CCG CCC CCC AAG AGG GCC CCC AGC ACC ACA CTG ACC CTG CGC TCC AAG
   R R A P P P P K R A P S T T L T L G R S K
721 TCC ATG ACA GCT GAG CTC GAG GAA CTT G|CCTCC ATT CGG AGA AGA AAA GGC G|AGAAG CTG
   S M T A E L E E L A S I R R R K G E K L
781 GAC GAG ATG CTG GCA GCC GCC GCA GAG CCA ACG CTG CGG CCA GAC ATC GCA GAC GCA GAC
   D E M L A A A A A E P T L R P D I A D A D
841 TCC AGA GCC GCC ACC GTC AAA CAG AGG CCC ACC AGT CGG AGG ATC ACA CCC GCC GAG ATT
   S R A A T V K Q R P T S R R I T P A E I
901 AGC|TCA TTG TTT GAA CGC CAG GGC CTC CCA GGC CCA GAG AAG CTG CCG GGC TCC TTG CGG
   S L F E R Q G L P G P E K L P G S L R
961 AAG GGG ATT CCA CGG ACC AAG TCT GTA G|CTAAA GGA TCT CAT ACG TTG ATG GAC ATG TGG
   K G I P R T K S V A K G S H T L M D M W
1021 GGA TTA GGC CTT CCC CAA CCC AGA GCT TTC CCC CGG CAG CCG ACA CTC CCT GTC CAG TGG
   G L G L P Q P R A F P R Q P T L P V Q W
1081 GCA CCG CCC CCC ATC GCC TCA TCC CTC CCA TGG GCA GTC TCA TCC CTG TCC CCA GCT GCC
   A P P P I A S S L P W A V S S L S P A A
1141 ACT CCC TGT CCA CTG GGC ACC CCC ACC TCC CCA TCA CCT CTC ATC CTT CCC ATG GGC AGC
   T P C P L G T P T S P S P L I L P M G S
1201 CTC ATC CCT GTC CCC AGC TGC CAC TCC CTG TCC ACT GGG CAC CCC CAC CTC CCC ATC ACC
   L I P V P S C H S L S T G H P H L P I T
1261 TCT CAT CCT TCC CAT GGG CAG CCT CAT CCC TGT CCC CAG CTC AGC TGC CTC CAT CGC GGT
   S H P S H G Q P H P C P Q L S C L H R G
1321 TGC TCC CTT GCA GCC CAA GTG CAT GTG AAG TTT CTG ACC CTC AAA CCC CCT GAA CTT GCC
   C S L A A Q V H V K F L T L K P P E L A
1381 TCT CCC CTA CTT CTC CAT GTA TTG CTG TCT CCC CTT ACA GCC AGT TTT CCC CAA AAA
   S P L L L H V L L S P P F T A S F P Q K
1441 GTC ACC TCT ACT GTC CAT CTT GTC TCC CGT CCC CCA CCG GCT CCT CAG CCT GCG GCA GAC
   V T S T V H L V S R P P P A P Q P A A D
1501 TTC CTC TCT CCA TAC CCA AAT TGA AAC TGC TCC CAC CAG GGT CAC CGG CGG CCT CCA GGG
   F L S P Y P N *
1561 CCT TCC TCC CCA TGG ATG CAT GGC CAG CAC TCC TGC TGA CCT CAG CTG GCA TTT GGT TTC
1621 TTT GCT CCT TCT TCC GTG AAA CGC TCC TCC CCA AAG CTT TAG GAC AAC ACT TAC TGG TTT
1681 TCC CCT CCT TTC TCA GAT GGC ACC TAT TTC AAC CCT GCC CCA GGC TCC CAA ATA TGG
1741 GAC CTT TCA AGG CTG TGT CTT CAG CCT CCT TCC CCT CTG TGC CTC ACC ACT CCC CAG CCA
1801 CCA TCA CAC CCC ACA CTC ACC GTC TCC TTT CCT GAC CCA GAT GGG GAC ATG GTA ACA AGC
1861 CCT GCC CAC TGT GCA TCC CTC CAG ACA TCC ACT CTC CCG TCC CCT GCC CAC CTG TGC ATC
1921 TGT CCA CCC ACA CAC CTG CCC ACT TGA CCG TCT GTC CAT ACA AAC CCC AGC ACA CCT
1981 CCC ATC TGT TAC ACA TAG GAG TAA ACC TCT CAG CAC TT

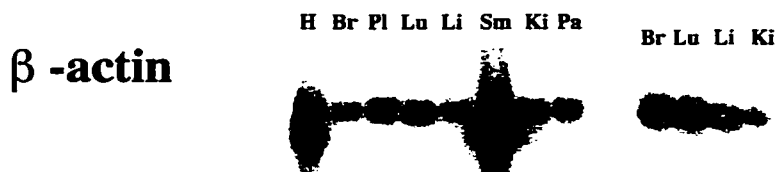
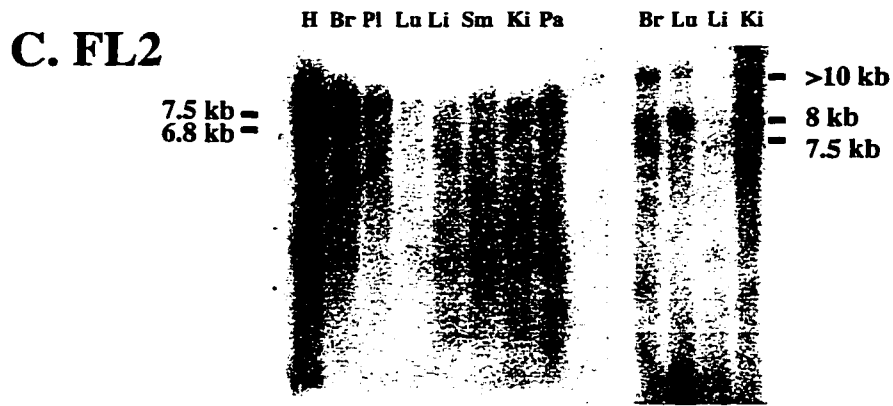
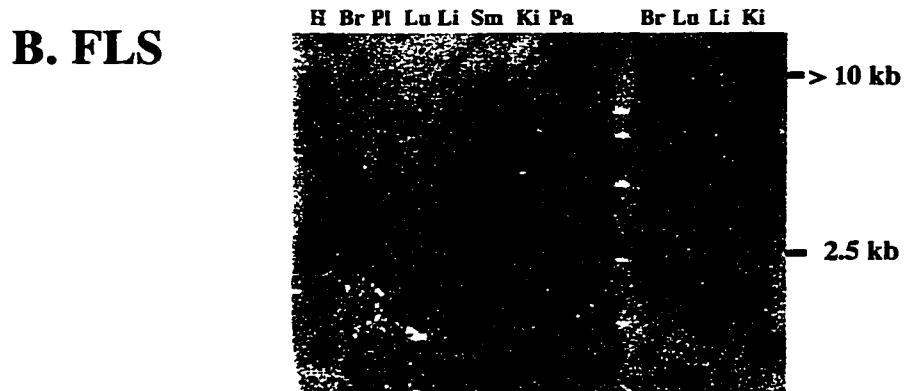
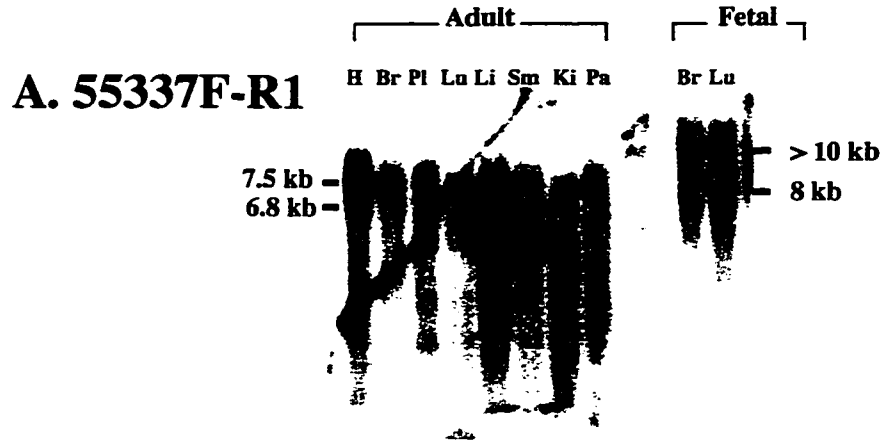
```

Figure 13 Nucleotide sequence of cDNA clone FLS. The clone is a 3' untranslated region so no open reading frame and exon boundaries are present. The polyadenylation signal is marked in bold.

>FLS, Soares fetal liver spleen 1NFLS clone 121540, 524 bases.

```
1   CAACGTG TTC  AGAACTTAAG  GACTTTGCAG  GTCTTACAAA  GGCCTGGCCA
51  TTCTACCTTC  TTTAGTTCAG  GATTCAAAAAG  ACAGGTAGGA  GCTTGGGAAA
101 CTCATGAGGC  CTCTCCTAAG  GTCCCGGGAT  GCTGCCTCCA  GCTCCTGTCA
151 TCCTGGGGAA  TTGCTCTGGG  GTCCTCTCCC  CTTTTAGCCT  TTTCCAACTC
201 TCAGCCAAAC  TGGAAAGCCC  TCTTCCCAGC  AGATGCAGTG  TTGAAGGTGC
251 CCGTAGAATG  GGTGTTATAA  TCAGAGTGAG  CAGCCTGGTC  CTAGGCCTCT
301 GTACAGGACC  AGACCCCTGA  GGCTGGGGTC  TCCTGACCCA  CACCTGACCA
351 GCCCCCATCT  TCCCTCTCTG  CTTCTCCCTC  CGCTCTTCTC  TGCCTCTTGG
401 TCTTGATGAA  AATCAAAGCC  ATTTTAAAAA  GTGCATAGCA  CAGTGCCTGG
451 CCTGGTTCGG  GCCCTCAATA  AACATTTCTT  AAATGGATGA  AAGAACAAAG
501 CAAAAAAAAA  AAAAAAAAAA  AAAA
```

Figure 14 Northern blot analysis of ALPR. The same human multiple tissue Northern blot pair was probed with 55337F-R1(A), FLS(B), and FL2(C). Adult mRNA of heart (H), brain (Br), placenta (Pl), lung (Lu), liver (Li), skeletal muscle (Sm), kidney (Ki), and pancreas (Pa) comprised one blot; and fetal mRNA of brain (Br), lung (Lu), liver (Li), and kidney (Ki) were included on the second Northern blot. 5533F-R1 and FL2 showed the same banding pattern. Both probes detected a common 7.5 kb band in adult heart, brain, and placenta. A specific 6.8 kb band is seen in brain (A and C). A large (> 10 kb) band in fetal brain is detected by three different probes (A, B, and C) but is unlikely to be fetus-specific. This band is most likely missing from the adult tissue due to poor transfer of large bands on that Northern blot. 55337F-R1 and FL2 show an alternative 8 kb band in fetal tissues (A and C). A weak 7.5 kb band was seen in fetal tissues when the Northern blot was probed with FL2 (C). FLS shows a 2.5 kb fetal liver specific band (B). The Northern blots were also hybridized with a control probe β -actin to determine the relative concentration of mRNA loaded on the gel (Bottom panel). The probing of β -actin was done by Dr Valérie Trichet.



section), it showed a large band (> 10 kb) in the fetal brain lane, which was the same band found on a Northern blot probed with 55337F-R1 (Fig. 14 A). Third, when a primer from the 3' end of genscanex1 (M1314F) and a primer from the FLS sequence (OFLS3) were used for RT-PCR, a 584 bp partial cDNA fli (Table 3, Fig. 10 and Fig. 15) was amplified. This result indicates that the putative full transcript of FLS contains the genscanex1. Since both I511 and genscanex1 are homologous to rat cortactin binding protein 1, and the putative full transcript that contains FLS has both I511 and genscanex1 exons, FLS probably represents an alternative 3' end of ALPR.

Genscan identified another exon at the position 69592 to 70173, which I called "Last Exon" (Fig. 10), because its ORF showed protein homology with the carboxyl end of C33B4.3 and the rat cortactin binding protein 1 (see below). When a primer from Last Exon (Ank R5, Table 4) was used to start reverse transcription, PCR from this reverse transcription product could also amplify F1-R1 and F2-R1, suggesting this sequence is also a part of ALPR (Table 3).

Approximately 850 bp distal to Last Exon, an EST contig was identified. One EST clone FL2 was chosen for further studies. FL2, a fetal lung cDNA, is probably another alternative 3' end of ALPR. Like FLS and Last Exon, the reverse transcription product which was synthesized by using a FL2 primer (OFH2) could be used as a template to amplify F1-R1 and F2-R1. When FL2 was hybridized to a Northern blot, bands from adult heart, brain, placenta, fetal brain and fetal lung (Fig. 14C) matched those found on the Northern blot probed with H55337-R1 (Fig. 14 A). Figure 16 shows the sequence and ORF of FL2. A mouse brain EST clone MB, which later proved to be a rat brain EST (<http://www.ncbi.nlm.nih.gov/dbEST/warning.html>), shares homology with FL2. When it was used as a probe to screen a mouse brain cDNA library (Table 6), a cDNA clone mbl101 was isolated. Multiple alignment analysis showed that human clone FL2, rat clone MB, and mouse clone mbl101 were very similar (Fig.17). FL2 and mbl101 shared 82.7% DNA identity, FL2 and MB were 81.9% identical, whereas the two rodent clones (MB and mbl101) shared 90.9% identity.

Figure 15 Nucleotide sequence and the putative open reading frame of cDNA clone fli. The exon boundary is marked by a vertical line in the sequence. The genscanex1 sequence starts from the beginning of the sequence and ends at the exon boundary. The FLS sequences begins at position 284 to the end of the sequence.

>fli, 584 bp

```

1   CCT GAA GAC GAC AAA CCA ACT GTG ATC AGT GAG CTC AGC TCC CGC CTG CAG CAG CTG AAC
   P   E   D   D   K   P   T   V   I   S   E   L   S   S   R   L   Q   Q   L   N
61  AAG GAC ACG CGT TCC CTG GGG GAG GAA CCA GTT GGT GGC CTG GGC AGC CTG CTG GAC CCT
   K   D   T   R   S   L   G   E   E   P   V   G   G   L   G   S   L   L   D   P
121 GCC AAG AAG TCG CCC ATC GCA GCA GCT CG|GTCT CCC CTC TCC TCT TTG GGT CTG GGG GGG
   A   K   K   S   P   I   A   A   A   R   S   P   L   S   S   L   G   L   G   G
181 TGG TAT GTG GAT GCC ACC TCT TGA CTC CTG CTT CTT GCT GCC TGG AAG ACC AAC CTA GAG
   W   Y   V   D   A   T   S   *
241 GGC CCC GTA CTG TCA GCC TTG GAG GAC AGA GTT CAC AGC GTA GCA ACG TGT TCA GAA CTT
301 AAG GAC TTT GCA GGT CTT ACA AAG GCC TGG CCA TTC TAC CTT CTT TAG TTC AGG ATT CAA
361 AAG ACA GGT AGG AGC TTG GGA AAC TCA TGA GGC CTC TCC TAA GGT CCC GGG ATG CTG CCT
421 CCA GCT CCT GTC ATC CTG GGG AAT TGC TCT GGG GTC CTC TCC CCF TTT AGC CTT TTC CAA
481 CTC TCA GCC AAA CTG GAA AGC CCT CTT CCC AGC AGT GCA GTG TTG AAG GTG CCC GTA GAA
541 TGG GTG TTA TAA TCA GAG TGA GCA GCC TGG TCC TAG GCC TCT G

```

Figure 16 Nucleotide sequence and the putative open reading frame of cDNA clone FL2. The exon boundary is marked by a vertical line in the sequence. The polyadenylation signal is marked in bold.

>FL2, Soares fetal lung MbHL19W clone 306553, 1160 bp

```

1   TGGGGCCGTGGGGCCTCAGTGTGATCTGGACTCAGCCTCTTCAGCGTGGCTGCTGGAGG
   W G R G G L S V I W T Q P L Q R G C W R
61  TGFTCGTGGGTGACGGTGCCTGGTGAAGTATCATGTGTTAGAGG|GAGGGGGCTAGCTTGG
   C S W V T V P G E V S C V R G R G L A W
121 TCCCCATGCTCTTGGGCAACTACAGCAGAGAAGCCTCCCTGCCTTGGACCCCAAAGTCTC
   S P C S W A T T A E K P P C L G P Q S L
181 CTGTCCTGCCCTTTATGTGTGTGGGTGAAACTGGGTGCGTCTGAGCAGTGGGAGCCGTG
   L S C P L C V W V K L G A S E H V G A V
241 TGTGTCCCTGATTACTGAGTGGCCACCAGGGGCCGCTCTGGACTAGCGCGGGGCCGTGGA
   C V P D Y *
301 GCGGTGCACCGTGTGCATGCGTGGGGTGTACCTGTGAGAGCACCCCTGTCTCCTCTTCAA
361 AGAAAGTCAGAGGCCATCCTGCACCCCTGGGTCCAGCTGTTGCCCAGCCTGTCTTCCAG
421 AGCCTCACCCAGCCTGAGCGGGTTCCTTGGTGAATCCCTGCTGCTTGGGGAGGCCCAA
481 GGGCCCTTGGAGGCAGCGCCCCACCTTGGGCTTCTGAGGGCATCATAGGGGGACCCCT
541 AGAGTCAGTTCACCACAGGCCCTGGGAGAGTCAAAGACCCCGAGGGTGCCAGCCCC
601 CACACTGTGACTCCTCACACTCAGCGATGACCTGTGGGGTGGGGGGCCCTGGGACGTTTT
661 TAAACCTAGGGTTTGGAGTCTGGACTAAGCTCCATCCACGTCACTCACAAGTTTCTGTTT
721 ATATTTCTAGCTTTTTTAATAAAATAAAAAAAAAAAGAAAACAGAAGTTTTCACAACCC
781 AGGGGCCTGGCACGCCGGTCTGTGCCTGCCCGCCCCGCCCTGGCCCACCGCCCCACTCC
841 CTGGGCACAGAGTCACCCCACTATCCTTCCGCCAACAGTCCAGGTCACACAGCAGCAG
901 TCACTGTAACAGACTGCCACATACACTCGGTCTCACACTCACCTGTGGGTTTTGGTTC
961 CGTCAATTTGGGTTTTTAACTTTACAGGGTCAGTCCCGCTTACCTCCTTTGTATGGA
1021 GTTCCATCCGGGGGTTTCAACCCCTGCTCCAGTCCCTGAGGCCTCCTGACCCCTGACGTTG
1081 TGATACGCCCCACAGAGATCTATGTTTCTTATATTATTATTATTGATAATAATTATTATA
1141 ATATTATTATGTAATAAATT

```

Figure 17 Multiple nucleotide sequence alignment of FL2 against its rat (MB) and mouse (mb1101) homologs using program ClustalW 1.7. Identical nucleotides are labeled in red.

```

1 MB -----
2 mb1101 -----
3 FL2  TGGGGCCGTGGGGGC CTCACGTGTATCTGG ACTCAGCCCTTTCAG CGTGGCTGTGGAG TGTTCGTGGGTACG GTGCCCTGTGAAGTA -----
0
0
90

1 MB -----
2 mb1101 -----
3 FL2  TCATGTGTTAAGAGG AGGGGGCTAGCTTGG TCCCCATGTCTTGG GCAACTPACAGCAGAG AAGCCTCCCTGCCTT GGACCCCAAGTCTC -----
0
0
180

1 MB -----ACGCCCTGC GTCGTGTGTGA-AAA TTGGGTTGTGGCTGAG CGCANTGGGTGCCCTG TANTGTCTGATTTGT GGAGTGTGTCCCAGG
2 mb1101 -----
3 FL2  CTGTCCCTGCCCCTTTA TGTGTGTGGGTGAAA CTGGGTGCCCTGAG CACGTGGAGCCGTTG TGTGTGCCCTGATTTAC TGAATGGCCACACAGG -----
82
0
270

1 MB -----GGCTGTCTTGATGG GTGGGAGGTTGAGGA AGCTTGCACAGGGGT GCANTGATGGGTGTG TGCCCTGTGAAAAGGC CCTGTCTTCTC---C
2 mb1101 -----
3 FL2  GGCCGCTCTGGACTTA GCGGGGGGCCCGTGA GCGCTGCACCCGTG-T GCANTGCGTGGG-GTGG TACCTGTGAGAGCAC CCTGTCTCTCTTCC -----
169
0
358

1 MB -----AAAGAAAGGCTG--- -TCCCTGCTCT-TG GGTCCCTGCTGTTTTC TCAGCCTGTCTTCCC TGAACCTGCACCCAGC TTAAGCAGGGGTTCT
2 mb1101 -----
3 FL2  AAAGAAATCAGAGG CCATTCCTGCACCCCTG GGTCCAGCTGTTTTC CCAGCCTGTCTTCC AGAGCCTCACCAGC CTGAGCGGGGTTCCC -----
252
0
448

1 MB -----TGTTGAATCCCTTTCA GCTTTTGGGAGGCGCTC AAGGGCTCCCGTGA GGCAGCACCCCT--- TTGGGCTTCTAAGGG AATTGT--GGGACC
2 mb1101 -----
3 FL2  TGTTGAATCCCTTGT GCTTTGGGGAGGCCCC AAGGGC-CCCTTGA GGCAGGCCCCCCACC TTGGGCTTCTGAGGG CATCATAGGGGACC -----
337
0
537

1 MB -----ACTAAAAATCAGGCCA CAACAGCCCTTGGAG AGAGGCAAAAGACTCC TGAGGGTAACCCITGGC CCCCCTTACTGTGA- TCTTCACAATTCAGC
2 mb1101 -----
3 FL2  CCTAGAGTCAAGTCA CCACAGGCCCTTGGGG AGAGTCAAAAGACCCC CGAGGGTGGCCAG-C CCCCCACACTGTGAC TCCTCAGACT-CAGC -----
426
0
625

```

Figure 17 (continued)

```

1 MB  AATGACCTGTGGGGC GGGGGGCCCTGGGGC ATTTTAAACATPAGG GTTTGGAGTCTGGAC TAAGCTCCATCCAGG TCACTCACAGTPTTC
2 mb1101 -----GGC A-----CGAGCGG GTTTGGAGTCTGGAC TAGGCTCCATCCAGG TCACCTCACAGTPTTC
3 FL2  GATGACCTGTGGGGT GGGGGGCCCTGGGGC GTTTTAAACCTPAGG GTTTGGAGTCTGGAC TAAGCTCCATCCAGG TCACCTCACAGTPTTC

1 MB  TGTTCCTAATTTCTAG CTTTTTTTTAATPAAA AATAAATAAATAAT AAATPATAAATAAA TATATATATATPAAA AAGACAGAAAACAGG
2 mb1101 TGTTCCTAATTTCTAG CTTTTTTAATPAAA AATAAATAAATAAT AAATPATAAATAAA TATATATATATPAAA AAGACAGAAAACAGG
3 FL2  TGTTCCTAATTTCTAG CTTTTTTAATPAAA -----TA-----TA-----AAA AAAAAAGAAAACAGG

1 MB  TGTTCCTAATTTCTAG GGGGCTTGGCAGCGCC GGTTCGTGCCCCACCC GCCCCGCCCCACCTG GCCCACCAGGCCCAT TCCTTAGACACAGAG
2 mb1101 TGTTCCTAATTTCTAG GGGGCTTGGCAGCGCC GGTTCGTGCCCCACCT GCCCCGAC-----CCTG GCCCACCAGGCCCAT TCCTTAGACACAGAG
3 FL2  AGTTTTCACAACCCA GGGGGCCCTGGCAGCGCC GGTCTGTGCTGCCCC GCCCCGC-----CCTG GCCCACCAGGCCCAT TCCTTAGACACAGAG

1 MB  TCACGCCCCACTAACC CTTTTAACCAACAGAG CAGGTACACACACACA GCAGCGGTCACTGTGA ACAGACTGCCACATA CACA-----GTCTCAC
2 mb1101 TCACACCCGACTAACC CTCTCACCAACAGAG CAGGTACACACACACA GCAGCGGTCACTGTGA ACAGACTGCCACATA CACA-----GTCTCAC
3 FL2  TCACACCCGACTCATC CTTCCGCCAACACAGATC CAGGTACACACA-----GCAGCAGTCACTGTGA ACAGACTGCCACATA CACACTGCGGTCTCAC

1 MB  ATTTAACCTGTGGGTT TTTGGTTCGTGTTGAG TTTGGGTTTTPTAAC TTAACAGGGTCAAGTTC CGCTTCATCCCCC ---TTTGTATGAGAG
2 mb1101 ATTTAACCTGTGGGTT TTTGGTTCGTGTTGAG TTTGGGTTTTPTAAC TTAACAGGGTCAAGTTC CGCTTCATCCCCC CCCTTTTGTATGAGAG
3 FL2  ACTCACCTGTGGGTT TT--GGTTCGTTCAAA TTTGGGTTTTPTAAC TTAACAGGGTCAAGTTC CGCTTCATCCCTC ---TTTGTATGAGAG

1 MB  TTCCATNTCCGGGG-C TTTCAACCCCTGTGCT CCAGTCCGTGAGGCGCT CCTGACCCTGACGTT GTGATACACCCCCACA GAGATCTAATGTTTCT
2 mb1101 TTCCATNTCCGGGGG TTTCAACCCCTGTGCT CCAGTCCGTGAGGCGCT CCTGACCCTGACGTT GTGATACACCCCCACA GAGATCTAATGTTTCT
3 FL2  TTCCATNTCCGGGGG-G TTTCA-CCCCGTGCT CCAGTCCGTGAGGCGCT CCTGACCCTGACGTT GTGATACACCCCCACA GAGATCTAATGTTTCT

1 MB  TATATTTAATTAATTT AATTAATAATTAATTAAT AATATTAAT---GTAA TAAATTTAATAAGAAA TG 991
2 mb1101 TATATTTAATTAATTT AATTAATAATTAATTAAT AATATTAAT---GTAA TAAATTTAATAAGAAA TG 556
3 FL2  TATATTTAATTAATTT GATTAATAATTAATTAAT AATATTAATTAATGTAA TAAATTT-----TG 1160

```

Northern analysis of ALPR

Three partial cDNA clones of ALPR were separately probed to the same Northern blot. 55337F-R1 is a RT-PCR product that contains six Grail exons in the I511 cDNA (Fig. 10). It detected a 7.5 kb band in adult heart, brain, and placenta. In addition, an alternative 6.8 kb transcript was seen in brain. In fetal tissues, the clone detected a large (> 10 kb) and an alternative 8 kb band in brain and lung. There might also be a 7.5 kb transcript in these tissues but the high non-specific background signal made it difficult to see (Fig. 14 A).

FLS is an alternative 3' end of ALPR. It only detected a large (> 10 kb) band in fetal brain and a 2.5 kb band in fetal liver (Fig. 14 B). The large band was also seen in the Northern blot that was probed with 55337F-R1. The fetal Northern blot that was hybridized with 55337F-R1 was contaminated by fungus, which caused the fetal liver and kidney lanes to be degraded. A follow up study showed that the I511 probe, which includes all the exons in 55337F-R1, show the same 2.5 kb band in fetal liver as on the Northern blot probed with FLS (Dana Shkolny, unpublished data). This result further proved that I511 and FLS are in the same gene but have two different 3' ends. No signal was detected by probe FLS in adult tissues (Fig. 14 B). However, faint signal could be masked by the high background signal on the Northern blot with adult tissues. It is therefore inconclusive whether FLS is also expressed weakly in adult tissues.

FL2 is another alternative 3' end of ALPR. When it was probed to the Northern blots, it showed the same hybridization signal as in the 55337F-R1 probe. A common 7.5 kb band was found in the adult heart, brain, and placenta. This band showed faintly in other adult tissues. This weak expression was confirmed in a second Northern blot (Dana Shkolny, unpublished data). A specific 6.8 kb band was seen in brain. Fetal tissues showed the large (> 10 kb) band, 8 kb and 7.5 kb bands in brain, lung and kidney (Fig. 14 C). This banding pattern is also seen in the Northern blot probed with 55337F-R1. These results show that these two partial cDNAs come from the same gene. Subsequent experiments with different Northern blots have shown that the >10 kb band is also seen in

adult tissues (Dana Shkolny, unpublished data). It is most likely missing from this Northern due to poor transfer of large fragments.

Construction of an ALPR cDNA contig

Several cDNA clones and putative exons were identified to be part of ALPR: the starting exon is identified by its homology to the *C. elegans* protein C33B4.3. Many proximal exons of ALPR are included in sc24 and I511, which is overlapping at the exon H55337. The alternative 3' ends (FLS, Last Exon, and FL2) have been related to ALPR by showing all of these transcripts contain at least four I511 exons (Fig. 10). However, there are two gaps in the cDNA contig between these cDNA clones. One is between the starting exon and sc24, and one between I511 and the other alternative 3' ends. In order to construct a cDNA contig that includes the full sequence of ALPR, two approaches were used to isolate cDNA clones that contain different gene fragments. The first method was cDNA library screening. Probes from sc24 (PCR product F3-R7, which contains the first 5 exons of sc24), FLS, FL2, and Last Exon (PCR product F5-R3, which contains the most conserved region of Last Exon) were used to screen different cDNA libraries (Table 6). Some putative cDNAs isolated had sizes smaller than the probes. Some cDNA clones were artifacts of the cDNA library, which was evident by the presence of introns. They are probably genomic DNA fragment contaminants in the cDNA libraries. Therefore, no cDNA was isolated from the cDNA libraries that could close the two gaps in the cDNA contig.

When primers from different alternative 3' ends (FLS, Last Exon, and FL2) were used for first cDNA strand synthesis, such reverse transcription product could be used to amplify exons within I511 (PCR product F1-R1 and F2-R1). However, RT-PCR was done numerous times (>10) to amplify a product that contained all proximal exons and different 3' ends, but no product was amplified. In one such RT-PCR, the first strand cDNA synthesis started from the alternative 3' end FLS. Primers from exons on I511 and genscanex1 (R1F and M1314R respectively, Fig. 10) amplified a 454 bp product. (Fig. 18) The sequencing of the product revealed that it contained two full exons of I511, and one "broken" exon of I511, which was evident by the absence of any splicing donor

signal after the "breakpoint". The broken exon was fused with another broken exon, which was a part of genscanex1 (Fig. 18). At the broken ends of both exons, they shared a common sequence GCCGGGC. Therefore, this PCR product (R1F-M1314R) was illegitimate because it missed a significant portion of genscanex1 and approximately half of the I511 exon.

Determining the G-C content of putative ALPR exons

Genscan predicted exons (genscanex1 and Last Exon) share significant homology with rat cortactin binding protein 1 (see below), yet RT-PCR failed to clone those exons in full length. Different factors could account for the failure to clone cDNAs by RT-PCR. One common cause is the high G-C content of the cDNA. Depending on the distribution of the G-C rich sequence in the whole cDNA, difficulty can be encountered if the G-C rich sequence is short (Orlicky and Nordeen, 1996). It has been found that particular regions can not be cloned if the G-C rich sequence formed a strong secondary structure (Soreq et al. 1990), or the transcript is not detectable by RT-PCR if the whole cDNA is G-C rich (Bondy et al. 1996). Genscanex1 and Last Exon have high overall G-C content (72% and 73% for genscanex1 and Last Exon respectively). In order to determine whether the G-C rich sequences cluster on particular regions, or if they are distributed equally throughout the exons, the GCG program WINDOW was used. Figure 19 shows the G-C content in different ALPR sequences. The WINDOW analysis showed that the entire sequence of genscanex1 and Last Exon are G-C rich when they are compared with the two other ALPR clones sc24 and I511, which have lower overall G-C content (both of them are ~62%). The WINDOW program also showed that the G-C content of the two exons are similar to the DRD4 gene, whose transcript was not detectable by RT-PCR due to the high G-C content (Bondy et al. 1996).

Homology between ALPR and other proteins in the public databases

Since all ALPR gene fragments showed protein homology with either the *C. elegans* protein C33B4.3 or the rat cortactin binding protein 1, the ORFs of those gene

Figure 18 The illegitimate RT-PCR product R1F-M1314R. *A.* A diagram shows the exons within the product. Two exons with irregular ends represent the "broken" exons. *B.* The sequence of R1F-M1314R. The vertical lines in the sequence mark the boundaries of the exons. The consensus sequence of the "breakpoint" of the two broken exons is boxed.

A



B

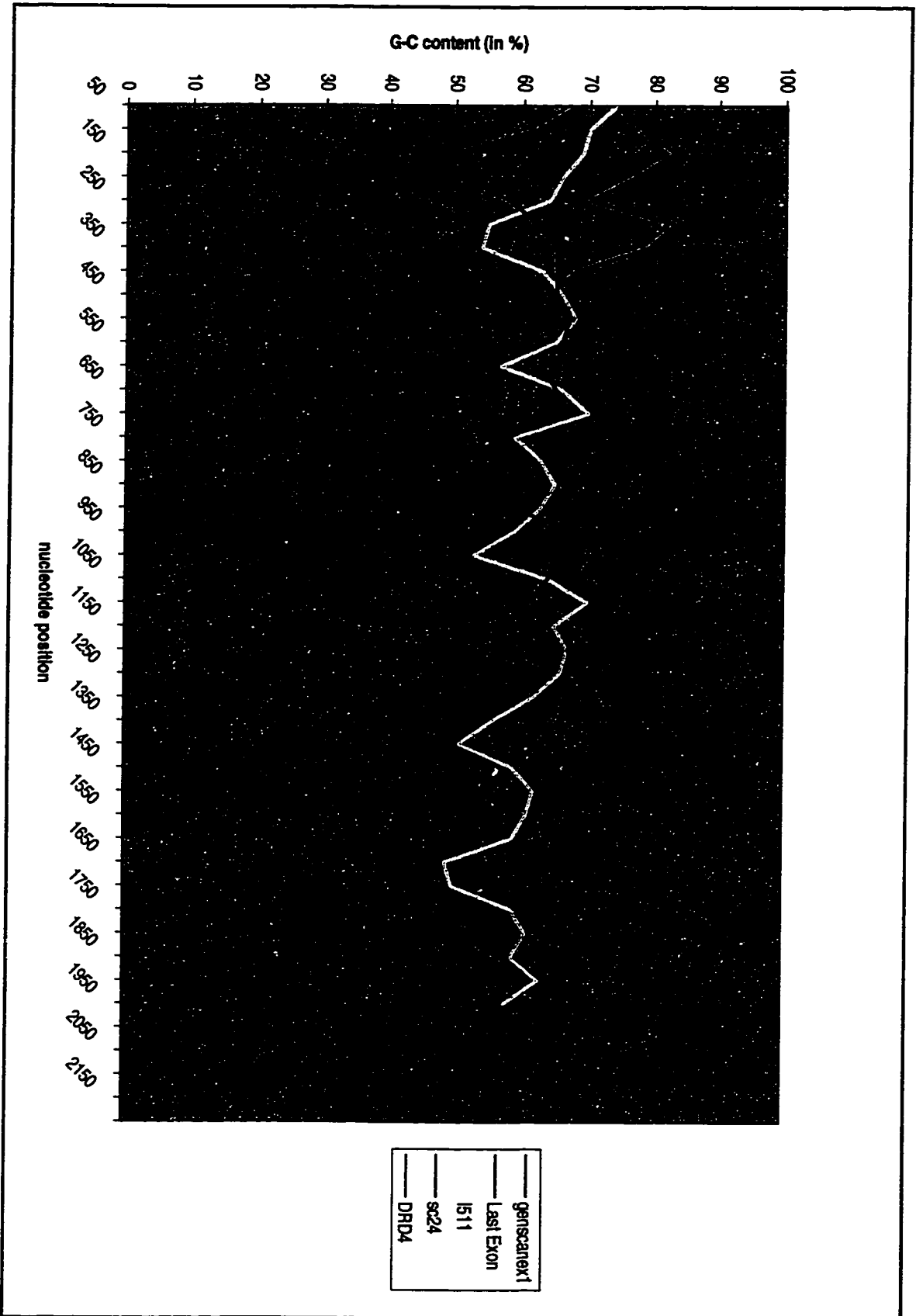
>R1F-M1314R, 454 bp

```

1   CCAAGTCCATGACAGCTGAGCTCGAGGAACCTG | CCTCCATTCGGAGAAGA
51  AAAGGGG | AGAAGCTGGACGAGATGCTGGCAGCCGCCGAGAGCCAACGCT
101 GCGGCCAGACATCGCAGACGCAGACTCCAGAGCCGCCACCGTCAAACAGA
151 GGCCCACCAGTCGGAGGATCACACCCGCCGAGATTAGC | TCATTGTTTGAA
201 CGCCAGGGCCTCCCAGGCCAGAGAAGCT TGCCGGG CCTGCCCGCCCTCGCT
251 ACCTCTTCCAGAGAAGGTCCAAGCTATGGGGGGACCCGTGGAGAGCCGGG
301 GGCTCCCTGGGCCTGAAGACGACAAACCAACTGTGATCAGTGAGCTCAGC
351 TCCCGCCTGCAGCAGCTGAACAAGGACACGCGTTCCCTGGGGGAGGAACC
401 AGTTGGTGGCCTGGGCAGCCTGCTGGACCCTGCCAAGAAGTCGCCCATCG
451 CAGC

```


Figure 19 Determination of the G-C content of the putative ALPR exons (genscanex1 and Last Exon) by the GCG program WINDOW. The window size was set at 100 bp with shifts of 10 bp. In order to compare their relative G-C content, two ALPR cDNA clones (sc24 and I511) and a G-C rich gene dopamine D4 receptor gene (DRD4; GenBank accession number L12398) were also put through the program using the same criteria. The percentage of G-C nucleotides in the window was shown along the sequences of the genes. The nucleotide position refers to the median of each window.



fragments were joined to construct a putative protein for ALPR. This hypothetical ALPR contained the ORFs from the starting exon, sc24, I511, genscanex1, and Last Exon. sc24 contains an out of frame exon H5537. This may represent an alternative splicing of ALPR that generates a premature termination in sc24. Therefore, the out-of-frame H5537 exon in sc24 was removed, and replaced by the in-frame one in I511. The ORF from the terminal exon of I511 was also removed, since the protein should not terminate at that exon if it includes genscanex1 and Last Exon. The resulting ORF was aligned against C33B4.3, cortactin binding protein 1, and other homologous proteins. Figure 20 shows the multiple alignment of those proteins, and Figure 21 shows the protein identities in different positions of the proteins.

ALPR was characterized by the presence of an ankyrin repeat domain (amino acid 168-238, Fig. 21) and a proline rich region (amino acid 749 to 779, Fig. 21). Such features were also found in C33B4.3. The ankyrin repeat domain shared ~42% protein identity or ~60% protein similarity with human ankyrin brain variant 2 (accession number X56958). The proline rich region was also found in the cortactin binding protein 1. However, ALPR did not show a continuous match with C33B4.3 or cortactin binding protein 1. The junctions between sc24 and I511, and between I511 and genscanex1 showed no homology with either proteins. Gaps were found when genscanex1 was aligned against cortactin binding protein 1. ALPR also matched human EST clones 208081, AR, and a fetal brain cDNA DS17. None of these clones have been completely sequenced. AR did not have an uninterrupted ORF. The ORF of AR showed in Figure 20 was a combination of three conceptual open reading frames. The regions where they matched to ALPR are shown in Figure 21. ALPR, C33B4.3, and cortactin binding protein share high homology in two regions. The first one is within I511, where ALPR shows 37% identity to C33B4.3 and 80% to cortactin binding protein 1. The second homologous region is the carboxyl ends of the proteins, where ALPR is 46% identical to C33B4.3 and 66% to cortactin binding protein 1. Other than C33B4.3 and cortactin binding protein 1, the carboxyl end of ALPR also shows homology with the carboxyl ends of some kinases. For example, it matches putative diacylglycerol kinase ETA (EC 2.7.1.107) (40% identity, $p\text{-value} = 2.2 \times 10^{-8}$), non-receptor tyrosine kinase spore lysis A (EC 2.7.1.112)

Figure 20 Multiple alignment of the proteins that share homology with ALPR. The amino acid sequences were first aligned by the pairwise program Gap Blast, the alignments were then combined into a multiple alignment matrix. Identical residues are marked in black boxes, whereas similar amino acids are in gray boxes. The amino acid sequence of the ALPR hypothetical protein is composed of the open reading frames of two cDNA clones (sc24, I511), and three putative exons (starting exon, genscanex1, Last Exon) (see text). The cDNA clone AR does not have an uninterrupted ORF, the amino acid sequence of AR is a combination of three conceptual open reading frames. The consensus sequence of proline rich ligand that is specific for cortactin is +PPΨPXKPXWL, where “+”, “Ψ” and “X” stands for basic, aliphatic, and any amino acid residue respectively (Sparks et al. 1996). It is highlighted in red at the position 1343-1351 of ALPR hypothetical protein, and position 946-956 of rat cortactin binding protein 1.

Figure 20 Multiple alignment of proteins that share homology with ALPR

ALPR hypothetical	D G P G A S A V R G D Q Q T C R	25
C33B4.3	N Q E E D T - N G F D N V R F A	24
Cortactin binding		
cDNA 208081		
AR		
DS17		
ALPR hypothetical	D D A A P V A A Q V C A N H S Q D	50
C33B4.3	T Q N D F D V R K L A T P Q A P Q	49
Cortactin binding		
cDNA 208081		
AR		
DS17		
ALPR hypothetical	L Q S R D L Q	75
C33B4.3	F L C D L D T	74
Cortactin binding		
cDNA 208081		
AR		
DS17		
ALPR hypothetical	P N L D T P F R A Q N	100
C33B4.3	- - F T D C V L K K K M L	97
Cortactin binding		
cDNA 208081		
AR		
DS17		
ALPR hypothetical	L D F A K L A N L	125
C33B4.3	N E L K A M G Q L Q	122
Cortactin binding		
cDNA 208081		
AR		
DS17		
ALPR hypothetical	H S T D A R E L D K P P D S	150
C33B4.3	K N N E E K M C S Q A A Q -	146
Cortactin binding		
cDNA 208081		
AR		
DS17		
ALPR hypothetical	C Q D A T D L K V K N	175
C33B4.3	T G P N R A V I G	171
Cortactin binding		
cDNA 208081		
AR		
DS17	C L	11

Figure20 (continued.....1)

ALPR hypothetical C33B4.3	T [REDACTED] [REDACTED] C T R Q [REDACTED] A A [REDACTED] N S [REDACTED] Q [REDACTED] F L S [REDACTED] F E N [REDACTED] K	200 196
Cortactin binding cDNA 208081 AR DS17	A [REDACTED] [REDACTED] [REDACTED] C A [REDACTED] H C L [REDACTED]	36
ALPR hypothetical C33B4.3	[REDACTED] [REDACTED] [REDACTED] S [REDACTED] [REDACTED] [REDACTED] Y [REDACTED] M [REDACTED] [REDACTED] [REDACTED] [REDACTED] P I [REDACTED]	225 221
Cortactin binding cDNA 208081 AR DS17	A [REDACTED] G [REDACTED] T [REDACTED] R [REDACTED] F [REDACTED] T [REDACTED]	61
ALPR hypothetical C33B4.3	- G G G [REDACTED] A L C C [REDACTED] H [REDACTED] Q [REDACTED] [REDACTED] [REDACTED] T A D S [REDACTED] D Q V A [REDACTED] R [REDACTED] A [REDACTED] D [REDACTED] V [REDACTED] M [REDACTED]	249 246
Cortactin binding cDNA 208081 AR DS17	[REDACTED] T X X C K [REDACTED] X	11
ALPR hypothetical C33B4.3	[REDACTED] Q [REDACTED] [REDACTED] F [REDACTED] H [REDACTED] [REDACTED] A [REDACTED] [REDACTED] N H [REDACTED] [REDACTED] N [REDACTED] L T [REDACTED] [REDACTED] [REDACTED] G	274 271
Cortactin binding cDNA 208081 AR DS17	[REDACTED] X [REDACTED] [REDACTED] X [REDACTED] X [REDACTED] [REDACTED] X	36
ALPR hypothetical C33B4.3	[REDACTED] G [REDACTED] [REDACTED] [REDACTED] [REDACTED] [REDACTED] [REDACTED] [REDACTED] Q [REDACTED] D [REDACTED] [REDACTED] V [REDACTED] [REDACTED] S P [REDACTED] [REDACTED] N [REDACTED] P E [REDACTED]	299 296
Cortactin binding cDNA 208081 AR DS17	[REDACTED] S X [REDACTED] [REDACTED] W [REDACTED] I [REDACTED] [REDACTED] D [REDACTED]	61
ALPR hypothetical C33B4.3	[REDACTED] [REDACTED] [REDACTED] N [REDACTED] [REDACTED] [REDACTED] Q [REDACTED] [REDACTED] [REDACTED] [REDACTED] [REDACTED] D H L A V [REDACTED] K Q G [REDACTED] L H [REDACTED] S H	324 321
Cortactin binding cDNA 208081 AR DS17	[REDACTED] [REDACTED] [REDACTED] X X [REDACTED] [REDACTED] E L K [REDACTED] [REDACTED] P [REDACTED] X [REDACTED]	86
ALPR hypothetical C33B4.3	[REDACTED] [REDACTED] [REDACTED] T [REDACTED] [REDACTED] [REDACTED] [REDACTED] E [REDACTED] [REDACTED] V [REDACTED] P G [REDACTED] [REDACTED] D [REDACTED] V G A [REDACTED] N P K S S [REDACTED] [REDACTED] G [REDACTED]	349 346
Cortactin binding cDNA 208081 AR DS17	[REDACTED] X [REDACTED] [REDACTED] [REDACTED] N [REDACTED] [REDACTED] [REDACTED]	104

Figure 20 (continued.....2)

ALPR hypothetical C33B4.3 Cortactin binding cDNA 208081 AR DS17	<p>S K G P G A S P P Q - Q T - - T T R R S S G S</p>	373 369
ALPR hypothetical C33B4.3 Cortactin binding cDNA 208081 AR DS17	<p>- - A S N - - - - L E A Q P A G I C S Q V Y R T P Q S V R P M S A P S</p>	390 394
ALPR hypothetical C33B4.3 Cortactin binding cDNA 208081 AR DS17	<p>L R L P H Q L L L Q R L Q E E D R D R D A R S R T T I T P S E Y G T M R S G M D S M</p>	415 419
ALPR hypothetical C33B4.3 Cortactin binding cDNA 208081 AR DS17	<p>D Q E S N S G P L A G R A G O S K S V L S I R G G G M A G H E T N I A I L - - - - -</p>	440 440
ALPR hypothetical C33B4.3 Cortactin binding cDNA 208081 AR DS17	<p>E G F W E G T V K G R T G W F P A D C V E E V Q - - - - -</p>	465 440
ALPR hypothetical C33B4.3 Cortactin binding cDNA 208081 AR DS17	<p>M R Q H D T R P E T R E D R T K R F R H T V - - - - - M M S V P G G G A A T V M T G N N</p>	490 440 20
ALPR hypothetical C33B4.3 Cortactin binding cDNA 208081 AR DS17	<p>S D S L S Y V D V A H - - - - - G V R P R N L Y C E T V N</p>	515 443 45

Figure 20 (continued.....3)


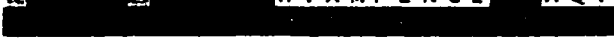




















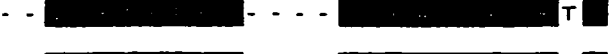
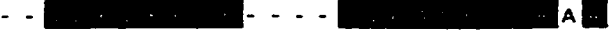





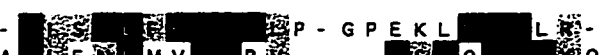

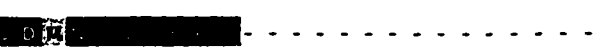



ALPR hypothetical C33B4.3		540
Cortactin binding cDNA 208081		467
AR		70
DS17		6
		16
ALPR hypothetical C33B4.3		565
Cortactin binding cDNA 208081		493
AR		95
DS17		31
		41
ALPR hypothetical C33B4.3		590
Cortactin binding cDNA 208081		518
AR		120
DS17		56
		63
ALPR hypothetical C33B4.3		614
Cortactin binding cDNA 208081		537
AR		145
DS17		81
ALPR hypothetical C33B4.3		639
Cortactin binding cDNA 208081		551
AR		169
DS17		105
ALPR hypothetical C33B4.3		664
Cortactin binding cDNA 208081		561
AR		189
DS17		125
ALPR hypothetical C33B4.3		687
Cortactin binding cDNA 208081		587
AR		213
DS17		149
ALPR hypothetical C33B4.3		709
Cortactin binding cDNA 208081		606
AR		238
DS17		158

Figure 20 (continued.....4)

ALPR hypothetical	-- K [REDACTED] R - [REDACTED] K S V E D [REDACTED] L [REDACTED] L G [REDACTED] F	731
C33B4.3	M Q Y D Q E S L N G G Y [REDACTED] S K [REDACTED] Y N [REDACTED] V S [REDACTED] Q M [REDACTED] R	631
Cortactin binding	P F [REDACTED] [REDACTED] [REDACTED] [REDACTED] [REDACTED] [REDACTED] D [REDACTED] F L S G I	263
AR	-- [REDACTED] [REDACTED] [REDACTED] [REDACTED] [REDACTED] G K [REDACTED]	175
DS17		
ALPR hypothetical	-----	731
C33B4.3	R K G Q [REDACTED] N V [REDACTED] A S S A G L N R S T F E Q A A P T	656
Cortactin binding	T E E E [REDACTED] Q F [REDACTED] A P P M L K F T R S L S M P D T S	287
ALPR hypothetical	-----	731
C33B4.3	T S T F E Y N C S C R S T P Q L S R M D S F D S F	681
Cortactin binding	-----	287
ALPR hypothetical	-----	731
C33B4.3	D D E D E M P A P P A S Y I S P D L Q R D S S M	706
Cortactin binding	-----	287
ALPR hypothetical	-----	731
C33B4.3	Q R S E Y S R P F R P T S R P K T P P P P P M Q	731
Cortactin binding	-----	287
ALPR hypothetical	----- [REDACTED] T [REDACTED] M Q D [REDACTED]	741
C33B4.3	H Q N H Q N H Q Y Q Q Q H P S L [REDACTED] A [REDACTED] T P Q [REDACTED]	756
Cortactin binding	-----	287
ALPR hypothetical	V [REDACTED] E [REDACTED] G [REDACTED] [REDACTED] [REDACTED] Q T A [REDACTED] [REDACTED] [REDACTED] [REDACTED] A P Y	763
C33B4.3	[REDACTED] Q Q Q Q S [REDACTED] [REDACTED] [REDACTED] P P P [REDACTED] [REDACTED] H C E [REDACTED] T M V	781
Cortactin binding	-----	287
ALPR hypothetical	[REDACTED] F [REDACTED] D [REDACTED] S G [REDACTED] [REDACTED] A F [REDACTED] P [REDACTED] [REDACTED] [REDACTED] G R A Y D [REDACTED] V R S	788
C33B4.3	[REDACTED] H V [REDACTED] [REDACTED] F T [REDACTED] [REDACTED] S T S [REDACTED] [REDACTED] [REDACTED] [REDACTED] L P [REDACTED] I [REDACTED] S S G	806
Cortactin binding	-- [REDACTED] D I [REDACTED] P Q [REDACTED] [REDACTED] S [REDACTED] [REDACTED] P S [REDACTED] [REDACTED] T Y N	311
ALPR hypothetical	[REDACTED] S F [REDACTED] K [REDACTED] G L E A R L [REDACTED] [REDACTED] A G A A G L Y E P G A A L G	813
C33B4.3	[REDACTED] [REDACTED] P [REDACTED] P P P P G [REDACTED] M H V A A S A [REDACTED] L M S	832
Cortactin binding	C [REDACTED] E S [REDACTED] T [REDACTED] R V Y [REDACTED] T [REDACTED] K P S F N Q N [REDACTED] - A K	335
ALPR hypothetical	P L [REDACTED] Y P E R Q K R [REDACTED] R S [REDACTED] I L Q D S A [REDACTED] E S G	838
C33B4.3	N S K G I S [REDACTED] A [REDACTED] K S V Q K [REDACTED] A - - E [REDACTED] T	854
Cortactin binding	V P [REDACTED] A T R [REDACTED] T [REDACTED] M [REDACTED] R E [REDACTED] G M F Y R [REDACTED] L	360
ALPR hypothetical	[REDACTED] P R P P P A A T P P E R [REDACTED] K R R P R P P G P D	863
C33B4.3	S [REDACTED] A [REDACTED] V S [REDACTED] N N N N N [REDACTED] S T T D F Q M D L [REDACTED] N	879
Cortactin binding	[REDACTED] R F [REDACTED] F D S E D V Y [REDACTED] R [REDACTED] A P Q A A F R T [REDACTED] R	385
ALPR hypothetical	S P Y A N - - - - - L G [REDACTED] F [REDACTED] A [REDACTED] [REDACTED] [REDACTED] [REDACTED]	881
C33B4.3	A L A K R R - - - - - S [REDACTED] [REDACTED] - H D V D E D E D R	898
Cortactin binding	G Q M P E N P Y S E V G [REDACTED] [REDACTED] [REDACTED] - K [REDACTED] V [REDACTED] [REDACTED]	409

Figure 20 (continued.....5)

ALPR hypothetical	Q █████ S P █████ L Q █████ Q █████ A A - - -	902
C33B4.3	E █████ F E █████ S █████ E T V R E N V V █████ G K G █████ Q	923
Cortactin binding	█████ █████ S N █████ P █████ K T C S █████ P	434
ALPR hypothetical	- - - L A █████ G S █████ G P G - - - - - - - - - G █████ F A R	918
C33B4.3	N I G █████ N █████ K D █████ G Y T █████ S █████ █████ Q █████ S █████ M E T	942
Cortactin binding	I P T █████ E █████ T S █████ G █████ █████ Q █████ S █████ M E T	459
ALPR hypothetical	█████ P █████ H R █████ P R P G G L D Y G A G D G P G L █████	943
C33B4.3	█████ E █████ E K D H █████ H F █████ D - H █████ N V Q R	966
Cortactin binding	D █████ Q A █████ P █████ Q L █████ D D █████ T V S █████ F A A █████	484
ALPR hypothetical	F G █████ P G P A █████ E █████ R - █████ T V █████ V	967
C33B4.3	V T L I S █████ H L █████ D N Y G Q █████ D █████ M █████ V █████ S S	991
Cortactin binding	A █████ A V █████ D █████ A █████ - █████ P █████	508
ALPR hypothetical	G - A I █████ G S A P █████ A D L █████ S █████ - - - - -	984
C33B4.3	A S S S S T █████ D █████ T K █████ G C F █████ V █████ H V I P P V	1016
Cortactin binding	D L G D █████ D █████ G █████ P █████ A █████ R █████ K F P E E G	533
ALPR hypothetical	- S █████ E	990
C33B4.3	D █████ D █████ P █████ G T G D S D G E I R C S E I S - -	1038
Cortactin binding	G █████ G █████ E █████ E Q P L L P T P G A A P █████ E █████ N	558
ALPR hypothetical	█████ L █████ T █████ P T █████ R D L L L P S █████ V █████ L K - -	1012
C33B4.3	█████ █████ █████ █████ █████ █████ █████ █████ █████ █████ █████	1038
Cortactin binding	H F █████ G █████ E - █████ A Q G E A G G █████ L █████ S T S K A	582
ALPR hypothetical	- - - - - - - - - - P L V S G P █████ L G P █████ G S T █████	1031
C33B4.3	- -	1039
Cortactin binding	K G P E S G P A A A L K S █████ S P A █████ P E N █████	607
ALPR hypothetical	█████ █████ P █████ █████ █████ █████ █████ █████ A S	1055
C33B4.3	█████ █████ █████ █████ █████ █████ █████ █████ █████ █████	1039
Cortactin binding	█████ █████ R L █████ █████ █████ █████ █████ █████ M Q E	631
ALPR hypothetical	Q A P █████ S P T P V H █████ D █████ R P G █████ █████ V	1080
C33B4.3	- -	1039
Cortactin binding	- - - █████ Q Q G H K G E A █████ K █████ L N K █████ █████ T	653
ALPR hypothetical	█████ A █████ D P E R G S L A S P A █████ S █████ R S P A W I P V	1105
C33B4.3	- -	1039
Cortactin binding	█████ M █████ P S - - - - - V E S G █████ P █████ V T R Q N T R G	674
ALPR hypothetical	█████ A █████ E A E G V P R █████ - - - - - E R K S P E █████	1126
C33B4.3	- -	1039
Cortactin binding	█████ L █████ Q E T E N K Y █████ T D L S █████ D █████ R A █████	699

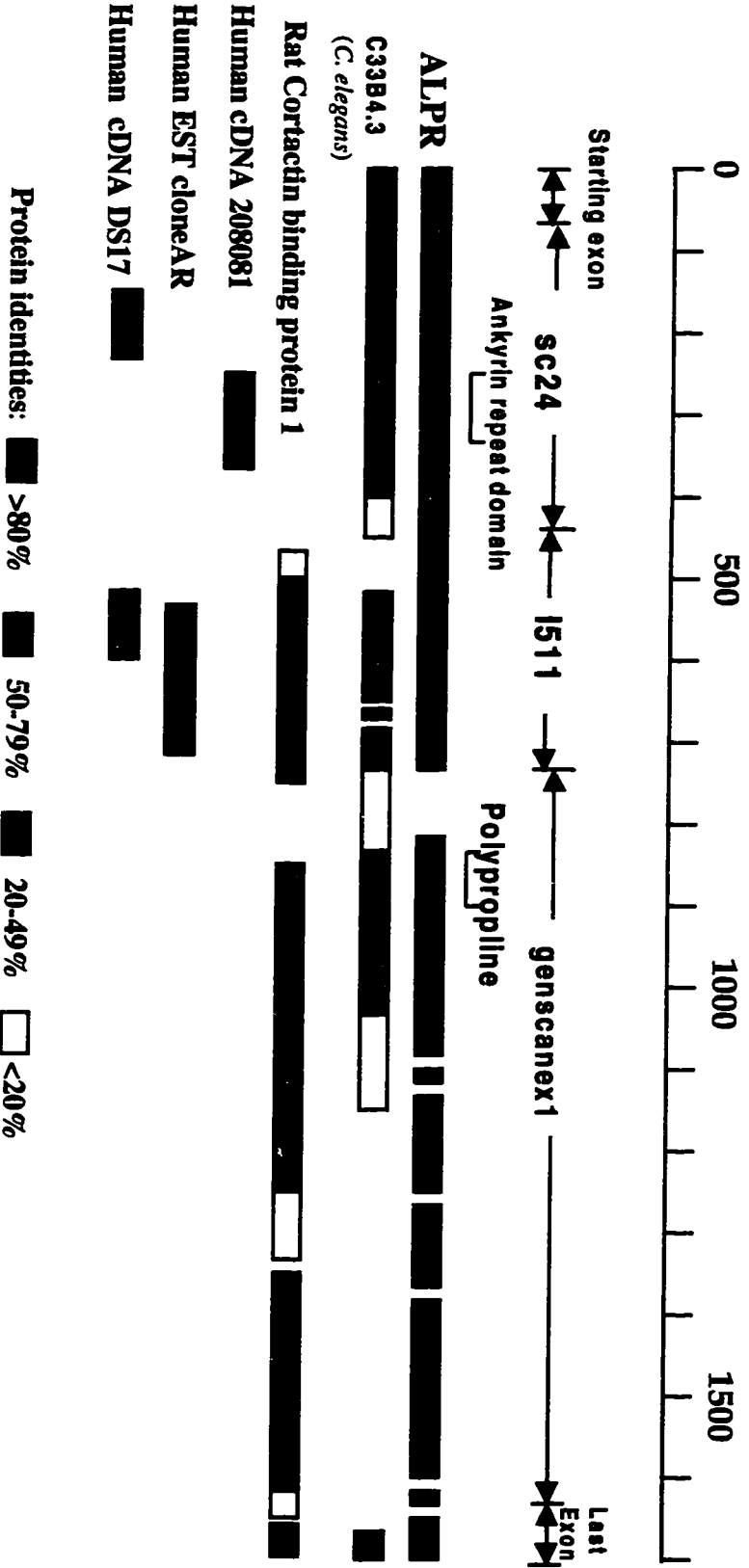
Figure 20 (continued.....6)

ALPR hypothetical C33B4.3	S ██████████ L P ██████████ A T S N	1151
Cortactin binding	██████████ Q S ██████████ T V D I	724
ALPR hypothetical C33B4.3	G Q E P S R ████████ G G A ████████ E R P ████████ - - - - -	1166
Cortactin binding	P V A G P P ████████ E E E ████████ R E D ████████ D T K P D H S P S	749
ALPR hypothetical C33B4.3	- T ████████ L A ████████ - - - A P ████████ S ████████ V A ████████ L P ████████	1188
Cortactin binding	T V ████████ G V ████████ K T E G A ████████ I ████████ A P ████████ A A ████████	773
ALPR hypothetical C33B4.3	R A Q P P G G T P A D A G P G Q G ████████ S E E ████████ P E ████████	1213
Cortactin binding	- - - - - G R T I V A ████████ G S V ████████ E A ████████	788
ALPR hypothetical C33B4.3	██████ F A V N ████████ A Q ████████ S ████████ E ████████ T R ████████ L A R I	1238
Cortactin binding	██████ L P F R ████████ P P ████████ V ████████ L ████████ - - - D F L F T	811
ALPR hypothetical C33B4.3	G L ████████ E ████████ G ████████ L ████████ A T P L ████████ G P G ████████	1263
Cortactin binding	E P ████████ L ████████ S ████████ D ████████ P D D R ████████ A S V ████████	836
ALPR hypothetical C33B4.3	P T T V - - - - -	1266
Cortactin binding	L A D L V K Q K K S D T P Q P P S L N S S Q P A N	861
ALPR hypothetical C33B4.3	- - - - - P S P ████████ G K ████████ E P ████████ A P ████████	1284
Cortactin binding	S T D S K K P A G I ████████ C L ████████ F L ████████ P - ████████	885
ALPR hypothetical C33B4.3	- - - A A ████████ A ████████ P ████████ R ████████	1307
Cortactin binding	F D A V T ████████ V ████████ H ████████ E ████████	910
ALPR hypothetical C33B4.3	██████████ ██████████ S ████████ L T ████████ H ████████ S	1332
Cortactin binding	██████████ ██████████ G ████████ S M ████████ C ████████ V	935
ALPR hypothetical C33B4.3	██████ H T ██████████ ██████████ S P ████████ G	1357
Cortactin binding	██████ Q A ██████████ ██████████ P I ████████ H	960
	+ P P P P X K P X W L	
ALPR hypothetical C33B4.3	██████ G P V T ████████ P ████████ L K Q S ████████ S E ████████ A Q Q H H	1382
Cortactin binding	██████ S N A L ████████ T ████████ P E E D ████████ G F ████████ P P P A P	985

Figure 20 (continued.....7)

ALPR hypothetical	A A S A L S A G P A R P R Y F Q R	1407
C33B4.3	- - - - -	1039
Cortactin binding	P P P P S Q G V A K V - - - Q P T	1007
ALPR hypothetical	P V S G - - P E D D P T	1430
C33B4.3	- - - - -	1039
Cortactin binding	V T V S P I S - - - A N	1029
ALPR hypothetical	R D T R S L E E P V G L S	1455
C33B4.3	- - - - -	1039
Cortactin binding	A I - - - - K S V K P E L	1049
ALPR hypothetical	L D P A K P I A A L F S L G E	1480
C33B4.3	- - - - -	1039
Cortactin binding	E P V G A A N L P - - - P E V	1071
ALPR hypothetical	A Q P G G P G G G A S G R Y	1505
C33B4.3	- - - - -	1039
Cortactin binding	G T T - - - - V G T S Q	1091
ALPR hypothetical	A - - - - - - - - - - K	1514
C33B4.3	- - - - -	1039
Cortactin binding	T L Q S R P P D Y E S R T S G P V S	1116
ALPR hypothetical	A S E V E G L G A G G G R - - - -	1533
C33B4.3	- - - - -	1039
Cortactin binding	T E S E I L P T P P A A A S S P T L S	1141
ALPR hypothetical	- - - - - - - - - - T P T I L K	1545
C33B4.3	- - - - -	1039
Cortactin binding	D V F S L P S Q S P A G D L N A G R S R	1166
ALPR hypothetical	S L P H E K E V R F V V R S V S A R S R S	1570
C33B4.3	- - - - -	1039
Cortactin binding	P P L Q Q - - - - - - - - - -	1175
ALPR hypothetical	P S P S P L P S P A S G P G P G A P G P R R Q	1595
C33B4.3	- - - - -	1042
Cortactin binding	- - - - - - - - - - I S N K T	1183
ALPR hypothetical	Q K D V D I G S S Y T	1620
C33B4.3	H - - - - -	1067
Cortactin binding	T H T P A N	1207
ALPR hypothetical	R E H A T M	1645
C33B4.3	P A R S Q R N R C R Q C D R S R T Q	1091
Cortactin binding	E T M N D N Q E L	1231
ALPR hypothetical	- - - - - D G S	1664
C33B4.3	A Q I S G Q	1110
Cortactin binding	- - - - - D R	1252

Figure 21 Protein identities between ALPR and its homologous proteins. Protein sequences are represented by the bars. Split bars indicate gaps introduced between the sequences to allow alignment. Protein identities are measured as the percentage of amino acid sequences that are identical to ALPR in a 100 amino acid window. For the amino acid sequences that are less than 100 residues, percentage of protein identity is measured as the number of identical amino acids which are shared with ALPR divided by the total number of the amino acids. The scale above the proteins indicates the number of amino acid residues.



(37% identity, p -value = 3.7×10^{-5}), and EPH receptor tyrosine kinase (31% identity, p -value = 1×10^{-4}). However, the carboxyl ends of those kinases do not contain the kinase catalytic domain. This result therefore does not suggest that ALPR is a kinase

Characterization of C33B4.3, the *C. elegans* homolog of ALPR

Life cycle and basic anatomy of C. elegans

C. elegans is a powerful tool for genetic research because it is small (≤ 1.5 mm), easy to grow in laboratory conditions, and has a rapid life cycle (3 days). Furthermore, its genome can easily be manipulated by microinjection of DNA or RNA into the gonad of a hermaphrodite. The phenotypic effect can then be observed in the large brood she produces.

C. elegans has a simple life cycle. A fertilized egg takes ~ 14 hours to complete embryogenesis. The larva that hatches from the egg is called the L1 larva (L1). With sufficient food in the environment, the L1 goes through three more molts (L2 – L4) before it develops into an adult. If the food resource is depleted, a nondeveloping dauer stage will be formed at the second larval molt. The dauer can survive up to 20 weeks. When food resources are available again, the dauer will resume development and form the adult. Organs that are specific for the adult stage begin to develop in various larvae stages (reviewed by Riddle et al. 1997). For example, in the hermaphrodite, the vulva is an opening for receiving sperm from the male and laying eggs. It begins to develop at the L3 stage. The organs that are specific for male copulation develop during the L4 stage. Hermaphrodite and male are the two sexes in *C. elegans*, which is determined by the number of sex (X) chromosomes. The hermaphrodite has two X chromosomes (XX) whereas the male has one (XO). The hermaphrodite can reproduce by either self-fertilization or crossing with the male. Since almost all sperm and eggs produced by a hermaphrodite carry an X chromosome, self-fertilization can only produce hermaphrodite progeny.

C. elegans has a simple body plan which is known as a triploblast (three germ layers). The ectoderm consists of the outermost cell layers (hypodermal cells) and nervous system. Mesoderm consists of all of the organs in between the ectoderm and

endoderm, including muscles of the body wall, the pharynx that pumps food into the intestine, and gonad. Endoderm or gut is the innermost germ layer.

Confirmation of the polyproline sequences

The putative gene C33B4.3 was identified by the gene prediction program GENEFINDER in AceDB (A *C. elegans* database). The gene includes 11 exons that span 5850 bp within a *C. elegans* cosmid genomic clone C33B4 (GenBank accession number Z48367). The putative protein is very rich in proline residues between amino acids 721-816 (Fig. 22). In order to test whether the sequences are actually expressed and not an artifact of the prediction program, RT-PCR was used to amplify the proline rich region (PCR product B431HR1-B4.332 and B4.3FP-B4.322, Table 3). The sequencing of two PCR products confirmed that the polyproline region is part of an expressed sequence.

Reporter gene fusion studies of C33B4.3

PCR was used to amplify a 6129 bp *C. elegans* genomic DNA product, which contained ~4.7 kb of the putative promoter region and the first 4 predicted exons (Table 3, Fig. 22). Since the genes in the *C. elegans* genome are small, the promoter region could be as small as 500 bp from the transcription start point (Morris Maduro, personal communication). I cloned a large upstream sequence to ensure *cis* regulatory elements (such as enhancers) were included in the expression construct. Four exons were included in the expression construct so that I could confirm the authenticity of those exons. Those exons were only putative exons predicted by the GENEFINDER program in AceDB (a *C. elegans* database). If reporter gene expression occurred when the exons fused in-frame with the lacZ gene, I could then confirm those exons and their ORFs were real. As a control, those exons were also fused with an out-of-frame construct in order to eliminate the possibility of ectopic expression of the reporter gene. The PCR was done using Precision *TaqPlus* polymerase (Stratagene), which is a mix of *Taq* and *Pfu* polymerases to reduce the mutation rate in long range PCR. The product was fused with reporter genes

Figure 22 The amino acid sequence of the *C. elegans* protein C33B4.3. The polyproline sequences are marked in red. The sequence in blue represents the ORF of the four exons that were fused with *gfp* and *lacZ* for expression studies. The underlined sequence is the ORF of the partial cDNA that was used as a template for synthesizing double stranded RNA by *in vitro* transcription, which was used in the double stranded RNA interference study.

C33B4.3, 1100 amino acids

1 MNQEEDTVNLQIFVPELNVKFLAVTQNDFIWDVKRKLATLPQALPOAFNYGLELPPCD
61 GRAGKELLEDRITIRDYPFTDCVPYLELKYKRVYKMLNLDEKOLKAMHTKGOLKKFMDYV
121 OOKNNEKVEKMC SOGLDANFHDAOGETPLTLAAGIPNNRAVTVSLIGGGAHVDFRNSEGO
181 TAMHKA AFLSSEENVKTLIELGASPNYRDP IGLTPLYNMLTADSNDOVAEILLREAA DI
241 GVTDMHG NHEIHOACKNGLTKHVEHLLYFGGOIDAENVNGNSPLHVC AVNNRPECARVLL
301 FRGADHLAVNKOGOTALHVSHIVGNPGVADVVOAHNPKSSVPYRGTP OYSTRRRLSSTIT
361 RRRSMSOSSICSODVYRTPOSVRKGPMSAAPS PPSRSRSTPTITPSEYGTMR RSGMDSMR
421 GGGMIAAGHETNIARILVIPRGVKGFILRGAKHVAMPLNFEPTAQVPALQFFEGVDMS
481 GMAVRAGLRPGDYLL EIDGIDVRRCSHDEVVEFIQQAGDTITLKVITVDVADMSRGGTIV
541 HRPPTDTHDAHGV DYYAPNEIRNAYSES RHASVRQRPGSGRRISAAELENLMVRQ RVPSV
601 QGSPYQM QYDQESLNGGYSSKKYNSVSDMKRRKQQRNVVASSAGLNRSTFEQAAPTSTF
661 EYNCSSRSTPQLSRMDSFDSFDEDEMPAPPPASYISPDLQRDSSMQRSEYSRPF RPTSR
721 PKTPPPPPMQHQNHQNHQYQQQHPSLPRSASTPQPIQQQQSSIPPPPPPPPHCEPTM
781 VHVEFTPPSTSSVPPPPPLPPISSGAPPPPPPPGGLMHVAASAPVLM SNSKGISADA
841 LKSVQLKKAEPRETSAA SVSNNNNNNNNSTTDFQMDLKNALAKRRSKVAHDVDEDEERES
901 RFEGLSLRET VRENVVERGKGIQNI GIVNKKDSGYTSSRTSLEPSESEEKDHRPHFSLDH
961 SPNVQRVTLISQHLEDNYGQKDNMSVASSSTASSSSTVDLTKPGCFVVP SHVIPVDYDD
1021 DPDSGTGSDS GEIRCSEISFEHKKVDVWSVDDVIGWLSLHLSEYTPAFRSQRINGRCLR
1081 QCDRSRFTQLGVTRIAHRQII ESALRGLLQ

gfp and *lacZ*. In each fusion protein construct, four clones were pooled to reduce the possibility that a non-functional construct due to PCR-induced mutation would interfere with gene expression. The reporter gene constructs were co-injected into the gonads of *unc-119* mutant hermaphrodites with a plasmid that carries the wildtype *UNC-119* gene (MMO16B) by Dave Hansen or Angela Johnson. Wild type progeny indicated that the extrachromosomal arrays were transmitting to the next generation. The transgenic worms were produced by self-fertilization of the hermaphrodite, which would only give hermaphrodite progeny. In order to study the gene expression in the males, the *unc-119* mutant strain also carries a *him* (high incidence of male) mutation, which gives rise to some male progeny.

For the *gfp* transgenic worms, the fusion protein expression was examined under the fluorescence microscope. Expression of *gfp* was found in the pharynx and rectum of the worms in both sexes. There was weak expression found in the vulval cells in hermaphrodites (data not shown). After X-gal reaction staining, *lacZ* transgenic worms were examined under a bright-field microscope. There were more tissues showing *lacZ* expression than found for the *gfp* transgenic worms. These results suggest that C33B4.3 produced more expression signal in *lacZ* than *gfp* reporter gene constructs. However, the level of expression in these two reporter gene systems usually varies from gene to gene. The expression patterns of *lacZ* were not consistent in all animals, presumably due to the mosaic transmission of the extrachromosomal arrays in different cells. Therefore an attempt was made to integrate the extrachromosomal arrays into the genome by UV irradiation as described in Materials and Methods. Although no line showed 100% transmittance (i.e. all wildtype progeny), there were two lines which showed a high incidence of transmittance (>90% were wildtype). Their *lacZ* expression seemed more stable than those with no UV treatment. These two lines were named AW95-03C7 and AW95-03J10. The *lacZ* expression of these two lines was studied further.

The C33B4.3::*lacZ* fusion protein was expressed in single to a few cells during the very early embryonic stage. Since the expression was still mosaic, it was difficult to trace the cell lineage (data not shown). However, during the two-fold stage of the embryo clear expression was found in the seam cells, which are the specific hypodermal cells arranged as rows of ten cells running along each lateral line. (Fig. 23. A). The identity of

the seam cells was confirmed by Dr Dave Pilgrim. He co-stained the embryos with anti- β -galactosidase antibody and MH27, a monoclonal antibody that recognizes belt desmosomes which serves as a marker for hypodermal cells (Pries and Hirsh, 1986). The MH27 staining overlapped with that the *lacZ* expression (Fig. 23 B and C). The seam cell expression continued to the L1 stage, and only the six posterior cells showed the expression (Fig. 25 A and B). These six cells may correspond to the V1-V6 cells in the ten seam cells (Fig. 25 C). At the three-fold stage when the embryos were ready to hatch, C33B4.3 started to express in the pharynx (Fig. 24 A and B). The pharynx expression continued in the larvae (Fig. 25 A) and adult (Fig. 26 A + C). Expression in the rectum appeared at the larval stage (Fig. 25 B), and continued throughout adulthood (Fig. 27 A and B). Figure 26 A shows the whole adult hermaphrodite. *lacZ* activity staining was found in the head, tail, and vulval region. Closer examination shows that the fusion protein is expressed in the vulval cells that form the structure, and not the muscle cells that control the opening of the vulva (Fig. 26 B). A magnified view revealed that other than the pharynx muscle, the fusion protein was also expressed in cells surrounding the pharynx (Fig. 26 C). The identity of those cells has not been determined.

At the tail region, C33B4.3::*lacZ* was expressed differently between the hermaphrodite and the male. By using anti- β -galactosidase antibody stain, Dr Dave Pilgrim showed the protein was expressed in the six posterior intestinal cells in the hermaphrodite (Fig. 27 B). However, with *lacZ* staining only four intestine cells showed a reaction (Fig. 27A), which presumably was a result of mosaicism. At the rectal region, staining showed four cells surrounding a ring shape structure (Fig. 27 A). The ring corresponded to the anal sphincter muscle which wraps around the intestinal -rectal valve and acts to close it. Among the four surrounding stained cells, the one at the upper left corner sends a process to the intestinal -rectal junction, which resembles the H-shaped anal depressor muscle. The identity of the rest of the three cells was difficult to determine. However, based on the position of the cells, they were probably the B, U, F, or Y postembryonic blast cells. This assumption is further supported by the expression pattern in male tails. The B, U, F and Y postembryonic blast cells do not further divide in

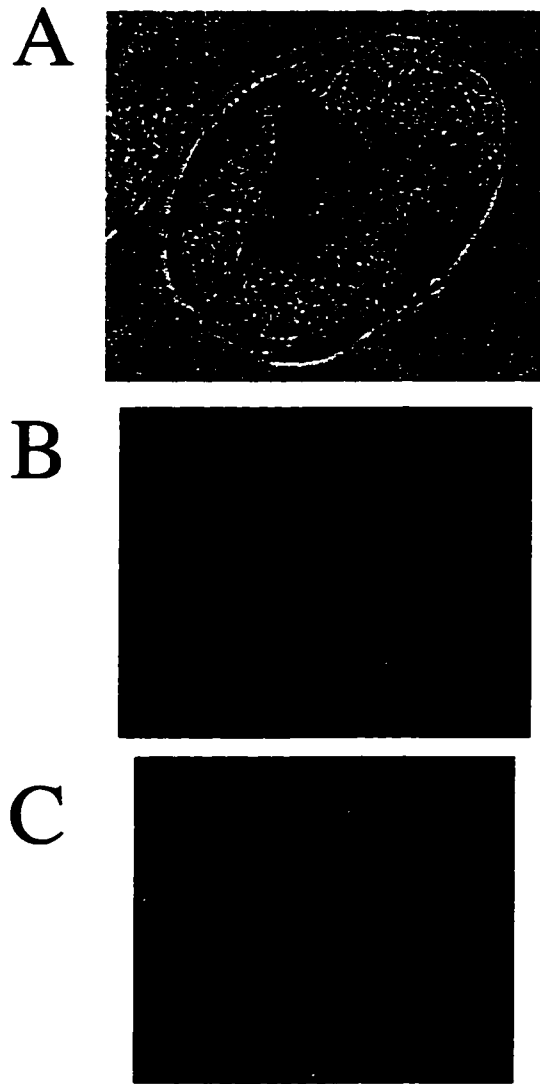


Figure 23 The reporter gene fusion study of C33B4.3. Transgenic worms with C33B4.3::lacZ extrachromosomal arrays were examined under bright field microscopy after X-gal reaction staining, or fluorescence microscopy after immuno staining with anti- β -galactosidase antibody. *A*. A two fold embryo stained for X-gal. *B*. 2-fold embryo stained with anti- β -galactosidase antibody, and the same embryo using MH27 monoclonal antibody stain (*C*). *A* and *B* show the staining of the seam cells. It is confirmed by the overlapping expression pattern of the cells in *B* and the belt desmosomes which serves as a marker for hypodermal cells (*C*). *B* and *C* are provided by Dr Dave Pilgrim and published here with permission.

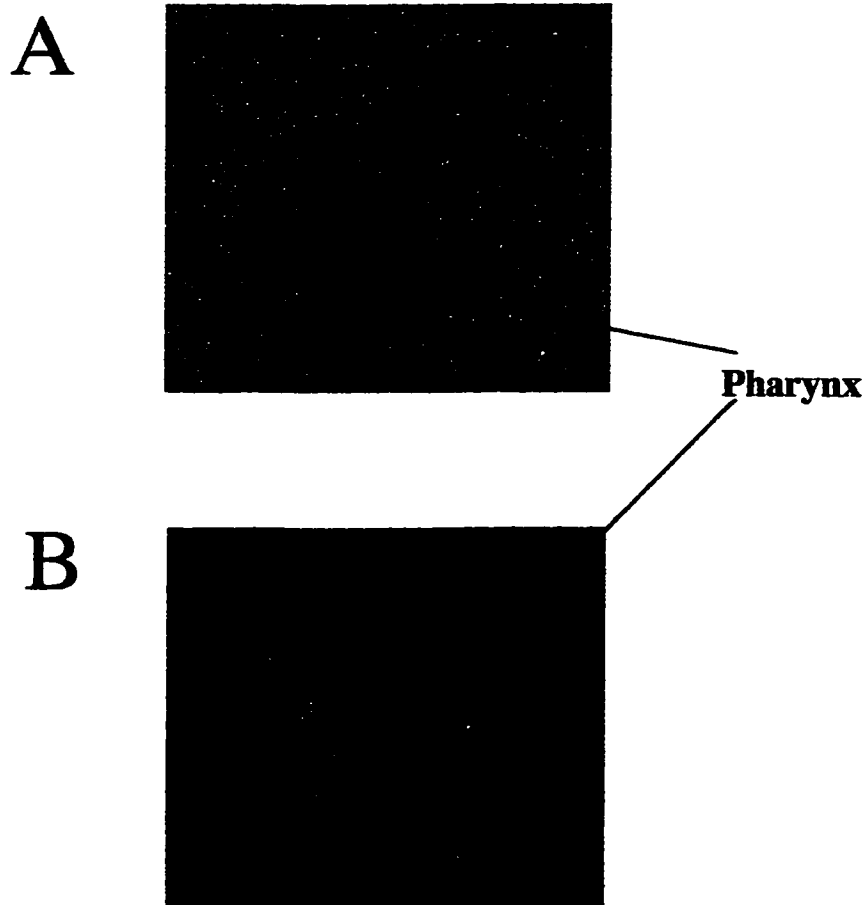


Figure 24 The reporter gene fusion study of C33B4.3. Transgenic worms with C33B4.3::lacZ extrachromosomal arrays were examined under bright field microscopy after X-gal reaction staining, or fluorescence microscopy after immuno staining with anti-β-galactosidase antibody. *A*. A three fold embryo stained for X-gal shows the pharynx expression, as does a three fold embryo using anti-β-galactosidase antibody stain (*B*). *B* is provided by Dr Dave Pilgrim and published here with permission.

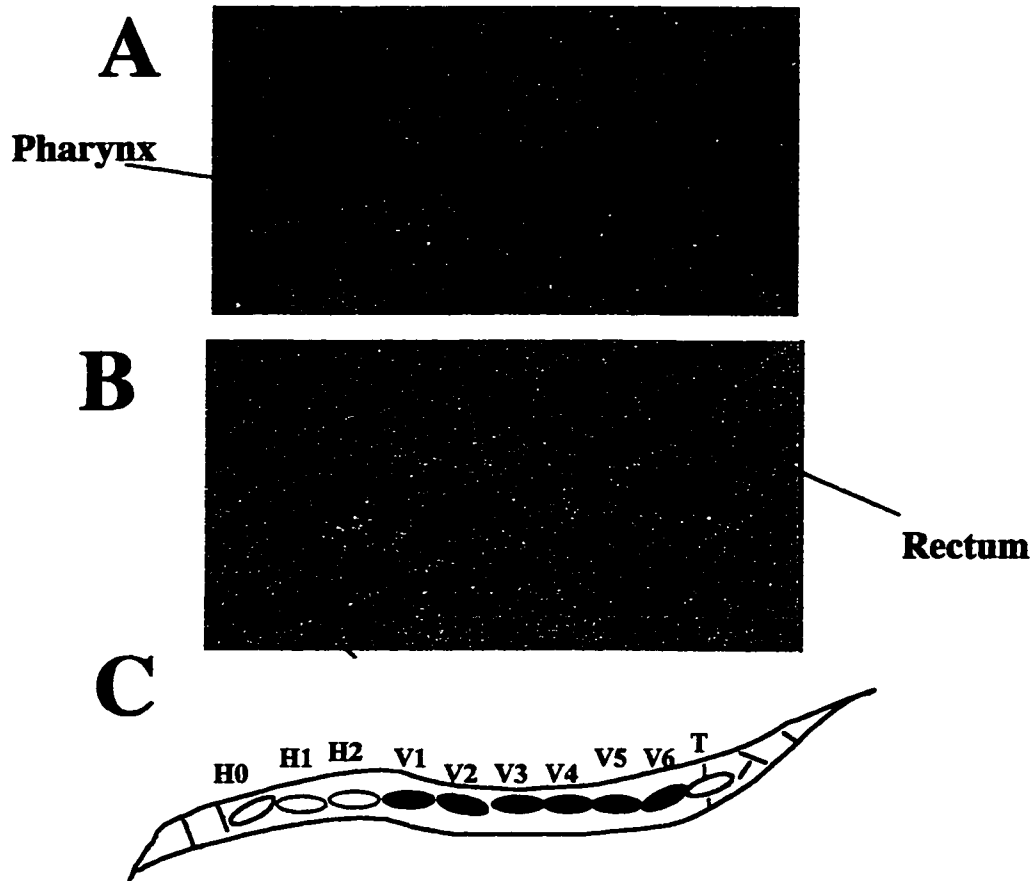


Figure 25 The reporter gene fusion study of C33B4.3. Transgenic worms with C33B4.3::*lacZ* extrachromosomal arrays were examined under bright-field microscopy after X-gal reaction staining. The anterior (A) and the posterior (B) end of the larvae shows ten seam cells that are arranged in a row of ten on each lateral line. Only the six posterior seam cells show the staining. (C) Graphic representation of ten seam cells (from H0 to T) on one lateral line. Only V1 to V6 show staining (marked in blue) (Redrawn from Emmons and Sternberg 1997).

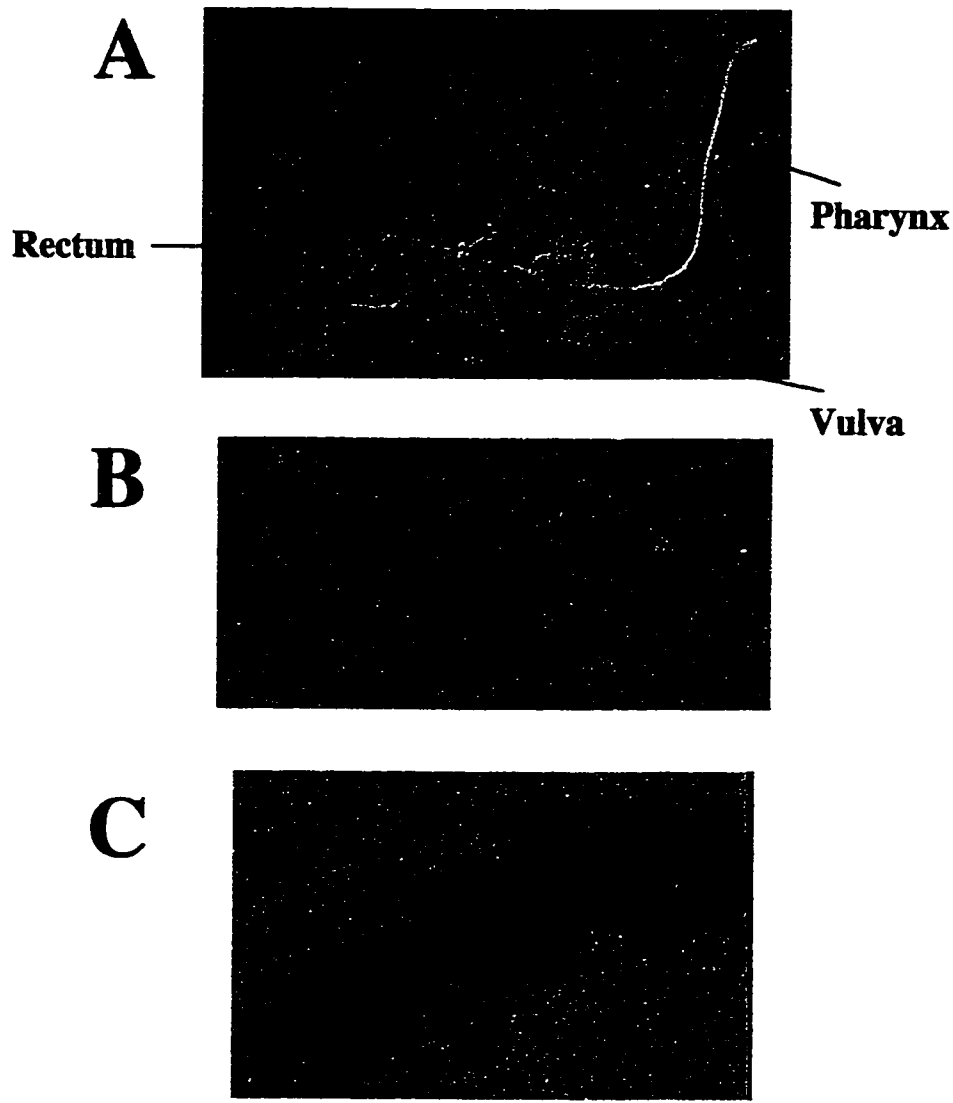


Figure 26 The reporter gene fusion study of C33B4.3. Transgenic worms with C33B4.3::*lacZ* extrachromosomal arrays were examined under bright-field microscopy after X-gal reaction staining. *A*. C33B4.3 expression is shown in the pharynx, vulva, and rectum in an adult hermaphrodite. *(B)* A lateral view of the vulva showing that the expression is in the vulval cells. *C*. A magnified view of the head region showing expression in cells other than the pharynx muscle cells.

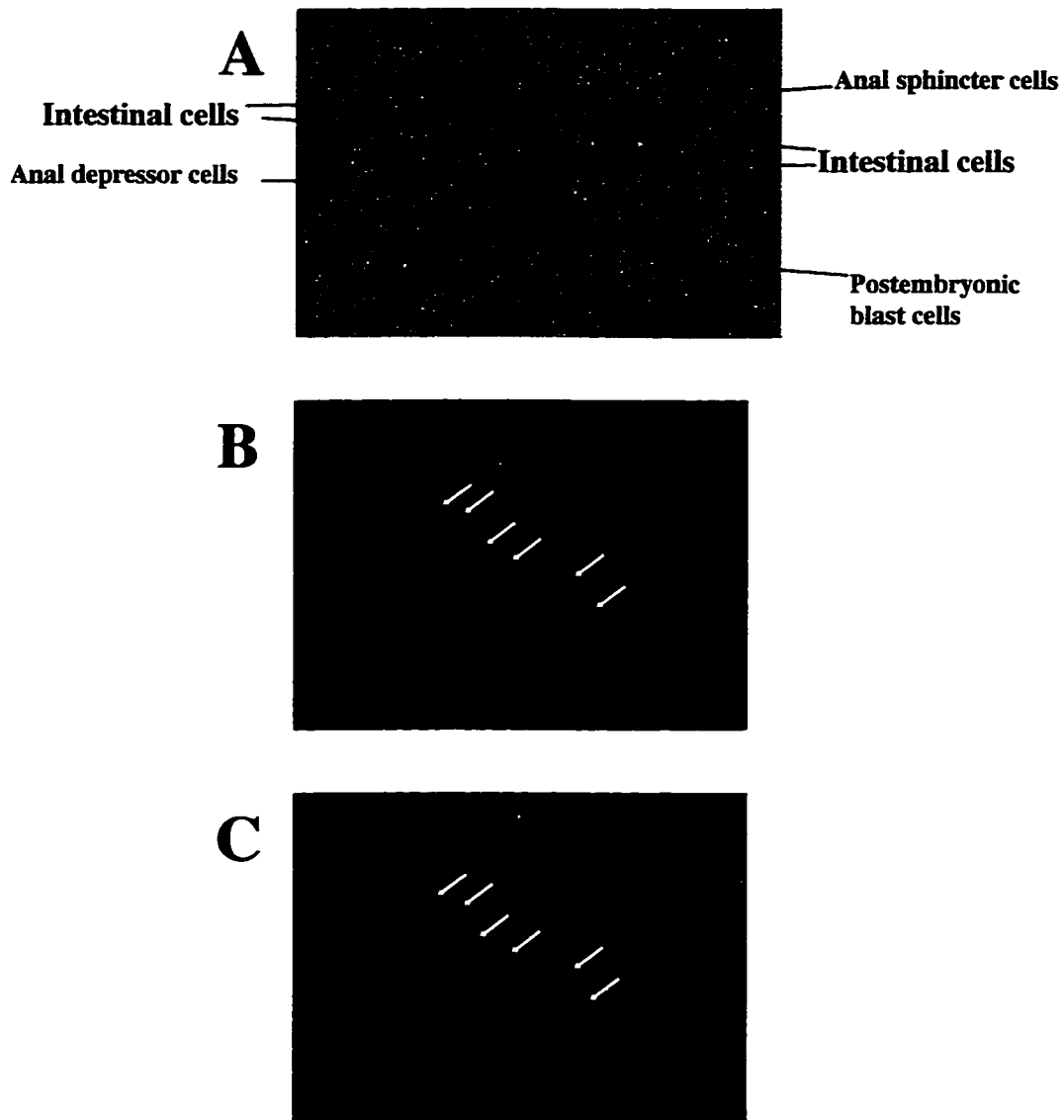


Figure 27 The reporter gene fusion study of C33B4.3. Transgenic worms with C33B4.3::*lacZ* extrachromosomal arrays were examined under bright-field microscopy after X-gal reaction staining, or fluorescence microscopy after immuno staining with anti- β -galactosidase antibody. Rectal and intestinal cells expression is found at the tail of the hermaphrodite using X-gal stain (A), but the expression in the six posterior intestinal cells (in arrows) is clear using anti- β -galactosidase antibody stain (B). The positions of those cells (arrows) can be identified using DAPI stain, which stains all nuclei (C). (B) and (C) are provided by Dr Dave Pilgrim and published here with permission.

the hermaphrodite, but they continue dividing in the male to form the proctodeum (the male rectum) and other male specific structures at the tail. Figure 28 A and B shows numerous cells with X-gal staining in the male tail, which presumably are the descendants of the postembryonic blast cells. Due to the intense staining, it was difficult to identify different cell lineages in the male tails. The expression seemed to occur in lines of cells that extended anteriorly (Fig. 28 B). Even though the intestinal cells are not clearly shown due to the orientation of the tail in Fig. 28 A and B, expression was found in the intestinal cells of other male tails (data not shown).

By anti- β -galactosidase antibody staining, Dr Dave Pilgrim found the C33B4.3::lacZ expression in ventral nerve cord (Fig. 29 A and B). Such expression was not found with the X-gal staining. Due to the low expression level in ventral nerve cord, it may not be detectable by X-gal staining alone.

Double stranded RNA interference study

In order to produce a knockout phenotype for C33B4.3, a partial cDNA B4.3HP was cloned by RT-PCR (Table 3, Fig. 22). It was used as a template to synthesize the sense and antisense mRNA by *in vitro* transcription. Two single stranded RNAs were annealed, and microinjected by Angela Johnson into the gonads of wildtype hermaphrodites. The hermaphrodites were transferred to a new plate each day after microinjection for up to four days, so that the fertilized eggs before microinjection were laid on one plate, while the new eggs after microinjection were laid in subsequent plates. The progeny from twelve different microinjected hermaphrodites were examined under the dissecting microscope. No observable phenotype was seen in those progeny. Those F₁ progeny produced normal F₂ generations. As a positive control, other wildtype hermaphrodites were microinjected with *apx-1* double stranded mRNA. *apx-1* controls the head formation of the embryo. The loss of function mutation results in lethality in the embryo. No egg hatched after microinjection of *apx-1* double stranded mRNA in four hermaphrodites.

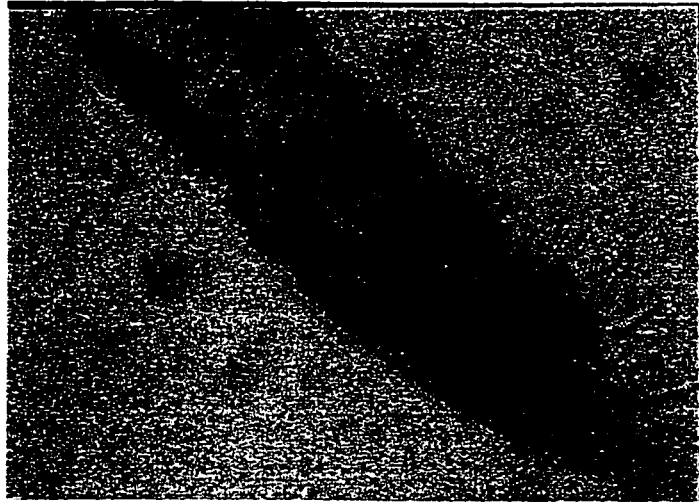
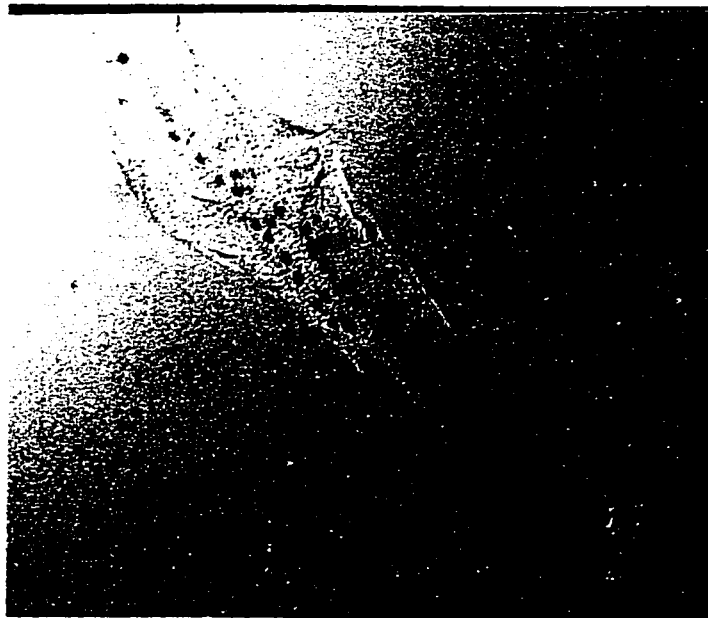
A**B**

Figure 28 The reporter gene fusion study of C33B4.3. Transgenic worms with C33B4.3::*lacZ* extrachromosomal arrays were examined under bright-field microscopy after X-gal reaction staining. *A*. C33B4.3 shows expression in male tail. Rows of cells that show staining seem to extend anteriorly (*B*).

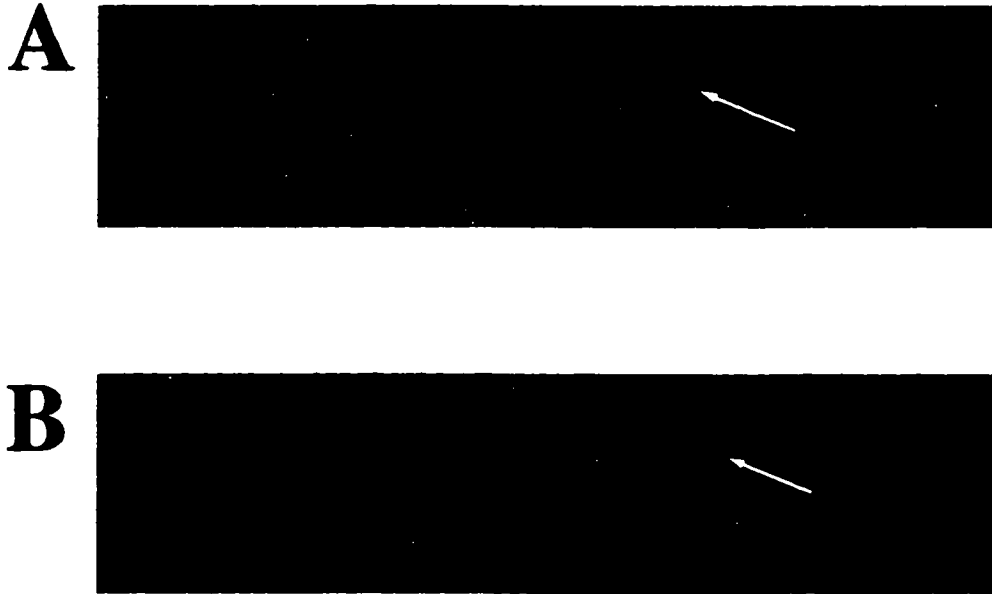


Figure 29 The reporter gene fusion study of C33B4.3. Transgenic worms with C33B4.3::*lacZ* extrachromosomal arrays were examined under fluorescence microscopy after immunostaining with anti-β-galactosidase antibody and DAPI. *A*. Ventral nerve cord cells show expression using anti-β-galactosidase antibody stain (arrow). *B*. The position of those cells can be determined using DAPI stain (arrow). These figures are provided by Dr. Dave Pilgrim and published here with permission.

RAB-Like gene (RABL)

Identification of the RABL locus

A putative gene was identified by an EST contig located at position 106843-122467. The Blast search gave more than 100 entries. Appendix lists the first 50 entries. An EST clone pul (a pregnant uterus long cDNA clone) (Appendix) spanned the whole contig, therefore it was chosen for further analysis.

Duplication of the RABL locus on chromosome 2q and 22q

The RABL locus is located on cosmid N1G3, at ~34.5 kb proximal to the putative 22q telomere (Fig. 9). When N1G3 was used as a probe for FISH, hybridization signals were found on many different chromosome ends as well as a region that is close to the centromere of 2q (data not shown). This result was further confirmed by Ning et al. (1996). By using cosmid C202 (Fig. 3) as a probe for FISH analysis, Ning et al. (1996) showed that the genomic region where RABL is located was duplicated on chromosome 2q13, the fusion site of two ancestral chromosomes (Ijdo et al. 1991). When the pul cDNA clone was hybridized with a monochromosomal hybrid panel, hybridization signals were shown on chromosome 2 and 22 (data not shown). In order to further characterize the genomic region on chromosome 2 and 22 where RABL is located, pul was hybridized to a partial hybrid panel which contained human chromosome 2, 22, total human and hamster DNA digested with HindIII, BamHI, and EcoRI. The banding pattern between the chromosome 2 and chromosome 22 copies were the same over these three restriction enzymes (Fig. 30). The sizes of the restriction fragments on the partial hybrid panel matched those in restriction maps constructed by the GCG program MAPSORT (Fig. 31) Therefore, the RABL gene was localized on both chromosome 2q13 and 22q13.3 resulting from a recent genomic duplication. These two RABL genes were named RABL2 (the copy on chromosome 2) and RABL22 (the copy on chromosome 22).

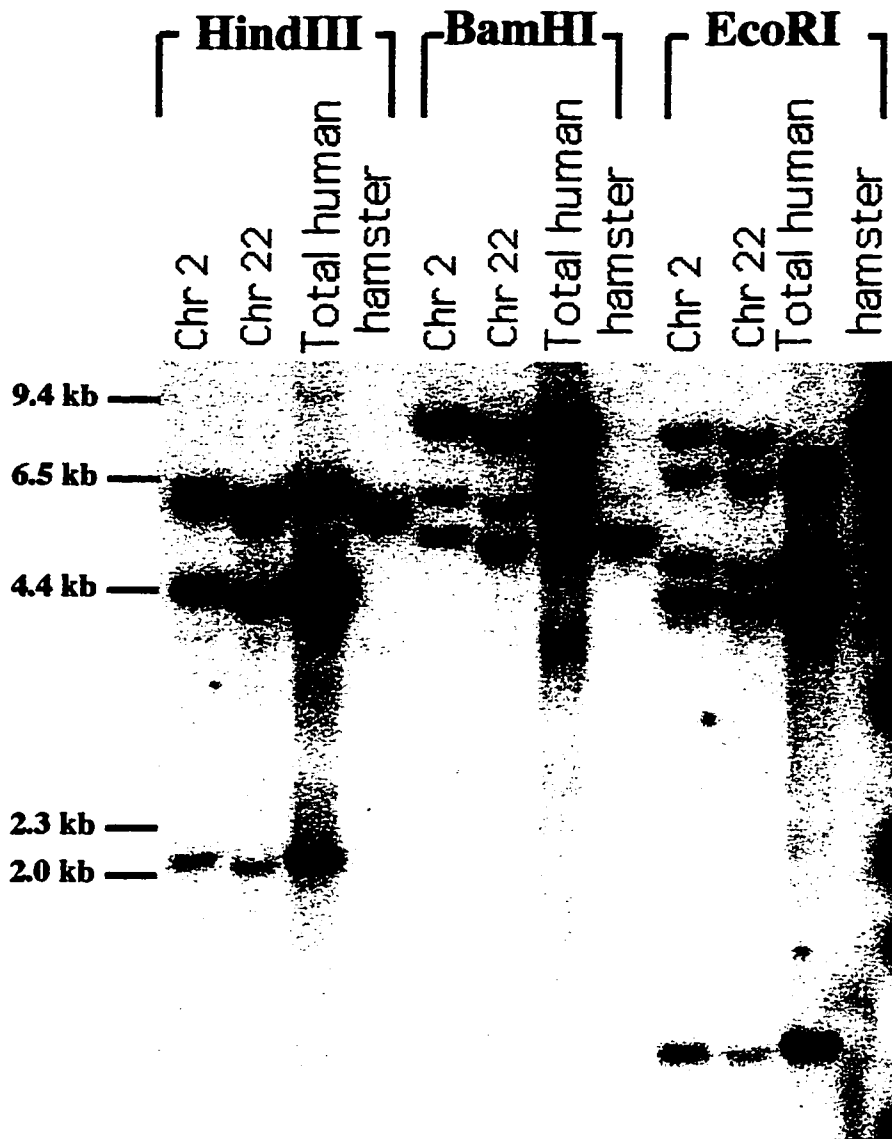


Figure 30 Partial hybrid panel probed with RABL cDNA probe pul. Genomic DNA from chromosome 2, and 22 hybrid cell lines, normal human, and hamster were digested with HindIII, BamHI, and EcoRI, electrophoresed, and transferred to a Southern blot. Slight differences in band size between lanes digested with the same enzyme are due to an electrophoresis artifact

Sequence analysis of RABL

The whole 1056 bp cDNA clone pul was sequenced and aligned against the AWcontig. It showed that the cDNA contains 9 coding exons (Fig. 31). There were some mismatches between pul and AWcontig. To determine whether they were sequencing errors or pul whether was actually RABL2, genomic fragments from chromosome 2 were amplified by PCR using the chromosome 2 hybrid cell line DNA as the template (PCR product RABLF1-R1, Table 3). The sequence of the chromosome 2 genomic DNA matched that of pul, indicating that pul was indeed the RABL2. When RT-PCR was done with the chromosome 22 hybrid cell line total RNA as a template, a partial RABL22 was amplified using RABLF1 and E0.91F as primers. This result indicates that both RABL2 and RABL22 are expressed. When the cDNA sequence of RABL2 was compared with the genomic DNA of RABL22, there are 5 nucleotides differences within the coding region that account for 3 amino acids sequence changes (Fig. 32). The first change is at amino acid sequence position 68, where a lysine (on RABL2) is changed to an arginine (on RABL22). The second change is at position 103, where leucine is substituted with valine. The last change is at position 223, where the valine on RABL2 is substituted with alanine in RABL22. All of them are conservative amino acid changes. Both RABL2 and RABL22 have an ORF that contains 228 amino acids, which shares similarity with other RAB proteins. RAB is a subfamily of the RAS superfamily, which currently has 40 family members in mammalian cells (reviewed by Lazar et al. 1997; Novick and Zerial, 1997; and Olkkonen and Stenmark, 1997). RAB proteins are small GTPases that control vesicular trafficking, either protein transport between different organelles within a cell, or transport in and out of the cell through endo- and exocytosis respectively. Common RAB conserved sequence motifs are found in the two RABL genes, which include the phosphate/Mg²⁺ binding motifs (G1 and G3), the effector region (G2), and one of two guanine base-binding motifs (G4) (Fig. 32). The third guanine base-binding motif, EXSA, is absent in the RABL genes. Another difference between RABLs and the other RAB proteins is the absence of cysteine residues at the carboxyl end in RABL. The

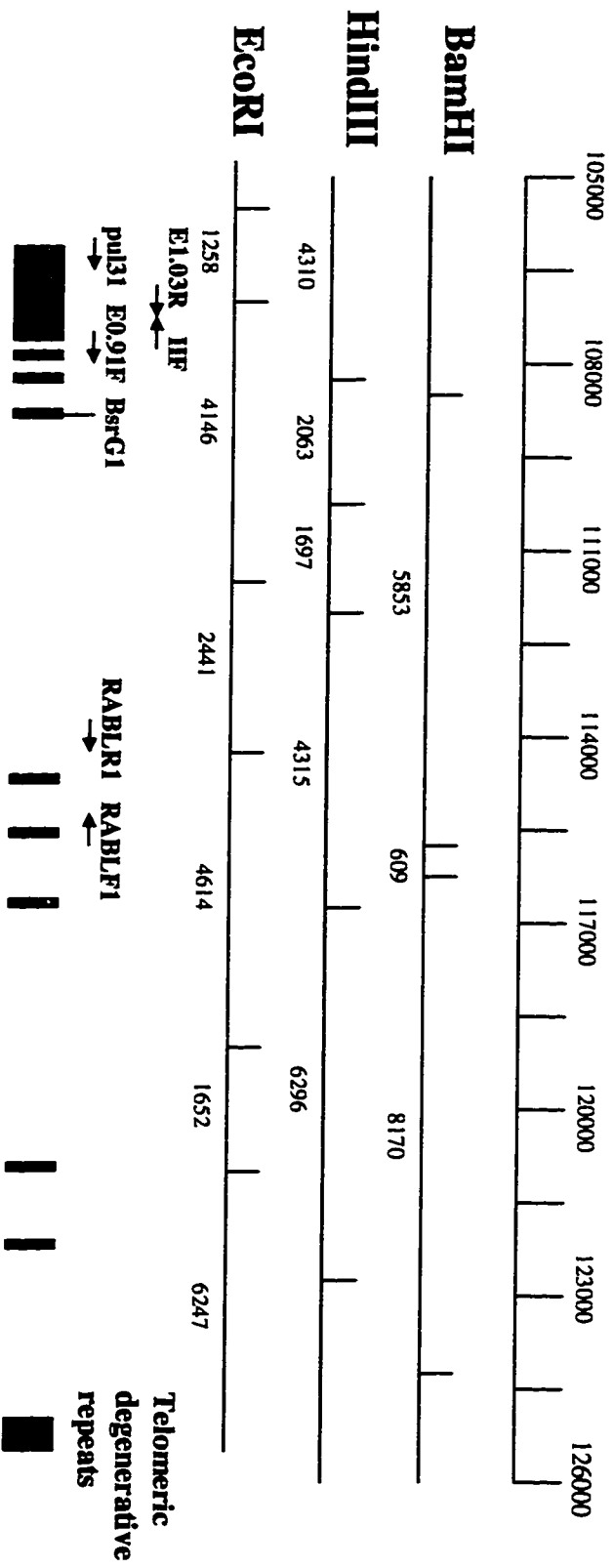


Figure 31 Genomic organization of RABL22. The scale indicates the coordinates in AWcontig. Restriction maps of three enzymes are shown, with the vertical lines indicates the restriction sites and the expected sizes of the fragments (in bp) below each map. RABL22 exons are in black bars, with the internal intron at the 3'UTR represented by a green rectangle. The telomeric degenerative repeats are shown in red rectangle

Figure 32 The cDNA sequence of RABL2 with its putative open reading frame. Some nucleotides in RABL22 are different than RABL2 in the coding region, which are shown as small letters highlighted by a yellow background above the sequence. The corresponding changes in amino acid are shown below the amino acid residues. One change of nucleotide in RABL22 creates a BsrG1 site (TGTACA, in red). The amino acid sequences in boxes are the functional domains of RAB proteins. They include two phosphate/Mg²⁺ binding motifs (G1 and G3), the effector region (G2), and one guanine base-binding motif (G4). The underlined sequence represents the internal intron (with the splicing donor and acceptor sites labeled in blue), which is spliced out in the 1.4 kb alternative transcript but is present in the 2.5 kb transcript. RABL does not have a polyadenylation signal. The fact that many cDNAs end at the same position (for example, pul, h5, h7, h8 and other ESTs listed on the appendix) indicates that the end of the sequence is authentic.

1 GCACGAGGTTTCGAGAGCGCGCAGAGTCCAGACTGGCGGCAGGGCCCGAGGGCCGACCCG
 61 CAGCGTCCCTGGTCTCTCCAGCCCTCACTCGGAACCGCACTGACAATACCCTCCCCTCCC
 121 TTGGGCTGGACCCCTCTCTACAGCTAGGAGCCAATGGCAGAAGACAAAACCAAACCGAGT
 M A E D K T K P S
 181 GAGTTGGACCAAGGGAAGTATGATGCTGATGACAACGTGAAGATCATCTGCCTGGGAGAC
 E L D Q G K Y D A D D N V K I I C L G D
 241 AGCGCAGTGGGCAAAT CCAAACCTCATGGAGAGATTTCTCATGGATGGCTTTCAGCCACAG
 S A V G K S K L M E R F L M D G F Q P Q
 g 1 a g
 301 CAGCTGTCCACGTACGCCCTGACCCCTGTACAAGCACACAGCCACGGTAGATGGCAAGACC
 Q L S T Y A L T L Y K H T A T V D G K T
 R
 361 ATCCTTGTGGACTTT TGGGACACGGCAGGCCAGGAGCGGTTTCGAGAGCATGCATGCCTCC
 I L V D F W D T A G Q E R F Q S M H A S
 g 2 c g 3 g
 421 TACTACCACAAGGCCCATGCCTGCATCATGGTGTGGTATATACAGAGGAAAGTCACCTAT
 Y Y H K A H A C I M V F D I Q R K V T Y
 v
 481 AGGAACCTGAGCACCTGGTATACAGAGCTTCGGGAGTTCAGGCCAGAGATCCCATGCATC
 R N L S T W Y T E L R E F R P E I P C I
 541 GTGGTGGCCAAATAAAATTTGATGACATAAACGTGACCCAAAAAAGCTTCAATTTTGCCAA
 V V A N K I D D I N V T Q K S F N F A K
 g 4
 601 AAGTTCTCCCTGCCCTGTATTTTCGTCTCGGCTGCTGATGGTACCAATGTTGTGAAGCTC
 K F S L P L Y F V S A A D G T N V V K L
 661 TTCAATGATGCAATTCGATTAGCTGTGTCTTACAACAGAACTCCAGGACTTTCATGGAT
 F N D A I R L A V S Y K Q N S Q D F M D
 721 GAGATTTTTTCAGGAGCTCGAGAACTTCAGCTTGGAGCAGGAAGAGGAGGACGTGCCAGAC
 E I F Q E L E N F S L E Q E E E D V P D
 c
 781 CAGGAACAGAGCAGCAGCATCGAGACCCCATCAGAGGAGGTGGCCTCTCCCCACAGCTGA
 Q E Q S S S I E T P S E E V A S P H S
 A
 841 GGGGCTGGGGCTAGGGGTGGGTGGAGCCCTTTTAAAAATACCCTTCCCTTCAACAACCTCTC
 901 CAGCTCTGAATGGAGAACTCTCTAGGCCATCCCCTCTTCTACCTCCTGCAACCCACCCA
 t
 961 TCCTATTAGCCTCCACATTC AAGGCCCGTGATACAGGGATGAGGT CAGCACCAGCAAAC
 g
 1021 TCTGGACTGGTGGAGAATTC CCCACCAGATCTCCTTGAAGCAGAATTAGGGATCAGCAT
 g
 1081 CATTAACACCTTCCCCACCCCTCCCCCAGGCAGACAGTGAAGAGAATCAGAAAACATG
 g t t t t t
 1141 ATTATGTGTCACTTTAATACAGGAAATTTAGGTGTTTTTTGGT - - - - - GTTTTTGT
 1201 TGTTTTTGT TTTCTTTCCAAAGCTCACCTCGGGGACAATTCCTTGGGCTTCTCCTGAGGT
 t
 1261 AATGATT - ACCCCCCACCCACAGCTGAGTCTGTGAGGCCCCATCCTTTCCCTACGTTTT
 1321 CTCCCATCTTTTTCTCTTCAGTCTCCAGTCATCTGGTTTTGTTTTCTTTGTTTCGT
 t
 1381 CCTGAGACGGAGTCTCGCTCTGTCGCCAGGCTGGAGTGCAGTGGCGCAGTCTCGGCTCGC
 1441 TGCAACCTCTGACTCCCTGGTTCAAACGATTCTCCTGCCTCAGCCTCCCGAGTGGCTGGC
 1501 ATCACCACGCCAGCTAATTTTTGTAATTTTTAGTAGAGACGGGGTTTTACCATGTTGCC
 1561 AGGATGGTCTCGATCTCCTCACCTCGTGATCCGCCCCCTCGGCCTCCCAAAGTGTGGG
 1621 ATTACAGGCATGAGCCACCGCGCCCGGCCCAATCATCTGTTTTTAAACAATCGTTTTTTG
 t
 1681 AGCAGATAGCTATTTCATTCCAGATTTTTCGTGTACCCACTCTGTTTTCAGGAGCTCTTCTAG
 1741 GTAAGCTGAGATCACAGGAACAGCAGGTGACAGGCCTAGCTATAGTTAGGAATACACAA
 c
 1801 GCGGTAATAATCGAGTCTTACAGCCATACCACAAGGTACGTCCATTTGGACTACAAGAAG
 g
 1861 AGCTTCTTTAAAGTTCCCTATTTTCAGCATAAAGAGGCTGTCTTTTTTTTTTAGGAATAG
 1921 TTTGGACCTTGTGCCTCTGTGGGAGGCTGAGACTGCAAGAGGAGAGCTAGCAGATATG
 1981 CCTGTTCAACCCCTCTCTGGTACTTGTGGCTTGCTAGTATGTTTTTATGATAATCTCGGGC
 2041 ATTGTTTTGCATTGTGTTTTATTAATAGGGTTTTGTTTTTATTGTTTTCTTTTTTACAGTAA
 t
 2101 AGGCTGAATGACAT

cysteine containing motifs are subjected to isoprenylation. The hydrophobic isoprenyl groups added on cysteines are important for the binding of RABs onto the membranes of vesicles. Since these cysteines are absent in the RABLs, they may not bind to the membrane and therefore may not function in the same way as other RAB members.

Comparison of the relative expression level between RABL2 and RABL22

Since the cDNA sequences between RABL2 and RABL22 are nearly identical, Northern hybridization with one cDNA sequence could not be used to distinguish the relative transcription level between the two RABL genes. Therefore, an RT-PCR approach was used to determine the relative abundance of these two transcripts. The basic principle of the RT-PCR experiment was to use a polymorphic restriction endonuclease site (BsrG1, TGTACA) at position 451 of RABL22 (Fig. 32) to differentiate the transcripts of RABL2 and RABL22. This RT-PCR method was first tested by a control experiment. RABL2 and RABL22 cDNAs were mixed in different ratios for PCR template as described in the Materials and Methods. Partial cDNA of the two RABLs were amplified by PCR using primers RABLF1 and E0.91F (Table 4, Fig. 31), and digested with BsrG1. The band intensity between RABL2 (492 bp, the non-cleaved product) and RABL22 (368 bp, the BsrG1 cleaved product) were compared by eye. The results showed that the ratio of the band intensity of the PCR products matched that of the starting ratio of the templates (data not shown). Therefore, reverse transcription was done in various tissues using the gene specific primer pul31 (Table 4, Fig 31). A partial cDNA was amplified using the RT products from different tissues as template. After BsrG1 digestion, the RT-PCR products were run on a 1% TAE gel. The results from several experiments showed that both the 492 bp RABL2 product and the 368 bp RABL22 product were present in all tissues tested (Fig. 33). Although the ratios of band intensity between the PCR reaction done in triplicate from the same tissue were relatively even, the ratios varied somewhat when the PCR was repeated using the same RT product for the same tissue. Among all the tissues tested, the expression of the two RABL genes in adult heart, skeletal muscle, small bowel, tonsil, thymus, fetal brain and fetal lung were consistently even in two experiments. Placenta showed more RABL22 expression in two

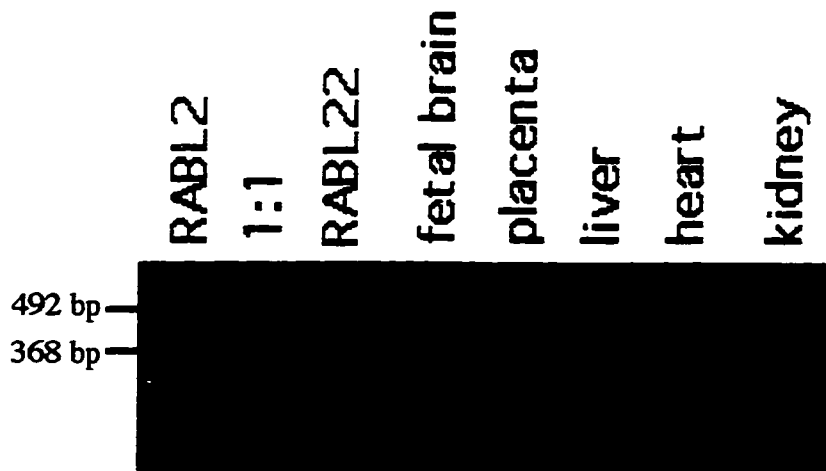


Figure 33 Comparison of the relative expression level between RABL2 and RABL22. RT-PCR was done in different tissues to amplify the RABL cDNAs. The products were then digested with BsrG1 and electrophoresed on 1% agarose gel. Three controls were included in the same RT-PCR experiment to monitor the possible PCR contamination by the residual products which were amplified in the previous experiments. They were the RABL2 only cDNA, RABL2 and RABL22 cDNA mixed in 1:1 ratio, and the RABL22 only cDNA as templates. Two bands are shown in each tissue: the upper band (492 bp) is the undigested RABL2 cDNA, whereas the lower band (368 bp) corresponds to the BsrG1-digested RABL22 cDNA. PCR reactions have been done in triplicate using the same reverse transcription product of each tissue as template. No subtle variation of RABL2 and RABL22 concentration was found within the triplicates. Therefore, only one RT-PCR reaction from each tissue was shown on this gel. This figure is provided by Dana Shkolny and published here with permission.

experiments, but it showed more RABL2 expression in a third experiment. Adult liver showed low RABL2 expression consistently in three experiments (data not shown). Figure 33 shows that adult kidney has a very low level of RABL22 expression, but in another experiment the RABL22 expression was higher than that of RABL2 (data not shown). This indicates that the RT-PCR method could not be used to show subtle differences of transcription level between RABL2 and RABL22. However, this experiment confirmed that both RABL2 and RABL22 are expressed in all tissues tested.

Northern analysis of the RABL gene

When Northern blots with multiple human tissues were hybridized with pul, a 2.5 kb signal was found in all adult and fetal tissues. This 2.5 kb transcript has a higher expression level in adult heart, brain, kidney, and pancreas than in placenta, lung, liver, and skeletal muscle. In fetal tissues, it has a stronger expression in brain than in lung. A ~1.4 kb alternative transcript was shown in adult heart and skeletal muscle. This transcript shows a stronger expression than the 2.5 kb transcript in both tissues. Faint 4.4 kb and high molecular weight bands appeared in fetal tissues (Fig. 34 A).

Cloning the 1.4 kb alternative transcript of RABL by 3' RACE

In order to clone the alternative 1.4 kb transcript found in heart and skeletal muscle, 3' RACE was done as described in Materials and Methods. Using adult heart total RNA as a template, 3' RACE amplified one ~1.7 kb and one ~0.7 kb product that corresponded to the size of the common 2.5 kb and the alternative 1.4 kb transcripts. The products were subcloned into pGEM-T Easy vector. Colonies that contained the clones of the 3' RACE products were hybridized to pul. Positive hybridization signals were found in 25 clones from the 1.7 kb product and 2 clones from 0.7 kb product. One 1.7 kb product clone (hc5) and two 0.7 kb product clones (hc7 and hc8) were completely sequenced. The results showed that the 1.7 kb clone has an extra ~1 kb of sequence at the 3' untranslated region (Fig. 32), which is not found in the 0.7 kb product. Intronic donor and acceptor splice sites are present at the ends of this 1 kb sequence. Therefore, the

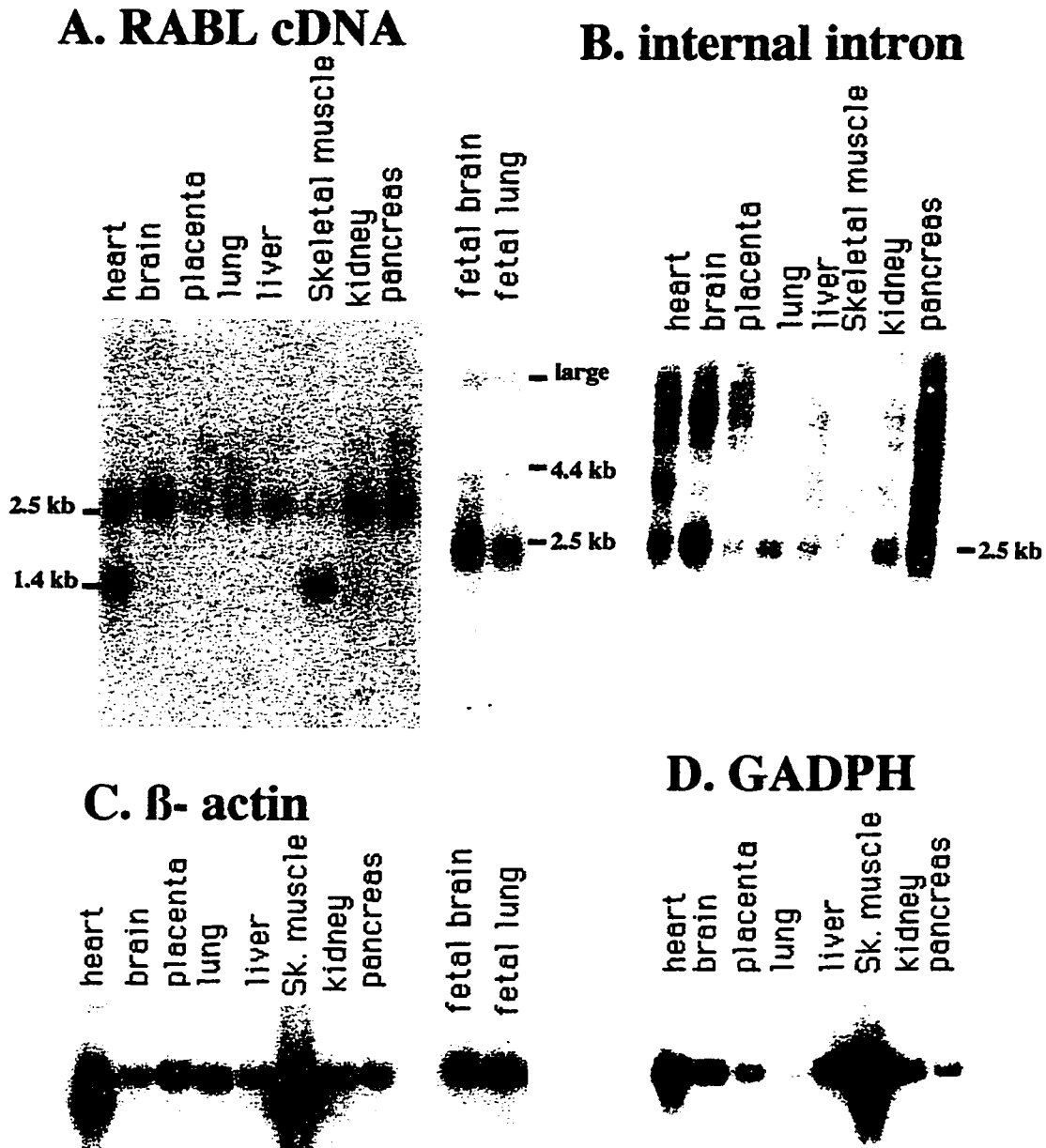


Figure34 Northern analysis of RABL. *A.* The human multiple tissues Northern blots were probed with RABL cDNA probe pul. One 2.5 kb signal is present in every adult tissue, whereas another 1.4 kb signal is only present in heart and skeletal muscle. The fetal tissues show a 2.5 kb, 4.4 kb and a large band. *B.* The human multiple adult tissues Northern blot probed with internal intron at the 3' UTR. Only the 2.5 kb signal is found on the autoradiography. *C.* The same Northern blots in *A* probed with β -actin (control probe, done by Dr Valerie Trichet). *D.* The same Northern blot in *B* probed with GAPDH gene (control probe, done by Dana Shkolny).

alternative 1.4 kb transcript is the result of the splicing out of an internal intron in the 3'UTR. This result suggests that the size difference between 1.4 and 2.5 kb transcripts is probably the result of the presence or absence of 1 kb internal intron in 3'UTR. This was confirmed by the hybridization of the internal intron (PCR product "internal intron" amplified by IIF and E1.03R primers, Table 3, Fig. 31) to a Northern blot. The internal intron probe produced a 2.5 kb band but no 1.4 kb band in any adult tissues (Fig. 34 B). The band intensity of this 2.5 kb transcript is similar to that on the Northern blot probed with pul: adult heart, brain, kidney, and pancreas have a higher expression level than that in placenta, lung, liver and skeletal muscle. Including the internal intron sequence, a 2114 bp of RABL cDNA has been cloned, leaving ~400 bp of 5'UTR in the common 2.5 kb transcript uncloned.

Discussion

Cloning of the NT Microdeletion region

Patient NT has a microdeletion at the terminal region of chromosome 22q. Based on studies done to date, this microdeletion overlaps within the deletion region of the 22q13.3 deletion syndrome. D22S163 is deleted in all 22q13.3 deletion patients tested, and the most distal cosmid clone C202 (Fig. 3) is deleted in those tested to date (Heather McDermid, unpublished data). Therefore, the NT microdeletion presented an opportunity to delineate the large (≥ 5 Mb) critical region of the 22q13.3 deletion syndrome. Since NT only represents one microdeletion case, it is possible that the microdeletion is not related to the child's phenotype of the 22q13.3 deletion syndrome. However, since the two clinical features shown in NT (i.e. mental retardation and expressive speech delay) are also major features of the deletion syndrome patients, it is likely that NT represents a subset of the 22q13.3 deletion syndrome phenotype. As a result, I studied the size and location of the microdeletion. I also mapped the genes within the microdeletion which may be involved in the neural development implicated by the phenotype of NT.

The NT microdeletion is spanned by two overlapping cosmid/P1 contigs with a size >150 kb, which include the two known loci D22S163 and acrosin (ACR) (Fig. 3). By Southern analysis, the microdeletion breakpoint mapped close to or within the D22S163 locus. Based on the restriction mapping data, there is ~ 130 kb between D22S163 and the 22q telomere (Fig. 3). The sequencing of the cosmid contig revealed that this distance is actually closer to 140 kb (Fig. 9). However, small gaps in the sequence of N66C4 still exist. Therefore, the NT microdeletion encompasses the last 140 kb or more of chromosome 22q.

There are still some uncertainties about the position of the 22q telomere relative to the cosmid/P1 contig. The P1 clone at the end of the contig (Fig. 3) is believed to contain genomic DNA that is close to the 22q telomere. The clone was obtained by screening a P1 genomic library with a TelBam3.4 probe (Ning et al. 1996), which is usually located immediately adjacent to the telomere (Brown et al. 1990). However, a recent report showed the presence of a TelBam3.4 sequence 60 kb away from the 20p

telomere adjacent to an internal stretch of telomere repeats. This suggests that the presence of the TelBam3.4 telomere associated sequence alone cannot be used to define the position of the telomere (Chute et al. 1997). As a result, there may be additional sequence past TelBam3.4 at the end of the P1. However, such additional material is unlikely to contain genes, because it is likely composed of subtelomeric associated repeats (see below). In order to determine whether the TelBam3.4 sequence is adjacent to the 22q telomere, one can do a BAL31 exonuclease analysis. The chromosome 22 hybrid cell line DNA can be digested with BAL31, electrophoresed, and Southern blotted. The resulting blot will be probed with TelBam3.4. If the TelBam3.4 fragment is sensitive to BAL31 digestion, the location of the telomere will then be confirmed.

Two pieces of evidence support that the NT deletion is terminal rather than interstitial. First, the EcoRV-digested rearrangement fragment is sensitive to the BAL31 exonuclease digestion, indicating the rearrangement fragment is close to the chromosome end instead of mapping at an interstitial position on the chromosome. Also, the characteristically smeared rearrangement bands seen with HindIII, SmaI, and Sau3AI suggest that the broken chromosome was healed by the addition of telomeric repeat (TTAGGG)_n with heterogeneous length. The rearrangement bands were detected by the D22S163 probe, which suggests that the D22S163 locus is close to the deletion breakpoint. The breakpoint has now been cloned by my collaborator Dr Jonathan Flint from Oxford University. He used a D22S163 locus specific primer and primer with three telomeric repeats to amplify the breakpoint junction sequence from NT genomic DNA (Wong et al. 1997). A comparison between the breakpoint sequence (Fig. 35A) and the corresponding normal chromosome 22q sequence (Fig. 35B) revealed that the breakpoint is located within the D22S163 locus and can be identified by the presence of telomeric repeats (Fig. 35A) substituted for the normal sequence (Wong et al. 1997). D22S163 consists of two distinct minisatellite arrays, MS607A and MS607B (Armour and Jeffreys 1991). The flanking sequence surrounding the MS607A minisatellite array has been determined for 500 bp proximally (accession number X58043) and 866 bp distally (accession number X58044). The location of the breakpoint is 26 bp distal to the 3' end of the 866 bp flanking sequence. Therefore, the breakpoint is 892 bp (866+26 bp) distal

#2:3' AA AAUC5'

#1:3' AUC AA 5'

A 5' AGGGGGTGGAGAGGGGGG GGT AGA ttagggtagggtta 3'

B 5' AGGGGGTGGAGAGGGGGGTGGAGGGGGTGGTGGCACAGGGG 3'

Figure 35 Sequence at the NT microdeletion breakpoint, compared to that of a normal chromosome. *A.*, Breakpoint sequence of NT. The 5' end of the sequence starts from the first base after the 866bp 3' flanking sequence of the MS607A minisatellite at the D22S163 locus. The breakpoint is within D22S163, identified by the presence of telomeric repeats (ttaggg, in small letters) substituted for the normal sequence. *B.*, Sequence of a normal individual. The normal and breakpoint sequences were cloned by Dr Jonathan Flint and published here with permission (Wong et al. 1997). The possible anchor sites (#1 and #2) for telomerase RNA template are shown above the sequences *A* and *B*. Nucleotides highlighted in blue show complementarity between the telomerase RNA template and the breakpoint sequence. The possible site for the breakpoint are marked in boxes, which would agree with the anchor position #2 of the telomerase RNA template. The substitution of the adenine in the breakpoint sequence for the guanine in the normal sequence (highlighted in yellow) is probably due to polymorphism.

to the MS607A. The distance between minisatellite MS607A and MS607B is not known, but an upper size is estimated to be ~1 kb (J.A.L. Armour, personal communication). If the NT deletion breakpoint is 891 bp distal to MS607A, it is either at the 5' end or within the MS607B minisatellite.

The addition of telomeric repeats onto the breakpoint suggests that the deleted chromosome was healed by telomerase activity. This could be done if the RNA template of the telomerase recognized the sequence that is adjacent to the breakpoint. The sequence of the human telomerase RNA template has been cloned, and was shown to encompass 11 nucleotides (5' CUAACCCUAAC) which are complementary to the human telomere repeat (TTAGGG)_n. In Fig. 35, two predicted anchoring positions of the telomere RNA template onto the breakpoint are shown. The first prediction (#1) is based on the fact that the first base added to the breakpoint is an adenine, therefore the first unpaired nucleotide on the RNA template would be an uracil. The vertical blue bars show the complementary nucleotides between the RNA templates and the sequence before the breakpoint. However, no more than 2 nucleotides that directly precede the breakpoint are complementary to the RNA template in position #1. Flint et al. (1994) found 3 to 4 nucleotides complementary to the RNA template at the breakpoints of 5 out of 6 deletion patients who showed terminal deletion on 16p13.3. The remaining case (referred as patient IC) had two additional nucleotides preceding the breakpoint which did not match to either normal sequence or telomeric repeat sequence. Therefore, the IC deletion is likely to be a result of illegitimate recombination (Flint et al. 1994). However, the NT deletion is not analogous to IC because the sequence on the deleted chromosome matches to that of the normal copy up to the point where the substitution of the telomere sequence occurred. Therefore, I hypothesize that the breakpoint is on the first thymine residue which substituted the guanine in the normal sequence (boxed nucleotides at Fig. 35). The adenine that substituted the guanine residue (both highlighted by yellow background) in the normal sequence could be the result of polymorphism. In order to test this hypothesis, one would have to sequence the breakpoint region of NT's father to see whether this adenine exists in either paternal allele of the two normal sequences because NT carries a paternal deletion.

The cosmid/P1 contig spanning the NT microdeletion serves to anchor the physical map of chromosome 22. PFGE analysis done by Dr Heather McDermid indicates that this contig lies ≤ 120 kb from ARSA (Wong et al. 1997), previously the most distally mapped locus on chromosome 22 (Dumanski et al. 1991; Collin et al. 1995). The order of known loci at the end of 22q is therefore ARSA-D22S163-ACR-telomere. This confirms the subtelomeric location of the VNTR at D22S163 used for detecting cryptic 22q terminal rearrangements (Flint et al. 1995).

Genomic Organization of the NT Microdeletion Region

The sequence spanned by the AW contig can be divided into two sections: a unique sequence that is specific to 22q, and the subtelomeric region that contains repeats shared by other chromosomes. I have mapped three real genes within the AWcontig, together they occupy ~ 84 kb, or 60% of the contig. Two genes (ALPR and ACR) are on the unique region, whereas RABL is at the proximal end of the subtelomeric region which usually shares homology with a few chromosome ends (Flint et al. 1997b and see below). Furthermore, a pseudogene U2 SnRNP specific protein A' appears to lie near the boundary of the unique sequences and the subtelomeric region (see below).

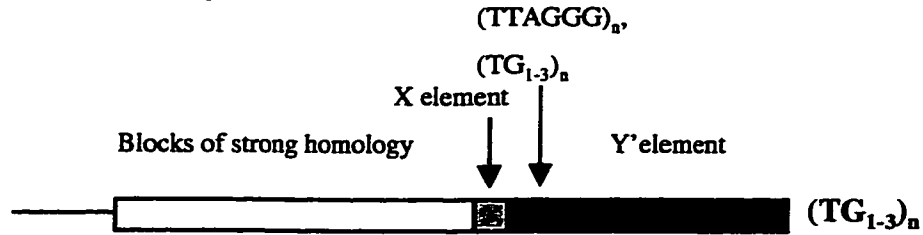
The region that is close to the telomere is commonly described as the subtelomeric region. The genomic organization of the subtelomeric region of several human chromosomes has been studied. For example, the subtelomeric region of chromosome 16p has been studied extensively by Flint et al. (1997a). Since I concentrated on the transcription mapping within the AWcontig, I provided the 22q subtelomeric sequence to Dr Jonathan Flint, who then compared it to that of 16p and 4p (Flint et al. 1997b). He found that the organization of the subtelomeric region is similar among the three chromosome ends. Flint et al. (1997b) show that the subtelomeric region can be divided into two distinct domains called proximal and distal subtelomeric domains. The distal subtelomeric domain is adjacent to the telomeric repeat (TTAGGG)_n. The sequence of this distal domain is homologous to many different chromosome ends, but the matches are interrupted. On the other hand, the sequence of the proximal subtelomeric domain of one chromosome only shares homology with a few

chromosomes, and the homology is strong and continuous (Flint et al. 1997b). The two subtelomeric domains are divided by an internal degenerative telomeric repeat sequence (Fig. 36 B).

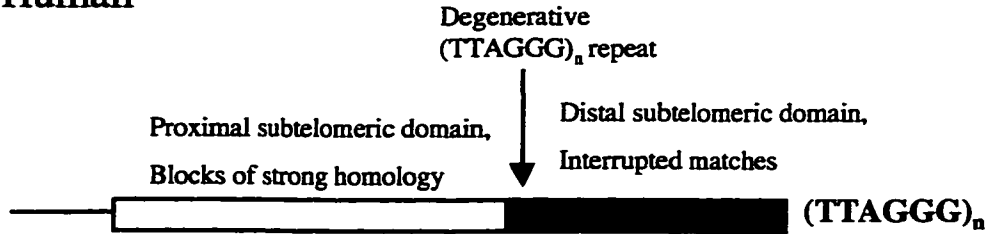
The overall organization of the subtelomeric region shares similarity to that of the budding yeast *Saccharomyces cerevisiae* (Fig. 36 A, Flint et al. 1997b). The distal subtelomeric domain of yeast contains variable numbers of repeats of Y' elements, which have a size of either 5.2 kb or 6.7 kb and 2 overlapping ORFs. This region shares interrupted homologies with different chromosome ends. The boundary element that divides the proximal and distal domains consists of X elements (a 475 bp sequence that contains an autonomously replicating sequence [ARS]), internal yeast telomeric repeats $(TG_{1-3})_n$, and human canonical telomeric repeat sequence $(TTAGGG)_n$ (Fig. 36A). Like its human counterpart, the proximal subtelomeric domain of yeast is a block of DNA that shows strong homology and continuous matches to a few chromosome ends. Such similarity of genomic organization in the subtelomeric region between yeast and human suggests a specific function of the boundary element between the proximal and distal subtelomeric domain. During interphase in yeast, the telomeres appear to be clustered near the nuclear periphery (Palladino et al. 1993). This anchoring of the telomere clusters to a nuclear structure facilitates the exchange of sequences between non-homologous chromosomes, resulting in the interrupted homologies in the Y' element region. The telomeric repeats (both human-like repeat and yeast internal repeat) and X elements act as a barrier to limit the spread of the non-homologous recombination to the proximal subtelomeric domain (Pryde and Louis 1997). Human telomeres are not clustered during interphase. However, the telomeres are clustered to form the chromosomal bouquet during prophase I in meiosis (Scherthan et al. 1996, Fig. 1), which also facilitates the exchange of the non-homologous chromosomes at the distal subtelomeric domain.

The results of my independent sequence analysis support the findings by Flint et al. (1997b). The terminal 19.7 kb of AWcontig contains the distal domain, which shares interrupted matches to the subtelomeric regions of 1q, 4p, 7p, 13q, 14q, 15q, 17q, 18p, 21q, and Xq (Appendix). This homology does not spread to the proximal domain of the subtelomeric region, which is proximal to the degenerative telomeric repeats located at

A. *Saccharomyces cerevisiae*



B. Human



C. Human chromosome 22q

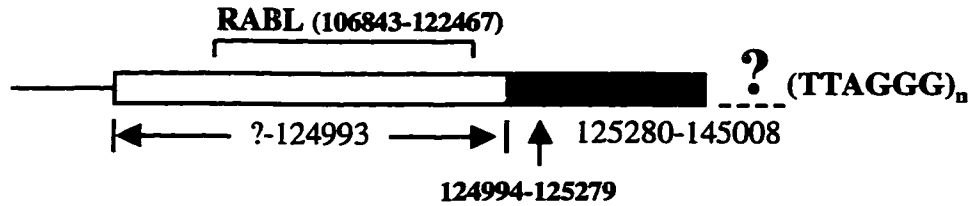


Figure 36 The genomic organization of the subtelomeric region in *A. yeast (Saccharomyces cerevisiae)*, *B. human* (Flint et al. 1997b) and *C. human chromosome 22q* (Flint et al. 1997b, this study). In *S. cerevisiae*, a boundary sequence composed of X element, yeast internal telomeric repeats $(TG_{1-3})_n$, and human telomeric repeats $(TTAGGG)_n$ separate two domains. The distal domain (the dark gray box) mainly consists of the Y' repeats that share split matches to many other telomeres. The proximal domain (the open box) has a block of continuous match and strong homology to a few telomeres. The single line preceding the open box represents the unique sequence of the chromosome. The human subtelomeric region shows similar organization, and has a degenerative telomeric repeat that divides the proximal and distal domain. The proximal domain has strong homology to a few chromosome ends, and the distal domain has interrupted matches to many different telomeres. The last 19.7 kb of AWcontig (position 125280-145008) is the distal subtelomeric domain, and the degenerative telomeric repeat is at position 124994-125279. RABL is within the proximal subtelomeric region, which shares high sequence identity to a region at 2q13, an interstitial region that contains an ancestral telomere fusion site. The distance between the end of AWcontig and 22q telomere is not known (the region with dot line and a question mark) but estimated to be at least 12 kb.

the position 124944-125279 of the AWcontig (Fig. 36 C). However, there are matches to 19q distal subtelomeric domain sequences within an intron of ALPR at the position 49491-50226 of AWcontig, which is located on cosmid N85A3. This result is unexpected because position 49491-50226 is in the unique region on 22q (Ning et al. 1996), and therefore it should not contain subtelomeric repeat sequences. Both FISH (Ning et al. 1996; Wong unpublished observations) and my hybrid panel analysis showed that the region centromeric to the degenerative telomeric repeat is homologous to 2q13, the site where the fusion of two ancestral ape chromosomes occurred (Ijdo et al, 1991). Therefore, the proximal subtelomeric domain of 22q is homologous to an ancestral proximal subtelomeric domain at 2q13. The RABL gene (described below) maps within this proximal subtelomeric domain on both chromosomes (Fig. 36C). The partial hybrid panel results show that a total of 15/15 restriction sites tested are conserved between the genomic region of RABL2 and RABL22 (4 BamHI sites, 5 HindIII sites, and 6 EcoRI sites, see Fig. 30 and Fig. 31). Therefore, the sequence divergence within the 15 kb RABL locus is no greater than 0.56% ($1/[15 \text{ restriction sites}][6 \text{ bp recognition site}][2 \text{ sequences}]$)(Van Arsdell and Wiener 1984). This result also agrees with the findings of Flint et al. (1997b), with the proximal subtelomeric domain of 22q showing strong homology to only one other chromosome. The boundary between the proximal subtelomeric domain and the unique sequence on chromosome 22q is difficult to access. Dosage analysis using the end fragment of N94H12 as a probe, which is a 1.5 kb fragment that contains the distal ~700 bp of the LINE/SINE repeat cluster between ACR and N1G3 (Fig. 9), did not show a 2:1 ratio between normal and NT genomic DNA (data not shown). This suggests that the end fragment of N94H12 is still within the proximal subtelomeric domain, because there are four copies of the N94H12 end fragment in the genome: two copies from chromosome 2q and two from chromosomes 22q. The deletion of the end fragment in NT would therefore give a 4:3 ratio between normal and NT genomic DNA instead of the 2:1 ratio, although the dosage analysis is unable to confirm this ratio. The U2 snRNP specific protein A' pseudogene (Fig. 9) is also considered to be in the proximal subtelomeric region. Bands in both the chromosome 2 and 22 hybrid cell lines were detected when the U2 snRNP specific protein A' gene probe was hybridized to the monochromosomal hybrid panel (data not shown), although localization to 2q13 was

not done. Therefore, the subtelomeric/unique sequence junction is most likely narrowed down to the region between ACR, which is unique to chromosome 22, and the U2 snRNP specific polypeptide A' pseudogene (Fig. 9).

In this study, the human interspersed repeats were divided into two groups: the first group consists of Short Interspersed Nuclear Elements (SINEs) and Long Interspersed Nuclear Elements (LINEs), and the second group consisted of the retrotransposons and retrovirus-like elements which are characterized by the flanking of long terminal repeats (LTR) on both sides. The LINE/SINEs are concentrated in the non-coding region, the biggest cluster occupies 16 kb between ACR and RABL (Fig. 9). The LTR type repeats seem to be concentrated in the distal subtelomeric domain (Fig. 9), which is the last 19.7 kb of AWcontig. It contains approximately 10 LTR type repeat elements which occupy 35% of the region. There is no LTR in the proximal subtelomeric domain or the unique sequence within AWcontig. The localization of the LTR type elements in the subtelomeric domain may have an evolutionary significance: in yeast, a LTR-type retrotransposon Ty5-1 is located at the subtelomeric region of the left arm of chromosome III (Voytas and Boeke 1992). The Ty5-1 element lies distal to the X element, close to the boundary with the distal subtelomeric domain. Vega-Palas et al. (1997) showed that the Ty5-1 element is subjected to telomere position effect. The Ty5-1 expression was analyzed in a wildtype strain; a null mutant of the SIR3 gene, which causes the loss of transcription silencing near the telomeres (Aparicio et al. 1991); and a strain that contains the SIR3 overexpression plasmid, which enhances the transcription silencing effect (Renauld et al. 1993). The Ty5-1 expression in the wildtype strain was low and was almost undetectable in the SIR3 overexpression strain, but it showed a high level of expression in the sir3 null mutant (Vega-Palas et al. 1997). These results suggest that the transcription silencing by telomere position effect regulates the expression of Ty5-1. High levels of Ty5-1 retrotransposition introduce deleterious mutations in the host strain. Therefore, the insertion of Ty5-1 into a silenced domain limits its own expression, so that it avoids the co-extinction of the host and the retrotransposon (Vega-Palas et al. 1997). The integration of mammalian LTR-retrotransposons (MaLRs) into the human subtelomeric domains may reflect a similar transcription control in the mammalian system. Such elements could collect at the chromosome ends without posing any

deleterious effect to the host. Although the active MaLRs appear to be extinct in higher primates (Smit 1993), the MaLRs in mouse may still continue to amplify (Smit 1996). The inactive MaLRs in human subtelomeric domains may be molecular fossils that record that such transcription regulation occurred in the ancestral mammalian genomes.

Transcription Mapping within the NT Microdeletion

There are two novel genes and one known gene found in the NT microdeletion region. The two new genes were named after the functional domains found in the putative proteins. ALPR represents “ankyrin-like, proline rich”, whereas RABL is “RAB-like”. Acrosin (ACR) was previously mapped to the same region as ARSA on chromosome 22q by somatic hybrid panel mapping (Budarf et al. 1996). My results confirm this finding and determine that ACR is distal to ARSA on chromosome 22q.

1. Localization of the acrosin (ACR) gene

I mapped the acrosin gene (ACR) within the microdeletion region. The locus occupies a ~7 kb genomic region between the two new genes ALPR and RABL (Fig. 9). Sequence analysis showed that the intron/exon boundaries of the five exons are identical to those in the Genbank database (accession number X66188, X54017, S40014, M77380, X54019, X54018, M77379, X54020, and M77378). Densitometric analysis showed that NT and a 22q13.3 deletion syndrome patient (FB) have a deletion ACR (Fig. 8).

ACR is a serine protease present in the acrosome of the sperm head (Klenn et al. 1991). Protease inhibition studies suggest that ACR plays a key role in the acrosome-mediated binding of sperm to the zona pellucida and the penetration of sperm into the oocyte (Liu and Baker 1993; Takano et al. 1993). It is highly unlikely that the deletion of ACR is related to any present visible phenotype of NT. The reduction of ACR activity in sperm has been associated with male infertility (Koukoulis et al. 1989; Mohsenian et al. 1982), suggesting that the deletion of ACR may affect the fertility of NT and male 22q13.3 deletion patients. However, it is still under debate whether the reduction of acrosin activity is a cause of male infertility. While Kennedy et al. (1989) showed that the successful rate for *in vitro* fertilization using sperm with normal acrosin activity level is

significantly higher than that with reduced activity, Yang et al. (1994) showed that such difference is insignificant. Furthermore, homozygous mouse knock-outs for the acrosin gene are still fertile, indicating that acrosin is not essential for fertilization in the mouse (Baba et al. 1994).

2. a) Structure of ALPR

ALPR was cloned by a combination of sequence analysis and molecular studies of the putative expressed sequences using RT-PCR and Northern blot analysis. Putative expressed sequences were determined by exon prediction programs, and by EST partial cDNA sequences. The exon prediction program identified a putative gene. Its ORF shows homology with a predicted *C. elegans* protein C33B4.3. This suggests that the human putative expressed sequence is likely to represent a novel gene. Both C33B4.3 and the putative human ORF have an ankyrin repeat domain and polyproline sequences. I therefore named it “ankyrin like, proline rich”. I constructed a cDNA contig of ALPR, which includes three putative exons (starting exon, genscanex1, and Last Exon), two RT-PCR products (sc24, and fli), one cDNA from fetal brain cDNA library (I511), and two EST contigs (FLS and FL2)(Fig. 10). Together they span a ~60 kb region that includes the deletion breakpoint of NT (Fig. 9). Since the deletion truncates ALPR, it is likely that one copy of ALPR is not transcribed in NT. If ALPR is haploinsufficient, this could produce the abnormal features of NT.

Northern blot analysis showed that there are multiple transcripts of ALPR. A common ~7.5 kb transcript is found in adult heart, brain and placenta when probed with FL2 and 55337F-R1 (Fig. 14 A and C). These two probes also detected an alternative transcript in adult brain with a size of ~6.8 kb (Fig. 14 A and C). In fetal tissue, the same two probes detected three transcripts (a large transcript >10 kb, one close to 8 kb, and one ~7.5 kb) in fetal brain and kidney. Fetal lung has a strong signal for the 8 kb band, but weak for the large band. Due to the difficulty of measuring band sizes between blots, especially with large fragments, I hypothesize that the 7.5 kb transcript in the adult tissues is the same as the 8 kb transcript in the fetal tissues. Likewise, 6.8 kb specific adult brain transcript is probably the same as the 7.5 kb transcript in fetal tissues. This

hypothesis is based on two observations. 1) Even though the adult 7.5 kb band is only strongly visible in the heart, brain, and placenta, faint expression can be found in other tissues. This suggests that this 7.5 kb transcript is commonly expressed in all adult tissues, even though the signal is too faint to see (Fig. 14 C). The 8 kb fetal transcript also seems to express in all four fetal tissues tested. It is visible in fetal brain and kidney. This signal is particular strong in fetal lung, and is not visible in fetal liver (Fig. 14 C). 2) Subsequent Northern blot analyses showed that the adult 7.5 kb transcript appears in all adult tissue tested (the same 8 adult tissues in Fig. 14, different blot), with a particular strong signal in brain when the Northern blot was probed with sc24 (Dana Shkolny, unpublished results). Probe sc24 also detected the 8 kb fetal transcript in the four fetal tissues tested. The expression level of sc24 in fetal lung and kidney are higher than that in fetal brain and liver. This result is consistent with the expression pattern of FL2 (Fig. 14 C).

The large (> 10 kb) transcript is not specific to the fetal tissue. Probe sc24 detected the same large band on all adult tissues. (Dana Shkolny, unpublished data). It is most likely missing from the Northern blot in Fig. 14 due to poor transfer of large fragments.

FLS detected the same large band in the fetal blot and one ~2.5 kb band in fetal liver (Fig. 14 B). It is interesting to note that this 2.5 kb fetal liver specific transcript was also present when the Northern blot was probed with I511, but not present on the Northern blot probed with sc24 (Dana Shkolny, unpublished data). The summation of the sizes of FLS (0.5 kb), genscanex1 (2.2 kb), and I511 (not including the 3' UTR, 1.0 kb) gives a size that is larger than the 2.5 kb (~4 kb). This suggests that the fetal liver specific transcript cannot contain any sc24 exon, which can account for lack of the band with sc24. It also implies that genscanex1 cannot be included in the 2.5 kb transcript, since it has a size of 2.2 kb. In order to determine what exons are present in this 2.5 kb band, one could do a 5' RACE from fetal liver RNA. Nevertheless, the absence of sc24 exons in the 2.5 kb transcript suggests that it does not contain the ankyrin repeat domain. Since the ankyrin repeat domain implies that the protein is a membrane bound (see below), this transcript may produce an alternative cytosolic protein.

The summation of various sections of the gene (sc24 +I511 + genscanex1 + FL2) gives a transcript size of ~6.7 kb. The size of the common transcript on a Northern blot appears to be ~7.5- 8 kb. This predicts that ~1.0 kb of the common transcript remains to be identified. The uncloned part of the transcript presumably corresponds to the 5'UTR of the gene. All the other transcripts would be a result of alternative splicing of the mRNA. The large transcript (>10 kb) may contain more 5' UTR sequence. This large transcript was detected by FL2, the most distal clone of ALPR that contains a 3'UTR (Fig. 10). Therefore another possibility is that the uncloned region of the long transcript may extend from the 3' end of ALPR, ignoring the polyadenylation signal in the FL2 clone and extending beyond.

It is of interest to note that the adult and fetal brain (and faintly in fetal kidney) contains the ~6.8-7.5 kb alternative transcript that is not found in other tissues. This suggests that ALPR may have a specific function in the brain. The phenotype of NT (i.e. mental retardation and expressive speech delay) reveals that the most affected tissue in this patient is the brain. Therefore, the deletion of ALPR may be responsible for producing the abnormal phenotype.

ALPR exhibits a complex organization. Some of the different transcripts shown in the Northern analysis are produced from the alternative splicing of at least three different 3' ends. These 3' ends are associated with I511, the EST clone FLS and FL2. It is not uncommon that a gene has different 3' ends, which produce different protein isoforms. Each isoform has a different carboxyl end, which gives them a specific function. For example, the immunoglobulin IgA has two different isoforms, one is membrane bound and one is secreted from the B cells. Whether the IgA molecule is membrane bound or secreted is determined by the presence of two different carboxy ends, which is a result of alternative splicing at the 3' end of the mRNA (Seipelt and Peterson 1995). The presence of the three 3' ends in ALPR may give a different function for each transcript.

I found two rodent homologs (MB and mb101) of one 3' end (FL2). These three clones share strong nucleotide homology (Fig. 17). This suggests that the sequence of this 3' UTR may be important for the stabilization or the function of the transcript. Two A-T rich sequences are found in the 3'UTR of FL2. One A-T rich track is located at position 732-757 of FL2, whereas another A-T rich sequence is adjacent to the poly-A tail (Fig.

17). A-T rich sequences are known to destabilize mRNA, thus reducing the half-life of the transcripts (Shaw and Kamen 1986). The half-life of the FL2 transcript is not known. However, if the A-T rich sequences are recognition signals for targeted RNA degradation, the short-lived nature of the transcript could explain why mouse embryo whole mount *in situ* hybridization did not reveal any signal when the mouse cDNA mbl101 was used as a probe (data not shown).

The Last Exon (Fig. 10) is named after its homology to the last exon of the *C. elegans* gene C33B4.3 (43% protein identity). This exon is also homologous to the 3' end of another ALPR homolog, the rat cortactin binding protein 1 (66% protein identity) (Fig. 21). This suggests that Last Exon is likely to be a part of ALPR. Last Exon was identified by the exon prediction program Grail 2 as a terminal exon without a splicing donor site at the 3' end. It is located within the intron of FL2 (Fig. 10). RT-PCR results demonstrated that the transcript that contains Last Exon would also contain other ALPR exons, because PCR from a reverse transcription product initiated by the Last Exon specific primer (Ank R5, Table 4) could amplify a group of I511 exons using forward primer AnkF1 and reverse primer AnkR1 (F1-R1, Fig. 10). However, I could not amplify a clone that includes the I511 exons and Last Exon by AnkF1 forward primer and a Last Exon specific reverse primer (Ank51). This may be due to its high G-C content (Fig. 19), or the presence of *genescanex1*, which appears to prevent reverse transcription (see below). Since the Last Exon is predicted as a terminal exon, it is probably represents another 3' end. However, the Grail program usually has a higher accuracy rate in predicting the exons which have coding sequence and splicing donor and acceptor sites than those having a 3' UTR. For example, the Grail exon prediction program cannot predict the terminal exons of I511, FLS, and FL2. Therefore, it is also possible that Last Exon can splice with FL2. RT-PCR using primers specific to Last Exon (Ank F3, Fig. 10) and FL2 (OFH2 as reverse transcription primer, OFH3 as reverse primer for PCR, see Table 4) in fetal brain and spleen did not amplify any product (data not shown). I also tried to clone the 3'UTR of the transcript that carries the Last Exon by 3' RACE using a Last Exon specific primer (F3, Fig. 10) in multiple tissues (heart, fetal brain, fetal lung, spleen, and thymus). However, no authentic PCR product was cloned (data not shown). Since the tissues I chose for the RT-PCR or 3' RACE have ALPR expression by Northern analysis,

this suggests that if Last Exon does extend 3', then it could either be part of a specific transcript in a tissue not tested, or a transcript which has a specific temporal expression. In order to find the tissue specific transcript that contains Last Exon, one could hybridize it to a multiple tissue Northern blot with more tissues, and then choose the tissues that showed hybridization signals for the 3' RACE.

A complete ORF can be predicted by sequence analysis programs, and I have confirmed most of this by cDNA cloning and RT-PCR. However, there are still two major problems in the existing cDNA contig which need to be resolved: there is a premature termination of translation in the ORF of sc24, and the genscanex1 exon seems to be unclonable. The premature termination of translation occurs at the nucleotide position 1255 of sc24 (Fig. 11). There are several possible explanations for this premature translation termination. First, it could be due to sequencing error. The sequencing of the cosmid clone N66C4 is still unfinished. The error rate in the "sequencing in progress" stage is presumably higher than that in the finished sequence. Even when sequencing is finished, sequencing errors still exist. For example, one sequencing error has been found at the ACR locus. At the beginning of exon 4, an extra cytosine insertion at position 82932 of AWcontig causes a frameshift at exon 4. Except for this error, the genomic sequence matches identically to the cDNA sequence of ACR. As shown by densitometric analysis (Fig. 8), there should be only one ACR locus in the human genome, making it unlikely that this locus is a pseudogene. Thus, the insertion of the cytosine in exon 4 of ACR is mostly likely a sequencing error rather than the presence of an ACR pseudogene on 22q. The stop codon in the ALPR H55337 exon of sc24 (position 1155 to the end of the clone) is the result of a frameshift introduced by the previous exon (position 881 to 1154). I511 contains the in-frame translation of the H55337 exon (position 137 to 250 of I511 clone, Fig. 12). If the starting nucleotide of the H55337 exon (i.e. the thymine) is located at the second position of a codon (i.e. *gtg ctc ...*), the translation will be in-frame. In sc24 the starting nucleotide thymine is at the third position of a codon (i.e. *agt gct ...*). If the frameshift of the H55337 exon in sc24 is due to a sequencing error, the error could be either a deletion of two nucleotides or an insertion of an extra nucleotide. I sequenced three cDNA clones from the same RT-PCR product (sc24, sc31, and sc32. See Table 3). Although there are nucleotide

substitutions which presumably were introduced by PCR, there are no deletions or insertions found in the cDNA sequence when it is compared to the genomic sequence. Strong denaturing agents were added in the polyacrylamide gel to eliminate possible compressions in the sequencing (see the Materials and Methods), which usually appear in G-C rich regions and lead to apparent deletions of nucleotides in the read sequence. As a control, I have sequenced a chimeric cDNA that contains a ribosomal RNA gene at the 3' end, which has a stretch of 35 G-C bp. The denaturation gel resolved all bases (data not shown). Clone sc24 does not have any region that has more than 35 bp of G-C only sequence. Therefore, it is unlikely that the frameshift is due to sequencing error.

The second possible explanation for the premature translation termination relates to the post transcriptional modification of the mRNA. It is known that RNA editing changes the sequence of some mRNAs after transcription. For example, there are insertions (Feagin et al. 1988) and deletions (Cruz-Reyes and Sollner-Webb 1996; Seiwert et al. 1996) of uracil residues in the transcripts of mitochondrial genes found in kinetoplastid protozoa. Although the RNA editing that causes nucleotide changes in the transcripts has been found in humans (Hodges et al. 1991; Sharma et al. 1994; Rueter et al. 1995; Skuse et al. 1996), there is no report to show that RNA editing introduces insertions or deletions into human mRNA. Also, there are no insertions/deletions in the three cDNA clones sequenced, which makes it unlikely that the frameshift of the H55337 exon could be corrected by RNA editing. If there is any post-transcriptional modification, it should be shown in some of the transcripts.

The most likely explanation is that the premature translation termination in the sc24 transcript is real. There are many different ALPR transcripts which are produced by alternative splicing. Therefore, the stop codon introduced in the H55337 exon could be a result of alternative splicing which creates a truncated form of ALPR. This hypothesis is supported by two observations. First, the exon that is 5' to the preceding exon of H55337 exon (the P1 exon at Fig. 37 A) has the splicing donor site end at position 881 of sc24 (Fig. 11). The last nucleotide of this exon is at the first position of a codon. If this exon skips one exon and splices to the H55337, the first nucleotide of H55337 will be on the second position of a codon, which makes the translation in frame (Fig. 37 B). Therefore, alternative splicing could produce an in-frame translation. Second, the ORF of the exon

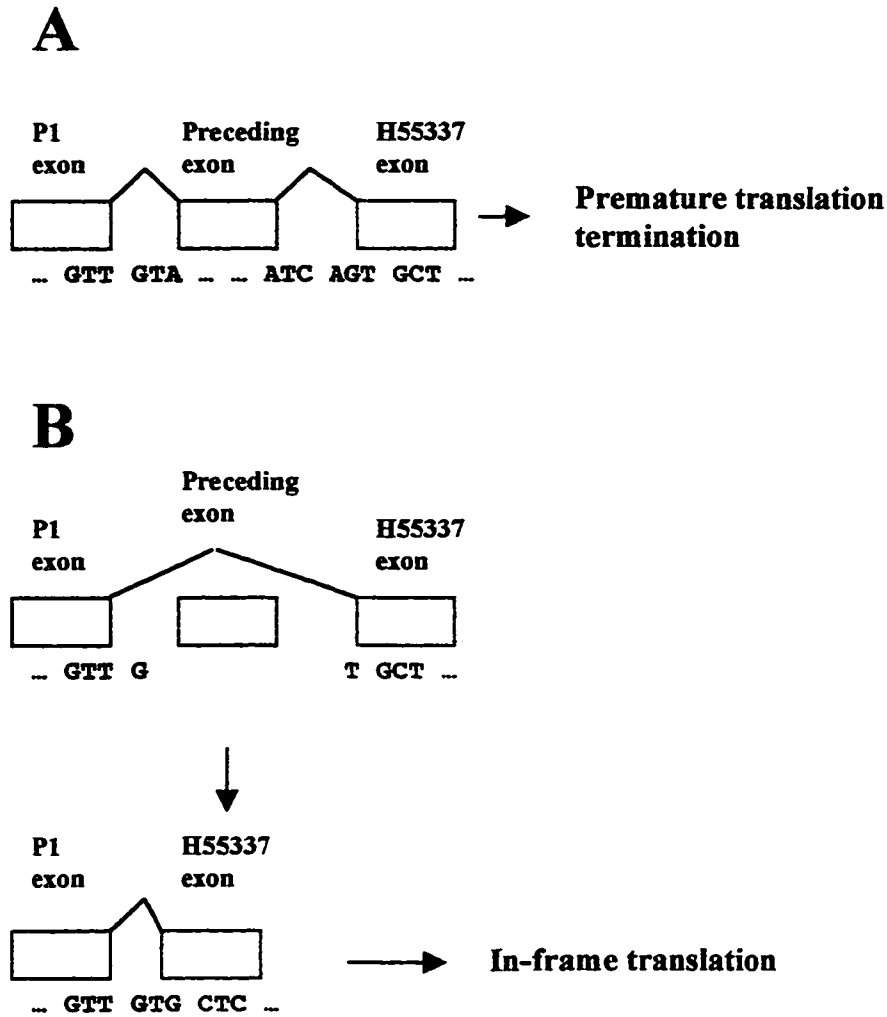


Figure 37 Alternative splicing to produce (A) a premature translation termination and (B) an in-frame translation through the H55337 exon. (A) When the transcript includes the preceding exon of H55337 exon, the first nucleotide of the H55337 exon (i.e. the thymine) will be located on the third position of a codon, which causes a frameshift and introduces a stop codon in the H55337 exon. (B) If the exon 5' to the preceding exon (i.e. the P1 exon) splices to the H55337 exon, the first nucleotide of the H55337 exon will be at the second position of the codon. The H55337 will have an in-frame translation.

that is 5' to H55337 exon (the preceding exon) does not share any homology to any protein that is homologous to ALPR (position 402-435 of the hypothetical ALPR protein, Fig. 20). This suggests that the full length ALPR may not contain this exon, and the transcript that contains this exon will create a truncated form of the ALPR. A truncated form of a protein that is created by a premature translation termination has been found for other proteins. For example, the striatum-enriched phosphatase (STEP) is a tyrosine phosphatase that is found in the central nervous system. It contains four different isoforms, two cytosolic and two membrane bound forms. Both cytosolic and membrane bound forms contain a full length and a truncated protein (Charbonneau et al. 1989). The truncated form of the membrane bound protein is produced by a transcript called STEP₃₈. STEP₃₈ contains a stop codon upstream of the nucleotide sequence which encodes the tyrosine phosphatase domain (Charbonneau et al. 1989). It is unlikely that the STEP₃₈ transcript is an artifact because it can produce a protein with the corresponding size, the STEP₃₈ protein is found in the membrane fraction of the rat brain tissue by subcellular localization study, and the protein does not have any phosphatase activity (Bult et al. 1997). Bult et al. (1997) suggested that the truncated form STEP₃₈ either presents the substrates molecules to the full length protein, or it protects the substrates from dephosphorylation by competitive binding of the truncated form with the full length phosphatase to the substrates. The truncated form of ALPR is similar to that in STEP₃₈. Both STEP isoforms contain a membrane binding domain, but a portion of the intracellular domain is deleted in STEP₃₈. In ALPR, the membrane binding domain is the ankyrin repeats domain (see below). The ankyrin repeats are found in sc24 (Fig. 21). Therefore both isoforms could be membrane binding but the truncated ALPR would be missing the polyproline region. To determine whether the stop codon on the H55337 is a result of alternative splicing, I would need to look for the alternatively spliced transcripts that put the H55337 in frame. This could be done by RT-PCR in different tissues to test whether the transcript that produces the full length protein is found in some tissues but not others. In Northern blot analysis, the sc24 probe detected the 7.5 kb band and a large band like the 55337F-R1 and FL2 ALPR probes (Dana Shkolny, unpublished data). In addition, it showed a small ~1.4 kb band in adult heart, brain, and skeletal muscle. This small transcript could be the truncated transcript. Alternatively, the partial cDNAs of the

human ALPR homologs (cDNA 208081, DS17. See Fig. 21) could be completely sequenced, and then the ORFs of those cDNAs could be compared with the ORF of sc24 to look for the homologous sequence of the preceding exon of H55337. Both cDNA 208081 and DS17 share strong homology with ALPR (Fig. 21). If both cDNAs contain an uninterrupted ORF and do not contain the sequence that is homologous to the preceding exon of H55337, this would indicate that the stop codon in H55337 is caused by the presence of the preceding exon.

The genscanex1 exon is very likely to be an exon of ALPR based on the following observations. First, it shares a polyproline sequence with C33B4.3 protein and a proline-rich ligand for the SH3 binding domain of rat cortactin binding protein 1 (see below). Second, the partial cDNA fli (Fig. 10, Fig. 15) contains the 3' end of the genscanex1 which splices to FLS. This confirms the 3' boundary of the genscanex1. However, I could not clone the exon in full length. This could be due to the high G-C content of the exon. The G-C content is similar to that of the dopamine D4 receptor gene (DRD4) (Fig. 19), whose transcript was not detectable by RT-PCR due to the high G-C content (Bondy et al. 1996). It is known that G-C rich sequence is a difficult region for PCR amplification. However, based on the sequence of the illegitimate PCR product R1F-M1314R (Fig. 18), it is more likely that the G-C rich content causes difficulty for reverse transcription. R1F-M1314R contains two full length I511 exons, and one I511 exon abnormally fused with a part of the genscanex1 (Fig. 18). G-C rich regions usually form secondary structure, which presumably causes a pause of reverse transcriptase synthesizing the first strand cDNA (Kotewicz et al. 1988) (Fig. 38 A). However, I hypothesize that the illegitimate PCR product R1F-M1314R could not be amplified if the reverse transcriptase paused before the secondary structure. It is also known that *Taq* polymerase is “jumpy”, which means the *Taq* polymerase may elongate using one template, and then “jump over” to another template to continue the elongation, thus splicing two transcripts together. This occurs when the template DNA is severely damaged (Pääbo et al. 1990). There is no evidence to suggest that the first cDNA strand for PCR was damaged. Therefore it is unlikely that the R1F-M1314R is an error due to PCR amplification. However, if the reverse transcriptase is also “jumpy”, so that it “jumped over” the secondary structure in the RNA template and continued the first strand

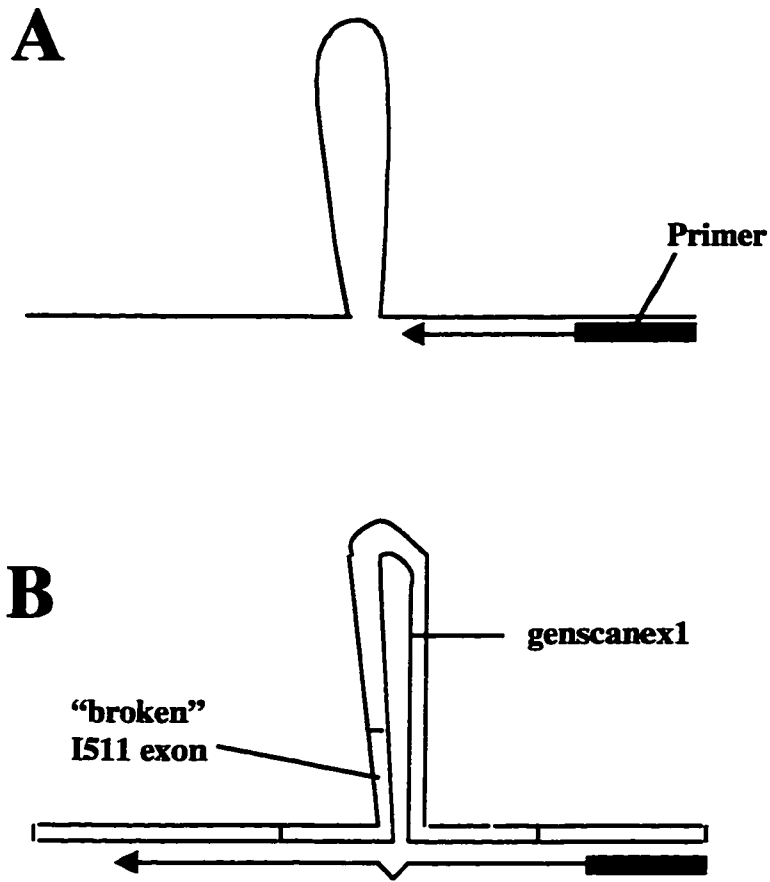


Figure 38 Two hypotheses for reverse transcription with an RNA template that contains a secondary structure represented by a loop in the template. The primer (green box) initiates the reverse transcription, and the line with an arrow represents the first strand cDNA and the direction in which it elongates. *A.* The reverse transcriptase pauses before the secondary structure. *B.* The reverse transcriptase jumps over the base of the secondary structure and the first strand cDNA continues to elongate.

synthesis (Fig. 38 B), the reverse transcriptase could synthesize a first strand cDNA that contained the first two I511 exons in full length as well as parts of the one I511 exon and genscanex1. The sequence within the secondary structure would be absent in the PCR product. This could explain why the I511 exon was abnormally fused with the partial genscanex1, if the fusion point is at the base of the secondary structure where the reverse transcriptase had to jump over to continue the strand elongation. The reverse transcriptase “jump over” would be a rare event; only one illegitimate product was amplified in more than ten RT-PCR experiments. The other RT-PCR reactions gave no amplification. However, this suggests that the secondary structure within genscanex1 exists and makes cloning difficult. Since reverse transcription is usually carried out at a low temperature (i.e. 42°C), it is difficult to melt the secondary structure within the RNA template during reverse transcription. I tried to set up a RT-PCR reaction using *Tth* polymerase, which works at a higher temperature (Myer and Gelfand 1991), but I still could not clone the genscanex1 in full length by that approach (data not shown).

2 b) Possible function(s) of the ALPR protein

ALPR is named after its two functional domains: the ankyrin repeat domain and the proline rich domain. Ankyrin repeats were originally found in the ankyrin molecule, which plays an important role in maintaining the membrane skeleton (Reviewed by Lambert and Bennett 1993). Ankyrin has many isoforms in erythrocytes and brain, but they all share a similar structure. The N terminus contains twenty-two tandem 33-residue repeats, which is a membrane binding motif that forms binding sites for various integral membrane proteins. Downstream to these repeats in ankyrin is the binding domain specific for spectrin, which is another membrane surface protein. The carboxyl end of ankyrin is the regulatory domain. The function of this domain is isoform-specific (reviewed by Lambert and Bennett 1993). The ankyrin repeat domain is found in various receptor molecules. This membrane binding domain anchors the receptor molecule on the membrane for the receptor to receive extracellular signals. For example, an ankyrin repeat domain is found in the LIN-12/Notch proteins, which are a family of receptor molecules with conserved structures and functions. Their family members include Notch,

which receives an inhibition signal from the neighbouring cells to specify the neurogenic cells to adopt either neural or epidermal fate (Atavanis-Tsakonas and Simpson, 1991). Another family member, LIN-12, plays a role in the vulval development in *C. elegans*. In the hermaphrodite, the anchor cell sends an induction signal through the *lin-3/let-23* pathway to P6.p, the middle of three vulval precursor cells (Hill and Sternberg 1992). This down-regulates the P6.p expression of LIN-12, which is a receptor molecule that receives an inhibitory signal. The inhibitory signal causes the cells to adopt the default fate (the secondary fate). Since less inhibitory signal is received by P6.p, it adopts the primary fate whereas the two flanking VPC cells, P5.p and P7.p, adopt the secondary fate (reviewed by Greenwald 1998).

Other than its membrane binding property, the ankyrin repeat domain is implicated in other protein-protein interactions. For example, the ankyrin repeat domain has been found in the 53BP2 protein, which binds to the p53 tumor suppressor protein (Gorina and Pavletich 1996). p53 is an important cell cycle checkpoint gene. When DNA damage occurs, the p53 protein induces the expression of the cyclin-dependent kinase inhibitor p21^{WAF1}, which in turn arrests the cell cycle (reviewed by Hansen and Oren 1997). However, unlike the null mutation of p53, p21^{WAF1} null mutant mice still retain some ability for checkpoint arrest (Deng et al. 1995). This suggests that there is another effector pathway of p53 that is independent of the p21^{WAF1} induction pathway. The binding of 53BP2 to p53 was found in a yeast two-hybrid system study (Iwabuchi et al. 1994). The 53BP2 binds to p53 by an ankyrin repeat domain and a SH3 domain which is usually bound to proline rich ligands (see below) The p53 amino acid residues at the surface that are in contact with 53BP2 are changed by point mutations in p53. Such changes abolish the binding between 53BP2 and p53, and the tumor suppressor ability is lost in those mutants (Gorina and Pavletich 1996). These results suggest that the tumor suppressor ability of p53 is conferred by the protein-protein interaction between p53 and 53BP2, which is mediated by an ankyrin repeat domain.

Another example of a protein-protein interaction through an ankyrin binding domain is the heterodimer of GA-binding proteins (GABP). GABP is a transcriptional regulator which is involved in the activation of genes in various gene families, such as the nuclear genes encoding mitochondrial proteins (Virbasius et al. 1993). GABP consists of

two subunits, namely, GABP α and GABP β . GABP α contains an ETS domain, which is commonly found in the members of the ets protein family. The ETS domain binds to DNA by recognizing the core motif 5'-GGAA/T-3' (Karim et al. 1990). GABP β contains an ankyrin repeat domain at the N-terminal. The ankyrin repeat domain forms loops of β -hairpins, and the tips of the loop insert into the α 1 helix of GABP α . This protein-protein interaction opens up the inhibitory configuration of the GABP α monomer, which in turn stabilizes the transcription activator activity of GABP (Batchelor et al. 1998). It is not clear whether the ankyrin repeat domain in ALPR acts as a membrane binding domain, or it is involved in other protein-protein interaction.

The second recognizable domain of ALPR is the proline rich domain. Various functional domains of proteins are rich in proline residues. For the following discussions, the term "proline rich" will indicate that the polypeptide sequence is mainly composed of proline residues, but it also contains other amino acid residues that interrupt a pure stretch of polyproline residues. The term "polyproline" will refer to a sequence of uninterrupted homoproline. The ALPR protein appears to have both domains.

Proline rich sequence is found in ligands which specifically bind to SH3 domains. Src homology 3 (SH3) domain was originally identified in a viral oncogene p47^{gag-ck}, which shares blocks of homology with a non-receptor class of tyrosine kinases at the non-catalytic region (Mayer et al. 1988). Since then the SH3 domain has been found in many proteins that are associated with the cytoskeleton (reviewed by Pawson 1995). The ALPR hypothetical protein shares extensive homology with rat cortactin binding protein 1 which contains a proline rich ligand for the SH3 domain (Fig. 21). As the name implies, this protein binds to the SH3 domain of cortactin, which is associated with the cytoskeleton. Cortactin was originally characterized in chicken embryo cells. It was named based on its localization in the cortical cytoskeletal network found in vertebrate cells (Wu and Parsons 1993) which is beneath the inner surface of the plasma membrane and is composed of a network of actin filaments and associated actin-binding proteins (Wu and Parson 1993). Cortactin is a substrate for the oncogenic tyrosine kinase pp60^{src}. It causes the redistribution of filamentous actin (F-actin) after phosphorylation (Wu et al. 1991), which in turn changes the cortical cytoskeleton structure that alters the cell morphology. Therefore, cortactin plays a role in the alteration of the cell morphology in

response to an extracellular signal. There is a human homolog of cortactin called EMS1. It also shows similar subcellular localization as cortactin in human transformed cells (Schuurin et al. 1993). The alteration in cell morphology after extracellular signaling is involved in different cellular processes, such as phagocytosis or cell differentiation. The amino terminal of cortactin contains a tandem repeat domain that binds to the actin filament, whereas the SH3 domain is found at the carboxyl end (Wu and Parson 1993). SH3 domains from different proteins bind to proline rich ligands with a specific consensus sequence. The consensus sequence of the ligand that is specific to cortactin SH3 domain was determined to be +PPΨPXKPXWL, in which “+”, “Ψ” and X stand for basic, aliphatic, and any amino acid respectively (Sparks et al. 1996). The ligand sequences in the ALPR hypothetical protein and rat cortactin binding protein 1 are identical, and are located at the position 1343-1353 of the ALPR hypothetical protein, or position 946-956 of the rat cortactin binding protein 1(Fig. 20).

The sequence of the rat cortactin binding protein 1 became available in GenBank database in June, 1998. By the time of writing the cloning of the rat cortactin binding protein 1 has not been published. As a result, the function of the protein is not known. The only information available is that it binds to the SH3 domain of cortactin. Cortactin binds constitutively to F-actin, regardless of whether it is in the phosphorylated or the non-phosphorylated state (Wu and Parson 1993). Also, the distribution of F-actin and cortactin overlap after pp60^{src} transformation (Wu et al. 1991). This indicates that the redistribution of the F-actin after transformation is not the result of the dissociation of the cortactin from F-actin. Therefore, the redistribution of the F-actin could be mediated by the protein-protein interaction between cortactin and its binding protein after phosphorylation. Thus, cortactin binding protein 1 could be involved in the F-actin remodeling. Its homolog ALPR will then have an implicated function in cytoskeletal regulation. In fact, the identical proline rich ligand sequence shared between the cortactin binding protein 1 and ALPR reveals that ALPR may be able to bind to cortactin. This makes it less likely that the ankyrin repeats of ALPR are involved in binding the extracellular side of the membrane.

Other than the proline rich ligand sequence, ALPR also shares strong homology with rat cortactin binding protein 1 in other regions, especially those close to the N-

terminal and the C-terminal (Fig. 21). However, ALPR does not seem to be the human orthologue of cortactin binding protein 1. First, the rat cortactin binding protein 1 does not have the ankyrin repeat domain (Fig. 21) which is present in both ALPR and C33B4.3. Second, the rat cortactin binding protein 1 is more similar to the retinal EST clone AR than ALPR (Fig. 20), indicating that AR is more likely to be the human homologue. However, AR does not have an uninterrupted ORF and a poly-A tail. Therefore AR is likely to be a pseudogene. When AR was used as a probe and hybridized to the monochromosomal hybrid panel, hybridization signals were seen on the chromosome 11 and 13 DNA (data not shown). These results suggest that there are two loci of AR in the human genome, one may be the real cortactin binding protein, and the other may be its pseudogene. Other than AR, ALPR also shows homology to two other human cDNAs. EST clone 208081 is homologous to ALPR at the ankyrin binding domain region. The 5' end of DS17 is homologous to the sc24 region of ALPR, and the 3' end is homologous to I511 and the 5' end of the rat cortactin binding protein 1 (Fig. 21). Therefore, these proteins may represent a new family, whose family member includes cortactin binding protein 1. A complete sequence for these clones might clarify this.

At the position 749 to 779 of the ALPR hypothetical protein, there are stretches of polyproline sequences (Fig. 20). The proline content is even higher in C33B4.3, which contains homoproline sequences with a maximum length of 10 residues (Fig. 22). The pyrrolidine rings in proline bend the polypeptide backbone of a protein, which is not favourable to form secondary structures such as α -helixes or β -sheets. However, polyproline forms a compact hydrophobic structure called type II poly-L-proline helix (PPII). This structure is a ligand for binding sites that expose hydrophobic amino acids on the protein surface (Mahoney et al. 1998). Polyproline has been implicated in a wide range of functions. For example, seven proline residues in the Epstein-Barr virus nuclear protein 2 are important for the transfection function of the protein (Yalamanchili et al. 1996). The increase in proline residues in the polyproline chain of lysozymes increases the bactericidal activity, presumably due to the increase of hydrophobicity and subsequent increases in the binding affinity to the outer membrane of the bacteria (Ito et al. 1997). The polyproline region in the zymogen form of acrosin (ACR) is cleaved off to

form an active mature enzyme (Baba et al. 1989). It is interesting to note that polyproline is also found in the ligand module of profilin, which plays an important role in actin filament assembly (Mahoney et al. 1997). Profilin is an actin monomer binding protein. It regulates the actin filament assembly by interacting with various cytoskeleton associated proteins that are involved in a various developmental and morphological processes (Mahoney et al. 1997). The protein-protein interaction is mediated by the binding of the profilin to the polyproline ligands with at least six residues of proline (Petrella et al. 1996). However, ALPR only contains two stretches of pentaproline (Pro₅, between position 749 to 779 of ALPR hypothetical protein; see Fig. 20). Therefore it is not clear whether ALPR is able to bind to profilin. The *C. elegans* homolog C33B4.3, however, contains long homopolymers of prolines that are the potential ligands of profilin. If ALPR is able to bind to profilin, it may be involved in cytoskeletal regulation by coupling the polymerization of the actin monomers to filaments, and redistribution of the filamentous actin under the regulation of cortactin.

Besides the ankyrin, polyproline and proline-rich sequences, there is another putative function domain at the C-terminal of ALPR, which is highly conserved in C33B4.3 and rat cortactin binding protein 1 (position 1594 to the end of ALPR hypothetical protein, Fig. 20). This region also shares moderate homology with the C-terminal of other kinases (see the results section “Homology between ALPR and other proteins in the public databases “). This conserved region was originally found in the C-terminal of the tyrosine kinases in the EPH receptor family, which is one of the four families of receptor tyrosine kinase (the other three families are EGF receptor, insulin receptor, and platelet-derived growth factor [PDGF] receptor; Lindberg and Hunter 1990). But it is also found in other non-receptor type tyrosine kinases, such as diacylglycerol kinase delta (Sakane et al. 1996). The C-terminal of the EPH receptor tyrosine kinase has been suggested to play a role in the kinase activity regulation by autophosphorylation of a tyrosine residue (Lindberg and Hunter 1990), which is not present in the ALPR and its homologs. Since the regulatory role of the C-terminal in EPH family tyrosine kinases has not been proven, it is not clear whether the C-terminal of ALPR and its homologs also have any regulatory function. However, as the C-terminal of these proteins is well conserved, it may play an important role in the protein function.

In summary, ALPR has a possible function in cytoskeletal regulation based on its homology to the rat cortactin binding protein 1. The presence of the ankyrin repeat domain in ALPR and C33B4.3 suggests that the proteins are either membrane bound, or they are involved in other protein-protein interaction. ALPR has different transcripts, including one that may produce a truncated protein due to the premature translation termination at sc24, and those that terminate at three different 3' ends. This suggests that ALPR may have multiple functions. For example, the C-terminal that is homologous to C33B4.3 and rat cortactin binding protein 1 comes from the sequence of the Last Exon. The proteins that are produced from the transcripts with the other 3' ends may function slightly differently to C33B4.3 or rat cortactin binding protein 1. The presence of three human homologs of ALPR (AR, cDNA 208081, DS17 in Fig. 21) suggests that ALPR may represent a member of a new protein family.

2 c) Functional analysis of the C33B4.3 in C. elegans

Model organisms are often useful to study gene function of human candidate disease genes. A common choice of model organism for studying human gene function is the mouse. The mutant phenotype for a gene can often be determined by the gene knockout approach. However, there are disadvantages to this approach. Gene knockout experiments involve difficult techniques, take a long period of time, and are expensive. Alternatively, *C. elegans* provides a powerful research tool for determining the basic function of a protein through genetic analysis. Many proteins involved in a biochemical pathway are well conserved throughout evolution. For example, apoptosis occurs during somatic development in *C. elegans*. There are three major genes that determine whether a cell is committed to programmed cell death: CED-3 and CED-4 promote cell death, whereas CED-9 inhibits the effect of CED-3/CED-4 (Chinnaiyan et al. 1997). The mammalian counterpart of CED-3 (the caspase family) and CED-9 (the Bcl-2 family) have been determined for some time (reviewed by Cohen 1997; Hengartner and Horvitz 1994). The mammalian homolog of CED-4, however, was missing. The induction of programmed cell death by caspase-9 in humans requires the processing of its zymogen form to mature protein (Zou et al. 1997). It resembles the activation of the CED-3

through CED-4 in *C. elegans*. (Chinnaiyan et al. 1997). Therefore, the *C. elegans* pathway predicted that a human homologue of CED-4 should exist, and eventually its homologue Apaf-1 was cloned (Zou et al. 1997). Studies showed that Apaf-1 binds to the zymogen form of caspase-9 to activate its apoptotic activity, and Bcl-X_L inhibits the caspase-9-induced apoptosis by interacting with Apaf-1 (Pan et al. 1998; Hu et al. 1998). Thus the pathway is well conserved from *C. elegans* to humans (Chinnaiyan et al. 1997). Therefore, besides helping to understand the function of a gene, studies in *C. elegans* can also help to identify other proteins involved in the same biochemical pathways as the candidate gene. As a first step towards understanding ALPR, I determined the expression pattern of its *C. elegans* homologue C33B4.3 by reporter gene fusion. I also tried to create a knockout phenotype by double stranded RNA interference.

C33B4.3 shares protein similarity to ALPR throughout the whole sequence. The presence of the ankyrin repeat domain reveals that C33B4.3 is probably involved in protein-protein interactions. The long stretches of polyproline sequences make it a potential ligand for profilin (Mahoney et al. 1997). Although it is homologous to rat cortactin binding protein 1 in some regions (Fig. 21), they are unlikely to have a similar function because the proline rich ligand sequence is absent in C33B4.3 (Fig. 20). C33B4.3 does not therefore interact with other proteins through the binding of a proline rich ligand to a SH3 domain as may be true for ALPR.

Analysis of the expression of the lacZ reporter gene showed that C33B4.3 is expressed in all three concentric germ layers of a metazoan. It is expressed in the hypodermal seam cells, the vulval cells that link to the hypodermal cells, and the ventral nerve cord (ectoderm), the pharynx, anal muscle cells and male reproductive organs (mesoderm), and the posterior six intestinal cells (endoderm).

C33B4.3 is expressed very early during embryogenesis. The expression appears in a single or a few cells. The first prominent structures where expression is found are the seam cells that run along the left and right lateral lines of the body (Fig. 23 A and B). The seam cells are specialized hypodermal cells, which form a specialized cuticle structure called alae at the L1, dauer and adult stage (Singh and Sulston 1978). In the L1 stage it appears that only the six posterior seam cells have C33B4.3 expression (Fig. 25 A and B). Those cells may be the V1 to V6 cells (Fig. 25 C), which undergo further cell division at

the end of that stage. In both sexes the division of the seam cells give rise to hypodermal nuclei. In male, the posterior V5, V6, and T cells undergo further division to generate rays structure. This specialization of cell fates in the seam cells is controlled by the *lin-22* inhibitory signal pathway. The anterior cells (V1 to V4) respond to the inhibitory signal from the posterior cells so that they do not commit to the ray sublineage (Emmons and Sternberg 1997). The early expression of C33B4.3 in these six seam cells suggests that it may play a role in this cell fate specialization.

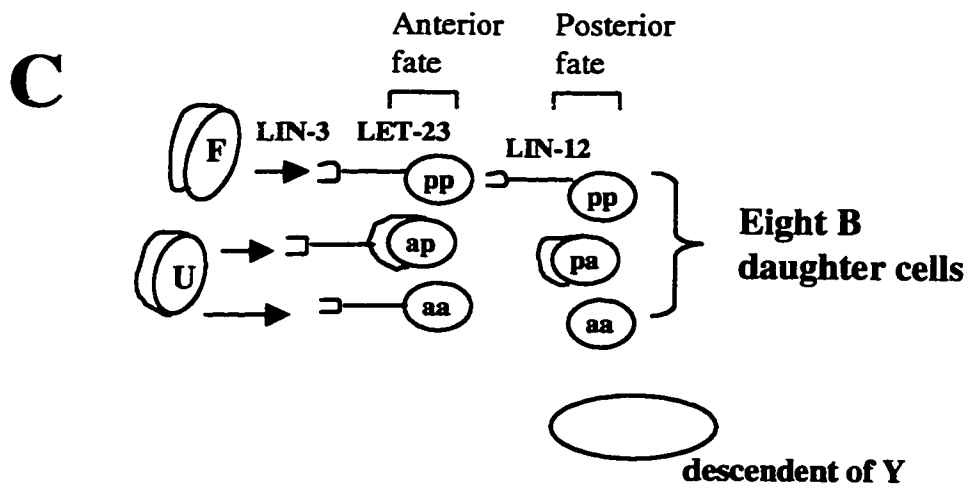
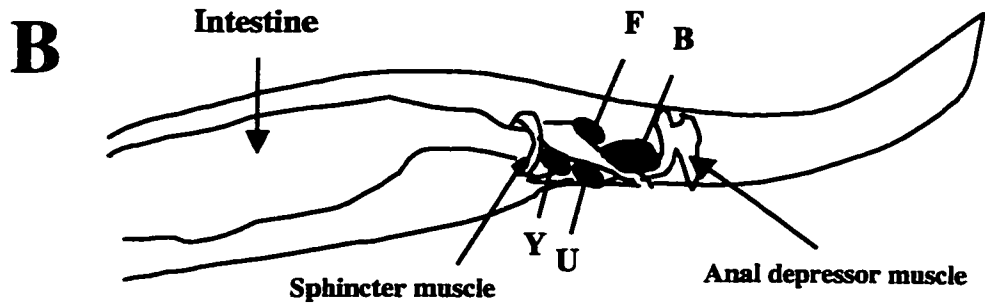
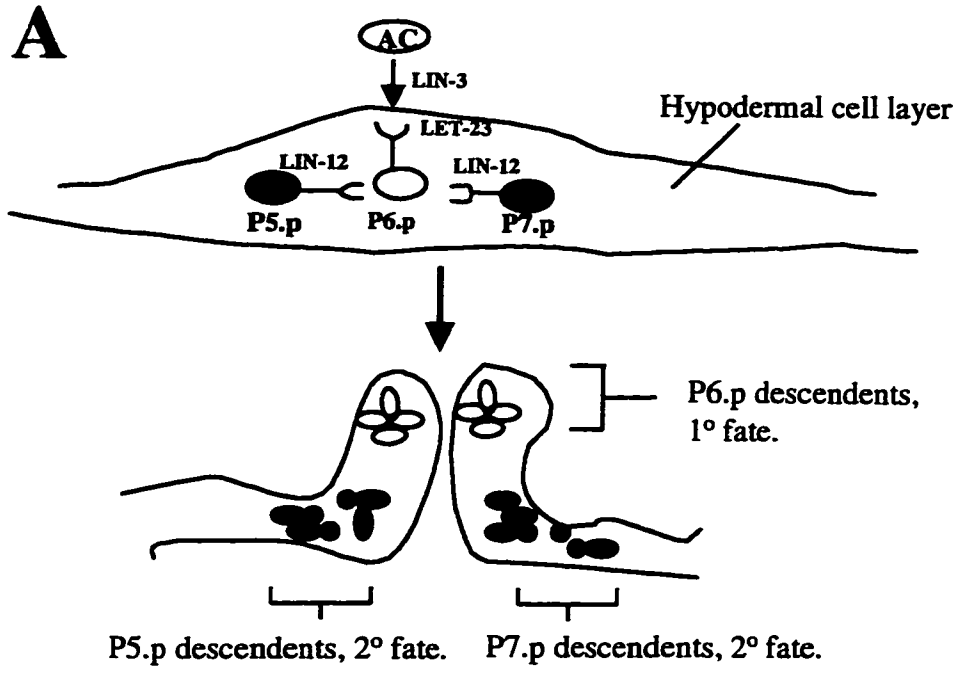
When the embryo is in the three-fold stage before it hatches to an L1 larva, the C33B4.3 expression is seen in the pharynx muscle cells. There may be also some rectal expression, although it is not clearly seen in the plane of focus (Fig. 24 A and B). However, the rectal expression is clear at the L1 larval stage (Fig. 25 B). During embryogenesis a transcription factor *pha-4*, which is a member of the fork head gene family, is expressed in the pharynx and rectum (Kalb et al. 1998). The PHA-4 protein regulates organogenesis of the digestive tract, so it is expressed in pharyngeal precursor cells at the early stage of embryogenesis. The expression persists in all stages of the life cycle (Kalb et al. 1998). C33B4.3 shows a similar expression pattern. Since the expression of C33B4.3 in the embryo is still mosaic, it is not clear whether it is also expressed in the same pharyngeal precursor cells as *pha-4*. However, the pharynx expression is maintained throughout adulthood. (Fig. 26 D). There are, however, differences between *pha-4* and C33B4.3 expression. Weak expression of *pha-4* is found in gut, but C33B4.3 expression is limited to the six posterior intestinal cells (Fig. 27 B). The rectal expression seems to be different between *pha-4* and C33B4.3. Expression of *pha-4* is found in the two rectal valve cells and the three rectal epithelial cells (Kalb et al. 1998), whereas C33B4.3 expresses in the postembryonic blast cells in hermaphrodites (Fig. 27 A, Fig. 39B) and a ring structure which corresponds to either the anal sphincter muscle or the two rectal valve cells that are wrapped by the anal sphincter muscle. Since the staining seems to be on a single ring rather than a two-cell structure, the latter is more likely to be anal sphincter muscle. Therefore, even though the expression pattern of *pha-4* and C33B4.3 is in similar regions, they are actually expressed in different cells. C33B4.3 in the muscle cells of pharynx and anus, versus *pha-4* expression is found in different cell types such as epithelial cells, muscle cells, marginal cells, gland cells and neuronal cells

in pharynx (Kalb et al. 1998). Although both C33B4.3 and *pha-4* express in the pharynx muscle cells throughout life, it is not clear whether they would interact with each other.

It is interesting to find the C33B4.3 expression in both vulval cells of the hermaphrodite and the cells in the male tails. The formation of these two structures is controlled by the same *lin-3/let-23* signal pathway (Fig. 39). The vulva is formed by the differentiation of three precursor cells: P5.p, P6.p, and P7.p. During vulval development, the anchor cell expresses the LIN-3 protein, which is an EGF receptor type ligand. It binds to LET-23 which is a EGF type receptor in the middle vulval precursor cells P6.p (Hill and Sternberg 1992) (Fig. 39 A). The *lin-3/let-23* signaling pathway down-regulates the P6.p expression of LIN-12, which is a receptor molecule that receives an inhibitory signal. The inhibitory signal specifies the cells to adopt the default fate (the secondary fate). As a result, the P6.p adopts the primary fate which forms the opening of the vulva, whereas the two flanking cells, P5.p and P7.p, adopt the secondary fate which forms the wall of vulva (reviewed by Greenwald 1998)(Fig. 39A). The lateral view of vulva in Figure 26 B shows that C33B4.3 is expressed in all descendents of the vulval precursor cells.

Four post embryonic blast cells called B, Y, U, and F lie at the walls of the rectum at the L4 stage (Fig. 39B). These cells do not divide further in the hermaphrodite but they undergo further divisions in male to generate male specific structures. The B cell lineage forms spicules, the two tubular structures that insert into the vulva of the hermaphrodite and transfer the sperm to the hermaphrodite during copulation. The Y lineage forms a group of neurons (postcloacal sensilla) that are close to the opening of the protodeum (the male rectum). U and F lineages form other structures in the protodeum. (Emmons and Sternberg 1997). During the spicule development, the B cells divide into four pairs of cells. The F and U cells are located anterior to the eight B daughter cells (Fig. 39C). They promote the anterior fate of the B daughter cells that are close to them through the *lin-3/let-23* signaling pathway. LIN-12 protein, on the other hand, inhibits the posterior cells from adopting the anterior fates (Chamberlin and Sternberg 1994). Therefore, the spicule development in males is very similar to the vulva development in the hermaphrodite. Both organ developments require the *lin-3/let-23* signaling pathway to specify the cell fates based on their relative positions to the origin of the signal. The

Figure 39 The *lin-3/let-23* signal transduction pathway is involved in vulval development in the hermaphrodite and spicule formation in the male. *A.* Vulva formation by the division of three vulva precursor cells in the hypodermis. The anchor cell (AC) specifies the vulva precursor cell P6.p to adopt the primary cell fate through the *lin-3/let-3* signalling pathway. The two vulva precursor cells that are adjacent to P6.p (P5.p and P7.p) receive an inhibitory signal through the LIN-12 receptor, which instructs them to adopt the secondary fate. C33B4.3 expression is found in all descendants of the three vulva precursor cells. *B.* The organization of the muscle and post-embryonic blast cells at the intestine-rectum junction of the hermaphrodite tail. C33B4.3 expression is found in the sphincter and anal depressor muscle cells as well as some of the post-embryonic blast cells (F, Y, U, and B). *C.* Cell fate specification during spicule development in the male. At the L4 stage, F and U divide once to give two descendent cells. B cells divided into four pairs of daughter cells. F and U promote the anterior fate for the four B daughter cells that are closest to them through the *lin-3/let-23* pathways. The inhibitory signal received by the LIN-12 receptor defines the posterior identity of the posterior pp cell. The figures are redrawn from Sulston and Horvitz (1977), White. (1988), Chamberlin and Sternberg (1994), and Emmons and Sternberg (1997).



lateral inhibition generated by *lin-12* will then specify the cells that are far away from the *lin-3/let-23* signaling source to adopt a secondary fate. The stained cells that are surrounding the opening of the rectum could be the postembryonic cells B, Y, U, and F (Fig. 27 A). The intense staining in the male tails makes it difficult to identify the cell lineage of the stained cells, (Fig. 28 A), but some of them could be derived from the postembryonic cells. It is not clear whether it is a coincidence that C33B4.3 expression occurs in the cells that are involved in the *lin-3/let-23* signal pathway. It is important to determine whether the C33B4.3 expression occurs in the vulval precursor cells at the L3 stage of the hermaphrodite, or the postembryonic blast cells in the L4 stage of the male, where the cell fates are specified in vulval or spicule development.

In order to investigate the function of C33B4.3 in the various cells showing expression, I tried to generate a gene knockout by double stranded RNA interference. This technique originated from the antisense knockout protocol. The idea was based on introducing antisense RNA into the gonad of the hermaphrodite. The antisense RNA will bind to its complementary mRNA, then the double strand RNA can be recognized by RNase and degraded. However, microinjection with sense RNA into the worms has the same effect as injecting the antisense RNA, indicating that the interference of gene expression by antisense RNA cannot be simply explained by the double stranded RNA degradation model (Fire et al. 1998). Fire et al. (1998) showed that microinjecting the worms with double stranded RNA gave a greater interfering effect than injecting either sense or antisense RNA alone. I tested whether double stranded RNA could produce a null mutation for C33B4.3. Although the double stranded RNA from the positive control gene *apx-1* produced the expected embryonic lethal phenotype, no observable phenotype was found when the double stranded RNA of C33B4.3 was microinjected into the worms. There are two possible explanations for this result. First, the loss of function mutation of C33B4.3 may not have a physiological effect on the worm. Because there are two ALPR homologs found in human, there could also be other *C. elegans* family members which would compensate for the loss of C33B4.3 although the sequencing of the *C. elegans* genome is near complete and no homology to another gene was found. Alternatively, the technique did not work for the C33B4.3 gene. Even though Fire et al. (1998) showed that double stranded RNA has a high and specific interfering effect on gene expression, the

mechanism for such interference effect is still unknown. The results of antisense knockout experiments showed that the technique had a higher success rate for maternally expressed genes than genes expressed in the zygote (<http://www.ummed.edu/pub/c/cmello/index.html>). In successful experiments Worms lose the transcription of the target gene after the microinjection of the double strand RNA (Fire et al. 1998). To test whether the double strand RNA causes any interference with C33B4.3 expression, one can examine the transcription level of this gene after the microinjection of double strand RNA. At this stage it is still inconclusive whether the double strand RNA can knockout the expression of C33B4.3 gene.

In summary, C33B4.3 is expressed in various tissue types in *C. elegans*. It is interesting to note the sexually dimorphic expression pattern of C33B4.3. This dimorphic expression appears to be due to the lack of cell division occurring in the B, Y, U, and F postembryonic cells in adult hermaphrodites, whereas these cells divide further in male tails to produce various male-specific structures. C33B4.3 expression is found in cells that are involved in two different signal pathways. The *lin-22* inhibition pathway controls the formation of the rays from the posterior seam cells in the male. The *lin-3/let-23* pathway controls vulva development in the hermaphrodite and spicule development in the male. C33B4.3 may have a function in those signal transduction pathways that control development. The human ALPR may be involved in cytoskeleton remodeling. It is not clear whether C33B4.3 also has a similar function. The polyproline sequence in C33B4.3 is a potential ligand for profilin, which also plays a role in cytoskeletal regulation. However, profilin is involved in assembling actin monomers to filaments, rather than the redistribution of filamentous actin in the cortical cytoskeleton network. Also, it is not known whether cytoskeleton remodeling is required for those cells which undergo differentiation during development. Nevertheless, this finding suggests that ALPR and its homologs may play a role in a signal transduction pathway. Components of such a pathway often show haploinsufficiency (Table 2), making ALPR a potential candidate gene for the 22q13.3 deletion syndrome.

3 a) Structure of RABL

RABL was identified by an EST contig located at position 106843 – 107128 of AWcontig. One representative cDNA from the EST contig was fully sequenced. Sequence analysis showed that it is homologous to the genes in the RAB family. Thus, the gene is named as RAB-like or RABL.

RABL is located in the proximal subtelomeric domain of 22q and oriented telomere to centromere. Although the 5' UTR of RABL has not yet been fully determined, RABL is adjacent to the boundary of the distal subtelomeric domain, which presumably does not contain any expressed sequences (Flint et al. 1997b). Therefore, RABL is probably the last gene on chromosome 22q. A proximal subtelomeric domain usually shows strong homology and continuous matches with a few chromosome ends (Fig. 36 B, Flint et al. 1997b). In the case of 22q, there is one extended region of homology. The genomic region where RABL is located is homologous to 2q13, the fusion site of two ancestral ape chromosomes (Ijdo et al. 1991). The genes encoded by the RABL locus on chromosome 2 (RABL2) and chromosome 22 (RABL22) only differ by three amino acids, and all of the differences represent conservative changes (Fig. 32). Both RABLs are expressed in all the tissues tested (Fig. 33), although subtle differences in the degree of expression may exist in some tissues.

3 b) Possible function of RABL

RABLs share similarity with RAB proteins, which belong to a sub-family of the RAS superfamily. To date mammalian cells have been found to have more than 40 members of the RAB family. RAB proteins are small GTPases that control vesicular trafficking in the cells (reviewed by Lazar et al. 1997; Novick and Zerial 1997; and Olkkonen and Stenmark 1997). RAB proteins are involved in transporting vesicles from one subcellular compartment to another, or transporting proteins in the vesicle in and out of cells through endo- or exocytosis respectively. The transportation of vesicles is

regulated by a cycle of GDP/GTP exchange by the RAB protein. The inactive form of the RAB protein is bound to GDP. RAB is activated by the exchange GDP to GTP catalysed by guanine nucleotide exchange factor (GEF) (Fig. 40A). The active form of RAB-GTP binds at the carboxyl end to the membrane of a vesicle. The carboxyl end contains a cysteine rich domain which is subjected to prenylation by geranylgeranyl transferase. The 20-carbon geranylgeranyl group makes the carboxyl end hydrophobic so that it can bind to the membrane of the vesicle (Fig. 40B). The RAB-GTP then moves the vesicle to the target site. It docks on the effector molecule which is at the surface of the target site (Fig. 40C). The vesicle fuses with the target membrane and the contents inside the vesicle is released. This is mediated by the hydrolysis of GTP to GDP in the presence of GTPase-activating protein (GAP)(Fig. 40D). The GDP dissociation inhibitor (GDI) removes the RAB-GDP from the membrane (Fig. 40E). In the cytosol, the GDI is dissociated from the RAB-GDP. The free RAB-GDP complex is available to bind to GEF, which restarts the vesicle transport cycle (Fig. 40F).

An Online Mendelian Inheritance in Man database search

(<http://www.ncbi.nlm.gov/omim/searchomim.html>) did not show any known human disorder that is associated with RAB gene mutations. However, there are two RAB gene-related human genetic diseases. Mutations in RAB escort protein-1 (REP-1) cause atrophy of the choroid of the eye, resulting in choroideremia, an X-linked recessive disorder. REP-1 shares protein homology with GDI. It assists the geranylgeranyl transferase in the process of prenylation. Furthermore, REP-1 shares some functional properties of GDI, including the removal of RAB-GDP from the membrane and inhibiting the release of GDP from RAB (Alexandrov et al. 1994). The target RAB for the REP-1 assisted prenylation is Ram/Rab27, which shows the absence of prenylation in patients with choroideremia (Seabra et al. 1995). Another RAB related disease is non-specific X-linked mental retardation (MRX). Mutations in the GDI1 gene, which encodes the α GDI protein, is believed to be the cause of this disorder. α GDI binds to RAB3, which regulates the Ca^{2+} -dependent synaptic-vesicle fusion at the synaptic terminals of dendrites (D'Adamo et al. 1998). The mutations of GDI1 in male patients abolish the binding between α GDI and RAB3-GDP, which may result in an increased GTP/GDP

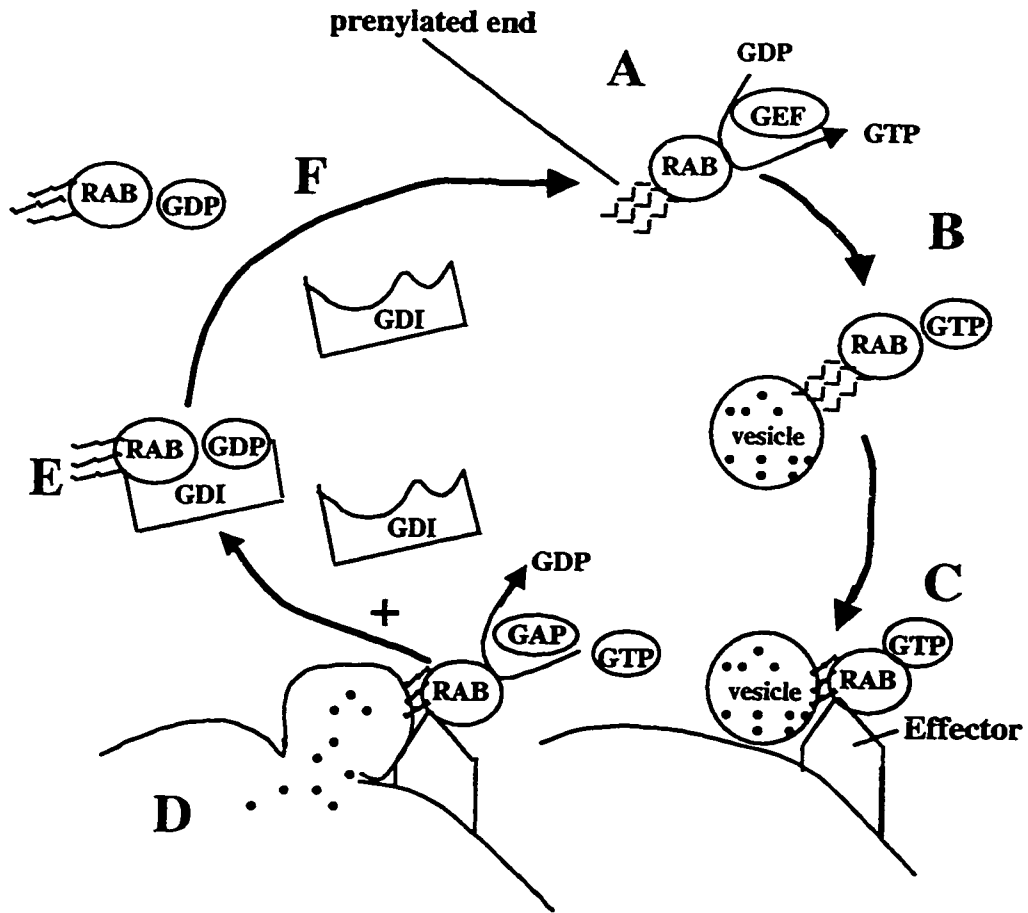


Figure 40 The cycling of a RAB protein vesicle transportation. *A.* The guanine nucleotide exchange factor (GEF) exchanges the RAB bound GDP for GTP. *B.* The RAB-GTP binds to the target vesicle. The vesicle attaches to the prenylated carboxyl end of RAB. *C.* The vesicle is transported to a membrane surface, which can be either the plasma membrane or the membranes of different organelles. RAB docks on a specific effector molecule. *D.* GTPase-activating protein catalyzes the hydrolysis of GTP to GDP. This results in membrane fusion and releases the vesicle contents. *E.* The GDP dissociation inhibitor (GDI) binds to RAB-GDP and removes it from the membrane. *F.* GDI dissociates from RAB-GDP, which makes the RAB-GDP accessible to bind to GEF to restart the cycle of the vesicle transport. Redrawn from Novick and Zerial (1997) and Lazar et al. (1997).

ratio and the constitutive binding of the RAB3 to the membrane. This defect occurs in the central nervous system, resulting in mental retardation in males.

Similar to all the GTPases in the RAS superfamily, RABs are small proteins with an average length of 200 – 300 amino acid residues. A typical RAB protein contains five conserved domains: two guanine base-binding motifs (designated as G4 and G5 by Lazar et al. 1997), two phosphate-Mg²⁺ binding motifs (G1 and G3) and one effector region (G2). RABLs are 228 amino acids in length, which is within the average size of a RAB protein. However, RABLs lack one of the two conserved guanine base-binding domain (the G5 motif with the conserved sequence EXSA), and the cysteine- rich domain at the carboxyl terminal which is important for the binding of RABs onto the membrane of vesicles (Fig. 32). The absence of the cysteine- rich domain suggests that RABLs may be unable to bind to the membranes of vesicles. Therefore, RABLs are unlikely to function in the same way as other RAB proteins. In fact, it is even questionable whether RABLs are GTPases because they do not have one of the two conserved guanine binding motifs. The effector region, G2, determines the specificity of the RAB molecules. Each specific effector sequence interacts with different GTP activating proteins (GAP), or other putative RAB effector sequences that acts downstream to RAB (reviewed by Olkkonen and Stenmark 1997). The effector sequences are different in the two RABLs (DGKTILVDF in RABL2 versus DGRITLVDF in RABL22). If the RABL effector region functions in a similar way to RABs, the change of the effector sequence may give the two RABLs different functions. However, since the difference in the amino acid is conservative (lysine and arginine), these two RABL proteins are likely to have similar if not identical functions.

RABL has a 2.5 kb transcript expressed ubiquitously and a 1.4 kb transcript with specific expression in adult and skeletal muscle (Fig. 34A). A 3' RACE result shows that the size difference between the 1.4 and 2.5 kb transcripts is probably the result of the presence or absence of ~1 kb internal intron in the 3'UTR (Fig. 32). The hybridization of the internal intron sequence to a Northern blot confirms that the 2.5 kb transcript carries the internal intron, whereas the 1.4 kb muscle specific transcript does not (Fig. 34B). Since the coding sequences of these two different sized transcripts are the same, the absence of the internal intron in the 3'UTR specifically in muscle is difficult to explain.

In general, the 3' UTR sequence can be involved in various functions, including stabilization of transcripts (Crowley and Hazelrigg 1995; Miyamoto et al. 1996; Wilson et al. 1998), subcellular localization of transcripts (Ainger et al. 1997; Macdonald and Kerr 1997; Mahon et al. 1997), post-transcriptional regulation (Kern et al. 1997; Savitsky et al. 1997), translational repression (Hel et al. 1998), translational modification (Kollmus et al. 1996), or even tumor suppression (Jupe et al. 1996). The internal intron occupies 1255 bp of the 3' UTR, and splicing out of this intron will leave only 220 bp of 3'UTR in the 1.4 kb transcript. If the length of the 3'UTR determines the stability of the transcript, the short 3' UTR transcript may be more stable than the long one, so that more RABL will be produced in muscle tissues which may have a higher demand for the protein. If the short 3' UTR transcript is less stable, it would be unclear why the muscle cells need short-lived transcripts if they already have the stable long 3' UTR transcripts. The subcellular localization model is attractive. If a particular specific subcellular structure in the muscle cells specifically needs RABL function, the alternative splicing in the 3' UTR may localize the corresponding transcript to that subcellular structure. It should be noted that some RAB proteins are found in more than one subcellular location of a cell (reviewed by Olkkonen and Stenmark 1997). If RABL functions in a similar way to RAB proteins, it may need a mechanism for the distribution of the protein to a specific subcellular location. Other models are possible but it is difficult to explain how they could fit in the scenario. For example, if the translation of either one of the transcripts is repressed, it is unclear why the cells are obligated to transcribe the mRNA. To test which model is correct, one could determine the stability or the subcellular location of these two transcripts. Different lengths of the 3'UTR generated by splicing out of the internal intron have also been found in the transducin beta3 subunit gene in dog (Akmedo et al. 1997), but the functions of the two different UTRs is not known.

Other than the common 2.5 kb transcript, RABL has two more faint transcripts (4.4 kb, > 10 kb) seen in the fetal tissues (Fig. 34). The putative 5' end of RABL is only ~2 kb away from the degenerative telomeric repeat sequences (Fig. 31), the boundary between the proximal and distal subtelomeric domain. If there is more 5' sequence for RABL that corresponds to the sizes of the two fetal transcripts, RABL would span the proximal and distal subtelomeric region. The 5' end would then contain the subtelomeric

repeat sequences shared by other chromosomes. There is only 1.6 kb of unique sequence centromeric to the 3' end of the RABL cDNA; after that the sequence runs into a 16 kb cluster of the LINE/SINE repeats between ACR and RABL (Fig. 9). Centromeric to the 16 kb LINE/SINE repeat cluster is the minisatellite repeat sequence 3'AR. There is only a total of 2.5 kb of unique sequence between ACR and 3'AR. As a result, if RABL extends 5' or 3' of the present location, it will run into repeat sequences. The fetal transcripts may have a large 3' UTR containing a larger number of repeats. Alternatively, the 4.4 and >10 kb transcripts may not represent RABL expression, but perhaps represent cross-hybridization with related genes.

In summary, the two RABL genes are homologous to RAB genes. One of the members of the RAB family (RAB3) has been associated with the non-specific X-linked mental retardation. This indicates that the mutations in RAB or related genes can possibly cause the abnormal phenotypic features seen in NT. However, RABLs do not seem to function in the same way as RAB proteins and they are not expressed specifically in brain. Further investigation of the novel function of these new RAB-like genes is necessary.

3c) Gene duplication in RABL

The genomic regions of RABL2 and RABL22 are highly homologous. The percentage of sequence divergence within the 15 kb RABL locus is only 0.56 (see the "Genomic Organization of the NT Microdeletion Region" section above). This suggests that the genomic sequence of RABL is recently duplicated. There are 23 nucleotide differences over the 2114 bp of the two RABL genes (~1% divergence). If the average rate of neutral nucleotide substitution in primates is between $1.5 - 2 \times 10^{-9}$ substitutions per nucleotide site per year (Régner et al. 1997), the two RABL genes would have begun to diverge from each other in the last 2.5 - 3 million years (Myr). The divergence date of the chimpanzee from human is 6.4 ± 2.6 Myr based on an estimation of the silent nucleotide substitution rate in the coding region (Sakoyama et al. 1987), or 4.5 Myr based on a calculation using a maximum likelihood method (Takahata and Satta 1997). Either estimate of the divergence date suggests that RABL was duplicated after the human

ancestral lineage branched off from the chimpanzee, which is the most closely related primate to human. This result would suggest that RABL is only duplicated in humans, and not other primates. Recently, we have attempted to estimate the number of RABL genes in various primates (Austin Chen, unpublished results). The primate DNAs were digested with BsrG1, which is the endonuclease site that can distinguish RABL22 from RABL2 (Fig. 30 and Fig. 31). The digested DNA was then electrophoresed and Southern blotted. A partial cDNA within the exon that contains polymorphic BsrG1 site was amplified by RT-PCR (RABLF1-E0.91, see Fig. 31), and used to probe the primate blot. This probe showed a common 3.2 kb band in orangutan, gorilla, chimpanzee and human. In human this band represents the RABL22 copy, as confirmed by comparing to the chromosome 22 containing somatic cell hybrid. Human also showed a ~7.5 kb band, representing the RABL2 locus, which does not have a BsrG1 site. This result suggests that other than humans, all the primates tested have only one copy of the RABL locus, the equivalent of RABL22. This matches the gene divergence data, which predicts that only human has a duplication in RABL. However, it is also possible that other primates have duplications in RABL, but that the BsrG1 site is conserved between the two RABL genes so that BsrG1 cannot differentiate between the two loci. To resolve this, FISH could be done on primate chromosomes using cosmid C202 (Fig. 3), which hybridizes to both human RABL loci. This has previously been done for the olfactory receptor genes, which are found on the subtelomeric region of 3q, 15q, and 19p (Trask et al. 1998). When the genomic DNA containing these genes were used as probes for FISH analysis, there was only one major signal found in the metaphase spreads of orangutan, chimpanzee, and gorilla (Trask et al. 1998), indicating that duplications had only occurred in human.

4 . Relevance to NT microdeletion

My research indicates that the NT microdeletion region contains three genes. The challenge is to determine if any of these genes are associated with the abnormal features of mental retardation and expressive speech delay found in this patient. It is unlikely that the deletion of acrosin, involved in the sperm penetration of oocyte, leads to the abnormal features of NT. Although RABL22 is homologous to RAB proteins, it lacks the

functional domain that is essential for the binding of RAB proteins to the membrane of vesicles. Also, the absence of one of the two guanine base binding domains makes it less likely to be a GTPase. Since the function of this novel protein is therefore unknown, it is difficult to predict whether it is involved in NT's abnormal phenotype. However, since both RABL2 and RABL22 are expressed, and very close in sequence, RABL may have the equivalent of 4 active copies. If this is the case, the deletion of one in four copies of RABL may not produce any phenotypic effect. This leaves ALPR. ALPR could be a protein involved in a signal transduction pathway, possibly the actin remodeling in response to a signal or specifying the position identity of a cell during development as in *C. elegans*. That makes ALPR a potential candidate to be affected by haploinsufficiency. There is also a specific transcript in fetal and adult brain. It is also possible that none of the genes are involved in the syndrome. In order to determine whether ALPR is the candidate gene, one may need to look for another microdeletion or cryptic rearrangements that disrupt the locus of ALPR, and compare the corresponding phenotype with that in NT. Further clinical comparison of the specific abnormalities shown by NT and patients with deletion 22q13.3 syndrome may also help to determine whether NT's features represent a subset of the syndrome.

Future Research

Construction of a full ALPR cDNA contig

There are still regions of ALPR which remain to be cloned. These are the 3'UTR of the Last Exon, genscanex1, and the 5'UTR. The 3' UTR of the Last Exon may be cloned by 3' RACE. One could hybridize a clone of the Last Exon to a Northern blot, and choose the appropriate tissue for the RACE experiment. The cloning of the sequence that is 5' to the Last Exon may be difficult, because that region also has a high G-C content (Fig. 19). Perhaps one could try to clone the mouse homolog of the ALPR that includes the Last exon. The G-C content of the region might be slightly lower in mice as it is in *C. elegans* where this region was cloned. ALPR may be more clonable in another organism.

The cloning of the *genscanex1* is also a difficult task. I experimentally found that *genscanex1* was refractory to cloning, perhaps because its mRNA forms a strong secondary structure. To prove it is a true expressed sequence, one could look for the mouse homolog of this gene if it is clonable. By comparing the ORF between human and mouse ALPR, one may be able to confirm the prediction of exon boundary.

Like the Last Exon and *genscanex1*, the 5'UTR is also G-C rich. Dr Bruce Roe's laboratory only sequenced 298 bp upstream of the starting exon of ALPR. This 298 bp contains 83% G-C content, and a few unresolved nucleotides. The sequence that is 5' to the gap is difficult to predict. I do not have any further cDNA sequence to compare the sequence that is next to this gap, I cannot determine the orientation of that sequence contig. A genomic BAC clone (799F10, GenBank accession number Z83345) overlaps with the AWcontig region and is being sequenced by the Sanger Centre, so eventually this gap will be closed. I tried twice to clone the 5' UTR by 5' RACE using a commercial cDNA library (fetal brain Marathon cDNA library, Clontech), but without success. The complete sequence of the region should be available soon, and then some of these problems may be easier to address.

The stop codon in sc24 suggests that there could be a transcript that produces a truncated form of the protein. The alternatively spliced transcripts that put the H55337 exon in frame could be cloned by RT-PCR. Alternatively, the two human ALPR homologs could be fully sequenced, and their ORFs compared to that in ALPR to see which exons could put the ALPR translation in frame.

Functional analysis of the mammalian ALPR gene

I have hypothesized that the clinical features of mental retardation and expressive speech delay in NT are possibly due to haploinsufficiency of ALPR. The mouse could be used as a model organism at least to study the possible role of ALPR in brain.

Experimental techniques such as whole mount and sectioned mouse embryo *in situ* hybridization could be used to determine if there is any specificity to embryonic/fetal brain. Since ALPR is probably involved in protein-protein interactions, the best approach to determining function might be to determine what ALPR interacts with. This could be

done by a immunoprecipitation or two-hybrid system. To see whether the deletion of ALPR will produce an abnormal phenotype in mouse, one could try a gene knockout approach. Although the mouse cannot be used to study speech development, it can show whether the deletion affects the function of the brain by doing behavioral analysis of the mutant animal.

Determine the basic function of ALPR using C. elegans

The gene expression data of C33B4.3 sets the stage for future work on this gene in *C. elegans* to determine its function in this organism. Further studies are needed to identify certain cell lineages that show C33B4.3 expression, especially the cells that show male specific expression at the tail of the worm. I made one attempt to determine the mutant phenotype by deletion library screening (Jensen et al. 1997) but without success. The basic idea for the deletion library is to mutagenize worms using chemicals that are known to generate deletions, and then arrange the mutagenized animals into pools which can be screened. A pair of primers flanking the target gene locus are then used for screening for a deletion by PCR in the mutagenized worm population (Jensen et al. 1997). It would be worthwhile to screen a new deletion library for a C33B4.3 mutant. Much information about the function of a gene could be revealed by a mutant phenotype. Once the mutant is isolated, one could try to rescue the mutant by introducing the human ALPR. If the human homolog can rescue the mutant phenotype, one can then conclude ALPR and C33B4.3 share functional homology. Furthermore, a yeast two hybrid system could be used to determine what proteins interact with C33B4.3. This would help to understand the biochemical pathway that the gene is involved in.

References

- Ainger K., Avossa D., Diana A.S., Barry C., Barbarese E., Carson J.H. (1997) Transport and localization elements in myelin basic protein mRNA. *J Cell Biol* 138: 1077-1087.
- Akhmedov N.B., Piriev N.I., Ray K., Acland G.M., Aguirre G.D., Farber D.B. (1997) Structure and analysis of the transducin beta3-subunit gene, a candidate for inherited cone degeneration (cd) in dog. *Gene* 194: 47-56.
- Alexandrov K., Horiuchi H., Steele-Mortimer O., Seabra M.C., Zerial M. (1994) Rab escort protein-1 is a multifunctional protein that accompanies newly prenylated rab proteins to their target membranes. *EMBO J* 13: 5262-5273.
- Altherr M.R., Bengtsson U., Elder F.F.B., Ledbetter D.H., Washmuth J.J., McDonald M.E., Gusella J.F., Greenberg F. (1991) Molecular confirmation of Wolf-Hirschhorn syndrome with a subtle translocation of chromosome 4. *Am J Hum Genet* 49: 1235-1242.
- Altherr M.R., Wright T.J., Denison K., Perez-Castro A.V., Johnson V.P. (1997) Delimiting the Wolf-Hirschhorn syndrome critical region to 750 kilobase pairs. *Am J Med Genet* 71: 47-53.
- Aparicio O.M., Billington B.L., Gottschling D.E. (1991) Modifiers of position effect are shared between telomeric and silent mating-type loci in *S. cerevisiae*. *Cell* 66: 1279-1287.
- Armour J.A.L., and Jeffreys A.J. (1991) STS for minisatellite MS607 (D22S163) *Nucleic Acid Res* 19: 3158.
- Armour J.A.L., Povey S., Jeremiah S., Jeffreys A.J. (1990) Systematic cloning of human minisatellites from ordered array charomid libraries. *Genomics* 8: 501-512.
- Artavanis-Tsakonas S., Simpson P. (1991) Choosing a cell fate: a view from the Notch locus. *Trends Genet* 7: 403-408.
- Aurias A., Rimbault C., Buffe D., Dubousset J., Mazabraud A. (1983) Chromosomal translocations in Ewing's sarcoma. *N Eng J Med* 309: 496-497.
- Baba T., Azuma S., Kashiwabara S.-I., Toyoda Y. (1994) Sperm from mice carrying a targeted mutation of the acrosin gene can penetrate the oocyte zona pellucida and effect fertilization. *J Biol Chem* 269: 31845-31849.
- Baba T., Watanabe K., Kashiwabara S.-I., Arai Y. (1989) Primary structure of human proacrosin deduced from its cDNA sequences. *FEBS Lett* 244: 296-300.

- Ballabio A., Bardoni B., Carozzo R., Andria G., Bick D., Campbell L., Hamel B., Ferguson-Smith M.A., Gimelli G., Fraccaro M., Maraschio P., Zuffardi O., Guiolo S., Carnerino G. (1989) Contiguous gene syndrome due to deletions in the distal short arm of the human X chromosome. *Proc Natl Acad Sci USA* 86: 10001-10005.
- Bamshad M., Lin R.C., Law D.J., Watkins W.S., Krakowiak P.A., Moore M.E., Franceschini P., Lala R., Holmes L.B., Gebuhr T.C., Bruneau B.G., Schinzel A., Seidman J.G., Seidman C.E., Jorde L.B. (1997) Mutations in human TBX3 alter limb, apocrine and genital development in ulnar-mammary syndrome. *Nat Genet* 16: 307-310.
- Barion A., Prat A., Caron F. (1987) Telomeric site position heterogeneity in macronuclear DNA of *Paramecium primaurelia*. *Nucleic Acid Res* 15: 1717-1728.
- Basson C.T., Bachinsky D.R., Lin R.C., Levi T., Elkins J.A., Soultis J., Grayzel D., Kroumpouzou E., Traill T.A., Leblanc-Straceski J., Renault B., Kucherlapati R., Seidman J.G., and Seidman C.E. (1997) Mutations in human cause limb and cardiac malformation in Holt-Oram syndrome. *Nat Genet* 15: 30-35.
- Batchelor A.H., Piper D.E., de la Brousse F.C., McKnight S.L., Wolberger C. (1998) The structure of GABP α/β : An EST domain-ankyrin repeat heterodimer bound to DNA. *Science* 279: 1037-1041.
- Bellugi U., Wang P.P., Jernigan T.L. (1994) Williams syndrome: an unusual neuropsychological profile. *In* Broman S.H. and Grafman J., eds. *Atypical Cognitive Deficits in Developmental Disorders: Implications for Brain Function*. Hillsdale, New Jersey: Erlbaum, pp23-56.
- Bennett C.B., Lewis A.L., Baldwin K.K., Resnick M.A. (1993) Lethality induced by a single site-specific double-strand break in a dispensable yeast plasmid. *Proc Natl Acad Sci USA* 90: 5613-5617.
- Berger R., Bernheim A., Weh H.-J., Flandrin G., Daniel M.-T., Brouet J.-C., Colbert N. (1979) A new translocation in Burkitt's tumor cells. *Hum Genet* 53: 111-112.
- Bernard O., Ganiatsas S., Kannourakis G., Dringen R. (1994) Kiz-1, a protein with LIM zinc finger and kinase domains, is expressed mainly in neurons. *Cell Growth Diff.* 5: 1159-1171.
- Beuren A.J., Apitz J., Harmjanz D (1962) Supravalvular aortic stenosis in association with mental retardation and a certain facial appearance. *Circulation* 26: 1235-1240.
- Biessmann H., and Mason J.M. (1997) Telomere maintenance without telomerase. *Chromosoma* 106: 63-69.

- Bondy B., de Jonge S., Pander S., Primbs J., Ackenheil M. (1996) Identification of dopamine D4 receptor mRNA in circulating human lymphocytes using nested polymerase chain reaction. *J Neuroimmunol* 71: 139-144.
- Brkanac Z., Cody J.D., Leach R.J., DuPont B.R. (1998) Identification of cryptic rearrangements in patients with 18q- deletion. *Am J Hum Genet* 62: 1500-1506.
- Brodsky F.M. (1988) Living with clathrin: its role in intracellular membrane traffic. *Science* 242: 1396-1402.
- Brooks J.K., Coccaro P.J. Jr., Zarbin M.A. (1989) The Rieger anomaly concomitant with multiple dental, craniofacial, and somatic midline anomalies and short stature. *Oral Surg Oral Med Oral Path* 68: 717-724.
- Brown W.R.A., MacKinnon P.J., Villasanté A., Spurr N., Buckle V.J., Dobson M.J. (1990) Structure and polymorphism of human telomere-associated DNA. *Cell* 63: 119-132.
- Budarf M.L. and Emanuel B.S. (1997) Progress in the autosomal segmental aneusomy syndromes (SASs): single or multi-locus disorders? *Hum Mol Genet* 6: 1657-1665.
- Budarf M.L., Eckman B., Michaud D., McDonald T., Gavigan S., Buetow K.H., Tatsumura Y., Liu Z., Hilliard C., Driscoll D., Goldmuntz E., Meese E., Zwarthoff E.C., Williams S., McDermid H., Dumanski J.P., Biegel J., Bell C., Emanuel B.S. (1996) Regional localization of over 300 loci on human chromosome 22 using a somatic cell hybrid mapping panel. *35: 275-288.*
- Bult A., Zhao F., Dirx R Jr., Raghunathan A., Solimena M., Lombroso P.J. (1997) STEP: a family of brain-enriched PTP. Alternative splicing produces transmembrane, cytosolic and truncated isoforms. *Eur J Cell Biol* 72: 337-344.
- Burge C., and Karlin S. (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 268: 78-94.
- Burn J., Wilson D.I., Cross I., Atif U., Scambler P., Takao A., Goodship J. (1995) The clinical significance of 22q11 deletion. *In* Clark E.B., Markwald R.R., Takao A. (eds) *Developmental Mechanism of Heart Disease*. Futura Publication, Armonk, New York pp 559-567.
- Chamberlin H.M., and Sternberg P.W. (1994) The *lin-3/let-23* pathway mediates inductive signalling during male spicule development in *Caenorhabditis elegans*. *Development* 120: 2713-2721.
- Chan D., Weng Y.M., Graham H.K., Silience D.O., and Bateman J.F. (1998) A nonsense mutation in carboxyl-terminal domain of type X collagen causes halpoin sufficiency in Schmid metaphyseal chondrodysplasia. *J Clin Invest* 101: 1490-1499.

- Charbonneau H., Tonks N.K., Kumar S., Diltz C.D., Harrylock M., Cool D.E., Kerbs E.G., Fisher E.H., Walsh K.A. (1989) Human placenta protein-tyrosine-phosphatase: Amino acid sequence and relationship to a family of receptor-like proteins. *Proc Natl Acad Sci USA* 86: 5252-5256.
- Chen X.N., and Korenberg J.R. (1995) Localization of human CREBBP (CREB binding protein to 16p13.3 by fluorescence *in situ* hybridization. *Cytogenet Cell Genet* 71: 56-57.
- Chinnaiyan A.M., O'Rourke K., Lane B.R., Dixit V.M. (1997) Interaction of CED-4 with CED-3 and CED-9: A molecular framework for cell death. *Science* 275: 1122-1126.
- Chitayat D., Babul R., Silver M.M., Jay V., Teshima I.E., Babyn P., Becker L.E. (1996) Terminal deletion of the long arm of chromosome 3 [46, XX, del (3)(q27-->qter)]. *Am J Med Genet* 61: 45-48.
- Church G.M., and Gilbert W. (1984) Genomic sequencing. *Proc Natl Acad Sci USA* 81: 1991-1995.
- Chute I., Le Y., Ashley T., Dobson M.J. (1997) The telomere-associated DNA from human chromosome 20p contains a pseudotelomere structure and shares sequences with the subtelomeric regions of 4q and 18p. *Genomics* 46: 51-60.
- Claeys I., Holvoet M., Eyskens B., Adriaensens P., Gewillig M., Fryns J.P., Devriendt K. (1997) A recognisable behavioural phenotype associated with terminal deletions of the short arm of chromosome 8. *Am J Med Genet* 74: 515-520.
- Cohen G.M. (1997) Caspases: the executioners of apoptosis. *Biochem J* 326(Pt 1): 1-16.
- Collins K., Kobayashi R., Gredier C.W. (1995a) Purification of Tetrahymena telomerase and cloning of genes encoding the two protein components of the enzyme. *Cell* 81: 677-686.
- Collins J.E., Cole C.G., Smink L.J., Garrett C.L., Leversha M.A., Soderlund C.A., Maslen G.L., Everett L.A., Rice K.M., Coffey A.J., Gregory S.G., Gwilliam R., Dunham A., Davies A.F., Hassock S., Todd C.M., Lehrach H., Hulsebos T.J.M., Weissenbach J., Morrow B., Kucherlapati R.S., Wadey R., Scambler P.J., Kim U.-J., Simon M.I., Peyrard M., Xie Y.-G., Carter N.P., Jurbin R., Dumanski J.P., Bentley D.R., Dunham I. (1995b) A high-density YAC contig map of human chromosome 22. *Nature Suppl* 377: 267-379.
- Comings D.E. (1978) Mechanisms of chromosome banding and implications for chromosome structure. *Annu Rev Genet* 12: 25-46.

- Conrad B., Dewald G., Christensen E., Lopez M., Higgins J., Piepont M.E. (1995) Clinical phenotype associated with terminal 2q37 deletion. *Clin Genet* 48: 134-139.
- Cooper J.P., Watanabe Y., Nurse P. (1998) Fission yeast Taz1 protein is required for meiotic telomere clustering and recombination. *Nature* 392: 828-831.
- Crowley T.E., and Hazelrigg T. (1995) A male specific 3' UTR regulates the steady-state level of the exuperantia mRNA during spermatogenesis in *Drosophila*. *Mol Gen Genet* 248: 370-374.
- Cruz-Reyes J., and Sollner-Webb B. (1996) Trypanosome U-deletional RNA editing involves guide RNA-directed endonuclease cleavage, terminal U exonuclease, and RNA ligase activities. *Proc Natl Acad Sci U S A* 93: 8901-8906
- Curran M.E., Atkinson D.L., Ewart A.K., Morris C.A., Leppert M.F., Keating M.T. (1993) The elastin gene is disrupted by a translocation associated with supravalvular aortic stenosis. *Cell* 73: 159-168.
- D'Adamo P., Menegon A., Lo Nigro C., Grasso M., Gulisano M., Tamanini F., Biennu T., Gedeon A.K., Oostra B., Wu S.-K., Tandon A., Valtorta F., Balch W.E., Chelly J., Toniolo D. (1998) Mutations in GDI1 are responsible for X-linked non-specific mental retardation. *Nat Genet* 19: 134-139.
- Daw S.C., Taylor C., Kraman M., Call K., Mao J., Schuffenhauer S., Meitinger T., Lipson T., Goodship J., Scambler P. (1996) A common region of 10p deleted in DiGeorge and velocardiofacial syndromes. *Nat Genet* 13: 458-460.
- De Jong P.J., Yokobata K., Chen C., Lohman F., Pederson L., McNinch J., van Dilla M., (1989) Human chromosome-specific partial digest libraries in κ and cosmid vectors (A2333). *Cytogenet Cell Genet* 51: 985.
- De la Chapelle A., Herva R., Koivisto M., Aula P. (1981) A deletion in chromosome 22 can cause DiGeorge syndrome. *Hum. Genet.* 57: 253-256
- De Lange T. (1998) Ending up with the right partner. *Nature* 392: 753-754.
- Deng C., Zhang P., Harper J.W., Elledge S.J., Leder P. (1995) Mice lacking p21^{CIP1/WAF1} undergo normal development, but are defective in G₁ checkpoint control. *Cell* 82: 675-684.
- Descartes M., Keppler-Noreuil K, Knops J., Longshore J.W., Finley W.H., Carroll A.J.(1996) Terminal deletion of the long arm of chromosome 4 in a mother and two sons. *Clin Genet* 50: 538-540.
- Devriendt S., Swillen A., Fryns J.P., Proesmans W., Gewillig M. (1996) Renal and urological tract malformations caused by a 22q11 deletion. *J Med Genet* 33: 349-352.

- Dib C., Fauré S., Fizames C., Samson D., Drouot N., Vignal A., Millasseau P., Marc S., Hazan J., Seboun E., Lathrop M., Gyapay G., Morissette J., Wissenbach J. (1996) A comprehensive genetic map of human genome based on 5,284 microsatellites. *Nature* 380: 152-154.
- DiGeorge A.M. (1968) Congenital absence of the thymus and its immunologic consequences: concurrence with congenital hypoparathyroidism. *Brith Defects Orig Art Ser IV* (1): 116-121.
- Disteche CM, Casanova M, Saal H, Friedman C, Sybert V, Graham J, Thuline H, Page DC, Fellous M (1986) Small deletions of the short arm of the Y chromosome in 46,XY females. *Proc Natl Acad Sci U S A* 83:7841-7844.
- Dobyns W.B., Curry C.J.R., Hoyme H.E., Turlington L., Ledbetter D.H. (1991) Clinical and molecular diagnosis of Miller-Dieker syndrome. *Am J Hum Genet* 48: 584-594.
- Dumanski J.P., Carlborn E., Collins V.P., Nordenskjöld M., Emanuel B.S., Budarf M.L., McDermid H.E. Wolff R., O'Connell P., White R., Lalouel J.-M., Leppert M. (1991) A map of 22 loci on human chromosome 22. *Genomics* 11: 709-719.
- Dutly F., and Schinzel A. (1996) Unequal interchromosomal rearrangements may result in elastin gene deletions causing the Williams-Beuren syndrome. *Hum Mol Genet* 5: 1893-1898.
- Eid J.E., and Sollnerwebb B. (1995) ST-1, a 39-kilodalton protein in *Trypanosoma brucei*, exhibits a dual affinity for the duplex form of the 29-base-pair subtelomeric repeat and its C-rich strand. *Mol Cell Biol* 15: 389-397.
- Ellison J.W., Wardak Z., Young M.F., Gehron Robey P., Laig-Webster M., Chiong W. (1997) PHOG, a candidate gene for involvement in the short stature of Turner syndrome. *Hum Mol Genet* 6: 1341-1347.
- Elsa S.H., Fritz E., Schoener-Scott R., Meyn S., Patel P.I. (1998) Gene for topoisomerase III maps within the Smith-Magenis syndrome critical region: analysis of cell-cycle distribution and radiation sensitivity. *Am J Med Genet* 75: 104-108.
- Emmons S.W., and Sternberg P.W. (1997) Male development and mating behavior. *In* Riddle D.L., Blumenthal T., Meyer B.J., Priess J.R. (eds) *C. elegans II*. Cold Spring Harbor Laboratory Press. pp. 295-361.
- Estabrooks L.L., Rao K.W., Donahue R.P., Aylsworth A.S. (1990) Holoprosencephaly in an infant with a minute deletion of chromosome 21(q22.3). *Am J Med Genet* 36: 306-309.

- Ewart A.K., Morris C.A., Atkinson D., Jin W., Sternes K., Spallone P., Stock A.D., Leppert M., Keating M.T. (1993) Hemizygoty at the ELN locus in a developmental disorder, Williams syndrome. *Nat Genet* 5: 11-16.
- Fajkus J., Kralovics R., Kovarik A., Fajkusova L. (1995) The telomeric sequence is directly attached to the HRS60 subtelomeric tandem repeat in tobacco chromosomes. *FEBS Lett* 364: 33-35.
- Feagin J.E., Abraham J.M., Stuart K. (1993) Extensive editing of the cytochrome c oxidase III transcript in *Trypanosoma brucei*. *Cell* 53: 413-422.
- Feinberg A.P., and Vogelstein B. (1983) A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. *Anal Biochem* 132: 6-13.
- Feng J., Funk W.D., Wang S.-S., Weinrich S.L., Avilion A.A., Chiu C.-P., Adams R.R., Chang E., Allsopp R.C., Yu J.H., Le S.Y., West M.D., Harley C.B., Andrews W.H., Greider C.W., Villeponteau B. (1995) The RNA component of human telomerase. *Science* 269: 1236-1241.
- Fields C. Adams M.D., White O., Venter J.C. (1994) How many genes in human genome? *Nat Genet* 7: 345-346.
- Fire A. (1992) Histochemical techniques for locating *Escherichia coli* β -galactosidase activity in transgenic organisms. *Genet Anal Tech Appl* 9: 152-160.
- Fire A., Xu S., Montgomery M.K., Kostas S.A., Driver S.E., Mello C.C. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391: 806-811.
- Flint J., A.O.M., Buckle V.J., Winter R.M., Holland A.J., McDermid H.E. (1995) The detection of subtelomeric chromosomal rearrangements in idiopathic mental retardation. *Nat Genet* 9: 132-140.
- Flint J., Craddock C.F., Villegas A., Bentley D.P., Williams H.J., Galandello R., Cao A., Wood W.G., Ayyub H., Higgs D.R. (1994) Healing of broken human chromosome by the addition of telomeric repeats. *Am J Hum Genet* 55: 505-512.
- Flint J., Rochette J., Craddock C.F., Dodé C., Vignes B., Horsley S., Kearney L., Buckle V.J., Ayyub H., and Higgs D.R. (1996) Chromosomal stabilisation by a subtelomeric rearrangement involving two closely related Alu elements. *Hum Mol Genet* 5: 1163-1169.
- Flint J., Thomas K., Micklem G., Rayham H., Clark K., Doggett N.A., King A., Higgs D.R. (1997a) The relationship between chromosome structure and function at a human telomeric region. *Nat Genet* 15: 252-257.

- Flint J., Bates G.P., Clark K., Dorman A., Willingham D., Roe B.A., Micklem G., Higgs D.R., Louis E.J. (1997b) Sequence comparison of human and yeast telomeres identifies structurally distinct subtelomeric domains. *Hum Mol Genet* 6: 1305-1314.
- Flomen R.H., Vatcheva R., Gorman P.A., Baptista P.R., Groet J., Barisic I., Ligutic I., Nizeti D. (1998) Construction and analysis of a sequence-ready map in 4q25: Rieger syndrome can be caused by haploinsufficiency of RIEG, but also by chromosome breaks ~90 kb upstream of this gene. *Genomics* 47: 409-413.
- Forney J.D., and Blackburn E.H. (1988) Developmentally controlled telomere addition in wild-type and mutant paramecia. *Mol Cell Biol* 8: 251-258.
- Fransgiskakis J.M., Ewart A.K., Morris C.A., Mervis C.B., Bertrand J., Robinson B.F., Klein B.P., Ensing G.J., Everett L.A., Green E.D., Proschel C., Gutowski N.J., Noble M., Atkinson D.L., Odelberg S.J., Keating M.T. (1996) LIM-kinase 1 hemizyosity implicated in impaired visuospatial constructive cognition. *Cell* 86: 59-69.
- Frints S.G., Schoenmakers E.F., Smeets E., Petit P., Fryns J.P. (1998) De novo 7q36 deletion: breakpoint analysis and types of holoprosencephaly. *Am J Med Genet* 75: 153-158.
- Fritz E., Elsea S.H., Patel P.I., Meyn M.S. (1997) Overexpression of a truncated human topoisomerase III partially corrects multiple aspects of the ataxia-telangiectasia phenotype. *Proc Natl Acad Sci USA* 94: 4538-4542.
- Funke B., Saint-Jore B., Puech A., Sirotkin H., Edelmann L., Carlson C., Raft S., Pandita R.K., Kucherlaqpati R., Skoultchi A., Morrow B.E. (1997) Characterization and mutation analysis of goosecoid-like (GSCL), a homeodomain-containing gene that maps to the critical region for VCFS/DGS on 22q11. *Genomics* 46: 364-372.
- Fusco JC, Fenwick RG Jr, Ledbetter DH, Caskey CT (1983) Deletion and amplification of the HGPRT locus in Chinese hamster cells. *Mol Cell Biol* 3:1086-1096.
- Galili N., Epstein J.A., Leconte I., Nayak S., Buck C.A. (1998) Gsc1, a gene within the minimal DiGeorge critical region, is expressed in primordial germ cells and the developing pons. *Dev Dyn* 212: 86-93.
- Gay C.T., Hardies L.J., Rauch R.A., Lancaster J.L., Plaetke R., DuPont B.R., Cody J.D., Cornell J.E., Herndon R.C., Ghidoni P.D., Schiff J.M., Kaye C.I., Leach R.J., Fox P.T. (1997) Magnetic resonance imaging demonstrates incomplete myelination in 18q-syndrome: evidence for myelin basic protein haploinsufficiency. *Am J. Med Genet* 74: 422-437.
- Glenn C. C., Nicholls R. D., Robinson W. P., Saitoh S., Niikawa N., Schinzel A., Horsthemke B., Driscoll, D. J. (1993) Modification of the DNA methylation imprint in unique Angelman and Prader-Willi patients. *Hum. Molec. Genet.* 2: 1377-1382.

- Goldberg R., Motzkin B., Marion R., Scambler P.J., Shprintzen R.J. (1993) Velo-cardio-facial syndrome: a review of 120 patients. *Am J Med Genet* 45: 313-319.
- Gong B.T., Norwood T.H., Hoehn H., McPherson E., Hall J.G., Hickman R. (1976) Chromosome 7 short arm deletion and craniosynostosis. A 7p-syndrome. *Hum Genet* 35: 117-123.
- Gong W., Emanuel B.S., Galili N., Kim D.H., Roe B., Driscoll D.A., Budarf M.L. (1997) Structural and mutational analysis of a conserved gene (DGS1) from the minimal DiGeorge syndrome critical region. *Hum Mol Genet* 6: 267-276.
- Goodship J., Curtis A., Cross I., Brown J., Emslie J., Wolstenholme J., Bhattacharya S., Burn J. (1992) A submicroscopic translocation, t(4;10) responsible for recurrent Wolf-Hirschhorn syndrome identified by allele loss and fluorescent *in situ* hybridization. *J Med Genet* 29: 451-454.
- Gorina S., and Pavletich N.P. (1996) Structure of the p53 tumor suppressor bound to the ankyrin and SH3 domains of 53BP2. *Science* 274: 1001-1005.
- Greenberg F. (1993) DiGeorge syndrome: an historical review of clinical and cytogenetic features. *J Med Genet* 43: 605-611.
- Greenberg F., Lewis R.A., Potocki L., Glaze D., Parke J., Killian J., Murphy M.A., Williamson D., Brown F., Dutton R., McCluggage C., Friedman E., Sulek M., Lupskji J.R. (1996) Multidisciplinary clinical study of Smith-Magenis syndrome (deletion 17p11.2) *Am J Hum Genet* 49: 1207-1218.
- Greenwald I. (1998) LIN-12/Notch signalling: lessons from worms and flies. *Gene Dev* 12: 1751-1762.
- Gueneri S., Bettio D., Simoni G., Brambati B., Lanzani A., Fraccaro M. (1987) Prevalence and distribution of chromosome abnormalities in a sample of first trimester internal abortions. *Hum Reprod* 2: 735-739.
- Gustincich S., Manfioletti G., Del Sal G., Schneider C. (1991) A fast method for higher-quality genomic DNA extraction from whole human blood. *BioTechniques* 11: 298-301.
- Hanish J.P., Yanowitz J.L., De Lange T (1994) Stringent sequence requirements for the formation of human telomeres. *Proc Natl Acad Sci USA* 91: 8861-8865.
- Hansen R., and Oren M (1997) p53; from inductive signal to cellular effect. *Curr Opin Genet Devel* 7: 46-51.

- Harrington L.A., and Greider C.W. (1991) Telomerase primer specificity and chromosome healing. *Nature* 353: 451-454.
- Heitzler P., and Simpson P. (1991) The choice of cell fate in the epidermis of *Drosophila*. *Cell* 64: 1083-1092.
- Hel Z., Di Marco S., Radzioch D. (1998) Characterization of the RNA binding proteins forming complexes with a novel putative regulatory region in the 3'-UTR of TNF- α mRNA. *Nucleic Acids Res* 26: 2803-2812.
- Hengartner M.O., and Horvitz H.R. (1994) *C. elegans* cell survival gene *ced-9* encodes a functional homolog of the mammalian proto-oncogene *bcl-2*. *Cell* 76: 665-676.
- Henikoff S., and Henikoff J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 89: 10915-10919.
- Hennekam R.C.M., Stevens C.A., Van de Kamp J.J.P. (1990) Etiology and recurrence risk in Rubinstein-Taybi syndrome. *Am J Med Genet Suppl* 6: 56-64.
- Hill R.J., and Sternberg P.W. (1992) The gene *lin-3* encodes an inductive signal for vulval development in *C. elegans*. *Nature* 358: 470-476.
- Hodges P.E., Navaratnam N., Greeve J.C., Scott J. (1991) Site-specific creation of uridine from cytidine in apolipoprotein B mRNA editing. *Nucleic Acids Res* 19: 1197-1201.
- Holger S., Kim U.-J., Slepak T., Blin N., Simon M.L., Shizuya H. (1996) Framework for a physical map of the human 22q13 region using bacterial artificial chromosome (BACs) *Genomics* 33: 9-20.
- Holmes S.E., Riazhi M.A., Gong W., McDermid H.E., Sellinger B.T., Hua A., Chen F., Wang Z., Zhang G., Roe B., Gonzalez I., McDonald-McGinn D.M., Zackai E., Emanuel B.S., Budarf M.L. (1997) Disruption of the clathrin heavy chain-like gene (CLTCL) associated with features of DGS/VCFS: a balanced (21;22)(p12;q11) translocation. *Hum Mol Genet* 6: 357-367.
- Hu Y., Benedict M.A., Wu D., Inohara N., Núñez G. (1998) Bcl-X_L interacts with Apaf-1 and inhibits Apaf-1-dependent caspase-9 activation. *Proc Natl Acad Sci USA* 95: 4386-4391.
- Hudson T.J., Engelstein M., Lee M.K., Ho E.C., Rubenfield M.J., Adams C.P., Housman D.E., Dracopoli N.C. (1992) Isolations and chromosomal assignment of 100 highly informative human simple sequence repeat polymorphisms. *Genomics* 13: 622-629.

- Hughes D.C., Legan P.K., Steel K.P., Richardson G.P. (1998) Mapping of the alpha-tectorin gene (TECTA) to mouse chromosome 9 and human chromosome 11: a candidate for human autosomal dominant nonsyndromic deafness. *Genomics* 48: 46-51.
- Ijdo J.W., Baldini A., Ward D.C., Reeders S.T., Wells R.A. (1991) Origin of human chromosome 2: an ancestral telomere-telomere fusion. *Proc Natl Acad Sci USA* 88: 9051-9055.
- Ito Y., Kwon O.H., Ueda M., Tanaka A., Imanishi Y. (1997) Bacterial activity of human lysozymes carrying various lengths of polyproline chain at the C-terminus. *FEBS Lett* 415:285-288.
- Iwabuchi K., Bartel P.L., Li B., Marraccino R., Fields S. (1994) Two cellular proteins that binds to wild-type but not mutant p53. *Proc Natl Acad Sci USA* 91: 6098-6102.
- Jacobs P.A., Browne C., Gregson N., Joyce C., White H. (1992) Estimates of the frequency of chromosome abnormalities detectable in unselected newborns using moderate levels of banding. *J Med Genet* 29: 103-107.
- Jensen G., Hazendonk E., Thijssen K.L., Plasterk R.H.A. (1997) Reverse genetics by chemical mutagenesis in *Caenorhabditis elegans*. *Nat Genet* 17: 119-121.
- Jones C., Penny L., Mattina T., Yu S., Baker E., Voullaire L., Langdon W.Y., Sutherland G.R., Richards R.I., Tunnacliffe A. (1995) Association of a chromosome deletion syndrome with a fragile site within the proto-oncogene CBL2. *Nature* 376: 145-149.
- Jones C., Slijepcevic P., Marsh S., Baker E., Langdon W.Y., Richards R.I., Tunnacliffe A. (1994) Physical linkage of the fragile site FRA11B and a Jacobsen syndrome chromosome deletion breakpoint in 11q23.3. *Hum Mol Genet* 3: 2123-2130.
- Jong M., Carey A., Stewart C., Rinchik E., Glenn C., Driscoll D., Nicholls R. (1993) The ZNF127 gene encodes a novel C3H4 zinc-finger protein and its expression is regulated by genomic imprinting. *Am J Hum Genet* 53: A697.
- Jorgenson R.J., Levin L.S., Cross H.E., Yoder F., Kelly T.E. (1978) The Rieger syndrome. *Am J Med Genet* 2:307-318.
- Jupe E.R., Liu X.T., Kiehlbauch J.L., McClung J.K., Dell'Orco R.T. (1996) Prohibitin in breast cancer cell lines: loss of antiproliferative activity is linked to 3' untranslated region mutations. *Cell Growth Differ* 7: 871-878.
- Kalb J.M., Lau K.K., Goszczynski B., Fukushige T., Moons D., Okkema P.G., McGhee J.D. (1998) *pha-4* is *Ce-fkh-1*, a fork head/HNF-3 α,β,γ homolog that functions in organogenesis of the *C. elegans* pharynx. *Development* 125: 2171-2180.

- Kamholz J., Spielman R., Gogolin K., Modi W., O' Brien S., Lazzarini R. (1987) The human myelin-basic protein gene: Chromosomal localization and RFLP analysis. *Am J Hum Genet* 40: 365-373.
- Karayorgou M., Morris M.A., Morrow B., Shprintzen R.J., Goldberg R., Borrow J., Gos A. Nestadt G., Wolynec P.S., Lasseter V.K., Eisen H., Childs B., Kazazian H.H., Kucherlapati R., Antonarakis S.E., Pulver A.E., Housman D.E. (1995) Schizophrenia susceptibility associated with interstitial deletions of chromosome 22q11. *Proc Nat Acad Sci USA* 92: 7612-7616.
- Karim F.D., Urness L.D., Thummel C.S., Klemsz M.J., McKercher S.R., Celada A., Van Beveren C., Maki R.A., Gunther C.V., Nye J.A., Graves B.J. (1990) The ETS-domain: a new DNA-binding motif that recognizes a purine-rich core DNA sequence. *Genes Dev* 4: 1451-1453.
- Kennedy W.P., Kaminski J.M., Van der Ven H.H., Jeyendran R.S., Reid D.S., Blackwell J., Bielfeld P., Zaneveid L.J.D. (1989) A simple, clinical assay to evaluate the acrosin activity of human spermatozoa. *J Androl* 10: 221-231.
- Keppler-Noreuil K.M., Carroll A.J., Finley W.H., Rutledge S.L. (1995) Chromosome 1p terminal deletion: report of new findings and conformation of two characteristic phenotypes. *J Med Genet* 32: 619-622.
- Kern J.A., Warnock L.J., McCafferty J.D. (1997) The 3' untranslated region of IL-1beta regulates protein production. *J Immunol* 158:1187-1193.
- Kilian A., Bowtell D.D.L., Abud H.E., Hime G.R., Venter D.J., Keese P.K., Duncan E.L., Reddel R.R., Jefferson R.A. (1997) Isolation of a candidate human telomerase catalytic subunit gene, which reveals complex splicing patterns in different cell types. *Hum Mol Genet* 6: 2011-2019.
- Kipling D., and Cook H.J. (1990) Hypervariable ultra-long telomeres in mice. *Nature* 347: 400-402.
- Kleckowska A., Fryns J.P., van den Berghe H. (1993) A distinct multiple congenital anomalies syndrome associated with distal 5q deletion (q35.1qter). *Ann Genet* 36: 126-128.
- Klemm U. Müller-Esterl W. Engel W. (1991) Acrosin, the peculiar sperm -specific serine protease. *Hum Genet* 87: 635-641.
- Kollmus H., Flohe L., McCarthy J.E. (1996) Analysis of eukaryotic mRNA structures directing cotranslational incorporation of selenocysteine. *Nucleic Acids Res* 24: 1195-1201.

- Korenberg J.R., Bradley C., Disteche C.M. (1992) Down syndrome: molecular mapping of the congenital heart disease and duodenal stenosis. *Am J Hum Genet* 50: 294-302.
- Kotewicz M.L., Sampson C.M., D'Alessio J.M., Gerard G.F. (1988) Isolation of cloned Moloney murine leukemia virus reverse transcriptase lacking ribonuclease H activity. *Nucleic Acids Res* 16: 265-277.
- Koukoulis G.N., Vantman D., Dennison L., Banks S.M., Sherins R.J. (1989) Low acrosin activity in a subgroup of men with idiopathic infertility does not correlate with sperm density, percent motility, curvilinear velocity, or linearity. *Fertil Steril* 52: 120-127.
- Kuwano A., Ledbetter S.A., Dobyns W.B., Emanuel B.S., Ledbetter D.H. (1991) Detection of deletions and cryptic translocations in Miller-Dieker syndrome by in situ hybridization. *Am J Hum Genet* 49: 707-714.
- Labuda D., and Striker G. (1989) Sequence conservation in Alu evolution. *Nucleic Acid Res* 17: 2477-2491.
- Lamb J., Harris P.C., A.O.M., Wood W.G., Dauwerse J.G., Higgs D.R. (1993) De novo truncation of chromosome 16p and healing with (TTAGGG)_n in the α -thalassemia/mental retardation syndrome (ATR-16). *Am J Hum Genet* 52: 668-676.
- Lambert S., and Bennett V. (1993) From anemia to cerebellar dysfunction. *Eur J Biochem* 211: 1-6.
- Lammer E.J., and Opitz J.M. (1986) The DiGeorge anomaly as a developmental field defect. *Am J Med Genet Suppl* 2: 113-127.
- Langer L.O., Krassikoff N., Laxova R., Scheer-Williams M., Lutter L.D., Corlin R.J., Jennings C.G., Day D.W. (1984) The Tricho-Rhino-Phalangeal syndrome with exotoses (or Langer-Giedion Syndrome): Four additional patients without mental retardation and review of the literature. *Am J Med Genet* 19: 81-111.
- Laurie D.A., and Hulten M.A. (1985) Further studies on chiasma distribution and interference in human male. *Ann Hum Genet* 49: 203-214.
- Lazar T., Gotte M., and Gallwitz D. (1997) Vesicular transport: how many Ypt/Rab-GTPases make a eukaryotic cell? *TIBS* 22: 468-472.
- Ledbetter D.H. (1992) Cryptic translocation and telomere integrity. *Am J Hum Genet* 51: 451-456.
- Ledbetter D.H., and Cavenee W.K. (1989) Molecular cytogenetics: Interface of cytogenetics and monogenic disorders. In Scriver C.J., Beaudet A.L., Sly W., and Valle D. (eds) *The metabolic basis of inherited disease*, 6th ed. McGraw-Hill, New York, pp.343-371.

- Lee B., Thirunavukkarasu K., Zhou L., Pastore L., Baldini A., Hecht J., Geoffroy V., Ducey P., Karsenty G. (1997) Missense mutations abolishing DNA binding of the osteoblast-specific transcription factor OSF2/CBFA1 in cleidocranial dysplasia. *Nat Genet* 16: 307-310.
- Li L., Krantz I.D., Yu D., Genin A., Banta A.B., Collins C.C., Ming Q. (1997) Alagille syndrome is caused by mutations in human Jagged 1, which encodes a ligand for Notch 1. *Nat Genet* 16: 243-250.
- Ligutic I., Brecevic L., Petkovic I, Kalogjera T., Rajic Z. (1981) Interstitial deletion 4q and Rieger syndrome. *Clin Genet* 20: 323-327.
- Lindberg R.A., and Hunter T. (1990) cDNA cloning and characterization of eck, an epithelial cell receptor protein-tyrosine kinase in the eph/elk family of protein kinases. *Mol Cell Biol* 10: 6316-6324.
- Lindsay E.A., Harvey E.L., Scambler P.J., Baldini A. (1998) ES2, a gene deleted in DiGeorge syndrome, encodes a nuclear protein and is expressed during early mouse development, where it shares an expression domain with a Goosecoid-like gene. *Hum Mol Genet* 7: 629-635.
- Linger J., Cooper J.P., Cech T.R. (1995) Telomerase and DNA end replication: no longer a lagging strand problem? *Science* 269: 1533-1534.
- Liu D.Y., and Baker H.W.G. (1993) Inhibition of acrosin activity with a trypsin inhibitor blocks human sperm penetration of the zona pellucida. *Biol Reprod* 48: 340-348.
- Lock R.B., Ross W.E. (1990) Inhibition of p34^{cdc2} kinase activity by etoposide or irradiation as a mechanism of G₂ arrest in Chinese hamster ovary cells. *Cancer Res* 50: 3761-3766.
- Louis E.J., Naumova E.S., Lee A., Naumov G., Haber J.E. (1994) The chromosome end in yeast: its mosaic nature and influence on recombinational dynamics. *Genetics* 136: 789-802.
- Luderus M.E., van Steensel B., Chong L., Sibon O.C., Cremers F.F., de Lange T. (1996) Structure, subnuclear distribution, and nuclear matrix association of the mammalian telomeric complex. *J Cell Biol* 135: 867-881.
- MacDonald H.R. and Wevrick R. (1997) The necdin gene is deleted in Prader-Willi syndrome and is imprinted in human and mouse. *Hum Mol Genet* 6: 1873-1878.
- Macdonald P.M., Kerr K. (1997) Redundant RNA recognition events in bicoid mRNA localization. *RNA* 3: 1413-1420.

- Mahon P., Partridge K., Beattie J.H., Glover L.A., Hesketh J.E. (1997) The 3' untranslated region plays a role in the targeting of metallothionein-I mRNA to the perinuclear cytoplasm and cytoskeletal-bound polysomes. *Biochim Biophys Acta* 1358: 153-162.
- Mahoney N.M., Janney P.A., Almo S.C. (1997) Structure of the profilin-poly-L-proline complex involved in morphogenesis and cytoskeletal regulation. *Nat Struct Biol* 953-960.
- Makarov V.L., Hirose Y., Langmore J.P. (1997) Long G tails at both ends of human chromosomes suggest a C strand degradation mechanism for telomere shortening. *Cell* 88: 657-666.
- Maquat L.E. (1995) When cells stop making sense: effects of nonsense codons on RNA metabolism vertebrate cells. *RNA* 1: 453-465.
- Maraschio P., Tupler R., Barbierato L., Dainotti E., Larizza D., Bernardi F., Hoeller H., Garau A., Tiepolo L. (1996) An analysis of Xq deletions. *Hum Genet* 97: 375-381.
- Mayer B.J., Hamaguchi M., Hanafusa H. (1988) A novel viral oncogene with structural similarity to phospholipase C. *Nature* 332: 272-275.
- McClintock B. (1938) The fusion of broken ends of sister half chromatids following chromatid breakage at meiotic anaphase. *Mo Agric Exp Station Res Bull* 290: 1-48.
- McDermid H.E., Duncan A.M.V., Brasch C.R., Holden J.J.A., Magenis E., Sheely R., Burn J., Kardon N., Noel B., Schinzel A., Teshima I., White B.N. (1986) Characterization of the supernumerary chromosome in cat eye syndrome. *Science* 232: 6464-648.
- McFadden D.E., and Friedman J.M. (1997) Chromosome abnormalities in human beings. *Mutation Res* 396: 129-140.
- Mears A.J. (1995) Molecular characterization of cat eye syndrome. Ph. D Thesis. University of Alberta, Edmonton, Alberta.
- Mears A.J., Duncan A.M.V., Budarf M.L., Emanuel B.S., Sellinger B., Siegel-Bartelt J., Greenberg C.R., McDermid H.E. (1994) Molecular characterization of the marker chromosome associated with cat eye syndrome. *Am J Hum Genet* 55: 134-142.
- Meng J., Fujia H., Nagahara N., Kashiwai A., Yoshioka Y., Funato M. (1992) Two patients with chromosome 6q terminal deletions with breakpoints at q24.3 and q25.3 *Am J Med Genet* 43: 747-750.

- Menko F.H., Madan K., Baart J.A., Beukenhorst H.L. (1992) Robin sequence and a deficiency of the left forearm in a girl with a deletion of chromosome 4q33-qter. *Am J Med Genet* 44: 696-698.
- Meyerson M., Counter C.M., Eaton E.N., Ellisen L.W., Steiner P., Caddle S.D., Ziaugra L., Beijersbergen R.L., Davidoff M.J., Liu Q., Bacchetti S., Haber D.A., Weinberg R.A. (1997) hEST2, the putative human telomerase catalytic subunit gene, is up-regulated in tumor cells and during immortalization. *Cell* 90: 785-795.
- Michaelis R.C., Velagaleti G.V., Jones C., Pivnick E.K., Phelan M.C., Boyd E., Tarleton J., Wilroy R.S., Tunnacliffe A., Tharapel A.T. (1998) Most Jacobsen syndrome deletion breakpoints occur distal to FRA11B. *Am J Med Genet* 76: 222-228.
- Miller J.F., Williamson E., Glue J., Gordon Y.B., Grudzinkas J.F., Sykes A. (1980) Fetal loss after implantation: a prospective study. *Lancet* 33: 107-116.
- Miyamoto S., Chiorini J.A., Urcelay E., Safer B. (1996) Regulation of gene expression for translation initiation factor eIF-2 alpha: importance of the 3' untranslated region. *Biochem J* 315 (Pt 3): 791-798.
- Mohsenian M., Syner F.N., Moghissi K.S. (1982) A study of sperm acrosin in patients with unexplained infertility. *Fertil Steril* 37: 223-229.
- Morgan G.T. (1995) Identification in the human genome of mobile elements spread by DNA-mediated transposition. *J Mol Biol* 254: 1-5.
- Morin G.B. (1991) Recognition of a chromosome truncation site associated with α -thalassaemia by human telomerase. *Nature* 353: 454-456.
- Moyzis R.K., Buckingham J.M., Cram L.S., Dani M., Deaven L.L., Jones M.D., Meyne J., Ratliff R.L., Wu J.-R. (1988) A highly conserved repetitive DNA sequence, (TTAGGG)_n, present at the telomere of human chromosomes *Proc Natl Acad Sci* 85: 6622-6626.
- Murayama K., Greenwood R.S., Rao K.W., Aylsworth A.S. (1991) Neurological aspects of del(1q) syndrome. *Am J Med Genet* 40:488-492.
- Murnane J.P and Yu L.C. (1993) Acquisition of telomere repeat sequences by transfected DNA integrated at the site of a chromosome break. *Mol Cell Biol* 13: 977-983.
- Myer T.W., and Gelfand D.H. (1991) Reverse transcription and DNA amplification by *Thermus thermophilus* DNA polymerase. *Biochemistry* 30: 7661-7666.
- Nakayama J.-I., Saito M., Nakamura H., Matsuura A., Ishikawa F. (1997) TLP1: A gene encoding a protein component of mammalian telomerase is a novel member of WD repeats family. *Cell* 88: 1-20.

- Naritomi K., Izumikawa Y., Nagataki S., Fukushima Y., Wakui K., Niikawa N., Hirayama K. (1992) Combined Goltz and Aicardi syndromes in a terminal Xp deletion: are they a contiguous gene syndrome? *Am J Med Genet* 43: 839-843.
- National Institutes of Health and Institute of Molecular Medicine Collaboration (1996) A complete set of human telomeric probes and their clinical application. *Nat Genet* 14: 86-89.
- Nesslinger N.J., Gorski J.L., Kurczynski T.W., Shapira S.K., Siegel-Bartelt J., Dumanski J.P., Cullen R.F., French B.N., McDermid H.E. (1994) Clinical, cytogenetic, and molecular characterization of seven patients with deletions of chromosome 22q13.3. *Am J Hum Genet* 54: 464-472.
- Nickel R.E., Pillers D.A.E., Mahenis E.R., Driscoll D.A., Emanuel B.S., Zonana J. (1994) Velo-cardio-facial syndrome and DiGeorge sequence with meningomyelocele and deletions of the 22q11 region. *Am J Med Genet* 52: 445-449.
- Nickerson E., Greenberg F., Keating M.T., McCaskill C., Shaffer L.G. (1995) Deletions of the elastin gene at 7q11.23 occur in ~90% of patients with Williams syndrome. *Am J Hum Genet* 56: 1156-1161.
- Nimmo E.R., Pidoux A.L., Perry P.E., Allshire R.C. (1998) Defective meiosis in telomere-silencing mutants of *Schizosaccharomyces pombe*. *Nature* 392: 825-828.
- Ning Y., Rosenberg M., Biesecker L.G., Ledbetter D.H. (1996) Isolation of the human chromosome 22q telomere and its application to detection of cryptic chromosomal abnormalities. *Hum Genet* 97: 765-769.
- Nobukuni Y., Watanabe A., Takeda K., Skarka H., Tachibana M. (1996) Analysis of loss-of-function mutations of the MTF gene suggest that haploinsufficiency is a cause of Waardenburg syndrome type 2A. *Am J Hum Genet* 59: 76-83.
- Novak R. W., and Robinson H. B. (1994) Coincident DiGeorge anomaly and renal agenesis and its relation to maternal diabetes. *Am J Med Genet* 50: 311-312
- Novick P. and Zerial M. (1997) The diversity of Rab proteins in vesicle transport. *Curr. Opin. Cell. Biol.* 9: 496-504.
- Nowell P.C., and Hungerford D.A. (1960) A minute chromosome in human chronic granulocytic leukemia. *Science* 132: 1497-1499.
- Olkkonen VM and Stenmark H (1997) Role of Rab GTPases in membrane traffic. *Inter. rev. Cytology* 176: 1-85.

- Olson T.M., Michels V.V., Urban Z., Csiszar K., Christiano A.M., Driscoll D.J., Feldt R.H., Boyd C.D., Thibodeau S.N. (1995) A 30 kb deletion within the elastin gene results in familial supravalvular aortic stenosis. *Hum Mol Genet* 4: 1677-1679.
- Ono J., Harada K., Hasegawa T., Sakurai K., Kodaka R., Tanabe Y., Tanaka J., Igarashi T., Nagai T., Okada S. (1994) Central nervous system abnormalities in chromosome deletion at 11q23. *Clin Genet* 45: 325-329.
- Orlicky D.J., and Nordeen S.K. (1996) Cloning, sequencing and proposed structure for a prostaglandin F_{2α} receptor regulatory protein. *Prostaglandins Leukot Essent Fatty Acids* 55: 261-268.
- Ortigas A.P., Stein C.K., Thomson L.L., Hoo J.J. (1997) Delineation of 14q32.3 deletion syndrome. *J Med Genet* 34: 515-517.
- Osborne L.R., Martindale D., Scherer S.W., Shi S.M., Huizenga J., Heng H.H.Q., Costa T., Pober B., Lew L., Brinkman J., Rommens J., Koop B., 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.
- Osborne L.R., Soder S., Shi X.M., Pober B., Costa T., Scherer S.W., Tsui L.C. (1997) Hemizygous deletion of the Syntaxin 1A gene in individuals with Williams syndrome. *Am J Hum Genet* 61: 449-452.
- Overhauser J., Bengtsson U., McMahon J., Ulm J., Butler M.G., Santiago L., Wasmuth J.J. (1989) Prenatal diagnosis and carrier detection of a cryptic translocation by using DNA markers from the short arm of chromosome 5. *Am J Hum Genet* 45: 296-303.
- Overhauser J., Huang X., Gersh M., Wilson W., McMahon J., Bengtsson U., Rojas K., Meyer M., Wasmuth J.J. (1994) Molecular and phenotypic mapping of the short arm of chromosome 5: sublocalization of the critical region for the cri-du-chat syndrome. *Hum Mol Genet* 3: 247-252.
- Pääbo S., Irwin D.M., Wilson A.C. (1990) DNA damage promotes jumping between templates during enzymatic amplification. *J Biol Chem* 265: 4718-4721.
- Pace T., Ponzi M., Scotti R., Frontali C. (1995) Structure and superstructure of *Plasmodium falciparum* subtelomeric regions. *Mole Biochem Parasitol* 69: 257-268.
- Palladino F., Larcoche T., Gilson E., Axelrod A., Pillus L., Gasser S.M. (1993) SIR3 and SIR4 proteins are required for the positioning and integrity of yeast telomeres. *Cell* 75: 543-555.
- Pan G., O'Rourke K., Dixit V.M. (1998) Caspase-9, Bcl-X_L, and Apaf-1 form a ternary complex. *J Biol Chem* 273: 5841-5845.

- Pankratz M.J., Hoch M., Seifert E., Jäckle H. (1989) *Krüppel*- requirement for *knirps* enhancement reflects overlapping gap gene activities in the *Drosophila* embryo. *Nature* 341: 337-340.
- Pauls D.L., Leckman J.F. (1986) The inheritance of Gilles de la Tourette syndrome and associated behaviors: evidence for autosomal dominant transmission. *N Eng J Med* 315: 993-997.
- Pawson T. (1995) Protein modules and signalling networks. *Nature* 373: 573-580.
- Pedersen B., and Kerndrup G. (1986) Specific minor chromosome deletions consistently occurring in myelodysplastic syndromes. *Cancer Genet Cytogenet* 23:61-75.
- Petersen B., Strassburg H.M., Feichtinger W., Kress W., Schmid M. (1998) Terminal deletion of the long arm of chromosome 10: a new case with breakpoint in q25.3. *Am J Med Genet* 77: 60-62.
- Petrella E.C., Machesky L.M., Kaiser D.A., Pollard T.D. (1996) Structural requirements and thermodynamics of the interaction of proline peptides with profilin. *Biochemistry* 35: 16535-16543.
- Petrij F., Giles R.H., Dauwerse H.G., Saris J.J., Hennekam R.C.M., Masuno M., Tommerup N., van Ommen G.-J.B., Goodman R.H., Peters D.J.M., Breuning M.H. (1995) Rubinstein-Taybi syndrome caused by mutations in the transcriptional co-activator CBP. *Nature* 376: 348-351.
- Phipps M.E., Latif F., Prowse A., Payne S.J., Dietz-Band J., Leversha M., Affara N.A., Moore A.T., Tomie J., Schinzel A. (1994) Molecular analysis of the 3p- syndrome. *Hum Mol Genet* 3: 903-908.
- Pingault V., Bondurand N., Kuhlbrodt K., Goerich D.E., Préhu M.-O., Puliti A., Herbarth B., Hermans-Borgmeyer I., Legius E., Mathijs G., Amiel J., Lyonnet S., Ceccherini I., Romeo G., Clayton Smith J., Read A.P., Wegner M., and Goossens M. (1998) SOX10 mutations in patients with Waardenburg-Hirschsprung disease. *Nat Genet* 18: 171-173.
- Plaja A., Vidal R., Soriano D., Bou X., Vendrell T., Mediano C., Pueyo J.M., Labrana X., Sarret E. (1994) Terminal deletion of 6p: report of a case. *Ann Genet* 37: 196-199.
- Podruch P.E., Yen F.S., Dinno N.D., Weisskopf B. (1982) Yq- in a child with livedo reticularis, snub nose, microcephaly, and profound mental retardation. *J Med Genet* 19: 377-380.
- Pries J., and Hirsh D. (1986) *Caenorhabditis elegans* morphogenesis: the role of the cytoskeleton in the elongation of the embryo. *Dev Biol* 117: 156-173.

- Pröschel C., Blouin M.-J., Gutowski N.J., Ludwig R., Noble M. (1995) Limk1 is predominantly expressed in neural tissues and phosphorylates serine, threonine and tyrosine residues *in vitro*. *Oncogene* 11: 1271-1281.
- Pryde F.E., and Louis E.J. (1997) *Saccharomyces cerevisiae* telomeres. A Review. *Biochemistry (Mos)* 62: 1232-1241.
- Régnier V., Meddeb M., Lecointre G., Richard F., Duverger A., Nguyen V.C., Dutrillaux B., Bernheim A., Danglot G. (1997) Emergence and scattering of multiple neurofibromatosis (NF1)-related sequences during hominoid evolution suggest a process of pericentromeric interchromosomal transposition. *Hum Mol Genet* 6: 9-16.
- Reiter L.T., Murakami T., Koeuth T., Pentao L., Muzny D.M., Gibbs R.A., Lupski J.R. (1996) A recombination hotspot responsible for two inherited peripheral neuropathies is located near a *mariner* transposon-like element. *Nat Genet* 12: 288-297.
- Renauld H., Aparicio O.M., Zierath P.D., Billington B.L., Chhablani S.K., Gottschling D.E. (1993) Silent domains are assembled continuously from the telomere and are defined by promoter distance and strength, and by SIR3 dosage. *Genes Dev* 7: 1133-1145.
- Renauld H., Aparicio O.M., Zierath P.D., Billington B.L., Chhablani S.K., Gottschling D.E. (1993) Silent domains are assembled continuously from the telomere and are defined by promoter distance and strength, and by SIR3 dosage. *Genes Dev* 7: 1133-1145.
- Riddle D.L., Blumenthal T., Meyer B.J., Priess J.R. (1997) *C. ELEGANS II*. Cold Spring Harbor Laboratory Press
- Rieger H. (1935) Beitrage zur Kenntnis seltener Missbildungen der Iris: ueber Hypoplasie des Irisvorderblattes mit Verlagerung und Entrundung der Pupille. *Albercht von Graefes Arch Klin Exp Ophthal.* 133: 602-635.
- Robertson H.M. (1993) The *mariner* transposable element is widespread in insects. *Nature* 362: 241-245.
- Robinson M.S. (1994) The role of clathrin, adapters and dynamin in endocytosis. *Curr Opin Cell Biol* 4: 538-544.
- Robinson W.P., Waslynska J., Bernasconi M., Wang S., Clark D., Kotzot D., Schinzel A. (1996) Delineation of 7q11.2 deletions associated with Williams-Beuren syndrome and mapping of a repetitive sequence to within and to either side of the common deletion. *Genomics* 34: 17-23.
- Rougeulle C., Glatt H., Lalande M. (1997) The Angelman syndrome candidate gene, UBE3A/E6-AP, is imprinted in brain. *Nat Genet* 17: 14-15.

- Rouquier S., Taviaux S., Trask B.J., Brand-Arpon V., van den Engh G., Demaille J., Giorgi D. (1998) Distribution of olfactory receptor genes in the human genome. *Nat Genet* 18: 243-250.
- Rowley J.D. (1973) A new consistent chromosomal abnormality in chronic myelogenous leukemia identified by quinacrine fluorescence and Giemsa staining. *Nature* 243: 290-293.
- Rubinstein J.H., and Taybi H. (1963) Broad thumbs and toes and facial abnormalities. *Am J Dis Child* 105: 588-608.
- Rueter S.M., Burns C.M., Coode S.A., Mookherjee P., Emeson R.B. (1995) Glutamate receptor RNA editing in vitro by enzymatic conversion of adenosine to inosine. *Science* 267: 1491-1494.
- Saccone S., Cacciò S., Kusuda J., Andreozzi L., Bernardi G. (1996) Identification of the gene-richest bands in human chromosomes. *Gene* 174: 85-94.
- Saccone S., De Sario A., Della Valle G., Bernardi G. (1992) The highest gene concentrations in the human genome are in telomeric bands of metaphase chromosomes. *Proc Natl Acad Sci USA* 89: 4913-4917.
- Sakane F., Imai S.-I., Kai M., Wada I., Kanoh H. (1996) Molecular cloning of a novel diacylglycerol kinase isozyme with a pleckstrin homology domain and a C-terminal tail similar to those of the EPH family of protein-tyrosine kinases. *271*: 8394-8401.
- Sakoyama Y., Hong K.-J., Byun S.M., Hisajima H., Ueda S., Yaoita Y., Hayashida H., Miyata T., Honjo T. (1987) Nucleotide sequences of immunoglobulin ϵ genes of chimpanzee and orangutan: DNA molecular clock and hominoid evolution. *Proc Natl Acad Sci USA* 84: 1080-1084.
- Sambrook J., Fritsch E.F., Maniatis T. (1989) "Molecular Cloning: A Laboratory Manual". 2nd ed. Cold Spring Harbour Laboratory Press.
- Sandell L.L., Zakian V.A. (1993) Loss of a yeast telomere: arrest, recovery, and chromosome loss. *Cell* 75: 729-739.
- Sauer F., and Jäckle H. (1993) Dimerization and the control of transcription by *Krüppel*-. *Nature* 364: 454-457.
- Savitsky K., Platzer M., Uziel T., Gilad S., Sartiel A., Rosenthal A., Elroy-Stein O., Shiloh Y., Rotman G. (1997) Ataxia-telangiectasia: structural diversity of untranslated sequences suggests complex post-transcriptional regulation of ATM gene expression. *Nucleic Acids Res* 25: 1678-1684.

- Scherthan H., Weich S., Schwegler H., Heyting C., Härle M., Cremer T. (1996) Centromere and telomere movements during early meiotic prophase of mouse and man are associated with the onset of chromosome pairing. *J Cell Biol* 134: 1109-1125.
- Schmickel R.F. (1986) Contiguous gene syndrome: A component of recognizable syndromes. *J Pediatr* 109: 231-241.
- Schuuring E., Verhoeven E., Litvinov S., Michalides R.J. (1993) The product of the EMS1 gene, amplified and overexpressed in human carcinomas, is homologous to a v-src substrate and is located in cell-substratum contact sites. *Mol Cell Biol* 13: 2891-2898.
- Seabra M.C., Ho Y.K., Anant J.S. (1995) Deficient geranylgeranylation of Ram/Rab27 in choroideremia. *J Biol Chem* 270: 24420-24427.
- Seidner G., Alvarez M.G., Yeh J.-I., O'Driscoll K.R., Klepper J., Stump T.S., Wang D., Spinner N.B., Brinbaum M.J., and De Vivo D.C. (1998) GLUT-1 deficiency syndrome caused by haploinsufficiency of the blood-brain barrier hexose carrier. *Natl Genet* 18: 188-191.
- Seipelt R.L., and Peterson M.L. (1995) Alternative processing of IgA pre-mRNA responds like IgM to alterations in the efficiency of the competing splice and cleavage-polyadenylation reactions. *Mol Immunol* 32: 277-285.
- Seiwert SD, Heidmann S, Stuart K (1996) Direct visualization of uridylyte deletion in vitro suggests a mechanism for kinetoplastid RNA editing. *Cell* 84: 831-841.
- Semina E., Reiter R., Leysens N.J., Alward W.L.M., Small K.W., Datson N.A., Siegel-Bartelt J., Bierke-Nelson D., Bitoun P., Zabel B.U., Carey J.C., Murray J.C. (1996) Cloning and characterization of a novel bicoid-related homeobox transcription factor gene, RIEG, involved in Rieger syndrome. *Nature Genet* 14: 392-399.
- Sharma P.M., Bowman M., Madden S.L., Rauscher F.J. 3rd, Sukumar S. (1994) RNA editing in the Wilms' tumor susceptibility gene, WT1. *Genes Dev* 8: 720-731.
- Shaw G., and Kamen R. (1986) A conserved AU sequence from the 3' untranslated region of GM-CSF mRNA mediates selective mRNA degradation. *Cell* 46: 659-667.
- Shiang R., Bell G., Divelbiss J.E., Haskins-Olney A., Overhauser J., Wasmuth J., Murray J.C. Mapping of ADH3, EGF, and IL2 in a patient with Riegers-like phenotype and 4q23-q27 deletion. *Am J Hum Genet* 41: A185.
- Shore D. (1997) Telomere length regulation: getting the measure of chromosome ends. *Biol Chem* 378: 591-597.

- Shovlin C.L., Hughes J.M.B., Seidman C.E., Seidman J.G. (1997) Characterization of endoglin and identification of novel mutations in hereditary hemorrhagic telangiectasia. *Am J Hum Genet* 61: 68-79.
- Shprintzen R.J., Goldberg R.B., Lewin M.L., Sidoti E.J., Berkman M.D., Argamaso R.V., Young D. (1978) A new syndrome involving cleft palate, cardiac anomalies, typical faces, and learning disabilities: velo-cardio-facial syndrome. *Cleft Palate J* 5: 56-62.
- Singh R.N., and Sultston J.E. (1978) Some observations on molting in *Caenorhabditis elegans*. *Nematologica* 24: 63-71.
- Sirotkin H., Morrow B., Das Gupta R., Goldberg R., Patanjali S.R., Shi G., Cannizzaro L., Shprintzen R., Weissman S.M., Kucherlapati R (1996) Isolation of a new clathrin heavy chain gene with muscle-specific expression from the region commonly deleted in velo-cardio-facial syndrome. *Hum Mol Genet* 5: 617-624.
- Skuse G.R., Cappione A.J., Sowden M., Metheny L.J., Smith H.C. (1996) The neurofibromatosis type I messenger RNA undergoes base-modification RNA editing. *Nucleic Acids Res* 24: 478-485.
- Small S., Kraut R., Hoey T., Warrior R., Levine M. (1991) Transcriptional regulation of a pair-rule stripe in *Drosophila*. *Genes Dev* 1991 5: 827-839.
- Smit A.F.A. (1996) The origin of interspersed repeats in the human genome. *Curr Opin Genet Devel* 6: 743-749.
- Smit A.F.A., Toth G., Riggs A.D., Jurka J. (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J Mol Biol* 246:401-417.
- Smit AFA (1993) Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Res* 21: 1863-1872.
- Soreq H., Ben-Aziz R., Prody C.A., Seidman S., Gnatt A., Neville L., Lieman-Hurwitz J., Lev-Lehman E., Ginzberg D., Lapidot-Lifson Y., Zakut H. (1990) Molecular cloning and construction of the coding region for human acetylcholinesterase reveals a G + C - rich attenuating structure. *Proc Natl Acad Sci USA* 87: 9688-9692.
- Spangler E.A., Ryan T., Blackburn E.H. (1988) Developmentally regulated telomere addition in *Tetrahymena thermophila*. *Nucleic Acids Res* 16: 5569-5585.
- Sparks A.B., Rider J.E., Hoffman N.G., Fowlkes D.M., Quilliam L.A., Kay B.K. (1996) Distinct ligand preferences of Src homology 3 domains from Src, Yes, Abl, Cortactin, p53bp2, PLC γ , Crk, and Grb2. *Proc Natl Acad Sci USA* 93: 1540-1544.

- Speed R.M. (1988) The possible role of meiotic pairing anomalies in the atresia of human fetal oocytes. *Hum Genet* 78: 260-266.
- Spinner N.B., Eunpu D.L., Schmickel R.D., Zackai E.H., McEldrew D., Bunin G.R., McDermid H.E., Emanuel B.S. (1989) the role of cytological NOR variants in the etiology of trisomy 21. *Am J Hum Genet* 44: 631-638.
- Starling J.A., Maule J., Hastie N.D., Allshire R.C. (1990) Extensive telomere repeat arrays in mouse are hypervariable. *Nucleic Acids Res* 18: 6881-6888.
- Stern J.J., Cerrillo M., Dorfmann A.D., Coulam C.B., Gutiérrez-Najar A.J. (1996) Frequency of abnormal karyotypes among abortuses from woman with and without a history of recurrent spontaneous abortion. *Fertil Steril* 65: 250-253.
- Stewart G.D., Harris P., Galt J., Ferguson-Smith M.A. (1985) Cloned DNA probes regionally mapped to human chromosome 21 and their use in determining the origin of non-disjunction. *Nucleic Acids Res* 13: 4125-4132.
- Sulston J.E., and Horvitz H.R. (1977) Post-embryonic cell lineages of the nematode *Caenorhabditis elegans*. *Dev Biol.* 56: 110-156.
- Sutcliffe J.S., Nakao M., Christian S., Orstavik K.H., Tommerup N., Ledbetter D.H., Beaudet A.L. (1994) Deletions of a differentially methylated CpG island at the SNRPN gene define a putative imprinting control region. *Nature Genet* 8: 52-58.
- Takahata N., and Satta Y. (1997) Evolution of the primate lineage leading to modern humans: Phylogenetic and demographic inferences from DNA sequences. *Proc Natl Acad Sci USA* 94: 4811-4815.
- Takano H., Yanagimachi R., Urch U.A. (1993) Evidence that acrosin activity is important for the development of fusibility of mammalian spermatozoa with the oolemma: inhibitor studies using the golden hamster. *Zygote* 1: 79-91.
- Takao A, Ando M., Cho K., Kinouchi A., Murakami Y.(1980) Etiologic categorization of common congenital heart disease. *In* Van Praagh. R., Takao, A. (eds.) *Etiology and Morphogenesis of Congenital Heart Disease*. Futura Publishing Company, Mount Kisco, N. Y. pp. 253-269.
- Tassabehji M., Metcalfe K., Fergusson W.D., Carette M.J.A., Dore J.K., Donnai D., Read A.P. (1996) LIM-kinase deleted in Williams syndrome. *Nature Genet* 13: 272-273.
- Tatsuya K., Lalande M., Wagstaff J. (1997) UBE3A/E6-AP mutations cause Angelman syndrome. *Nature Genet* 15: 70-77.

- Taylor L.D., Krizman D.B., Jankovic J., Hayani A., Steuber P.C., Greenberg F., Fenwick R.G., Caskey C.T. (1991) 9p monosomy in a patient with Gilles de la Tourette's syndrome. *Neurology* 41: 1513-1515.
- Teebi A.S., Gibson L., McGrath J., Meyn M.S., Breg W.R., Yang-Feng T.L. (1993) Molecular and cytogenetic characterization of 9p- abnormalities. *Am J Med Genet* 46: 288-292.
- Trask B.J., Friedman C., Martin-Gallardo A., Rowen L., Akinbami C., Blankenship J., Collins C., Giorgi D., Iadonato S., Johnson F., Kuo W.-L., Massa H., Morrish T., Naylor S., Nguyen O.T.H., Rouquier S., Smith T., Wong D.J., Youngblom J., van den Engh G. (1998) Members of the olfactory receptor gene family are contained in large blocks of DNA duplicated polymorphically near the ends of human chromosomes. *Hum Mol Genet* 7: 13-26.
- Udwin O., Yule W., Martin N. (1987) Cognitive abilities and behavioral characteristics of children with idiopathic infantile hypercalcemia. *J Child Psychol Psychiatry* 28: 297-309.
- Underwood A.P., Louis E.J., Borts R.H., Stringer J.R., Wakefield A.E. (1996) *Pneumocystis carinii* telomere repeats are composed of TTAGGG and the subtelomeric sequence contains a gene encoding the major surface glycoprotein. *Mol Microbiol* 19: 273-281.
- Upcroft P., Chen N., Upcroft J.A. (1997) Telomeric organization of a variable and inducible toxin gene family in the ancient eukaryote *Giardia duodenalis*. *Genome Res* 7: 37-47.
- Urioste M., Arroyo I., Cilla A., Lorda-Sanchez I., Barrio R., Lopez-Cuesta M.J., Rueda J. (1995) Distal deletion of chromosome 13 in a child with the "opitz" GBBB syndrome. *Am J Med Genet*. 59: 114-122.
- Van Arsdell S., and Weiner A.M. (1984) Human genes for U2 small nuclear RNA are tandemly repeated. *Mol Cell Biol* 4: 492-499.
- van Deutekom J.C., Bakker E., Lemmers R.J., van der Wielen M.J., Bik E., Hofker M.H., Padberg G.W., Frants R.R. (1996) Evidence for subtelomeric exchange of 3.3 kb tandemly repeated units between chromosomes 4q35 and 10q26: implications for genetic counseling and etiology of FSHD1. *Hum Mol Genet* 5: 1997-2003.
- Vaux C., Sheffield L., Keith C.G., Voullaire L. (1992) Evidence that Rieger syndrome maps to 4q25 or 4q27. *J Med Genet* 29: 256-258.
- Vazquez-Levin M.H., Reventos J., Gordon J.W. (1992) Molecular cloning, sequencing and restriction mapping of genomic sequence encoding human proacrosin. *Eur J Biochem* 207: 23-26.

- Vega-Palas M.A., Venditti S., Di Mauro E. (1997) Telomeric transcriptional silencing in a natural context. *Nat Genet* 15: 232-233.
- Verooden L.R. (1963) Extended tables of critical values for Wilcoxon's test statistic. *Biometrika* 50:177-185.
- Viot-Szoboszalai G., Amiel J., Doz F., Prieur M., Couturier J., Zucker J.N., Henry I., Munnich A., Vekemans M., Lyonnet S. (1998) Wilm's tumor and gonadal dysgenesis in a child with the 2q37.1 deletion syndrome. *Clin Genet* 53: 278-280.
- Virbasius J.V., Virbasius C.A., Scarpulla R.C. (1993) Identity of GABP with NRF-2, a multisubunit activator of cytochrome oxidase expression, reveals a cellular role for an ETS domain activator of viral promoters. *Genes Dev* 7: 380-392.
- Volbrath D., Jaramillo-Babb V.L., Clough M.V., McIntosh I., Scott K.M., Lichter P.R., Richards J.E. (1998) Loss-of-function mutations in the LIM-homeodomain gene, LMX1B, in nail patella syndrome. *Hum Mol Genet* 7: 1091-1098.
- von Deimling A., Nagel J., Bender B., Lenartz D., Schramm J., Louis D.N., Wiestler O.D. (1994) Deletion mapping of chromosome 19 in human gliomas. *Int J Cancer* 57: 676-680.
- Voullaire L.E., Webb G.C., Leversha M.A. (1987) Chromosome deletion at 11q23 in an abnormal child from a family with inherited fragility at 11q23. *Hum Genet* 76: 202-204.
- Voytas D., and Boeke J. (1992) Yeast retrotransposon revealed. *Nature* 358: 717.
- Vu T.H., and Hoffman A.R. (1997) Imprinting of the Angelman syndrome gene: UBE3A, is restricted to brain. *Nat Genet* 17: 12-13.
- Wang Y.K., Samos C.H., Peoples R., Pérez-Jurado L.A., Nusse R., Francke U. (1997) A novel human homologue of the *Drosophila* frizzled Wnt receptor gene binds wingless protein and is in the Williams syndrome deletion at 7q11.23. *Hum Mol Genet* 6: 465-472.
- Wellinger R.J., Ethier K., Labrecque P., Zakian V.A. (1996) Evidence for a new step in telomere maintenance. *Cell* 85: 423-433.
- Wenger S.L., Boone L.Y., Surti U., Steele M.W. (1997) Terminal 2q deletion -- a recognizable syndrome. *Clin Genet* 4: 290.
- Wevrick, R.; Kerns, J. A.; Francke, U. (1994) Identification of a novel paternally expressed gene in the Prader-Willi syndrome region. *Hum. Molec. Genet.* 3: 1877-1882.

- White J. (1988) The Anatomy in Wood W.B. (ed) "The Nematode *Caenorhabditis elegans*". Cold Spring Harbor Laboratory Press.pp106.
- Wicking C., Shanley S., Smyth I., Gillies S., Negus K., Graham S., Suthers G., Haites N., Edwards M., Wainwright B., and Chenevix-Trench G. (1997) Most germ-line mutations in nevoid basal cell carcinoma syndrome lead to a premature termination of the PATCHED protein, and no genotype-phenotype correlations are evident. *Am J Hum Genet* 60: 21-26.
- Wild A., Kalff-Suske M., Vorkamp A., Bornholdt D., König R., Grzeschik K.-H. (1997). Point mutations in human *GLI3* cause Greig syndrome. *Hum Mol Genet* 6: 1979-1984.
- Wilkie A.O.M., Lamb J., Harris P.C., Finney R.D., Higgs D.R. (1990a) A truncated human chromosome 16 associated with α -thalassemia is stabilized by addition of telomeric repeat (TTAGGG)_n. *Nature* 346: 868-872.
- Wilkie A.O.M., Buckle V.J., Harris P.C., Barton N.J., Reeders S.T., Lindenbaum R.H., Nicholls R.D., Barrow M., Bethlenfalvai N.C., Hutz M.H., Tolmie J.L., Weatherall D.J., Higgs D.R. (1990b) Clinical features and molecular analysis of the α thalassemia/ mental retardation syndrome. I. Case due to deletions involving chromosome band 16p13.3. *Am J Hum Genet* 46: 1112-1126.
- Wilkie A.O.M., Higgs D.R., Rack K.A., Buckle V.J., Spurr N.K., Fischel-Ghodsian N., Ceccherini I., Brown W.R.A., Harris P.C. (1991) Stable length polymorphism of up to 260 kb at the tip of the short arm of human chromosome 16. *Cell* 64: 595-606.
- Wilkie A.O.M. (1993) Detection of cryptic chromosomal abnormalities in unexplained mental retardation: a general strategy using hypervariable subtelomeric DNA polymorphisms. *Am J Hum Genet* 53: 688-701.
- Williams J.C.P., Barratt-Boyes B.G., Lowe J.B. (1961) Supravalvular aortic stenosis. *Circulation* 24: 1311-1318.
- Wilson D. I., Burn J., Scambler P., Goodship J. (1993) DiGeorge syndrome, part of CATCH 22. *J. Med. Genet.* 30: 852-856.
- Wilson G.M., Vasa M.Z., Deeley R.G. (1998) Stabilization and cytoskeletal-association of LDL receptor mRNA are mediated by distinct domains in its 3' untranslated region. *J Lipid Res* 39: 1025-1032.
- Wong A.C.C., Ning Y., Flint J., Clark K., Dumanski J.P., Ledbetter D.H., McDermid H.E. (1997) Molecular characterization of a 130 kb terminal deletion at 22q in a child with mild mental retardation. *Am J Hum Genet* 60: 113-120.

- Wood W.B. (1988) "The Nematode *Caenorhabditis elegans*". Cold Spring Harbor Laboratory Press.
- Wu H., and Parsons J.T. (1993) Cortactin, an 80/85 kilodalton pp60^{src} substrate, is a filamentous actin-binding protein enriched in the cell cortex. *J Cell Biol* 120: 1417-1426.
- Wu H., Reymolds A.B., Kanner S.B., Vines R.R., and Parsons j.T. (1991) Identification and characterization of a novel cytoskeleton-associated pp60^{src} substrate. *Mol Cell Biol* 11: 5113-5124.
- Wu T.C., and Lichten M. (1995) Factors that affect the location and frequency of meiosis-induced double-strand breaks in *Saccharomyces cerevisiae*. *Genetics* 140: 55-66.
- Wu Y.-Q., Sutton R., Nickerson E., Lupski J.R., Potocki L., Korenberg J.R., Greenberg F., Tassabehji M., Shaffer L.G. (1998) Delineation of the common critical region in Williams syndrome and clinical correlation of growth, heart defects, ethnicity, and parental origin. *Am J Med Genet* 78: 82-89.
- Wydner K.L., Bhattacharya S., Eckner R., Lawrence J.B., Livingston D.M. (1995) Localization of human CREB-binding protein gene (CREB BP) to 16p13.2-p13.3 by fluorescence *in situ* hybridization. *Genomics* 30: 395-396.
- Xu Y., Einstein J.R., Mural R.J., Shah M., Uberbacher E.C. (1994) An improved system for exon recognition and gene modeling in human DNA sequences. *In Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA. pp376-384.
- Yalamanchili R., Harada S., Kieff E. (1996) The N-terminal half of EBNA2, except for seven prolines, is not essential for primary B-lymphocyte growth transformation. *J Virol* 70: 2468-2473.
- Yang Y.S., Chen S.U., Ho H.N., Chen H.F., Lien Y.R., Lin H.R., Huang S.C., Lee T.Y. (1994) Acrosin activity of human sperm did not correlate with IVF. *Arch Androl* 32: 13-19.
- Zou H, Henzel WJ, Liu X, Lutschg A, Wang X (1997) Apaf-1, a human protein homologous to *C. elegans* CED-4, participates in cytochrome c-dependent activation of caspase-3. *Cell* 90:405-413.

Appendix : Sequence annotations on AWcontig

Position on AWcontig	Description	Genbank accession #	% identities	P-value
623-797	CpG DNA clone 92a12	z63877	90	1 x e-62
3781-3928	L2 (LINE)			
4282-4578	AluSq (SINE)			
4736-4831	HY3 (scRNA)			
4835-4886	AluJo/FRAM (SINE)			
5046-5341	AluSx (SINE)			
5734-6179	LIM4 (LINE)			
6609-6911	AluSq (SINE)			
6976-7287	AluSx (SINE)			
7713-7893	MIR (SINE)			
7946-8246	AluY (SINE)			
8572-8684	MIR (SINE)			
9018-9311	AluJb (SINE)			
9457-9525	LIMB3 (LINE)			
9537-9594	HY1 (scRNA)			
9637-9932	AluSx (SINE)			
10094-10216	LIM4 (LINE)			
10219-10520	AluY (SINE)			
10521-10766	LIM4 (LINE)			
10775-10825	AluS (SINE)			
11210-11272	Starting exon of ALPR C. elegans protein C33B4.3	Z48367	43 in protein level	1 x e-90 in protein level
11616-11819	exon of ALPR clone sc24 C. elegans protein C33B4.3	Z48367	43 in protein level	1 x e-90 in protein level
13190-13261	exon of ALPR clone sc24 C. elegans protein C33B4.3	Z48367	43 in protein level	1 x e-90 in protein level
13631-13698	MIR (SINE)			
15153-15261	exon of ALPR clone sc24 C. elegans protein C33B4.3	Z48367	43 in protein level	1 x e-90 in protein level
	DS17T7		67 in protein level	
15337-15488	exon of ALPR clone sc24 C. elegans protein C33B4.3	Z48367	43 in protein level	1 x e-90 in protein level
	DS17T7		67 in protein level	

Appendix (continued.....1)

Position on AWcontig	Description	Genbank accession #	% identities	P-value
15587-15754	exon of ALPR clone sc24 C. elegans protein C33B4.3	Z48367	43 in protein level	1 x e-90 in protein level
	DS17T7		67 in protein level	
15692	beginning of 1NFLS clone 208081	H62649	72 in protein level	7 x e-4 in protein level
15745-21015	minisatellite MS607 (VNTR locus D22S163)	X58043 and X58044	100	0
15880-15996	exon of ALPR clone sc24 C. elegans protein C33B4.3	Z48367	43 in protein level	1 x e-90 in protein level
	exon of 1NFLS clone 208081	H62649	72 in protein level	7 x e-4 in protein level
20607-20684	exon of ALPR clone sc24 C. elegans protein C33B4.3	Z48367	43 in protein level	1 x e-90 in protein level
	exon of 1NFLS clone 208081	H62649	72 in protein level	7 x e-4 in protein level
23499-23583	MER81 (SINE)			
23793-23855	exon of ALPR clone sc24 C. elegans protein C33B4.3	Z48367	43 in protein level	1 x e-90 in protein level
	exon of 1NFLS clone 208081	H62649	72 in protein level	7 x e-4 in protein level
26727-27026	AluY (SINE)			
27035-27113	MIR (SINE)			
27525-27610	MIR (SINE)			
28112-29435	LIMB5 (LINE)			
29437-29737	AluSq (SINE)			
29740-30469	LIMB5 (LINE)			
30470-30571	U6 (snRNA)			
30572-31372	LIMB5 (LINE)			
31484-31782	AluSp (SINE)			
31808-31928	AluJb (SINE)			
31939-32243	AluSp (SINE)			
32244-32424	LIMEB5 (LINE)			
32425-32715	AluJb (SINE)			
32732-33399	LIMB5 (LINE)			
33400-33582	MER58A (SINE)			
33834-33941	MER81 (SINE)			
33983-34256	exon of ALPR clone sc24			
35168-35285	MIR (SINE)			
365620	Proximal end of N85A3			

Appendix (continued.....2)

Position on AWcontig	Description	Genbank accession #	%identities	P-value
37562-37675	Chromosome 22 exon exon of ALPR clone sc24 exon of ALPR clone I511	H55337	100	0
39034-39098	L2 (LINE)			
41178-41281	L2 (LINE)			
41301-41711	L2 (LINE)			
42732-42807	exon of ALPR clone I511			
43038-43120	exon of ALPR clone I511 Rat cortactin binding protein 1	AF060116	50 in protein level	1 x e-73 in protein level
43609-43733	exon of ALPR clone I511 Rat cortactin binding protein 1	AF060116	50 in protein level	1 x e-73 in protein level
	C. elegans protein C33B4.3	Z48367	47 in protein level	2 x e-18 in protein level
	DS17T3		82 in protein level	
	AR, Soares retina N2b4HR cDNA clone 191111	H41098	80 in protein level	1 x e-61
43835-43967	exon of ALPR clone I511 Rat cortactin binding protein 1	AF060116	50 in protein level	1 x e-73 in protein level
	C. elegans protein C33B4.3	Z48367	47 in protein level	2 x e-18 in protein level
	DS17T3		82 in protein level	
	AR, Soares retina N2b4HR cDNA clone 191111	H41098	80 in protein level	1 x e-61
44943-45023	exon of ALPR clone I511 Rat cortactin binding protein 1	AF060116	50 in protein level	1 x e-73 in protein level
	C. elegans protein C33B4.3	Z48367	47 in protein level	2 x e-18 in protein level
	AR, Soares retina N2b4HR cDNA clone 191111	H41098	80 in protein level	1 x e-61 in protein level
45612-46609	AL, Stratagene liver cDNA clone 77923	T61326 T61275		
45899-46498	AluSq (SINE)			
46938-47236	AluSq (SINE)			
47713-48014	AluJb (SINE)			
48053-48183	LIMEC (LINE)			
49491-49706	Human telomeric clone 19QTEL005	Z96444	100	1 x e-116
49661-50062	Human telomeric clone 19QTEL005	Z96444	89	4 x e-19
50127-50226	Human telomeric clone 19QTEL005	Z96444	92	1 x e-5

Appendix (continued.....3)

Position on AWcontig	Description	Genbank accession #	% identities	P-value
45486-50509	exon of ALPR clone I511 Rat cortactin binding protein 1	AF060116	50 in protein level	1 x e-73 in protein level
	AR, Soares retina N2b4HR cDNA clone 191111	H41098	80 in protein level	1 x e-61 in protein level
52657-52774	MIR (SINE)			
53787-53918	exon of ALPR clone I511 Rat cortactin binding protein 1	AF060116	50 in protein level	1 x e-73 in protein level
	AR, Soares retina N2b4HR cDNA clone 191111	H41098	80 in protein level	1 x e-61 in protein level
54540-54629	exon of ALPR clone I511 Rat cortactin binding protein 1	AF060116	50 in protein level	1 x e-73 in protein level
	AR, Soares retina N2b4HR cDNA clone 191111	H41098	80 in protein level	1 x e-61 in protein level
54772-55801	Terminal exon of ALPR clone I511			
55822-55932	MIR (SINE)			
56062-56187	L2 (SINE)			
59055-61308	genscan predicted exon genescanex1 Rat cortactin binding protein 1	AF060116	22 in protein level	2 x e-14 in protein level
60867-62902	TF, Nb2HF8 cDNA clone 760010	AA451718		
61109-61308	M1314, NbME13.514.5 cDNA clone 441452	AA008999		
61159-61308	ALPR clone Fli			
62052	Proximal end of N9412			
62269-62722	ALPR clone Fli			
62403-62902	FLS, 1NFLS cDNA clone 121540	T97800 T97699		
	1NFLS cDNA clone 109839	T85152 T88710		
62903-63618	Soare testis NHT clone 728630	AA435832 AA397596		
63910-63995	MIR (SINE)			
64311-65357	Soare testis NHT clone 757178	AA444132 AA443953		
66599-66804	exon of FL2, Soares fetal lung NbHL19W cDNA clone 306553	W31128 N91835		
	exon of FL, Soares fetal lung NbHL19W cDNA clone 306908	W21394 N79090		

Appendix (continued.....4)

Position on AWcontig	Description	Genbank accession #	%identities	P-value
69592-70173	gscan predicated "Last Exon" Rat cortactin binding protein 1	AF060116	66 in protein level	3.8 x e-21 in protein
	C. elegans protein C33B4.3	Z48367	43 in protein level	7.5 x e-9 in protein level
71015-72061	exon of FL2, Soares fetal lung NbHL19W cDNA clone 306553	W31128 N91835		
	exon of FL, Soares fetal lung NbHL19W cDNA clone 306908	W21394 N79090		
	FH, fetal heart NbHH1 cDNA clone 342586	W68487 W68304		
	AB, Soares breast 2NbHBst cDNA clone 155784	R72190		
	MB, Life Tech mouse brain cDNA clone 369089	W48990		
	mb1101, mouse brain cDNA			
73744-74026	AluJo (SINE)			
74980-75275	AluSq (SINE)			
77090-77183	1st exon of acrosin gene, ACR	Y00970	100	0
78142-78345	2nd exon of acrosin gene, ACR	Y00970	100	0
78565-78848	3rd exon of acrosin gene, ACR	Y00970	100	0
79196-79260	MIR (SINE)			
80468	distal end of N85A3			
80721-81015	AluSq (SINE)			
81134-81434	AluY (SINE)			
81587-81638	L2 (LINE)			
81906-82205	AluY (SINE)			
82932-83077	4th exon of acrosin gene, ACR	Y00970	100	0
83524-84169	last exon of acrosin gene, ACR	Y00970	100	0
86764-88145	minisatellite repeat clone 3'AR	S69626	100	0
88325-88408	LIME (LINE)			
88402-89019	LIME1 (LINE)			
89029-93315	LIPA2 (LINE)			
93328-94338	Human mRNA for U2 snRN? specific A' protien	X13482	98	0
94340-96193	LIPA2 (LINE)			
96197-96583	LIMC/D (LINE)			
96573-96791	LIM4 (LINE)			
96797-96942	LIMC/D (LINE)			
96956-97259	AluSq (SINE)			
97265-97594	LIMC/D (LINE)			
97599-97893	AluY (SINE)			
97903-98572	LIMA6 (LINE)			
98580-98706	LIM4 (LINE)			

Appendix 1 (continued.....5)

Position on AWcontig	Description	Genbank accession #	% identities	P-value
98861-98910	LIMC/D (LINE)			
98963-90012	LIM4 (LINE)			
99099-99428	LIM4 (LINE)			
99430-99731	AluSx (SINE)			
99742-100097	LIM4 (LINE)			
100139-100742	LIMC3(LINE)			
100764-100905	AluJo(SINE)			
100928-101167	LIMC4(LINE)			
101215-101303	LIMC3(LINE)			
101311-101606	AluSg (SINE)			
10614-101915	AluY (SINE)			
101917-102012	LIM4 (LINE)			
102031-102325	AluJo (SINE)			
102478-103086	LIMB8 (LINE)			
103104-103511	LIMA10 (LINE)			
103547-103853	AluY (SINE)			
103796	proximal end of N1G3			
103859-103937	LIM4 (LINE)			
103960-104731	LIM4 (LINE)			
105505	distal end of N94H12			
106379-107744	Terminal exon of RABL22			
106581-107627	Internal intron of RABL22			
106843-107128	AluSc (SINE)			
106379-121217*	pul, Soares pregnant uterus cDNA clone 504484	AA149886 AA150066		
	pus, Soares INFLS S1 cDNA clone 415405	W80394 W78972		
	Smt, Stratagene mouse testis cDNA clone 567614	AA183362		
	Bmer, Beddington mouse embryonic region cDNA clone 538279	AA116785		
	Stratagene neuron hNT cDNA clone 648339	AA210938 AA207204		
	Soares pregnant urterus cDNA clone 472080	AA036989		
	Jurkat T cell VI EST 182877	AA1312189		
	Placenta Nb 2HP cDNA clone 131603	R23729		
	Adult lung directed MboI cDNA clone HMGS02645	D45473		
	Soares INFLS cDNA clone 207564	H60176		
	Soares INFLS cDNA clone 241050	H80307 H80306		
	Soares INFLS cDNA clone 129994	R19279		

Appendix (continued.....6)

Position on AWcontig	Description	Genbank accession #	% identities	P-value
106379-121217*	Breast 2Nb HBst cDNA clone 155956	R72353		
	Soares INFLS cDNA clone 193843	H51752		
	Soares INFLS cDNA clone 210317	H65526 H65328		
	Fetal lung cDNA clone HUML1875	D31574		
	Soares INFLS cDNA clone 128770	R09999		
	Soares Infant brain INIB cDNA clone 49429	H15347		
	6 week I embryo cDNA clone EST34770	AA330982		
	Stratagene fetal spleen cDNA clone 71766	T51405		
	Melanocyte 2NbHM cDNA clone 267260	N32345		
	HSC172 cells II cDNA clone EST87454	AA375193		
	Bone marrow stromal fibroblast cDNA clone HBMSF2B11-REV	AA545802		
	Umbilical vein endothelial cells II cDNA clone EST10878	AA296394		
	Melanocyte 2NbHM cDNA clone 267219	N31889 N23988		
	Umbilical vein endothelial cells II cDNA clone EST10923	AA296220		
	Heart cDNA clone D226R , D226F	T20240 T20239		
	Stratagene ovarian cancer cDNA clone 594366	AA169674 AA169486		
	Total fetus Nb2HF8 9w cDNA clone 758744	AA436912		
	Prostatic intraepithelial neoplasia 2 NCI CGAP Pr2 cDNA clone 1010472	AA228316		
	Fetal heart NbHH19W cDNA clone 366489	AA026499 AA026422		
	Takeda pancreatic islet cells cDNA clone HUMHBC4569	D82203		
	DKrizman ovary NCI_CGAP_Ov2 cDNA clone 980661	AA525419		
	TNF alpha-treated aorta endothelial cells clone EST17155	AA304238		
	12 week I Embryo cDNA clone EST32016	T28211		

Appendix (continued.....7)

Position on AWcontig	Description	Genbank accession #	% identities	P-value
106379-121217*	DKrizman ovary NCI_CGAP_Ov2 cDNA clone 980646	AA525402		
	Adult lung, 3' directed MboI cDNA clone HUMGS02903	D45706		
	Fetal heart NbHH19W cDNA clone 342627	W68775 W68672		
	Soares INFLS cDNA clone 207983	H60545		
	Soares INFLS cDNA clone 773556	AA428182		
	Soares INFLS cDNA clone 247438	N54149 N58674		
	Multiple sclerosis 2NbHMSP cDNA clone 279828	N40970 N44978		
	Soares INFLS cDNA clone 205332	H62205		
	Soares INFLS cDNA clone 229380	H79283		
	Soares Infant brain INIB cDNA clone 20095	H29611		
	Soares INFLS cDNA clone 293422	N63690		
	testis NHT cDNA clone 728207	AA435681 AA393643		
	Fetal heart NbHH19W cDNA clone 342875	W68753 W68836		
	Soares INFLS cDNA clone 241050	H80306		
	Brain IV cDNA clone EST00634	M78486		
107911-107994	exon of RABL22			
108335-108419	exon of RABL22			
108775-108886	exon of RABL22			
109104-109405	AluSq (SINE)			
109621-109911	AluJo (SINE)			
109916-110154	AluSx (SINE)			
110162-110445	AluSx (SINE)			
110482-110779	AluJb (SINE)			
110873-111198	AluSx (SINE)			
112769-113064	AluY (SINE)			
113335-113357	H. sapiens mRNA (clone 1A4)	Z78283	97	1 x e-129
113389-113657	H. sapiens mRNA (clone 1A4)	Z78283		
113393-113657	Human BAC end clone R-23J23	AQ014266	99	1 x e-129
113658-113948	AluSq (SINE)			
113949-114032	Human BAC end clone R-23J23	AQ014266	99	1 x e-129
114643-114721	exon of RABL22			
115540-115619	exon of RABL22			
116291-116432	MER5B (SINE)			
116476-116587	MIR			
116822-116851	exon of RABL22			
117201-117374	AluSg (SINE)			

Appendix (continued.....8)

Position on AWcontig	Description	Genbank accession #	% identities	P-value
117420-117493	L2 (LINE)			
117534-117682	L2 (LINE)			
118572-118864	AluSx (SINE)			
119270-119357	AluJ/FRAM (SINE)			
119929-120232	AluY (SINE)			
121058-121217	exon of RABL22			
122371-122467	Starting exon of RABL22			
123054-123493	MER4A2 (SINE)			
123632-124041	MER4A2 (SINE)			
124042-124080	Soares testis NHT cDNA clone 1408432	AA868289		
	NCI CGAP Kid3 cDNA clone 1535374	AA919161		
124081-124245	L2 (LINE)			
124944-125279	degenerative telomere repeat			
125305-125454	Chromosome 20 telomere-associated repeat DNA	AF020783	96	2 x e-6
125305-125443	Homologous to human telomeric seq	X16162	91	1 x e-37
125306-125632	Human DNA sequence from 4PTEL, Huntington's disease region	Z95704	91	1 x e-117
125306-125632	Human chromosome 18p, 18pterP1	U32384	94	1 x e-117
125343-125618	telomeric DNA sequence, clone 13QTELO25	Z96268	92	4 x e-44
125633-125723	MIR (SINE)			
125724-126443	Human DNA sequence from 4PTEL, Huntington's disease region	Z95704	93	0
125724-126002	Human chromosome 18p, 18pterP1	U32384	94	1 x e-117
126156-126294	telomeric DNA sequence, clone 13QTELO25	Z96268	92	4 x e-44
126225-126435	Subtelomeric region of chromosome 10q	AF017467	94	1 x e-84
126238-126443	Subtelomeric region of chromosome 4q	AF017468	94	1 x e-81
126444-126663	MER58A (SINE)			
126664-126791	Human DNA sequence from 4PTEL, Huntington's disease region	Z95704	97	4 x e-56
126841-126992	Human DNA sequence from 4PTEL, Huntington's disease region	Z95704	96	1 x e-65
127618-127668	Soares testis NHT cDNA clone 1408432	AA868289		
	NCI CGAP Kid3 cDNA clone 1535374	AA919161		
127689-128016	Stratagene HeLa cell S3937216 cDNA clone 843176	AA488505		

Appendix (continued.....9)

Position on AWcontig	Description	Genbank accession #	% identities	P-value
127765-128016	Soares testis NHT cDNA clone 1408432	AA868289		
	NCI CGAP Kid3 cDNA clone 1535374	AA919161		
127787-128016	Soares fetal heart NbHH19W cDNA clone 346493	W74183		
127820-128016	Soares NFL T GBC S1 cDNA clone 1578907	AA961171		
127901-128009	Human DNA sequence from 4PTEL, Huntington's disease region	Z95704	93	0
128053-128223	Stratagene HeLa cell S3937216 cDNA clone 843176	AA488505		
	Soares fetal heart NbHH19W cDNA clone 346493	W74183		
	Soares NFL T GBC S1 cDNA clone 1578907	AA961171		
128053-129469	Human DNA sequence from 4PTEL, Huntington's disease region	Z95704	93	0
128250-129475	Soares infant brain cDNA clone HY18-214, 229	AA016323 AA007236		
	Soares infant brain cDNA clone 46431	H09648 H09685		
128857-129427	telomeric DNA sequence, clone 18PTEL027	Z96386		
129079-129471	NCI CGAP GCB1 cDNA clone 704320	AA279467		
129119-129575	Soares testis NHT cDNA clone 781987	AA429504		
129498-130704	Human DNA sequence from 4PTEL, Huntington's disease region	Z95704	91	0
130053-130532	Multiple sclerosis 2NbHMSP cDNA clone 277406	N57518		
130705-131149	MER4A2 (SINE)			
131152-131278	Human chromosome 18p, 18pterP2	U32385	92	4 x e-41
	telomeric DNA sequence, clone 18PTEL003	Z96376	92	1 x e-39
131193-131608	Stratagene fetal retina 937202 cDNA clone 629664	AA218713		
131420-131783	NCI CGAP Lu5 cDNA clone 1415875	AA832514		

Appendix (continued.....10)

Position on AWcontig	Description	Genbank accession #	% identities	P-value
131428-131897	Human clone 1 chromosome 15 subtelomeric sequences	AF035600	91	1 x e-162
	Human clone 1 chromosome 14 subtelomeric sequences	AF035598	93	0
	Human clone 1 chromosome 21 subtelomeric sequences	AF035602	98	0
131540-131666	Human telomeric DNA sequences, clone 7PTEL005	Z96671	84	2 x e-12
131621-132004	Soares 1NFLS cDNA clone 297084	N73768		
132030-132406	LIMC3 (LINE)			
132417-133370	LIMC3 (LINE)			
133374-133814	MSTB (LTR/MaLR)			
133818-133991	LIMC4 (LINE)			
134039-134222	LIMC4 (LINE)			
134225-134349	Human DNA sequence from 4PTEL, Huntington's disease region	Z95704	95	4 x e-47
134350-134459	MER31-internal (LTR/MER4)			
134463-134685	AluJb			
134725-135160	MER31-internal (LTR/MER4)			
135204-135565	LTR10C (LTR/retroviral)			
135631-135684	Human telomeric DNA sequences, clone 17QTEL049	Z96374	100	4 x e-19
	Human telomeric DNA sequences, clone 1QTEL011	Z96459	100	4 x e-19
135685-136476	MER31-internal (LTR/MER4)			
136777-136946	MER31-internal (LTR/MER4)			
136984-137471	MER31-internal (LTR/MER4)			
137472-138077	Human chromosome 1q subtelomeric sequences D1S553	U06155	98	0
137472-138028	Human genomic DNA. 21 region clone r1136BG46	AG008144	96	0
137514-138024	Human ribosomal protein	U43701	90	1 x e-174
137548-138024	Human clone 1 liver expressed protein 3' end	L13799	90	1 x e-163
138078-138524	MER31-internal (LTR/MER)			
138528-139523	minisatellite repeat, also found in 21q clone T1136BG46	AG008148		
139549-139616	MLT2CA (LTR /retroviral)			
139642-139709	MLT2CA (LTR /retroviral)			
139735-139833	MLT2CA (LTR /retroviral)			
139869-140230	THE1B (LTR/MaLR)			
140231-140581	MLT2CA (LTR /retroviral)			
140588-140952	MER31-internal (LTR/MER4)			

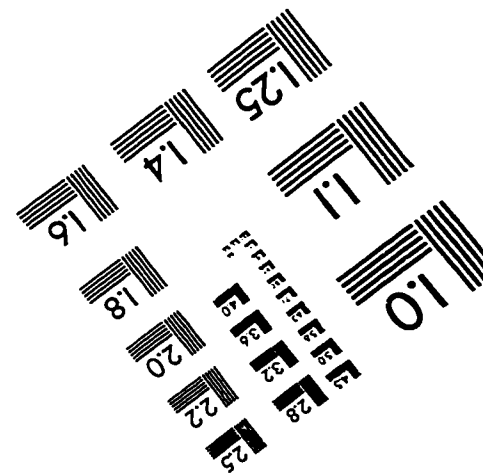
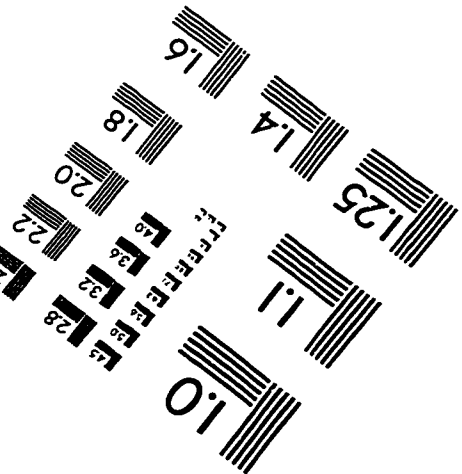
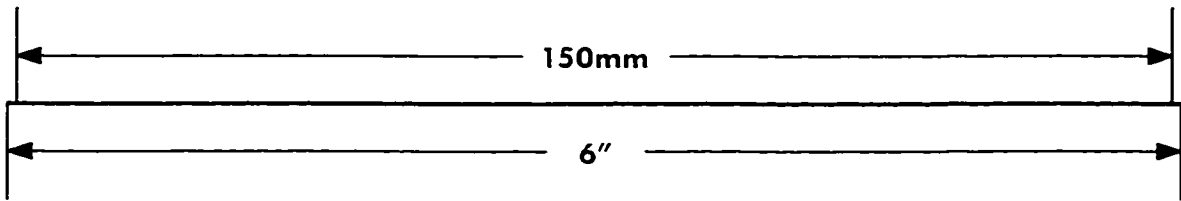
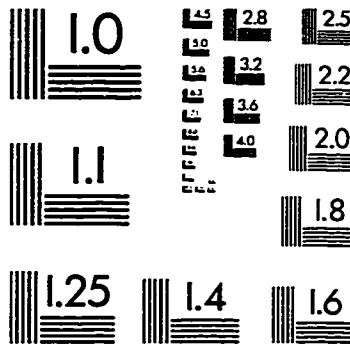
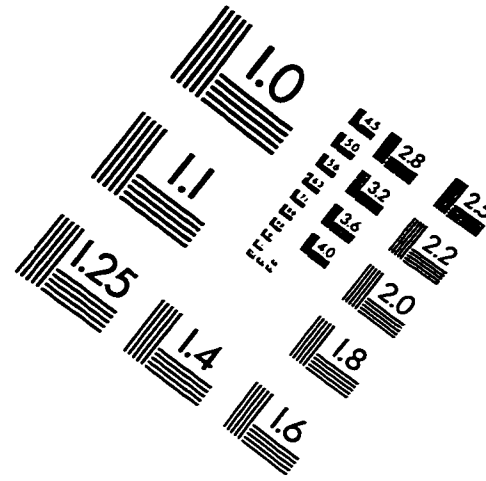
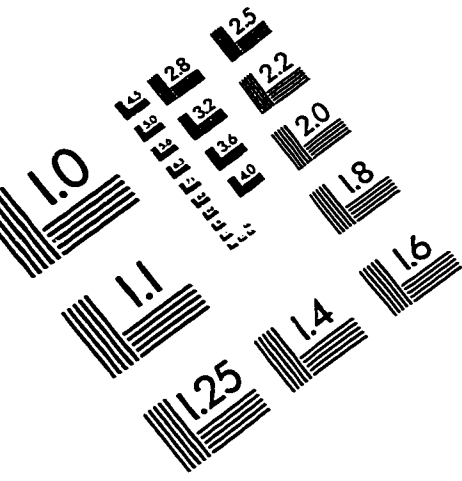
Appendix (continued.....11)

Position on AWcontig	Description	Genbank accession #	% identities	P-value
140928-141265	MER31-internal (LTR/MER4)			
141668-141718	MER72 (LTR/MER4-group)			
141730-142027	AluSq (SINE)			
142029-142667	MER72 (LTR/MER4-group)			
142668-142876	Human mRNA for subtelomeric repeat sequences	X92108	97	1 x e-161
142877-143064	LIMC4 (LINE)			
143065-143121	Human mRNA for subtelomeric repeat sequences	X92108	97	1 x e-161
143168-143511	Human subtelomeric repeat	X58156	99	0
143169-143511	Human mRNA for subtelomeric repeat sequences	X92108	97	1 x e-161
143593-143635	Human subtelomeric repeat	X58156	99	0
	Human mRNA for subtelomeric repeat sequences	X92108	97	1 x e-161
143169-143511	Human mRNA for subtelomeric repeat sequences	X92108	97	1 x e-161
143531-143592	MER34 (Unknown/MER21 g)			
143593-143635	Human subtelomeric repeat	X58156	99	0
	Human mRNA for subtelomeric repeat sequences	X92108	97	1 x e-161
143680-144244	LIMC4 (LINE)			
144261-144306	MER50 (LTR/MER4 -group)			
144307-144473	Human subtelomeric repeat	X58156	99	0
144359-144399	Human DNA from PAC30P20, chromosome Xq21.1-Xq21.3	Z95126	87	2 x e-21
144474-144781	AluYb8 (SINE)			
136947-136977	Human DNA from PAC30P20, chromosome Xq21.1-Xq21.3	Z95126	87	2 x e-21
144782-145008	Human DNA sequence from 4PTEL, Huntington's disease region	Z95704	89	8 x e-67
	Human subtelomeric repeat	X58156	99	0
	Human PGB4G7 gene	X56278	90	3 x e-69

The word in **Bold** indicates the name of the cDNA clone used in this study.

* More than 100 entries on the EST contig that spans the position 106379-121217. Only the first 50 entries are listed here. Not all of the listed EST clones span the whole region.

IMAGE EVALUATION TEST TARGET (QA-3)



APPLIED IMAGE, Inc
 1653 East Main Street
 Rochester, NY 14609 USA
 Phone: 716/482-0300
 Fax: 716/288-5989

© 1993, Applied Image, Inc., All Rights Reserved