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 OKLAHOMA

GRADUATE COLLEGE

SEQUENCE AND ANALYSIS OF CANCER RELATED GENES FROM THREE GENOMIC REGIONS; 9p21, 9p22, AND 3q26

A Dissertation

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

Dennis Burian
Norman, Oklahoma
1998

UMI Number: 9822819

Copyright 1998 by
Burian, Dennis Martin

All rights reserved.

UMI Microform 9822819
Copyright 1998, by UMI Company. All rights reserved.

This microform edition is protected against unauthorized
copying under Title 17, United States Code.

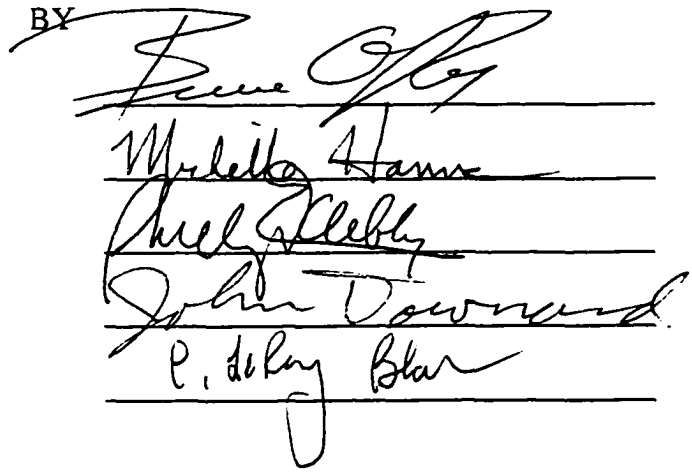
UMI
300 North Zeeb Road
Ann Arbor, MI 48103

© Copyright by Dennis Burian 1998
All Rights Reserved

SEQUENCE AND ANALYSIS OF CANCER RELATED GENES
FROM THREE GENOMIC REGIONS: 9p21, 9p22, AND 3q26

A Dissertation APPROVED FOR THE
DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

BY



Acknowledgements

I wish to thank my co-workers in Dr. Roe's lab for their encouragement and helpful discussion, especially Stephanie Chissoe and Angelika Bodenteich who got me off on the right foot and answered all the beginner's questions. Thanks especially to Sandy Clifton, Ying Ping Wang, Huaquin Pan, GuoZhong Zhang, Judy Crabtree, Fares Najar, and Angela Dorman for their advice, HongShing Lai, Steve Kenton, and James White for keeping the computers going and greatly expanding my computer knowledge, and finally to the gel room crew for all you do and to all the other graduate students, post-docs, and technicians in the lab for their fellowship and camaraderie.

I want to acknowledge committee members who have moved on to other pastures, Greg Reinhardt and Art Johnson, and those who currently serve, C. Leroy Blank, Michelle Hanna, Phil Klebba, and John Downard. Thanks to all of you for your time, support, advice, and when it was needed, the occasional swift kick.

A very special thanks goes to F. Paul Doerder for taking me on when I began this second career. His advice and encouragement during those early days has stayed with me throughout. Thanks also to my parents and parents-in-law. You have been very supportive of me and I appreciate it.

Thanks to my major professor, Bruce Roe, for his guidance and motivation over the last 5+ years. I appreciate all you have done for me and my family and I will do my best to make you proud.

Finally, my deepest gratitude goes to my wife Deborah and daughter Johanna. You have both been wonderful; through the difficult times you were always

understanding and encouraged me every step of the way. I love you both very much and treasure all you have done for me and your patience.

Table of Contents

Glossary of Terms

Alu - a human specific DNA repeat element derived from 7SL RNA

ALL – acute lymphoblastic leukemia

AML - acute myelogenous leukemia

BLAST - Basic Local Alignment Search Tool, an alignment program to compare a piece of sequence to GenBank or the protein databases

BLASTP – the variant of BLAST that compares protein sequence to the protein database of choice

CML - chronic myelomonocytic leukemia

contig - a continuous sequence fragment assembled from the endsequence reactions of several subclones

CpG islands – a region with the higher than statistically expected occurrence of the CG dinucleotide

dNTP - deoxynucleotide triphosphate gene

DNA - deoxyribonucleotide acid

E. coli - *Escherichia coli*

EST - expressed sequence tag

FastM – a transcription factor interaction prediction program

FISH - fluorescence *in situ* hybridization

GCG - the Genetic Computer Group package of DNA and protein analysis software, Madison, Wisconsin

GenScan – an exon and gene prediction program

GRAIL – the Gene Recognition and Analysis Internet Link program useful for gene predition

IFN – the interferon gene cluster used as a linkage marker for the cyclin dependent kinase inhibitor genes and af9

IPTG - isopropylthio-β-D-galactoside

kB - kilobase

kD - kilodalton

LINE - Long Interspersed Nucleotide Element, a type of DNA repeat

LOH – Loss of Heterozygosity

MatInspector – a transcription factor binding site prediction program

MDS – Myelodysplastic Syndrome

MIR – Mammalian-wide Interspersed Repeat element

MER – MEdium Reiteration repeat element

mRNA - messenger ribonucleotide acid

MTAP - the 5'methylthioadenosine phosphorylase gene used as a marker in 9p21deletion studies

NCBI - National Center for Biotechnology Information

NLM - National Library of Medicine

NNPP – Neural Net Promoter Prediction program

Northern blot – a probing of cellular RNA

NSCLC – Non-Small Cell Lung Cancer

ORF - Open Reading Frame

PCR – Polymerase Chain Reaction

PCR-SSCP – Polymerase Chain Reaction Single Stranded Conformation Polymorphism

phredPhrap - a DNA sequence assembly software package

pUC - the plasmid vector used to produce the shotgun sequencing library

rATP - ribonucleotide adenine triphosphate

RepeatMasker – a computer program that finds interspersed repeat elements in genomic DNA sequence

RT-PCR - Reverse Transcriptase Polymerase Chain Reaction

SAR – Scaffold Attachment Region

SCLC – Small Cell Lung Cancer

Southern blot – a probing of DNA

Sputnik – a computer program that finds tandem simple repeats in genomic DNA sequence

STS - Sequence Tagged Site

SwissProt - a protein database collection

Taq - *Thermus Aquaticus*

TBP – TATA element Binding Protein

TE - a buffer contains Tris base and EDTA salt

TFIID – a subunit of RNA polymerase II that includes TBP and binds the TATA element

TESS – a transcription factor binding site prediction program

TSSG – a promoter prediction program

Western blot – a probing of protein with antibody

X-gal - 5-bromo-4-chloro-3-indolyl-β-D-galactoside

Abstract

In an effort to gain a deeper understanding of the primary structure of human genomic DNA regions implicated in specific instances of tumorigenesis, over a third of a megabase of human genomic DNA has been sequenced from four cosmids from human chromosome band 9p22, four cosmids from human chromosome band 9p21, a P1 clone from human chromosome band 3q26, and a single cosmid from human chromosome band 9q34. The 9p22 cosmids include over 80% of the af9 gene that is involved in the leukemogenic translocation with the MLL gene on human chromosome band 11q23 and ten breakpoints have been mapped to the second of the three contigs formed by these four cosmids. The entire coding sequence of two genes, p16INK4a and p15INK4b, is found within the four cosmids in chromosome band 9p21 and this chromosomal region frequently is deleted in cancers originating in many different tissues. Chromosome band 3q26 also is involved in leukemogenic translocations and the P1 clone sequenced as part of the research presented here includes a single exon from the MDS1 gene. The TSC1 gene is found in human chromosome band 9q34 and although the single cosmid sequenced originally was thought to contain this gene, it was not found in this cosmid. This cosmid contains several Alu repeat elements and subclones from this cosmid were used in the latter portion of this dissertation research to investigate the ability of Alu repeat sequences to bind protein at the DNA level.

The regulatory regions of af9, p16INK4a, and p15INK4b were investigated in depth using computer algorithms designed to predict the locations of promoters and upstream transcription factor binding sites in DNA sequence. The evolution of these

regions also was investigated through analysis of the repeat elements found in the sequence data.

A shotgun sequencing strategy was employed to determine the unambiguous nucleotide sequences reported in this dissertation. Here, randomly generated fragments from each genomic clone were subcloned initially into M13mp18 and and the latter stages into pUCmp18. After isolation by a modified cleared lysate procedure, each subclone was end-sequenced employing cloning vector specific universal forward or reverse primers. The resulting nested fragement set was resolved by electrophoresis on ABI377 automated sequencers and the sequence data was assembled on Sun workstations using the phred and phrap software. Each cosmid was assembled into a single contiguous sequence and discrepancies resolved through primer-walking reactions with custom synthesized primers on subclones, PCR product, or the original genomic clone and all projects have reactions covering both orientations for over 95% of their length. The sequence of each cosmid was analyzed with GRAIL and GENSCAN to predict putative exons, RepeatMasker and Sputnik for determine the location of repetitive DNA, and BLAST to find regions with homology to other known genes.

In all the sequence determined, four genes were found and the regulation of three of these genes was investigated through computer predicted transcription factor binding sites. The presence of various computer predicted promoters and transcription factor binding sites confirmed and extended the known tissue specific expression patterns for the human proto-oncogenes encoded in the sequenced regions.

Introduction

General overview

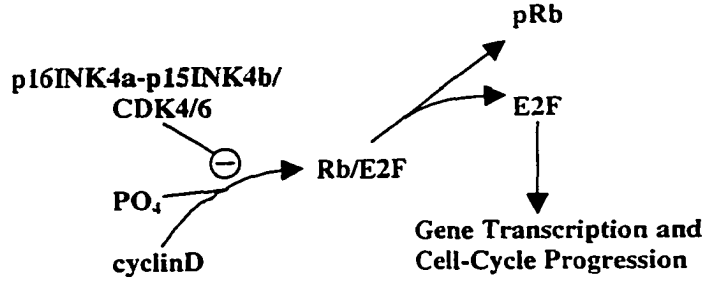
During the course of this dissertation research, over one third of a megabase of unique human DNA sequence was determined and has been deposited in GenBank. The sequencing work undertaken has been a directed search for the genomic organization of known oncogenes in 3 different genomic locations, 9p21, 9p22, and 3q26. The presence of heretofore unknown genes was determined with the gene prediction programs GRAIL and GenScan. Further investigations of the regulatory regions of three genes were performed using the available promoter and polyadenylation site prediction programs and binding sites for transcription regulation proteins were determined.

9p21, The cyclin dependent kinase inhibitor genes

Overview of the cell cycle

The cell cycle is a complex and tightly regulated series of events with retinoblastoma protein (pRb) playing a central role in progression of the cell cycle at the G1/S boundary and during cellular differentiation (Figure1).

Figure 1. Cell cycle control by p16INK4a or p15INK4b.



Binding of p16INK4a or p15INK4b to cyclin dependent kinase 4 or 6 (CDK4/6) prevents cyclinD/CDK4 or 6 complex formation and subsequent phosphorylation of Retinoblastoma protein (Rb). Phosphorylated Rb (pRb) would release the E2F transcription factors resulting in transcription of genes necessary for the cell cycle to proceed to S-phase.

Hypophosphorylated pRb binds to the E2F family of positive transcription regulators preventing transcription of several genes that play a role in DNA replication (reviewed in (1)). Phosphorylation of pRb by a cyclinD/cyclin dependent kinase (CDK) 4 or 6 protein complex results in the release of E2F and entrance of the cell into S phase of the cell cycle. CyclinD/CDK4 or 6 complexes are regulated at several levels. They are activated through phosphorylation of CDKs by CDK activating kinase whose activity is regulated by a class of molecules known as cyclin dependent kinase inhibitors (CDKI) (2-4). CDKIs are often included in the more general category of proteins known as tumor suppressors. The INK4 family consists of at least four CDKIs, p16INK4a/MTS1/CDKN2, p15INK4b/MTS2, p18INK4c, and p19INK4d. A common feature of this family of proteins is the presence of ankyrin repeats. This class of repeats is important in protein-protein interactions. The p16INK4a and p15INK4b proteins interact with CDK4 and 6 (5), but immunoprecipitation of CDK4 or 6 complexes does not co-precipitate cyclinD and p16INK4a or p15INK4b in the same complex (5).

Transforming growth factor-beta (TGFβ) is a protein that causes cell cycle arrest in G1 (2,3). Incubation of human keratinocytes in TGFβ increases p15INK4b gene expression and concurrently, increases the association of p15INK4b with CDK4 and 6 (5). A similar effect is seen in mammary tissues (6). In contrast, p15INK4b recently was shown to bind CDK4 or 6/cyclinD complexes both *in vivo* and *in vitro*, (7,8). Binding *in vivo* was dependent upon TGFβ activation of the murine lung cells and resulted in G1 cell cycle arrest. Mice nullizygous for p16INK4a have increased susceptibility to the development of sarcomas and lymphomas (9) when compared to wild type mice and are more susceptible to transformation by other oncogenes such as Ha-Ras. In addition,

fibroblasts (MEFs) from p16INK4a$^{-/-}$ mice grow faster than MEFs from p16INK4a$^{+/-}$ or p16INK4a$^{+/+}$ mice and are able to escape senescence. Immunoprecipitation of p16INK4a results in the co-precipitation of 70% of the CDK4 and 85% of the CDK6 in senescent human fibroblasts (10). Furthermore, p16INK4a transfection into U20S cells dramatically decreases the number of cells in S or G2/M phase (11). Phosphorylation of pRb decreases in a p16INK4a dose dependent manner. The P114L, D74N, R78P, and H98P mutations found in p16INK4a protein from tumors inactivate p16INK4a inhibition of pRb phosphorylation resulting in failure to arrest cell cycle progression at G1. A fifth mutation found in a pancreatic adenocarcinoma, H83Y, also fails to inhibit pRb phosphorylation as does deletion of the fourth ankyrin repeat (12). Expression of p16INK4a is increased in response to a feed back loop with phosphorylated pRb (1). Therefore, p16INK4a is induced in response to intracellular signals, whereas p15INK4b responds to extracellular stimuli.

Genetic evidence for involvement of 9p21 in cancer

Studies to determine if there is linkage between other tumor suppressor genes have been performed (13,14). Twenty-three glioma patients were tested for mutation of p15INK4b, p16INK4a, p53, and pRb. Two patients carried mutations in both p16INK4a and p53 (13). In the second study of 91 leukemia patients, four were harbored mutations in p15INK4b, p16INK4a, and pRb, two carried mutations in p15INK4b, p16INK4a, and p53, and one patient carried mutations in p53 and pRb (14). Taken together, the results from these two studies would indicate that there is little if any selection for mutation in more than one tumor suppressor gene. Furthermore, bearing in mind the stimulatory

effect of pRb effects on p16INK4a expression discussed above, mutation in any one of these proteins is sufficient for tumorigenesis in nervous and hematopoietic tissues.

Genetic mapping studies have demonstrated that interstitial deletions are linked to the interferon gene cluster and the 5'methylthioadenosine phosphorylase (MTAP) gene on chromosome band 9p21-p22 and are common in cancers of many different cell types and tissues (15-23). In lymphoblastic leukemias (ALL) of T- B- and nonB-nonT subtypes, two different studies found homo- or hemizygous deletion of all or part of the interferon gene cluster (15,16). The earlier study presented data mainly derived from primary cell lines, whereas the second study investigated primary cancers from 62 patients. These studies revealed that deletion of the IFNα genes is more common than deletion of the IFNβ1 gene indicating that the leukemogenic deletion involved either one the IFNα genes or a gene proximal to the IFN gene cluster. Lower MTAP activity is observed in some cell lines used in the first study (15) although precise mapping of the MTAP gene was not completed at that time.

In later mapping studies, fifteen glioma derived cell lines and 35 glioma patient samples were assayed for deletion of 9p markers including the IFN genes and MTAP (18). Ten of the cell lines and 13 of the patient samples were homo- or hemizygously deleted for either IFNα or IFNω genes by Southern blot or RFLP analysis. MTAP activity was observed in at least one of the IFN deleted cell lines indicating that MTAP is not a tumor suppressor in gliomas. The presence of IFNβ1 in cancers deleted for IFNα and IFNω indicates that IFNβ1 can not be a tumor suppressor and that in the minimally deleted region, the order of markers from telomere to centromere is IFNβ1, IFNα, IFNω, MTAP. The presence of IFNα and IFNω in cancers deleted for more proximal markers

suggests that none of the IFN genes functions as a tumor suppressor gene in gliomas, but that a tumor suppressor gene is closely linked to IFNω. Confirming this result was the observation that the IFN gene cluster was completely or partially homozygously deleted in 14 of 30 other glioma cell lines (21). During these studies, it also was noted that all homozygous 9p deletions in these cell lines included at least part of the IFN gene cluster.

Deletions of 9p in lung cancer (19,22) and mesothelioma (19,20) also have been observed. Olopade et al. (19) described the results of a survey of 47 cell lines derived from non-small cell lung cancers (NSCLC) and small cell lung cancers (SCLC). By Southern blot and MTAP enzyme activity, homo- or hemizygous deletion of MTAP and at least part of the IFN gene cluster was seen in 14 of 26 (54%) SCLC cell lines, 6 of 21 (29%) NSCLC cell lines, and 43% overall. Eight cell lines that showed no IFN rearrangement by Southern blot then were assayed by restriction fragment length polymorphism (RFLP) analysis of the 1400kB *NotI* fragment containing the IFN gene cluster. Two of these cell lines were homozygously deleted within this fragment. However, since MTAP activity was seen in cell lines deleted for IFN genes, it was concluded that MTAP is not a tumor suppressor gene in these lung cancers confirming earlier data from gliomas (18). Through deletion analysis of other markers in this region, the position of the MTAP gene was localized to a position proximal to the IFN gene cluster and distal to the D9S3 marker. PCR of 24 microsatellite markers in the region was used to determine that loss of heterozygosity (LOH) was seen in 27 of 40 patients tumors from SCLC, adenocarcinoma or large cell carcinomas (22). Here, the tumor suppressor gene was localized to a region of 4cM proximal to the IFN gene cluster and distal to the D9S171 microsatellite marker, distal to D9S3. The presence of the most

proximal of IFN genes in primary cancers deleted for the D9S171 marker and deletion of D9S171 in cancers positive for the IFN genes indicates that a tumor suppressor gene is located between the two markers but does not include either one.

A survey of 20 melanoma cell lines and 4 primary melanomas from three patients concluded that 86% of informative cell lines were hemizygously deleted for at least two of the markers they used (17). These markers, in order from 9pter to the proximal q arm were D9S33, IFNβ1, IFNα, D9S126, D9S3, D9S19, the glycoprotein 4b-galactosyltransferase gene, and the argininosuccinate synthetase pseudogene 3. D9S126 is proximal to the D9S171 microsatellite marker used by Merlo et al. (22). The physical distance between D9S126 and the IFN gene cluster was estimated to be 2-3Mb. Homozygous deletion was observed in two of their cell lines at the D9S126 locus. In addition, hemizygosity at neighboring loci was observed for all but one of the cell lines indicating that an unknown tumor suppressor was closely linked to this locus and that none of the known gene markers used in this study was a tumor suppressor gene. Two patient samples were hemizygously deleted in the same region but the third was difficult to interpret.

PCR analysis of polymorphic microsatellite markers in 29 invasive head and neck squamous cell cancers and 17 preinvasive lesions demonstrated LOH at similar rates in both classes of tumors (23). Since LOH for markers proximal to IFNα is observed in 72% of invasive tumors and 71% of preinvasive tumors, deletion of 9p21 is thought to be an early event in tumorogenesis in agreement with data from melanomas (17).

In over 100 melanomas, homozygous deletions were localized to a specific region within chromosome band 9p21 in a majority of cases (24). The partial sequence of a

single cosmid centered within this deleted region shows the presence of part of two related genes which they called multiple tumor suppressor 1 and 2 (MTS1 and MTS2). Sequencing of cDNAs for these two genes shows that showed that they are 93% identical in a stretch of 255bp starting at the beginning of the second exon of p16 (24). MTS1 was the previously characterized cyclin dependent kinase inhibitor p16INK4a (25) while MTS2 is the functionally related p15INK4b protein (5). Although the presence of two highly related genes within 40kB suggested to these investigators that a gene family may be located in this region, probing neighboring genomic clones with the second exon of p16 revealed no other related genes in this genomic region. Sequence tagged sites (STS) from both genes also were used to determine the extent of homozygous deletions in cell lines derived from twelve different kinds of cancer including astrocytoma, glioma, leukemia, melanoma, neuroblastoma, osteosarcoma, bladder, breast, colon, lung, ovarian, and renal cancers. Homozygous deletion of at least one STS was detected in all cancer types tested except colon cancer and neuroblastoma. Deletion of p15INK4b without deletion of p16INK4a was not detected in any of the cell lines tested. Homozygous deletion of p16INK4a was demonstrated by PCR of p16INK4a exon sequences and closely linked STSs in a survey of 46 cell lines derived from melanoma, glioma, leukemia, and lung cancers (26). A total of 46 cell lines were used in this investigation. They found that 61.5% of melanoma, 87.5% of glioma, 64.3% of leukemia, and 36.4% of lung cancer cell lines carried homozygous deletion of p16INK4a. Reverse transcriptase-PCR (RT-PCR) with a primer unique to the 3' end of p16INK4a confirmed these results on total poly-A mRNA from the above cell lines.

Deletion of 9p21 also has been linked to melanoma in primary tumors of sporadic melanoma (27). In contrast to previous findings indicating that deletion of this region was an early occurrence in the progression of many malignancies, homozygous deletion of p16INK4a is associated with progression of non-Hodgkins lymphoma from low to high grade (28). This study also reported that there was no correlation between mature T- or B-cell malignancy and deletion of p16INK4a, but 20% of B-cell precursor ALL (BCP-ALL) patients carried homozygous deletions of p16INK4a by Southern blot. Another large survey of 209 B-cell non-Hodgkins lymphoma patients found that homozygous deletion of both p16INK4a and p15INK4b occurs in 12 patients (29). However, none of these patients carried a deletion for only one of the INK4 genes from 9p21 nor were they able to find tumor specific point mutations by PCR-single stranded conformational polymorphism (PCR-SSCP) analysis in patients hemizygously deleted for these genes. Another large survey of primary leukemias observed that p16INK4a and p15INK4b were common deletion targets (14). Of 230 patients tested for 9p21 deletion, 45 were homozygously deleted at the p16INK4a locus, 42 were homozygously deleted at the p15INK4b locus, and 41 were homozygously deleted at both loci. Hemizygous deletion was observed in 22 and 20 patients at the p16INK4a and p15INK4b loci respectively. No point mutations in remaining p16INK4a alleles were detected by PCR-SSCP. Homozygous deletion of both p16INK4a and p15INK4b was detected in 6 of 13 bladder cancer cell lines (30) and 34 of 140 primary tumors (31) tested by PCR of STSs within the two genes. Although six tumors tested in the second study were homozygously deleted at the p16INK4a locus, two retained p15INK4b alleles. A study to determine the extent of 9p21 mutations in breast cancers found that half of 104 ductal and lobular breast

cancers lacked p16INK4a expression (32). Deletion of p15INK4b but not p16INK4a could be linked to pediatric BCP-ALL (33). Myelodysplastic syndrome (MDS) (34) and chronic myelogenous leukemia (CML) (33) have little or no linkage to alteration of p15INK4b or p16INK4a. Linkage to these loci also has not been observed in pituitary adenoma (35), although other 9p21 loci are deleted in these cancers.

Promoter methylation studies in the p15INK4b and p16INK4a genes

Hypermethylation of CpG islands in the promoters of both p15INK4b and p16INK4a has been demonstrated in cell lines and primary tumors in humans (36-39) and mice (40). Hypermethylation initially was observed in 67% of bladder tumor patients and cell lines within the CpG island upstream of the p16INK4a gene (36) and expression of p16INK4a was not observed in hypermethylated tissues. Interestingly, these same studies found that normal colonic mucosa also is methylated in the p16INK4a promoter, but colon cancers are not methylated and express p16INK4a at high levels. The p15INK4b promoter was not methylated. Hypermethylation of CpG islands was observed in the promoters of both p15INK4b and p16INK4a in 67% of multiple myeloma patients (39) and hypermethylation was associated with the severity of the cancer. The authors noted that deletion of either gene had never been observed in tumors of this classification, and thus, these genes might play a role in the disease state through a novel mechanism. Tumors from breast, colon, NSCLC, SCLC and leukemia were tested for hypermethylation (37). The promoters of both p15INK4b and p16INK4a in a third of breast cancer tumors and half of colon cancer tumors and NSCLC were hypermethylated, but hypermethylation of the p16INK4a promoter was not observed in SCLC or leukemia. However, the promoters of both genes from two of four primary gliomas were

hypermethylated and the remaining two were only hypermethylated in the p15INK4b promoter. This result also was observed in T-cell ALL (37,38), AML patients, and AML cell lines (37), where the p15INK4b promoter was hypermethylated but the p16INK4a promoter was not. In the larger study of leukemia patients, seventeen of 45 patients with T-ALL initially presented with hypermethylation of the p15INK4b promoter (38) and seven of 32 patients were shown to be hypermethylated at this locus upon relapse. These investigators calculated that expression of p15INK4b and p16INK4a were altered in expression either by deletion or methylation in 76% and 67% of T-ALL cases respectively. Herman et al. (37) showed that six of seven AML patients were hypermethylated in the p15INK4b promoter. Treatment with 5-aza-2'-deoxycytidine, a demethylating agent, resulted in increased expression of p15INK4b (37). Hypermethylation of CpG islands in these studies was measured using an assay that relies upon the methylation sensitivity of certain restriction enzymes such as *Bss*HI followed by PCR using primers that generate a fragment spanning the restriction enzyme cut site. Since the methylated sites are not cut, a PCR fragment is produced while the unmethylated sites are cut and no PCR product is made. A second method to test the methylation state of CpG islands, methylation sensitive PCR (MSP), takes advantage of the deamination/oxidation reaction that converts unmethylated cytodine residues to uridine in the presence of sodium bisulfite (41). PCR is performed on the resulting template using primers specific for either the converted bases or unconverted bases, the results of which are specific for the methylation of the cytodines in the priming site. The major advantage of MSP method is that it does not rely on the presence of restriction enzyme cut sites, but can be used to test the methylation of any CpG dinucleotide. This

method was used to demonstrate that p15INK4b was hypermethylated in adult onset AML, was a common occurrence in adult onset ALL and pediatric onset ALL and AML, but not CML (33). Taken together, CpG island methylation and subsequent gene inactivation are common at the p16INK4a and p15INK4b loci, and can be the sole mechanism of gene inactivation in association with some cancers (39).

As noted above, hemizygous deletion of 9p21 markers is a common feature in many cancers. To address the importance of these observations, Kamb et al. (24) sequenced cDNA clones from 34 melanoma cell lines. They found four silent point mutations, three frameshift mutations, five nonsense mutations and three missense mutations within p16INK4a. In summary, analysis of p15INK4b and p16INK4a inactivation can occur through three mechanisms. Homozygous deletion is the most common mechanism of inactivation of these two genes. Hypermethylation within gene promoters is relatively common and does result in lower p16INK4a and p15INK4b expression levels. Finally, inactivating point mutations are known but not common.

The $p19^{ARF}$ gene is located in 9p21

Recently, an alternative first exon for p16INK4a was found in mice (42). These investigators discovered that transcripts from this exon, called E1b, splice to exons two and three normally, but utilize an alternative reading frame such that the resulting protein has no sequence identity to the product of p16INK4a. Due to the larger size of the gene product of E1b, they called this gene $p19^{ARF}$. Transcripts from E1b also were observed in humans (43,44), but these investigators incorrectly assumed that the p16INK4a reading frame was conserved in the gene transcribed from the second promoter. If this were the case, the first methionine would be in the second exon and the resulting protein would be

smaller. However, the expected 10kDa protein product could not be detected on Western blots although transcripts could be detected with a probe specific for E1b (43). They observed transcription from E1a, the originally described first exon, in tumors with disruptions in E1b. They concluded that there are two promoters and the two gene products are not splice variants of a single transcript. Since expression of $p19^{ARF}$ in mammalian cells leads to cell cycle arrest in either G1 or G2, it is a functional cell cycle regulatory agent. Binding of $p19^{ARF}$ to any known cyclin or CDK has not been detected, therefore, although the exact mechanism of action is unknown, deletion of p16INK4a could lead to loss of not one but two genes with tumor suppressor function.

Deletions of 9p21 without deletion of either p15INK4b or p16INK4a also have been observed (35). Although homozygous deletion of either 9p21 CDKI gene were not observed in a survey of 57 pituitary adenoma patients, homozygous deletion of markers on both sides of p15INK4b and p16INK4a were found. Neither promoter methylation or gene expression were investigated in this study but the possibility exists that another tumor suppressor gene may be present in this region. As part of the research reported in this dissertation, the sequence from four cosmids totaling 121205 nt of unique DNA from 9p21 has been deposited in GenBank and the genomic organization of the p15INK4b and p16INK4a genes deduced.

9p22, The AF9 Gene Region

Leukemogenic translocations involving human chromosome band 11q23 are known to partner with chromosomes four, six, nine, eleven, nineteen, and X. Fusion genes result from these translocations and chimeric transcripts are detected (60,65). Acute leukemias arising from these translocations can involve either myeloid or

lymphoblastic cell types (45). The partner locus involved in the translocation has been implicated in the cell-type affected, and stage of cell differentiation, i.e., the FAB subtype (46) at which the leukemia is manifested. For example, a t(4;11)(q21;q23) usually is associated with acute lymphoblastic leukemia (ALL) of the L1 (46) subtype whereas t(9;11)(p22;q23) usually result in acute myeloid leukemias that involve monocytes and are classified as either M4 or M5 subtype (46). The *de novo* t(9;11)(p22;q23) is prevalent in leukemias in youth (47,48). The t(9;11)(p22;q23) also has been found in therapy related acute monocytic leukemias (t-AML) (49,50). These patients initially present with other cancers and are treated with DNA-topoisomerase II inhibitors as part of their treatment plan. These drugs inhibit the ligation of double stranded breaks formed by topoisomerase II, but a detailed mechanism of the recombination has yet to be elucidated.

Af9 transclocations involve the MLL gene in chromosome band 11q23

Mapping of the 11q23 region using biotinylated YAC and cosmid probes has demonstrated that a common locus is involved in a subset of translocations with chromosomes four, six, nine, and eleven (51). The involved locus is distal to both the IFN gene cluster and CDG3 (51). The chromosome 11 gene involved in these translocations has been identified by three different groups and variously named as ALL-1 (52), MLL (53), or HRX (54).

Messenger RNA from this gene is present in peripheral blood cells of both myeloid and lymphoid lineages, and cell lines of neural, epithelial, and hepatic origins (54). By Northern blot analysis, the same group of workers observed two transcripts of 15 and 13kB are present in all cell lines tested (54). Using the same assay, Yamamoto et al. (55) also found transcripts of 15, 14, 12, 7.5 and 5.0kB in peripheral blood

lymphocytes and similarly, McCabe et al. (56) showed a more complex pattern of expression. Normal pre-B cells had transcripts of 12.5, 12.0, 11.5kB that hybridize to any of three cDNA derived probes. The most 3' of these probes detects a 2.0kB transcript in normal pre-B cells and a 5.0kB transcript in monocytoid cell lines.

A 0.74kB cDNA derived probe used in the above studies also detected genomic rearrangements when used as a probe in Southern blots of DNA from cell lines or patients showing *de novo* rearrangements (56,57). Partner chromosome bands involved in the balanced translocations in these studies included 4q21, 9q23, 14q32, and 19p13. The same probe also could detect rearrangements when used in Southern blot experiments of DNA from patients with t-AML and therapy related myelodysplastic syndrome (t-MDS) because of 11q23 rearrangement (49). Karyotypes from patients in this study showed MLL translocations with chromosome bands 9p22, 6q27, and 19p13 after treatment for the original malignancy. Here, nine of the patients with 11q23 rearrangements initially had been treated for the primary tumor with topoisomerase II inhibitors.

The MLL breakpoints constitute a breakpoint cluster region

Data from Southern blots of patient's DNA shows that MLL translocations are limited to a specific region of the MLL gene which can be detected by the 0.74kB *BamHI* fragment from the MLL cDNA. This probe detects MLL exons 5, 6, 7, 9, 10, and 11 (58). Exon 8 is missing from this probe due to a frequently observed alternative splicing event (59-64). Reverse transcriptase PCR (RT-PCR) has been used by numerous groups to narrow the location of translocations within the MLL gene. For example, Negrini et al. (65) described a t(9;11) that places the der11 breakpoint in intron 8 of MLL. Yamamoto

et al. (60) surveyed 12 cell lines and 13 patients with 11q23 rearrangements and found that in all detectable cases the MLL breakpoint is in introns 6, 7, or 8. In 11 of the 12 cases where the breakpoint was 5' of exon 8, transcripts not including exon 8 were detected. These studies indicate that exon 8 is frequently spliced out of chimeric transcripts in the disease state as well as normal MLL transcripts.

Gu et al. (66) and Broeker et al. (58) independently sequenced the 8.3kB genomic *BamHI* fragment from exons 5 to 11 of MLL that includes the majority of MLL breakpoints. Each group independently determined that there are eight Alu sequences within this region, four in intron six, three in intron eight, and one in intron nine (58,66). By sequence analysis, Gu et al. (66) also determined that there is a single exactly matching topoisomerase II consensus binding site (67) within exon 9. If a single mismatch were allowed, 11 sites were found. Broeker et al. (58) noted the presence of seven topoisomerase II sites distal to exon 8 using the criteria set forth in Ebert et al. (68) of 70% percent match on one DNA strand and 40% match on the other strand. Since, topoisomerase II is a major scaffold attachment region (SAR) associated protein (69), the presence of topoisomerase II sites suggested the presence of a SAR in this region. Therefore, direct measurements were performed which demonstrated that the distal half of this region includes a SAR (58). Of the seven topoisomerase II binding sites, they found that six were within the SAR. Broeker et al. (58) summarized the results of 107 *de novo* breakpoints from the literature, including those of Gu et al. (66), and determined that roughly 75% of *de novo* breakpoints within the 8.3kB region are proximal to the SAR. They further point out that of the thirteen *de novo* breakpoints sequenced, eight were within an Alu sequence. Conversely, the breakpoints from 12 of 16 therapy related

AML patients fell within the SAR. The single t-AML breakpoint sequenced at the time did not fall within an *in vitro* topoisomerase II site, but occurs near a topoisomerase II site (58). Thus, Broeker et al. (58) proposed that the t-AML translocations occur as result of a treatment mediated change in the structure of the SAR. In *de novo* translocations, scaffolding proteins bound to DNA within the SAR are postulated to protect against breakage which could lead to translocations.

Trucation of MLL is leukemogenic

Truncation of MLL has been observed in least two instances (49,65). Negrini et al. (65) hypothesized that the leukemogenic event was not the translocation itself, but instead the resulting truncation of MLL. Evidence in support of this hypothesis includes the recent studies by Joh et al. (70) who transfected COS7 cells with MLL, a MLL-af9 fusion gene, and a MLL truncation gene. Protein product from all three constructs localized to the nucleus of these cells indicating that the nuclear localization signal was intact. In addition, in a second cell line capable of terminal differentiation in response to G-CSF, transfection with the MLL truncation gene resulted in an increased growth rate without terminal differentiation in response to G-CSF. Wild-type cells and MLL transfectants terminally differentiated into neutrophils, whereas MLL truncation transfectants differentiated into promyelocytes and myoblasts (70). However, other *in vitro* evidence indicates that this may not be the case, and that the translocation partner has an effect on the disease state (71). These workers retrovirally transfected mouse cell populations enriched in stem cells with a MLL-ENL fusion gene, and a MLL intron 7 truncation gene product. The fusion gene induced an increase in the number of cell divisions the stem cells could undergo while the cells transfected with the truncation

became senescent after only a few generations. Similar levels of G-CSF were used in these two studies however the cytokines and cells used differed. Mice transduced with a newly constructed cell line carrying the MLL-ENL fusion underwent a leukemogenic transformation (71). More recently, a MLL-af9 chimera has been used to create mice carrying a mosaic of embryonic stem cells (72). These mice developed acute myeloid leukemia by six months and 32 of 35 of these mice had died in less than one year. Interestingly, control mice that were mosaic for a MLL-myc fusion, which could function as a MLL truncation, did not present with any form of leukemia and were normal at the cellular level.

Proposed translocation mechanisms

The V(D)J recombinase enzyme has been implicated as the translocation mediator in two t(4;11) cell lines (73) and a t(9;11) (74). Recombinase normally has a role in the genomic rearrangements that accompany immunoglobulin gene maturation. The enzyme has a characteristic recognition signal sequence (RSS) consisting of the nonamer GGTTTTTGT followed by either a twelve (RSS12) or 23bp (RSS23) spacer and the heptamer CAC(A/T)GTG. The enzyme cuts 3' to the heptamer, but extra nucleotides often are found added at the junction by terminal deoxynucleotidyl transferase (75). Gu et al. (73) determined the sequence at the breakpoint junctions in the MV4:11 and RS4:11 cell lines and observed heptamer-nonamer like recombinase recognition sequences in close proximity to the breakpoints. These RSS' all had 23bp spacers except one breakpoint in MV4:11 where the nonamer was missing. Of the three breakpoints they describe, none contains a consensus heptamer or nonamer, and two of the heptamers have extra bases. Likewise, Negrini et al. (74) reported the presence of an RSS12 correctly

positioned at the MLL breakpoint in a t(9:11), but the heptamer differed from the consensus at the third position and the breakpoint on chromosome 9 lacked a nonamer. Mutation analysis of binding to the heptamer showed that mutation of the last four bases was the most detrimental (76). There also is evidence that recombinase can cleave a single RSS *in vivo*, both in mutant non-consensus substrates and in substrates containing a consensus RSS12/RSS23 pair (77). Because the result of such a cleavage could lead to a recombinogenic substrate, V(D)J recombinase can not be ruled out as a putative mediator of MLL translocation.

MLL has homology to *Drosophila* trithorax

Analysis of the MLL cDNA sequence reveals significant homology to the trithorax gene from Drosophila (54) including a putative zinc finger DNA binding domain high in cysteine residues. Trithorax has been shown to be important for normal segmentation pattern development in the larval fly (78). In addition, MLL shares sequence homology to two domains in HMG-I(Y) (54). The first homology domain is a DNA binding domain known as an AT hook through which the HMG-I(Y) protein binds DNA in the minor groove (79). The second is a binding site for cdc2 kinase, which regulates HMG-I(Y) binding to DNA. Interestingly, additional homology also was found to another minor groove binding motif, the "SPKK" motif, and to a DNA methyltransferase.

A functional study was performed to test the ability of MLL domains to activate or repress transcription (80). Here, MLL domains were fused to a GAL4 DNA binding domain and transcriptional activity was measured from a reporter plasmid with high basal activity. A repressor activity was observed between amino acids 1101 and 1400 while an

activating potential was found between amino acids 2340 and 3123. The repressor domain includes the region of sequence similarity to a DNA methyltransferase and is proline rich. Gel mobility shift assays were used to show that the AT hooks are capable of binding to a cruciform DNA structure that was not sequence dependent. Nam et al. (63) also demonstrated that depending on cell type, alternative splicing of MLL transcripts can result in RNAs lacking either exons 3, 6, 8, or 18. Splicing out these exons would result in the loss of the AT hooks, the repression domain, the zinc finger domain or an activation domain, respectively.

In summary, MLL has sequence homology to a protein known to be important for normal development in the fly, to a second protein known to be important in normal gene regulation, HMG-I(Y), and to two other possible DNA binding motifs. MLL transcripts of different lengths appear in differentiating peripheral blood cells, and the disease state is strongly associated with recombination involving MLL. Based on these observations, it is likely that MLL is important for normal development of myeloid and lymphoblastic cell lines.

Chromosome 9 loci involved in leukemogenic translocations with 11q23

At least two loci on chromosome 9 have been demonstrated to be involved in translocations with MLL, 9q33 (81), and 9p22 (47,48,82). Translocations involving 9p22 originally were linked to non-lymphocytic leukemias (48) and later were narrowed to myelomonocytic and monocytic leukemias (82), FAB classes M4 and M5. The chromosome 9 gene involved in these translocations was cloned and named LTG9 or MLLT3 by Iida et al. (57), or af9 (83).

Transcripts from the normal af9 gene were found at low levels in the spleen, and at high levels in thymus and peripheral blood cells, especially monocytic and megakaryocytic lineages (57). These workers detected transcripts of 7.4, 4.2, 3.9, 3.4, 2.6, 1.8 and 1.2kB in these blood cell types and reported a cDNA sequence of 1486nt from which they deduced a 342 amino acid protein sequence. Nakamura et al. (83) detected a transcript of 5kB in hematopoietic cell lines of unnamed types. Furthermore, they reported a cDNA sequence of 3376nt and a primary amino acid sequence of 568 amino acids. Their protein sequence included a nuclear targeting sequence and a protein that is very serine and proline rich, including a homopolymeric region of 42 serines that starts at amino acid 149. A search of the databases revealed sequence homology with ENL, one of two proteins involved in leukemogenic MLL translocations with chromosome 19. They calculated sequence similarity between the two proteins to be 68% and 56% identical. The greatest regions of homology were the N-terminal 140 amino acids and the C-terminal 67 amino acids where the two proteins are over 90% similar and 82% identical. Interestingly, the C terminal 90 amino acids of ENL have been shown to have the potential to activate transcription from reporter plasmids in pre-B and myeloid cells (84).

Sequencing of RT-PCR products from MLL-af9 gene fusions reveals that the resulting genes are translated in the same reading frame as both the MLL and af9 genes (57,61,83). The MLL-af9 gene fusion in the IMS-M1 cell line reported by Iida et al. (57) carries a fusion between exon 8 of MLL and amino acid 376 of af9. The two patients reported by Nakamura et al. (83) differed from each other and from IMS-M1. Patient CO carried a fusion between exon 8 of MLL and position 356 of af9. Patient FI carried an

unbalanced translocation resulting in the duplication of four amino acids from af9. The breakpoint on the der(9) chromosome was within an exon of af9 resulting in the fusion of amino acid 481 to exon 7 from MLL and the der(11) chromosome fused MLL exon 6 to af9 amino acid 477.

As part of this dissertation research, four cosmids totaling 150146 nt have been sequenced. From centromere to telomere, and coincidentally, in the direction of transcription of the af9 gene, the clones are 34a5, 92f5, c48, and 213a7. As a result, several of the known breakpoints now have been mapped and heretofore unmapped breakpoints now are being cloned by PCR and sequenced in an effort to learn more about these translocations.

3q26, The MDS1 gene region

Chromosome band 3q26 is involved in many interesting leukemogenic chromosomal rearrangements including t(3;3)(q21;q26), inv(3)(q21;q26), t(3;5)(q26;q21), t(3;8)(q26;q22), t(3;12)(q26;p13), t(3;21)(q26;q22) (85,86). FAB subtypes (46) arising from these rearrangements (M0, M1, M2, M4, and M7) indicate that they involve either immature myeloid cell types (87), monocytes, or megakaryocytes (88). Elevated platelet counts also have been observed in patients with 3q26 rearrangements (89) and rearrangements of 3q26 have been linked to blast crisis of Philadelphia chromosome positive chronic myelogenous leukemias (CML) (90,91). In a survey of therapy related leukemias, rearrangement of 3q26 in t-MDS was associated with loss of chromosomes 5q, 7, or 7q, while loss of chromosome 5 was seen equally in t-MDS and t-AML (86,92). Linkage of 3q26 translocations with the AML1 gene on 21q22

to therapy related AML and MDS has been demonstrated (92,93) and includes therapy with topoisomerase II inhibitors (93,94).

Known genes within chromosome band 3q26

Three genes map to chromosome band 3q26. The order from telomere to centromere is EAP, MDS1, and Evi-1. All three genes are transcribed telomere to centromere and together they span approximately 400kB (95). The leukemogenic breakpoints in 3q26 map 5' of EAP, between EAP and MDS1, between MDS1 and Evi-1, and 3' to Evi-1.

Transcripts from this region are complex. MDS1 and Evi-1 from normal tissues are detected both individually and as a fusion transcript. Northern blot analysis with a 1.5kB MDS1 probe derived from an AML1/MDS1 fusion detected transcripts of 5.1 and 6.4kB in lung, pancreas and kidney (95). Weak hybridization to transcripts of the same sizes from heart, placenta and skeletal muscle also has been observed. A more detailed study of the patterns of MDS1 expression in multiple human tissue blots found transcripts of 6.5kB, 2.0kB, 1.5kB, and 1.0kB, and a weak 5.8kB transcript (96). A probe for the Evi-1 gene also detected the 6.5 and 5.8kB bands as well as a unique 5.0kB band. These workers deduced that the 6.5kB band was a fusion transcript of MDS1 and Evi-1 and the 5.0kB band corresponded to Evi-1 alone. The three smaller bands represented MDS1 transcripts that included no Evi-1 sequence since they were not detected with the Evi-1 probe. Expression analysis in RNA blots from individual tissues yielded similar results to those seen previously. Hybridization of the 6.5kB band to the MDS1 probe demonstrated MDS1 expression in heart, placenta, lung, skeletal muscle, kidney and pancreas. A probe for Evi-1 hybridized to bands of the same size as above and a

ribonuclease protection assay performed on kidney RNA showed that the fusion transcript and the Evi-1 transcript were present in approximately equal quantities. Recent RT-PCR data demonstrate that the MDS1/Evi-1 fusion gene is expressed at low levels in control patients with no 3q26 rearrangement as well as patients positive for either inv(3) or t(3;3) rearrangements (97).

Sequence analysis of clones isolated from pancreas and kidney cDNA libraries revealed a unique feature in the MDS1/Evi-1 fusion gene. Transcription from the Evi-1 promoter yields a message that is translated from a start codon in the third exon of Evi-1 (98). In contrast, transcripts starting from the MDS1 promoter can be spliced such that a splice donor site within the MDS1 transcript is joined to the untranslated second exon of Evi-1 (96). Thus, 44 codons of the normal MDS1 gene are not present in the fusion transcript and are replaced by 63 codons from Evi-1. This is because the AML1/MDS1 fusion gene is not correctly spliced within MDS1 and translation of the fusion protein terminates after the addition of 14 amino acids from the MDS1 intron. From the cDNA sequence, MDS1 is predicted to be a hydrophilic protein of 169 amino acids that weighs approximately 19kDa. The MDS1/exon2-3-Evi-1 cDNA sequence has homology to the PR domain of the RIZ protein (96), which interacts with the retinoblastoma protein (99). An acidic patch common to transcriptional activators also is found on the C-terminal side of the second set of three zinc finger motifs (98). In addition, evidence has been presented indicating that alternative splicing of Evi-1 yields a protein lacking 315 amino acids including the last two zinc fingers of the N-terminal domain and a nuclear localization signal sequence (98).

Antisera against EVI1 specifically strongly detect a protein of 145kDa and weakly detect proteins of 135, 105 and 90kDa in Western blots from myeloid leukemia cell lines expressing Evi-1 (100). Furthermore, the 145 and 135kDa species are detected in nuclear fractions from the same cell lines (100). The same researchers showed that EVI1 is phosphorylated and that binding to double stranded DNA-cellulose columns is inhibited by metal chelating agents and is at least partially restored by addition of zinc. This finding is consistent with the prediction from cDNA sequence that Evi-1 encodes a protein that includes two different zinc finger DNA binding domains (98,101).

The first zinc finger domain includes seven zinc fingers at the N-terminus of the protein, with the first zinc finger separated from the remaining six consecutive zinc fingers by a 40 amino acid spacer. The second zinc finger domain includes three consecutive zinc fingers. Using a truncated protein that included only the first zinc finger domain, Perkins et al. (102) demonstrated binding to randomly generated ten-mer oligonucleotides with the consensus sequence TGACAAGATAA. Because there are seven zinc fingers in this domain, Delwel et al. (103) hypothesized that the binding site would be larger and expanded the work of Perkins et al. (102). Using random 35bp oligonucleotides, they determined that the consensus binding sequence for this domain was three copies of the five base pair repeating motif GAYAA in the form of the 15-mer GA(C/T)AAGA(T/C)AAGATAA (103). They further made a series of deletion mutants encoding different subsets of the N-terminal zinc fingers and determined that binding to the consensus sequence was dependent upon either zinc fingers four through six or five through seven. In addition, the full length domain including the first zinc finger and the

40 amino acid spacer conferred a strong preference for GACAA as the first repeated motif with some preference for either GATA or GACA flanking the core sequence.

In a similar series of experiments, the second zinc finger domain was shown to bind to the consensus sequence GAAGATGAG (104) and site directed mutations in the middle finger of this zinc finger domain abolished this binding. From the cDNA sequence, a potential acidic transactivation domain also is predicted on the C-terminal side of the second set of zinc fingers (101). Using CAT constructs downstream of consensus binding sites for the second zinc finger domain, no transactivation was detected (104). Subsequently, EVI1 was shown to bind to the consensus sequence GACAAGATAAGATAA[(N)1-28]CTCATCTTC (105), which combines the prior work by defining the spacer between the two previously determined consensus sequences for the separate zinc finger domains. This consensus sequence is the reverse and complement of that published by Funabiki et al. (104).

Morishita et al. (105) also showed that binding of the second set of zinc fingers is required for EVI1 to have transactivation activity. If only the consensus binding site for the first set of zinc fingers was present in their promoter construct, transactivation was not observed in either CAT or luciferase assays. Furthermore, this activity was specific for the binding sites and protein domains in question because no cross-reactivity was seen in binding assays. Excess oligonucleotide containing domain 2 binding sites did not compete for binding to domain 1 and vice versa. In other experiments, the MDS1/EVI1 fusion protein was shown to possess a transactivation activity that requires the second and the untranslated portion of the third exon of the Evi-1 gene (97). The transactivation activity of the MDS1/EVI1 fusion protein is repressed by expression of EVI1. EVI1

expression has been shown to repress GATA-1 induced maturation of erythroid cells in response to erythropoietin (106) but does not repress the cellular response to IL-3 was intact. EVI1 expression also has been shown to specifically block differentiation of granulocyte precursors in response to granulocyte colony stimulating factor (107). GATA-1 is required for maturation of both erythroid and megakaryoblastic cell lineages (108). Therefore, results of functional studies demonstrate that aberrant EVI1 expression may be a general block to hematopoietic cell differentiation, most likely by repressing the activity of GATA binding factors.

3q26 translocations

Rearrangements of 3q26 fall into two basic classes, those which result in aberrant expression of EVI1, and those which result in the fusion of the Evi-1 gene to another expressed gene resulting in EVI1 expression. The first type of rearrangements usually are either inv(3)(q21;q26) or t(3;3)(q21;q26) and the breakpoints for these rearrangements have been mapped both 5' and 3' to Evi-1 . Aberrant expression of Evi-1 is thought to arise as a result of the juxtaposition to Evi-1 of either an enhancer element or a locus control region upstream of the constitutively expressed ribophorin I gene located at 3q21 (109). However, Soderholm et al. (97) recently demonstrated that overexpression of Evi-1 was seen in only a subset of these rearrangements and the breakpoints of at least one of each of the inv(3) and t(3;3) Evi-1 overexpressing patients mapped to between MDS1 and Evi-1. From these observations, they suggest that leukemogenesis may arise due to uncoupling MDS1 from Evi-1.

The second type of rearrangements that results in aberrant EVI1 expression results from translocations with other chromosomes resulting in fusion of Evi-1 to

another gene. One of the more common translocations is with the AML1 gene on 21q22 (110,111). AML1 was originally isolated as a partner in a translocation with chromosome 8 resulting in an AML-M2 phenotype. Translocation breakpoints within AML1 are heterogeneous and result in the fusion of AML1 to EAP, MDS1, or EVI1 (95). Fusion of AML1 to EAP results in a transcript that is out of frame within EAP and results in the addition of 17 amino acids unrelated to EAP to the C-terminus of AML1 (110,112). Fusion of AML1 to MDS1 also has been detected in diagnoses of t-MDS, CML-BC, and t-AML (95) and in one of their t-AML patients, transcripts including either Evi-1 alone or MDS1 and Evi-1 were detected. The breakpoints within AML1 can be heterogeneous (87,95,110,112) but it is unclear whether differing AML1 breakpoints confer a phenotypic effect. AML1, also known as CBFα2, encodes one of a family of alpha subunits of the core binding factor (CBF), a heterodimer that includes a beta subunit (reviewed in (113)). The N-terminal half of the CBFα subunit which contains the DNA-binding activity and is involved in protein dimerization (113), has a domain with homology to the Drosophila protein runt, one of the pair-rule segmentation genes (114). CBF is a general transcription factor expressed in hematopoietic tissues including fetal liver, thymus, spleen, and bone marrow, especially erythroblasts and myeloblasts (113) but the mechanism of leukemogenesis associated with fusion of the runt domain of AML1 to EVI1 is unclear. However, the expression of an AML1/EVI1 fusion protein has been demonstrated in a cell line from a patient with CML in blast crisis (115) and Rat1 cells were transformed after transfection with a cDNA encoding the AML1/EVI1 fusion protein independent of the presence of the Philadelphia chromosome fusion transcript (116). These investigators showed that the second zinc finger domain was

necessary for transformation whereas the runt domain was not. Zent et al. (117) then used a luciferase construct driven by the CSF1R promoter to demonstrate that AML1/EAP and AML1/MDS1 fusion proteins repressed AML1 induction from this promoter in a dose-dependent manner. Since overexpression of CBFβ also repressed expression from the CSF1R promoter, they ruled out competition for CBFβ as the mechanism of repression by the fusion proteins. Interestingly, AML1/EAP was not transforming in Rat1 fibroblasts whereas AML1/MDS1 transfected cells lost contact inhibition and grew large colonies on soft agar. Furthermore, when transfected cells were injected into the flanks of athymic nude mice, only the AML1/MDS1 transfected cells grew into tumors and this effect was repressed by cotransfection with the AML1/EAP. Taken together, it is clear that the complex fusion of AML1 to EAP, MDS1, or EVI1 is leukemogenic through a complex mechanism, which includes repression of at least a subset of genes regulated by CBF.

In related studies, it has been shown that in some promoters, members of the *ets* family of transcription factors interact with CBF (113). Interestingly, at least one *ets* protein, ETV6 on 12p13, is involved in a translocation with 3q26 (118) and the resulting chimera includes the first two exons of ETV6 fused to MDS1/EVI1. Thus, it was proposed that since ETV6 is expressed in hematopoietic cells, the translocation would result in the inappropriate expression of EVI1. A subsequent survey of MDS and CML patients carrying the same translocation revealed that the translocation breakpoint on chromosome 3 is variable and can occur 5' to MDS1, between MDS1 and Evi-1, or 3' to Evi-1 (119). Fluorescent *in situ* hybridization also supported the conclusion that the chromosome 12 breakpoint was within the ETV6 gene in three of the patients. Recently,

it has been reported that the *ets* proteins function by dimerizing through a domain in the N-terminus (120), a domain included in the fusion with genes on 3q26. Therefore, although the implications of *ets* dimerization on the mechanism of leukemogenesis are unknown, it will be an interesting direction for further study.

As part the research in this dissertation study, a 64215 nt P1 clone that includes part of the MDS1 gene has been sequenced. Although additional data most likely will be required to reveal the mechanism of the translocations within this region, the observation that many 3q26 translocations are therapy related may make it possible to compare therapy related breakpoints from this and other regions, for example 11q23 and 9p22, to shed further light on this translocation.

9q34, Tuberous Sclerosis

Tuberous Sclerosis (TS) is a multiphenotype disease primarily affecting the nervous, renal, and pulmonary organ systems and the skin (reviewed in (121)). The disease is primarily characterized in the nervous system by cortical tubers or subependymal nodules. The subependymal nodules often are manifested as giant cell astrocytomas that can invade the third ventricle and may become calcified late in life.

Symptoms are extremely variable, even among family members, and because diagnosis is difficult, a three tiered system has been developed to aid practitioners in the diagnosis of TS (121). Here a "definite" case is defined as one exhibiting either one primary and two secondary, or one secondary and two tertiary symptoms. A diagnosis of "probable" is made if the patient exhibits one secondary and one tertiary, or three tertiary symptoms and a patient is diagnosed as "suspect" if only one secondary or two tertiary

symptoms are seen. Certain features are considered "diagnostic" because of their high incidence in TS patients as compared to the general population.

Retinal astrocytomas occur in 50% of TS patients and are considered diagnostic. Renal cysts and angiomyolipoma, highly vascularized cysts, are seen in two thirds of TS patients at autopsy and are considered a primary feature but occur in the general population at a high enough rate that they are not considered diagnostic for TS. Cardiac rhabdomyomas are seen in 80% of patients with TS and are considered diagnostic for the disease. They are one of the earliest signs of TS, often occurring in the fetus. Subungual fibromas in the skin also occur as a primary feature of TS. Secondary features of TS include retinal astrocytomas other than giant cell astrocytomas, achromic retinal patches, cerebral tubers that correlate with an increased frequency of seizures and mental retardation, and Shagreen patches, an elevated hamartoma of the lower back that appears to be greenish or reddish. Tertiary features include hypopigmented macules of the skin confetti skin lesions, randomly pitted dental enamel, hamartomatous rectal polyps, bone cysts, cerebral white matter "migration tracts" or heterotopias, gingival fibroma, infantile spasms, or hamartomas of other organs. Approximately 50% of children with TS are diagnosed as autistic and many have attention-deficit hyperactive disorder. The expected life span of TS patients is normal unless the brain is affected in such a way that the flow of spinal fluid is restricted, or there is kidney or heart failure as a result of functional constriction of the organ by the hamartia.

Genetic loci linked to Tuberous Sclerosis

The occurrence of TS has been estimated to be at least 1 in 10,000 but due the difficulty of diagnosis, it could be as high as 1 in 5800 (121). The disease can be

inherited as an autosomal dominant trait, although it has been estimated that 50 to 75% of all cases result as a result of a *de novo* mutation. Linkage to two loci has been established, TSC1 on chromosome band 9q34 (122-125), and TSC2 located at 16p13.3 (121,125). Approximately 50% of patients map to each locus and loss of heterozygosity is observed at both loci in TS patients (125-127). In work with a limited number of lesions, TS was shown to arise from a single cell (128) lending credence to the hypothesis that TSC genes normally function as tumor suppressors (126-128).

Recently, the TSC2 gene earlier was cloned and sequenced, and the gene product named tuberin (129). A 5.5kB gene transcript was either interrupted or deleted in all cases of TS available to these investigators indicating that TSC2 is a gene involved in TS. Although the function of the TSC2 gene is not yet established, tuberin has a region of homology to a GTPase activating protein (129).

In the case of the TSC1 gene locus, a yeast artificial chromosome (YAC), P1 artificial chromosome (PAC), and cosmid map consisting of three ordered contigs has been reported (130). Restriction enzyme cut sites frequently found in CpG islands were used to determine that there are at least seven CpG islands in the three contigs. Two of the CpG islands are in the central contig that also includes the ABO blood type gene, which they used as a marker in these mapping experiments. Since no recombination has been observed between ABO and TSC1, the two genes are very closely linked. Quite recently, the TSC1 gene was cloned and sequenced (131) and given the name hamartin. Heteroduplex analysis of PCR product from sixty TS patients was confirmed the identity of TSC1 as the target for mutation in the disease state. Hamartin is predicted to consist of 1164 amino acids with a predicted molecular weight of 130kDa. A search of GenBank

reveals homology to a hypothetical protein from *Schizosaccharomyces pombe* and the predicted structural features from the amino acid sequence include a putative transmembrane domain and a coiled-coil domain. An 8.6kB TSC1 transcript is detected at high levels in skeletal muscle and at lower levels in heart, brain, placenta, kidney and pancreas but the exact function of hamartin is unknown at this time.

Alu Repeats

As part of this dissertation research, a 35004bp cosmid, cosB1, from 9q34 was sequenced. Although this cosmid does not contain the TSC1 gene as originally thought (132), it is notable for the high Alu sequence content as determined by BLAST, an observation also confirmed by end sequencing a YAC/PAC/cosmid map of this region (130). At the DNA level, Alu sequences have been implicated in recombinogenic events at several loci (133-138), although TSC1 mutations do not include any large homozygous deletions or recombinations (131).

It is well established that Alu DNA sequences are a well-conserved set of interspersed repeats specific to the primate lineage. There are estimated to be 500,000 to one million copies of Alu repeats (139-141) in the human genome. The consensus Alu is 282bp consisting of two imperfectly repeated monomer units each followed by a poly-A tract. Alu sequences are derived from 7SL RNA (142), the RNA scaffold of the SRP that is involved in translation arrest and translocation of nascent polypeptides into the endoplasmic reticulum. Expansion of Alu sequences probably was the result of retroposition of a reverse transcribed DNA from an RNA intermediate (139). Evidence for the retroposition hypothesis comes from the observation that a direct repeat typically is found on either end of Alu repeats (139). Comparative analysis of syntenic regions in

individuals with and without newly inserted Alu elements not yet fixed in the human genome reveals that the direct repeats flanking each Alu insertion were single copy sequence before insertion and A-rich (143).

The evolution of Alu repeat elements

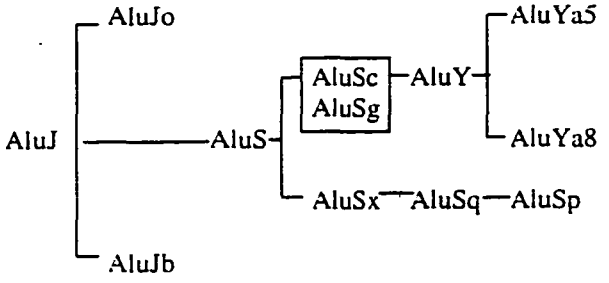
The oldest recognized Alu sequence known as the Free Alu Monomer (FAM), lacks the dimeric structure of the newer and more active dimeric Alu repeat (144). This unit gave rise to the Free Right Alu Monomer (FRAM) through a deletion. A deletion within the FRAM gave rise to a second monomer unit known as the Free Left Alu Monomer (FLAM) (145,146). Subsequent insertion of a FLAM upstream of a FRAM resulted in the dimeric Alu sequence (146). The dimeric Alu then expanded within the primate genome to its present day levels. Approximately the same number of Alu repeats has been found in the genomes of chimpanzees and humans leading to the conclusion that Alu expansion predated the divergence of the two species (140).

An intense area of research in this field of research has been the study of the evolution of Alu sequences. One hypothesis holds that Alu sequences arose as a result of transcription and retroposition of a small number of "master genes"(147) that become active and retropose within the genome. Over evolutionary time, mutations occurred and the gene cycled between active retroposition and inactivity. This hypothesis does not state whether the active gene is a newly retroposed repeat or is the original gene. Although the current evidence is inconclusive, at least one group of investigators favors the latter (148). The antithesis of the first hypothesis is that virtually any Alu sequence can serve as the template for its own transcription and subsequent retroposition. However, the bulk of the evidence points toward the "master gene" hypothesis (147-150)

because Alu sequences can be grouped into families and subfamilies based on sequence homology at "diagnostic" positions that have the same mutations when compared to other Alu sequence families (147,150,151).

Recently, the Alu sequence research community agreed upon a naming convention (152). The two major families are the S and J families and in the human genome, the S family occurs three times more often than the J family (151). As represented graphically in Figure 2, the J family is the oldest dimeric Alu family (147,150,151) and is further divided into the Jo and Jb subfamilies (153). The S family is divided into S, Sc, Y, (151) Sx, Sq, Sp, (150) and Sg subfamilies (153). Jurka reports Sx to be the progenitor of Sq, which then gave rise to Sp (153). He further shows a second major branching of the S family, the Sg subfamily, at the same time as the divergence of the Sx subfamily. Sg is the progenitor of Sc, which then gave rise to Y.

Figure 2. The Evolution of Alu Sequences.



From reference (153), the relationships between different Alu sequence subfamilies are shown. The ages of the AluSg and AluSc subfamilies vary depending on the assay (see text).

Using the percent divergence from the consensus 7SL RNA, Labuda and Striker estimated the age of the S subfamily to be 31 Myr, Sc is 24 Myr. and Y is 18Myr old (147). In contrast, a recent report used the percent similarity of individual sequences within a subfamily and evolutionary divergence to estimate the time at which the Alu subfamilies became active (154). Here, the FLAM was estimated to be 112 million years

(Myr) old placing it at the beginning of the mammalian radiation while the Jo and Jb subfamilies emerged 81 Myr ago and S, a descendent of J, arose 48 Myr ago. Thus, Kapitonov et al. (154) calculated Sq to have arisen 44 Myr old; Sx and Sp emerged 37 Myr ago. It should be noted that Kapitonov et al. (154) reverse the relative ages of Sg and Sc from that given in Jurka (153), estimating Sg to be 31 Myr old and Sc 35 Myr (154) while the Y family is estimated to be 19 Myr of age. In addition, Kapitonov et al. (154) also report that Ya5 and Ya8 arose four and three Myr old, respectively in the human genome. Although the J and S families account for >95% of the Alu repeat sequences in the human genome today (140), they are thought to be inactive at present time. In contrast, the Ya branch of the Y subfamily represents less than 1% of the Alu repeats and has been shown still to be retropositionally active at a low level (140,150).

Further evidence for the "master gene" hypothesis is the CpG dinucleotide content of the new families of Alu repeats. Older Alu sequences have undergone mutation of CpG dinucleotides to TpG or CpA and are retropositionally silent. However, newly inserted Alu sequences have maintained a high CpG content indicating that they most likely arose from one or a few "master genes" that are somehow protected from methylation within their CpG residues. If a random activation hypothesis were true, conservation of CpG's would not be expected.

The biology of Alu repeat elements

As discussed above, Alu repeats are dimers. The left half contains consensus A and B boxes for RNA polymerase III (RNA polIII). Interestingly, the B box is missing from the right half such that transcription of the right monomer alone can not begin from the still intact A box (155). A basal level of transcription resulting in full length Alu

transcripts is observed in a pluripotent embryonal carcinoma cell line (156) and HeLa cells (157). Subsequent sequence analysis of their full-length Alu RNAs revealed a mixture of transcripts that had a lower than expected number of mutations at the CpG dinucleotides and these Alu RNAs were enriched for the Sc and Y subfamilies. It was concluded that Alu transcription and retroposition are uncoupled and that selection for retropositionally active Alu sequences is post-transcriptional. Other reports indicate that Alu transcription is regulated in a tissue specific manner by methylation of C residues in CpG dinucleotides (158). By restriction enzyme analysis, Alu DNA in somatic tissues was greater than 90% methylated, but male derived germ-line tissues, i.e., sperm or seminoma, were hypomethylated. Treatment of HeLa cells with a demethylating agent increased Alu transcription (157). When a cloned Alu sequence was placed downstream of the 7SL RNA 5' flanking region it was transcribed more efficiently than when placed downstream of vector DNA, results that indicate insertion position can mediate the efficiency of Alu transcription (159). The same investigators also showed that a synthetic terminator consisting of 5-Ts increased the transcriptional efficiency, presumably through more efficient template clearance.

Alu RNA occurs in two forms. The first form is a polymorphic full-length transcript resulting from termination at fortuitous terminators in flanking sequence that is processed at the 3' end to end at the poly-A tail. The second form is a small cytoplasmic transcript (scAlu) that results from processing of full-length transcripts to include only the left monomer lacking a poly-A tail (160) and binds SRP14 (161) and SRP9 (162). Transcription of Alu sequences is induced in response to viral infection (163), heat shock, or treatment with chemicals that affect translation such as cycloheximide or puromycin

(164,165) yielding an approximately 30-fold increase in the levels of full-length transcripts. Not surprisingly, these transcripts are found associated with SRPs (164) and SRP9/14 (166). Levels of scAlu RNA show a moderate increase relative to the increase in full-length transcripts (164,166,167) and these RNAs were shown to arise from processing of full-length transcripts rather than from premature termination of transcription (164,166) in 293 cell lines. This effect was specific for Alu RNA as no increase in other translation related RNA or heat shock proteins in cycloheximide treated cells was observed (164,166). The binding affinity of SRP9/14 to the Alu left monomer was observed to be higher than to the right monomer (167) and it was concluded that protein binding to full-length Alu transcripts leads to processing of the full-length transcript to the scAlu form and prevents retroposition. In agreement with these results, biochemical and evolutionary data demonstrates that human SRP9/14 has a higher affinity than mouse SRP9/14 for Alu RNA due to expansion of a trinucleotide repeat within the coding sequence, an event postulated to have occurred concurrent with increased activity of Alu retroposition in the human genome (168). Binding affinities of SRP9/14 for scAlu RNA are highest to transcripts from older Alu subfamilies (167,169) and it is postulated that an increase in cellular levels of SRP9/14 (168) occurred concomitant with the primate Alu expansion. Taken together, these results suggest that increased SRP9/14 expression and higher Alu RNA affinity were concurrent events that served to protect the genome from excessive Alu insertion and disruption (168). In addition, scAlu RNA levels appear to be regulated indicating that there may be some biological function for scAlu RNA such as arrest of viral protein translation (163), or arrest of translation in cells under other stresses (165). Adding to the likelihood of a

function for these RNAs is the observation that almost 200 residues of the full-length Alu sequence have mutated at a lower than expected rate, an observation consistent with the hypothesis that there is pressure to maintain certain features of Alu repeats (170). A protein of unknown function that does not react with antibody to any SRP related protein binds to scAlu RNA in a region that does not interact with SRP9/14 (160).

Secondary structures of Alu transcripts also have been proposed based on energy minimization calculations (147) and by phylogenetic analysis of bases conserved from 7SL RNA (150) but the two proposed structures differ dramatically. Biochemical evidence indicates that each half of the Alu dimer RNA folds into a 7SL like structure (171) consistent with SRP9/14 binding to full-length Alu transcripts (164,166). Likewise, the structure proposed for scAlu RNA is consistent with 7SL RNA (165) and is reasonable considering the protein interactions known to take place with SRP9/14 (164,166,167).

At the DNA level, Alu repeats are known to mediate RNA polymerase II (RNA polII) directed transcriptional activity (reviewed in (172)) including promoter (138) and terminator structure, and alternative or new splice donor or acceptor sites (173). Furthermore, a role for Alu sequences in chromosome function has been postulated for an Alu sequence inserted at the pseudoautosomal boundary on the human Y-chromosome (174). In the Ewing Sarcoma gene region, an allele with an Alu repeat mediated deletion within an intron is thought to stabilize the allele and has been used to explain the decreased susceptibility of persons of African descent for Ewing's Sarcoma tumors (175). Alu repeats also have been shown to direct nucleosome positioning (176). Two nucleosomes interact with Alu DNA; one protects the 5' flank and the promoter region to

the proximal edge of the B-box. The second nucleosome is centered within the right monomer.

Chromosomal regions rich in Alu repeats also have been shown to be unstable by Southern blot and RFLP analysis of DNA from neoplastic cells (134). During the course of this work, extrachromosomal DNAs were shown to be rich in Alu repeats leading to the conclusion that Alu repeats may mediate some of the abnormal chromosomal rearrangements observed in neoplastic cells. Subsequently, deletions have been observed in the LDL gene resulting in familial hypercholesteremia (135,177,178). Alu mediated deletions of globin genes also have been observed in patients with δ-β-thalassemia (133), and α-thalassemia (137). Insertion of an Alu repeat resulted in loss of exon 6 and premature truncation of the *NF1* gene in a neurofibramatosis patient (179). A homologous recombination of Alu repeats has been found in a patient with Glanzmann thrombasthenia (180), a bleeding disorder. Finally, a familial breast and ovarian cancer has been linked to deletion of exon 17 of BRCA1 through an Alu mediated deletion (181).

In summary, Alu repeats are so common in the human genome that several different functions have been hypothesized for them, in particular a role in the cellular response to stress and translational control. In addition, it is clear that Alu sequences have played a role and continue to function in the evolution of the human genome although most of this evidence is retrospective.

Other interspersed repeats: L1, MER, MIR and THE1 elements

The consensus sequence for the L1 family of long interspersed nucleotide elements (LINE) has a consensus sequence over 6Kb long (182) and includes two open

reading frames (183). Approximately 10,000 copies of the L1 repeat are postulated to be present in the human genome (184). The first open reading frame has homology to transferrins, whereas the second has homology to yeast mitochondrial class II introns and viral RNA-dependent DNA polymerase. L1 elements have poly-A tails, but are heterogeneous at their 5' ends. The exact mechanism that preserves them in full-length form is unclear, but an RNA intermediate is probably is involved and, as stated above, L1 elements most likely encode their own reverse transcriptase. L1 elements have been classified into two major families primarily based on heterogeneity in their 3' UTR (184). Hybridization experiments have demonstrated that these two major families diverged at the time of the radiation of primates from other mammals. Accordingly, they are designated L1P and L1M followed by a numbering system to designate the subfamily within the major category and forty-seven subfamilies have been described (184). These investigators speculate that competition for the L1 encoded reverse transcriptase by the quickly expanding Alu repeat in the primate genome may have been the driving force behind the mutations that lead to the L1P class. The activity level of L1 elements in the primate genome would also explain the source of reverse transcriptase that made the rapid expansion of the Alu family of repeats possible.

Another class of interspersed repeats is the Medium Reiteration frequency sequence (MER). MERs do not share any of the features of retroposed elements since they do not encode RNA polIII promoters, poly-A tails, or have flanking direct repeats (185). Six MERs, numbered one through six, originally were described (185) but the repeat databases now include almost 50 with the MER42 family recently being described (184). The original six MER elements are heterogeneous and vary in length between 150

and 650 bps. Even within families, they are not homogeneous and have truncated ends and Alu repeats of the older families including J and Sx have been found inserted within MERs indicating that MERs probably predate the expansion of Alu sequences within the mammalian genome. Very little if any sequence homology is seen across MER family boundaries.

The Mammalian-wide Interspersed Repeat (MIR) is a more homogeneous class of repeat elements that has been estimated to be at least 130Myr old (186) and probably arose about the time of the mammalian radiation from birds. MIR elements are found at a higher frequency in marsupials and monotremes than in primates and contain a well-conserved 65 nt core sequence that was deduced by aligning 455 MIR elements from the databases (186). A 260 nt consensus has been deduced from an alignment of over 80 MIR sequences (187). These investigators believe they have deduced the entire consensus sequence of the MIR repeat because they found a tRNA gene at the 5' end and an A/T rich tract at the 3' end of the consensus sequence, structural features typical of tRNA derived SINEs (188).

Another relatively homogeneous class of repeats in the human genome is the Transposon-like Human Element (THE1) (189). THE1 repeats are about 2kB long and are flanked by 350 nt long terminal repeats. The three THE1 families originally described all share several common restriction sites, are approximately the same length, and cross hybridize with each other indicating that they a share considerable sequence homology. They are estimated to occur roughly 10,000 times in the human genome. The long terminal repeat appears separately ~30,000 times.

As a beginning to a determination of possible roles repeated sequence elements may have at the DNA level, as part of this present research, cosB1 was subcloned into the single stranded vector M13mp18 and single stranded and double stranded forms of two different subclones were isolated and bound to cellulose forming DNA-cellulose affinity columns (190). Buffy coat nuclear extracts were incubated with the columns and bound proteins eluted by varying the salt concentration and analyzed on denaturing polyacrylamide gels.

Deoxyribonucleic Acid: Chemical Structure

It is well established that deoxyribonucleic acid (DNA) is a polymer with each unit consisting of a deoxyribose sugar, a nitrogenous base, and a phosphate. Deoxyribose differs from ribose in that it is lacking a hydroxyl group at the 2' carbon. There are four nitrogenous bases used in DNA, adenine (A), cytidine (C), thymidine (T), and guanine (G). Ribonucleic acid (RNA) differs from DNA in two respects, the sugar ribose, which has a 2' hydroxyl, replaces deoxyribose and uridine (U) replaces T. The bases are bound to the 1' carbon of deoxyribose through a glycosidic and the phosphate group is esterified to the 5' carbon. A and G are purines, C and T are pyrimidines. In the cell, the DNA polymer is formed enzymatically by joining the alpha phosphate group of a new nucleotide triphosphate to the 3' hydroxyl group of the last unit in the nascent DNA strand.

The helical secondary structure of DNA is formed by hydrogen bond formation via Watson-Crick base-pairing between complimentary bases of two DNA strands oriented in an anti-parallel manner. An A forms two hydrogen bonds with a T, and a C forms three hydrogen bonds with a G. A similar set of base pairing interactions occurs in

RNA except that U replaces T and U/A base pairs are formed. Alternative base pairing interactions are known to occur in RNA, which can form G/U base pairs for example. The phosphate-sugar backbone takes on a characteristic helical twist that is stabilized by hydrophobic interactions, i.e. base stacking, between adjacent bases. The helix is typically right-handed although a left-handed form, Z-DNA, can be formed in sequence regions of alternating purines and pyrimidines. Two right-handed DNA forms are most common, the A and B forms. A-form DNA has 11 residues per turn and B-form DNA has 10 residues per turn and A-form DNA is more tightly wound than is B-form DNA.

The Central Dogma of Molecular Biology

The Central Dogma of Molecular Biology holds that DNA is transcribed to RNA, which in turn is the template for translation into protein. For most genes, this rule is true, with the exceptions being the genes that have RNA final products, the ribosomal, transfer, and other small RNA genes (e.g., U1 and related RNAs), and RNA viral genomes which are first reverse transcribed to DNA prior to transcription and translation. As discussed above, throughout evolution "genes" have moved into the genome by reverse transcription of their RNA product followed by insertion of the resulting complimentary DNA (cDNA), but these are the exception, not the rule. In eukaryotes, genes that have a protein product are transcribed by RNA polymerase II (RNA polII) whereas tRNA genes are transcribed by RNA polymerase I while rRNA and other small RNA genes are transcribed by RNA polIII. The primary transcript of RNA polII transcribed genes is processed by poly-adenylation of the 3' end, capping of the 5' end and excision of the introns to form messenger RNA (mRNA) in the nucleus. The mRNA is exported to the cytosol where it is translated into protein on the ribosome.

Sequencing Strategies

The shotgun sequencing strategy was employed throughout this dissertation research because its ease of use and reproducibility. Briefly, the strategy begins with random shearing of a large insert DNA clone into smaller fragments. The smaller fragments then are enzymatically treated to generate ends suitable for cloning into the sequencing vector of choice. Following the ligation and transformation, individual cells are colony purified and the subcloned DNA is isolated. Sequencing reactions are performed on the subclones using a variation of Sanger's dideoxynucleotide chain terminator method with fluorescent labeled products. The sequencing reactions are purified and loaded onto ABI automated sequencing machines for separation by electrophoresis and data collection. After the raw data is assembled into contiguous regions using computer algorithms that determine the overlap between sequencing reads, the individual contigs are joined using a directed strategy and proofread. Each base must be called by at least three gel readings and at least one gel reading must be in each orientation (Sanger's Rule of Three) yielding a final sequence accuracy of no more than one ambiguous base in 10,000.

Sequence Analysis Approaches and Strategies

During the course of this research, several computer-based tools were used to analyze finished sequence data. Several excellent alignment tools are available, but one of the most versatile is Crossmatch, a sequence comparison program which is part of Phil Green's sequence assembly and editing package. Crossmatch uses an advanced Smith-Waterman algorithm to quickly determine the overlap between two DNA sequences.

Promoter searches were performed using MatInspector (191), TESS (192), TSSG and TSSW (193) to predict transcription factor and enhancer binding sites. MatInspector and TESS use Transfac v3.2 (194) as the transcription factor database. NNPP (195-197) was used as a promoter prediction program. TSSG predicts promoter positions, then determines the transcription factors from the promoter.dat (198) database that may bind in that promoter. FastM (199) was used to predict the relationships between transcription factors found using MatInspector, and TESS. A FastM search is performed by entering the two transcription factors thought to interact and the maximum distance between them up to 200bp. Analysis to determine the transcription start site was begun by finding a predicted general transcription factor-binding site, i.e., TFIID, TATA-binding protein (TBP), or Sp1, upstream of the beginning of the cDNA sequence. All transcription factors predicted to bind on the same strand within 200bp then were investigated for potential protein interactions by FastM. PolyAH (200) was used to find polyadenylation signals at the 3' ends of genes. GRAIL (199,201), an exon prediction program, and GENSCAN (202,203) which predicts whole eukaryotic genes, were used with all sequence data.

Searches for repetitive DNA were performed with two programs, RepeatMasker (204) and Sputnik (205). RepeatMasker uses the Crossmatch algorithm to screen against a library that includes interspersed repeats such as SINEs and LINEs, and simple repeats such as mononucleotide and dinucleotide repeats. RepeatMasker also scans for low-complexity DNA that has no regularly repeating pattern but has a bias in nucleotide frequency. Sputnik detects repetitive DNA using a library of DNA sequence from GenBank that was labeled as repetitive by the submitter. One advantage of the Sputnik

program is that it can detect simple repeats with a longer repeating unit, i.e., five or six nucleotides. Sputnik also will only return output if the length of the repeat sequence exceeds eight nucleotides and although the repeat sequences might not be in the form of perfectly repeated tandem units, the matrix it uses is biased in that direction.

Materials and Methods

Protocols used during the course of this dissertation research are briefly summarized below. Unless otherwise noted, detailed expansion of these protocols is available in Protocols for Recombinant DNA Isolation, Cloning and Sequencing (206), DNA Isolation and Sequencing Essential Techniques (207), or from the Advanced Center for Genome Technology web-site at http://www.genome.ou.edu.

Large Scale DNA Isolation

Large scale genome level sequencing requires microgram quantities of the genomic clone to be sequenced. We have used two methods for large scale isolation of cosmid, PAC or BAC DNA. They differ in the purification method. Cesium chloride gradient ultracentrifugation has been used to purify supercoiled DNA in sufficient quantity for a sequencing project to be completed from a single isolation, but requires the use of concentrated ethidium bromide, a carcinogen. Our method of choice involves purification of supercoiled DNA on a diatomaceous earth matrix. This method is quicker than ultracentrifugation and does not require any hazardous materials.

The reagents required for this method are: two or four liters plus 50 mL and a 3mL aliquot of Lauria broth (LB), all containing the appropriate antibiotic to select for the clone of interest; 50 mM glucose, 25 mM Tris-HCl pH 8.0, 10mM EDTA (GET buffer); lysozyme; 200 mM NaOH, 1% sodium dodecylsulfate (SDS), (alkaline lysis buffer); 3M sodium acetate; 25 mg/mL RNase A; isopropanol; 100 mg/mL de-fined diatomaceous earth in 6M guanidine-HCl; 10mM:1mM Tris-HCl, pH 7.6, EDTA, 50% ethanol (wash buffer); acetone; and 10mM:1mM Tris-HCl, pH 7.6, EDTA (elution buffer).

A smear of several *E. coli* colonies was picked with a sterile toothpick, placed in three mL of LB, and grown for eight hours at 37°C with shaking at 250 RPM. The 3mL culture was decanted to the flask containing 50 mL of LB and grown for a further eight hours under the same conditions. Twelve and a half mL of the 50 mL culture was added to each liter of LB and again allowed to grow for eight hours at 37°C with shaking at 250 RPM. When the 24 hour growth phase was completed, the cells were harvested by centrifugation and the cell pellets combined to yield two centrifuge bottles each containing half of the cells. The cell pellets could be frozen at -70°C at this point.

The DNA was isolated from the cells by first gently resuspending each pellet in 35 mL of GET buffer by teasing the pellet off the side of the centrifuge bottle and swirling the buffer. A vigorous manner should be avoided at this point to avoid shearing the *E. coli* genomic DNA. Once the pellets were resuspended completely, lysozyme was added to a final concentration of 2 mg/mL and allowed to incubate at room temperature for 5 min. Cell lysis was achieved by adding and mixing 70 mL of alkaline lysis buffer to each bottle followed by incubation for an additional ten minutes at room temperature. The cellular proteins, membranes and genomic DNA were precipitated by adding 52.5 mL of 3M sodium acetate to each bottle, gently mixing the solutions, and incubating in ice water for thirty to sixty minutes. The bulk of the precipitate was cleared by filtration through a double layer of cheesecloth into clean bottles.

The filtrate was centrifuged to clear any small particulates that may not have been trapped by the cheesecloth. The supernatant was decanted to clean bottles and the RNA digested by adding RNase A to a final concentration of 40 µg/mL and incubation at 37°C for thirty minutes. The DNA was precipitated by adding an equal volume of isopropanol,

incubating the solution at room temperature and centrifugation. The resulting DNA pellet was resuspended on 10mM:1mM Tris-HCl:EDTA. De-fined diatomaceous earth (DE) was added and the nucleic acids allowed to bind by incubation at room temperature for five minutes with occasional mixing. The mixture was centrifuged to pellet the DE, the supernatant decanted, and the pellet washed by resuspending the pellet in wash buffer to remove the residual guanidine-HCl. The DE was again pelleted by centrifugation, the supernatant decanted, the pellet resuspended in acetone, and the mixture pelleted a third time by centrifugation. The acetone was decanted from the DE pellet and the pellet dried in a vacuum oven.

The dry pellet was resuspended in 10mM:1mM Tris-HCl:EDTA and the DNA eluted by incubation at 65°C with mixing for ten minutes. The mixture was centrifuged to separate the suspended DNA from the DE and the supernatant transferred to clean bottles. A second centrifugation may be required to remove residual DE. The DNA of interest was ethanol precipitated from the supernatant by adding 2.5 volumes of 95% ethanol containing 0.12 M sodium acetate. The mixture was incubated at -20°C for two hours, centrifuged to pellet the DNA, the DNA pellet washed with 80% ethanol, centrifuged, and the pellet dried. The DNA pellet(s) was resuspended in one milliliter of 10mM:0.1mM Tris-HCl:EDTA for every liter of original culture and assayed for yield and purity by agarose gel electrophoresis.

Physical Shearing, End Repair, Size Fractionation, and Ligation

Several methods have been developed to prepare the random fragments necessary for a shotgun DNA sequencing strategy, including the French Press, sonication, exonuclease digestion, and nebulization. Sonication and exonuclease digestion require large amounts of DNA for the reactions to optimize conditions and generate fragments of

the proper size. In addition, exonuclease digestion is laborious and time consuming. On the other hand, sonication preferentially cleaves DNA in A/T rich regions. Nebulization has proved to reproducibly generate random fragments regardless of clone size or G/C content without the need for laborious and wasteful optimization reactions. Furthermore, nebulization yields a higher percentage of fragments that are more nearly blunt ended on both ends resulting in more efficient end repair and ligation reactions.

Fifty µg of DNA was placed in a large Falcon tube with 500 µL of sterile glycerol, 200 µL of 10X TM buffer (500 mM Tris-HCl, pH 8.0, 150 mM $MgCl_2$) and a quantity of water sufficient for a total volume of 2000 µL. The contents were mixed by capping the tube and inverting the tube until the mixture appeared homogeneous. The mixture was then pipetted to the cup of a nebulizer (IPI Medical Products, Inc., Akron, OH) which had been altered such that the rim of the inner cone of the nebulizer had been cut off and the cone inverted to improve the recovery and turnover of sample within the nebulizer during use. The nipple in the lid of the nebulizer for the breathing apparatus was plugged with a polypropylene cap and the lid screwed to the nebulizer. The nebulizer then was connected to a high-pressure nitrogen gas source that had been adjusted to the desired gas pressure, typically 4 - 10 psi. for cosmids or larger clones. The gas was allowed to flow through the nebulizer for two and one half minutes, and the nebulizer briefly centrifuged to collect the solution at the bottom of the cup. The DNA then was precipitated by splitting the DNA into 400 µL aliquots and adding 1000 µL of cold 95% ethanol, 0.12M sodium acetate and 1 µL of glycogen, 20 mg/mL (Boerhinger-Mannheim) as a carrier. Because the nebulizers do not seal completely, 3-500 µL of liquid was typically lost during the nebulization process with

the result that the precipitation could be performed in four tubes. The DNA pellets were washed once with 500 µL of 80% ethanol and dried.

The sheared DNA fragments were pooled into two tubes by resuspending the dry DNA pellets in two aliquots of 35 µL. The end repair reactions were carried out by adding five µL of the polynucleotide kinase buffer supplied by the manufacturer (Boehringer-Mannheim), 1 µL 10mM rATP (Pharmacia), 7 µL of a mixture containing 0.2mM each dNTP (Pharmacia), 3 µL T4 DNA polymerase (3U/µL, New England Biolabs), 2 µL Klenow fragment of *E. coli* DNA polymerase 1 (5U/µL, New England Biolabs), 1 µL T4 polynucleotide kinase. The reactions were incubated for thirty minutes at 37°C and loaded onto a low melting point agarose gel for size fractionation of the sheared and end repaired fragments. Lambda cut with *Hind*III and phi-X 174 cut with *Hae*III were used as size markers.

Fragments between 1.3 and 4.0 kilobasepairs (Kb) were excised from the gel and the DNA fragments recovered by phenol extraction. The gel pieces were heated in a 70°C water bath to melt the agarose. An equal volume of phenol equilibrated against 10mM Tris-HCl pH 7.6 was added to the melted gel pieces, vortexed and centrifuged in a microcentrifuge for five minutes. The upper aqueous layers were removed to a clean microcentrifuge tubes and an equal volume of 1:1 phenol:chloroform was added to each tube. The tubes were vortexed, and centrifuged as above. The aqueous layers were again removed to clean tubes, an equal volume of chloroform added, the tubes shaken vigorously by hand, and centrifuged briefly to separate the phases. The upper layer was again pipetted to clean tubes and ethanol precipitated by adding 2.5 volumes of cold 95% ethanol, 0.12M sodium acetate. The DNA pellets were washed once with 80% ethanol and dried.

Typically, the DNA pellets were resuspended in a total of 20 µL of 10:0.1 Tris-HCl, pH 7.6:EDTA. Ligation reactions were performed on newly nebulized DNA using a range of DNA volumes to optimize the ligation reaction efficiency. One half, one, and two microliters of DNA were added to one microliter of ligase buffer containing 10mM rATP as supplied by the manufacturer (New England Biolabs), two microliters pUC18 cut with *SmaI* (Pharmacia), one microliter T4 DNA ligase (New England Biolabs), and a volume of sterile water sufficient for a total reaction volume of ten microliters. If initial reactions failed to yield a sufficient number of colonies, additional reactions were performed using other DNA volumes, or polyethylene glycol to a final concentration of five percent was included in the reactions. Future ligation reactions using the same batch of nebulized DNA would be performed using only the optimized DNA volume. Two sets of conditions were routinely used for ligation reactions with equal success, three hours at room temperature or overnight at 4°C.

Recombinant DNA molecules generated by the ligation reactions were amplified by electroporation into XL1-Blue MRF' cells and the cells spread on Lauria broth (LB) agar plates supplemented with 100 µg/mL ampicillin. Cells competent for electroporation were made by spreading cells from a partially thawed glycerol stock on LB plates supplemented with 20 µg/mL tetracycline. Colonies from the plate grown overnight were used to start a 3 mL culture which also was grown overnight. The 3 mL culture then was used to inoculate one liter of LB. This culture was grown until the A600 was between 0.4 and 0.5. The cells were collected by centrifugation and the cell pellet washed twice by resuspension in cold sterile water and centrifugation. The cells were washed once in 10% glycerol, and finally, the cells from one liter of culture were

resuspended in 4 mL 10 percent glycerol. The cells then were divided into 40µL aliquots and, at this point, could either be used immediately or frozen at -70°C.

Alternatively, competent cells could be prepared by growing a three hour culture at 37°C with shaking at 250RPM from a 1.3mL glycerol stock. The cells were recovered by centrifugation, the media decanted and the cells resuspended in 20mL of cold 50mM $CaCl_2$ and incubated for 20 minutes on ice. Following the incubation, the cells were recovered by centrifugation, the supernatant decanted, and the cells resuspended in 4mL cold 50mM $CaCl_2$. This method is less efficient than electroporation and, for the most part, has been discontinued.

An *E. Coli* Pulser (Bio-Rad) was used to electroporate competent cells. The unit was set for a 2500 kV pulse, which typically lasted five microseconds. Two microliters of the ligation mix was pipetted into a microcentrifuge tube containing 40 µL of competent cells that had been thawed on ice. The suspension was transferred to an electroporation chamber, the pulse applied, and 1mL of YENB medium (7.5gm yeast extract, 15gm Bacto-tryptone in 1L) quickly added to the chamber. The contents of the chamber then were pipetted to a small Falcon tube and were incubated at 37°C with shaking at 250 RPM for 15 minutes. The cell suspension then was briefly centrifuged to recover the cells, the media decanted, and the cells resuspended in 200 µL of fresh YENB, 25 µL isopropyl-thiogalactoside (IPTG, 25 mg/mL), and 25 µL 5-bromo-4-chloro-3-indolyl g-D-galactosidase (X-gal, 25 mg/mL in N,N-dimethylformamide). The resulting cell suspension was spread on LB plates supplemented with ampicillin (100 µg/mL), and incubated for 16 to 20 hours at 37°C.

Semi-Automated Isolation of Plasmid Subclones

One of the challenges of the Human Genome Project has been the automation of as many processes as possible. Success on this front has been achieved in many of the procedures used in the large scale sequencing of DNA including the isolation of the hundreds to thousands of subclones needed to complete a single project. The following protocol is an adaptation of the alkaline lysis plasmid isolation method that allows for the isolation of 384 subclones in five hours. The following reagents are needed for the isolation of pUC18 subclones on the Biomek2000 (Beckman Instruments, Inc.) workstation: Terrific broth consisting of 12 g Bacto-tryptone, 24 g yeast extract, and 4 mL glycerol in 900 mL water, and 2.31 g KH_2PO_4, 12.54 K_2HPO_4 in 100 mL water, autoclaved separately and mixed when cool; ampicillin (20mg/mL); 50:10 Tris-HCl, pH 7.6:EDTA, 40µg/mL RNaseA, (TE-RNaseA); 0.2 N sodium hydroxide, 1% SDS (alkaline lysis solution); 3M sodium or potassium acetate; 95% or 100% ethanol; and 80% ethanol.

Colonies grown overnight were picked with a sterile toothpick and placed in 1.5 mL of TB supplemented with 100 µg/mL ampicillin which was aliquoted to each well of a 96-well block (Beckman, 14504). Four blocks, 384 colonies, were typically isolated together. After the colonies were picked, the toothpicks were removed from the blocks and the blocks covered with a polystyrene lid (US Plastics, US10). The blocks then were placed in a shaking incubator and incubated for 21 - 22 hours at 37°C with shaking at 350 RPM. At the end of the incubation time, the blocks were centrifuged to collect the cells containing the plasmid subclones at the bottom of the blocks and the media decanted. The blocks were placed on the tablet of the Biomek2000 workstation according to the designed layout for this procedure. The workstation pipetted 200 µL of TE-RNaseA to each well of the blocks,

resuspending the cell pellets by aspirating and dispensing the solution 20 times. Cell lysis was achieved by the addition of 200μL of alkaline lysis solution to each well followed by mixing of the solutions by aspirating and dispensing the solution 10 times. Two hundred microliters of 3M sodium acetate then was added to each well to precipitate the cellular components. The sodium acetate was mixed manually by vortexing the blocks and the blocks then were placed in a 37°C shaking incubator for shaking at 250 - 350 RPM for 15 minutes to thoroughly mix the solutions and begin pelleting the precipitate in the center of the wells. After shaking, the blocks were placed at -20°C for one half hour to completely precipitate the cellular components. The precipitate was collected by centrifugation of the blocks, and the blocks placed back on the Biomek2000 where 400μL of the supernatant was transferred to clean 96-well blocks. The plasmid DNA was precipitated by addition of 1000μL of 95% ethanol 0.12M sodium acetate followed by incubation at -20°C for at least one hour. The DNA was pelleted by centrifugation and washed with 500μL of 80% ethanol. Yield and purity were assayed by agarose gel electrophoresis of representative random samples from each block.

Cycle Sequencing Reactions

The primary advantage of the ABI automated fluorescent sequencing machines is the ability to collect all the data from one sample in one lane of a gel rather than needing four lanes, one for each base, to collect the data from one sample. This is possible because the ABI machines detect sample based on fluorescent dye labels attached to the newly synthesized single stranded DNA product of the cycle sequencing reaction, which is an outgrowth of the Sanger dideoxy method. Further efficiency has been realized as a result of the dyes being attached to the terminating base rather than the primer, or dye primer

reactions. Dye primer reactions require that four reactions be performed, using the same template, in separate tubes. Each reaction includes the dye specific for the single dideoxynucleotide included in the reaction attached to the primer. After the reaction components have been incubated under the conditions necessary for the enzymatic reaction to proceed to sufficient yield to be detectable, the four reactions are pooled, precipitated and loaded on the sequencing gel. Therefore, a 96-well reaction plate is only sufficient for reactions from 24 samples.

On the other hand, dye terminators have the dye attached to the terminating base at the 2' carbon. One reaction per sample with a mixture of the labeled terminating bases is sufficient to generate the complete sequence because each newly synthesized strand is terminated with the labeled moiety.

The ability to isolate subclones in a 96-well format and subsequently perform a single reaction per clone using dye terminators in a 96-well format and the recent development of a microtiter plate format 384 well plate (Cycleplate384®, Robbins Scientific) for the Perkin-Elmer 9600 thermocycler lead us to develop a semi-automated protocol with which we could prepare 384 cycle sequencing reactions in approximately one hour. The Cycleplate384® has 96 wells divided into four quadrants by molded dividers. Typically, both the forward and reverse reactions from four blocks of isolated pUC subclones were prepared at one time using the Hydra96 from Robbins Scientific.

A volume of DNA sufficient to yield one milligram per reaction was aspirated from the first block by the Hydra96 and then dispensed to the same quadrant of each well of two different Cycleplate384®'s. The syringes of the Hydra96 were washed using double distilled water. The Hydra96 then was used to transfer DNA from the second block to the next

quadrant of the Cycleplate384®'s and the syringes washed as before. The process was repeated until DNA from four blocks had been pipetted to the respective wells of two Cycleplate384®'s.

The remaining reaction components consisted of 13uM forward or reverse primer, 150mM each of dATP, dCTP, and dTTP, 450μM dITP, 0.1μL per reaction of a 1:30 dilution from the PE/ABI PRISM Ready Reaction Dyedeoxy terminator kit (ABI, P/N 401384) as a source for dye-labeled ddNTPs, 1mM manganese sulfate, reaction buffer as supplied by the manufacturer (ABI), 10 μL KlentaqTR (1,000U/mL, Wayne Barnes), and sterile water sufficient to yield a final reaction volume of 20 μL for each of 384 reactions. The above components are mixed and a quantity sufficient for four reactions was aliquoted to each of 96 wells of a 96-well Cycleplate (Robbins Scientific) and the Cycleplate placed on the Hydra96. The entire volume of each well was aspirated and the correct volume of reaction components for one reaction dispensed to the side of each quadrant of each well of one of the Cycleplate384®'s containing the DNA to be amplified. An almost identical set of reaction components was prepared, the exception being that the primer was the one not previously used. The second set of reaction components was added to the second Cycleplate384® and both sets of samples briefly centrifuged to collect and mix all the reaction components. The reactions were incubated in Perkin-Elmer9600 thermocyclers using a 2-step cycling program consisting of a denaturing step at 95°C for five seconds and a primer annealing/extension step at 60°C for 150 seconds.

Closure reactions were performed with the TaqFS kit (ABI) on shotgun clones using universal primers to extend gel readings or performed either on shotgun clones or directly on

cosmid DNA with custom synthesized primers. Reactions using TaqFS enzyme were prepared in a manner similar to the KlentaqTR reactions with the following changes: DNA quantity was decreased to 100 ng; the primer concentration was decreased to 6.5 mM; two microliters of the TaqFS mix as furnished by the manufacturer (ABI) and containing all the reaction components except DNA and primer were used; and the reaction volume was decreased to 5μL. In addition, dimethyl sulfoxide was often included at concentrations up to 10%. Incubation conditions were as recommended by the manufacturer.

Following the incubation of the reactions, unincorporated terminators and buffer salts were removed by gel filtration through columns made with Sephadex G-50 medium (Pharmacia) in a 96-well filter plate (Millipore, MADVN6550). This procedure, first developed at The Washington University Genome Center, involved manually pipetting 400μL of hydrated column matrix to each well of a filter plate. One filter plate was needed for each set of 96 reactions performed. The filter plates then were taped on top of microtiter plates and centrifuged for 3-5 minutes at 1500-2000 RPM. The completed reactions were briefly centrifuged to collect condensation from the sides of the reaction vessel and then the first set of 96 reactions aspirated by the Hydra96. The reactions then were dispensed to the tops of the first filter plate and the syringes washed as above. The process was repeated until all the sets of 96 reactions were transferred from the Cycleplate384®'s to the tops of the columns in each respective filter plate. The filter plates were taped to the tops of clean V-bottom microtiter plates and the purified reaction products collected in the microtiter plate by centrifugation as above. The microtiter plates containing the reaction products were dried under vacuum.

Sample Loading and Automated Data Collection

Dried samples were resuspended in 1μL of a loading buffer consisting of 25μL of loading dye and 100 μL deionized formamide. Loading dye was 25mM EDTA, and 50 mg/mL Blue Dextran. Samples were heated for three minutes at 100°C and loaded on polyacrylamide sequencing gels. During the shotgun phase of a project, 5.0% polyacrylamide, 8.0M urea, 36cm short gels typically were used, whereas closure reactions were electrophoresed on 5.3% polyacrylamide, 8.0M urea, 48cm long gels for better sample separation and longer read lengths. Short gels were electrophoresed for three hours at 3kV. Long gels were electrophoresed for ten hours at 1kV.

Data was automatically collected and analyzed by the Data Collection and Data Analysis software that is included with the ABI373 and ABI377 sequencing machines. Analyzed data was transferred to a Unix platform LAN for project assembly and analysis. Automated assembly was performed using the Phred and Phrap companion programs from Phil Green at the University of Washington. At the beginning of a project, assembly parameters were optimized to generate the most accurate assembly and facilitate the gathering of preliminary sequence analysis information through database searches and homology to other known sequence.

Analysis of finished sequence data was performed by homology searches of GenBank using the Basis Local Alignment Search Tool (208) algorithm through the PowerBLAST interface. PowerBLAST was written by Jinghui Zhang and is available by anonymous FTP from the National Center for Biotechnology Computing. PowerBLAST first performed a search of the repeats database, screened homologies to that database from the sequence of interest, and then executed a GenBank homology search. The output was

returned in three formats, a graphical interface file, an alignment file, and an ASN.1 format file. The graphical interface allowed the GenBank entries to homologous sequences to be called up to the screen of the workstation. The alignment file showed the alignment of the sequence of interest to the GenBank homologous sequence.

Exon prediction was accomplished with XGRAIL, Xwindows interface GRAIL, and GENSCAN.

Polymerase Chain Reaction (PCR)

PCR reactions were performed starting from 10 to 20 ng of the DNA to be amplified. Reaction components were 100 picomoles of each primer, 20 picomoles of each deoxynucleotide, 50mM KCl, 10mM Tris-HCl, pH8.5, 1.5mM $MgCl_2$, and 5U Taq polymerase. Reaction conditions typically were 25 to 30 cycles of template denaturation at 95°C for 15 seconds, primer annealing at 55 to 65°C for 15 to 30 seconds, and extension at 72°C for 60 seconds. Regions that proved difficult to amplify under these conditions could usually be amplified by varying the magnesium concentration, or including single long denaturation or extension steps at either the beginning or end, respectively, of the reactions. Reaction products were visualized by agarose gel electrophoresis on 0.7% agarose gels stained with ethidium bromide. Purification of reaction products was performed by electrophoresis through low-melting point agarose, excision of bands and recovery by phenol extraction and ethanol precipitation.

Primer sequences for both PCR and gap closure were determined with the aid of the Oligo Selection Program. Primers were synthesized on Beckman 1000M eight column oligo synthesizers. Cleavage from the columns was accomplished by adding 200μL of AMA (Beckman) to a screwcap microcentrifuge tube, screwing the column to the tube and

affixing a 10mL syringe to the top of the column. All joints were sealed with Parafilm. The tubes were placed in a 70°C waterbath for ten minutes, and then allowed to cool. The apparatus then was disassembled, and the primer eluted from the columns by affixing a 1mL syringe containing 0.5mL of sterile water to the tops of the columns. The columns were inverted so the top of the syringe was down, and enough water to fill the column dispensed from the syringe. The columns were incubated for three minutes then clean screwcap microcentrifuge tubes were affixed to the columns. The columns were inverted so the tubes were down, and the columns were flushed with the remaining water from the syringe and the water containing the primer collected in the tubes. Primer concentrations were determined by ultraviolet absorption at 260nm (A_{260}). Primer stocks of 40μM and 13μM were made from the original stock based on calculations that relied on a conversion factor of 37μg per A_{260} unit.

Generation and Large Scale Isolation of cosB1 Subclones in M13mp18

Ten micrograms of the Alu sequence enriched cosmid clone B1 was restriction digested with 10U of *Bam*HI and *Bgl*II in separate reactions. Fragments were gel purified by electrophoresis through low-melting point agarose. Bands of interest were recovered from the agarose by phenol extraction and concentrated by ethanol precipitation. M13Rf was restriction digested with *BamHI* and digestion efficiency initially determined by agarose gel electrophoresis using uncut M13Rf as a control. A more sensitive control was performed during the competent cell transformations below. Calf intestinal phosphatase was used to remove the 5' phosphate groups from the vector.

Ligation reactions to clone the B1 fragments into the M13 cloning vector were performed in a manner similar to that detailed above. A negative control reaction that

included no insert DNA was performed to determine the efficiency of both the restriction digest and the phosphatase reaction. Ligation reactions were exposed to JM101 cells made competent with $CaCl_2$ as detailed above and the cells spread on lambda plates containing no antibiotic. Using a sterile toothpick, clear plaques from the reactions that included B1 DNA were picked into 1.5mL of 2XTY containing a one hour culture of JM101 in a small Falcon tube. One hour cultures of JM101 cells were made by thawing a 1.3mL glycerol stock of cells and decanting it into 50mL of 2XTY followed by a one hour incubation at 37°C. The 1.5mL culture containing phage from the plaque was incubated an additional 5.5 to 6 hours at 37°C with shaking at 250RPM.

The cell culture was centrifuged to pellet the cells. One milliliter of the supernatant was pipetted to a clean microcentrifuge tube and 200µL of 20% polyethylene glycol 8000, 2.5M NaCl (PEG/NaCl) was added, vortexed to mix, and incubated at room temperature for 15 to 30 minutes. The tubes were centrifuged to pellet the phage and the supernatant decanted. Any remaining drops of the supernatant were removed by vacuum aspiration through a finely drawn Pasteur pipette. The pellet was resuspended in 100µL of Tris-HCl, pH7.6 by vortexing. The suspension was extracted by adding 50 µL 10mM Tris-HCl, pH8.0, 1mM EDTA saturated phenol (TE saturated phenol) vortexing to mix and centrifuged to separate the layers. The top, aqueous, layer was removed to a fresh tube and extracted with 50 µL of a 1:1 mixture of TE saturated phenol:chloroform. The top layer was again removed to a fresh tube and extracted a third time with 50 µL of chloroform. The aqueous layer was pipetted to a fresh tube and ethanol precipitated by addition of 2.5 volumes of 95% ethanol containing 0.12M sodium acetate, incubated on ice for 15 minutes and centrifuged to pellet the single stranded phage DNA. Yield and purity were determined

by agarose gel electrophoresis. Clones were sequenced with forward primer to determine their location within the B1 clone.

Two of the recombinant clones were determined to have large inserts of greater than 4kb by agarose gel electrophoresis and by the position within the known cosmid sequence of the end sequence. Single and double stranded DNA from these two recombinant clones and a non-recombinant clone were isolated from a two liter culture to generate sufficient DNA for protein affinity studies. Detailed below is the method to isolate a single clone.

A two hour culture of JM101 cells were grown from 1.3mL of a thawed glycerol stock incubated in 50mL of 2XTY without shaking. A sterile toothpick was used to pick single plaques of each clone into two tubes, each containing 1.5mL of the two hour culture of JM101 cells, and allowed to grow at 37°C for eight hours with shaking at 250RPM. Concurrently, from the same two hour culture, four 50mL cell cultures were started in 2XTY, each from 1.5mL of the original culture. Two of these were grown for two hours and placed in the cold room, the other two were incubated for eight hours at 37°C with shaking at 250RPM. At the end of the eight hour incubation, the two hour 50mL cultures were taken from the cold room and the 1.5mL cultures containing the phage plaque were added to them. This cell culture was placed in a 37°C incubator and allowed to grow for eight hours with shaking at 250RPM. The untransfected eight hour 50mL cultures were used to inoculate 1L each of sterile 2XTY and incubated with shaking at 250 RPM two hours at 37°C and then placed in the cold room. When the transfected 50mL cell cultures had grown for eight hours, the 1L cultures were taken from the cold room, and each was inoculated with 50mL of the transfected cells that had been growing for eight hours. The

newly inoculated 1L cultures were allowed to grow at 37°C for an additional eight hours with shaking at 250RPM.

The cells were pelleted by centrifugation and saved for isolation of the double stranded Rf clone. The single stranded DNA containing phage were precipitated by addition of one fifth volume PEG/NaCl and room temperature incubation for 30 minutes. The phage were recovered by centrifugation and the supernatant discarded. The phage pellet was resuspended in 105mL of 10mM Tris-HCl, pH7.6, 0.1mM EDTA. The suspension was successively extracted with 50mL of TE saturated phenol, 1:1 TE saturated phenol:chloroform, and chloroform. The single stranded phage DNA was precipitated with 2.5 volumes of 95% ethanol containing 0.12M sodium acetate, washed with 20mL 80% ethanol and assayed for yield by A_{260}. Double stranded DNA was recovered from the cell pellet by means of the same protocol detailed above for the large-scale isolation of cosmid or BAC DNA.

Preparation of DNA-Cellulose Affinity Columns

Acid-washed cellulose was prepared and DNA from the m13mp18 clones bound to it by the method of Litman (190). Fifteen grams of Sigmacell® Type 50 cellulose (Sigma) was washed in 300mL of 1N HCl with stirring for 10 minutes. The cellulose was filtered on Whatman No. 1 filter paper in a Buchner funnel, recovered from the filter paper, and stirred again with 200mL of 1N HCl. After filtering, the cellulose was washed with sterile water until no trace of acid remained in the filtrate as assayed by pH paper. The cellulose was spread out in petri dishes to dry overnight. The product of this procedure is termed acid-washed cellulose.

Single stranded and double stranded DNA from the cosB1 subclones was diluted to 2mg/mL in 1mM NaCl. Six milliliters of the resulting solution was mixed with 0.75g of acid-washed cellulose in a 250mL beaker. A spatula was used to break up the cellulose and evenly mix the DNA/cellulose slurry. The slurry was spread over the sides of the beaker to dry by tipping the beaker and allowing the solution to run up the side. The dried cellulose was scraped from the side of the beaker, resuspended in 20mL of 100% ethanol, and transferred to a 150mL beaker. The slurry was mixed in the dark room and exposed to ultraviolet radiation from a Mineralight®UVS-11 (Ultra-Violet Products, Inc., San Gabriel, CA) for 15 minutes. The alcohol was removed from the cellulose by filtration through Whatman No. 1 filter paper in a Buchner funnel. The cellulose was washed with three successive rounds of mixing for 10 minutes in 1mM NaCl followed by filtration as above. DNA binding was assayed by measuring the absorbance at 260 nm of the filtrate from the washes. The product of this procedure was termed DNA-bound cellulose.

DNA-cellulose columns were prepared by resuspending 0.5g of DNA-bound cellulose in 20mL of 10mM phosphate buffer, pH6.8, and pouring the slurry into a 1.3cm column (part # PD-10, Pharmacia) that had been plugged with glass wool. The columns were washed with an additional 30mL of 10mM phosphate buffer, pH6.8, before use.

Isolation of Nuclear Extracts

A buffy coat was purchased from the Oklahoma Blood Institute. Cells from human cell lines were the gift of Charles Esmon at the Oklahoma Medical Research Foundation. The method of Dignam et al. (209) was used to isolate nuclear extracts from these tissues and cell lines. Three buffers are used in this procedure, Buffer A, Buffer B, and Buffer C. Buffer A was 10mM HEPES, pH7.9 at 4°C, 1.5mM $MgCl_2$, 10mM KCl, and 0.5mM

dithiothreitol (DTT). Buffer B was 300mM HEPES, pH7.9 at 4°C, 25% (v/v) glycerol, 420mM NaCl, 1.5mM $MgCl_2$, 0.2mM EDTA, 0.5mM phenylmethylsulfonyl fluoride (PMSF), and 0.5mM DTT. Buffer C contained 20mM HEPES, pH7.9 at 4°C, 20% (v/v) glycerol, 100mM KCl, 0.2mM EDTA, 0.5mM PMSF, and 0.5mM DTT. Buffer solutions were made fresh the day the isolation was performed.

Cells were thawed on ice in the cold room and suspended in five packed cell volumes phosphate buffered saline (210). Cells were collected by centrifugation at 2200 RPM for 10 minutes in a Beckman GS-6R centrifuge. Pellets were suspended in five packed cell volumes of Buffer A and allowed to stand for 10 minutes. Cells were collected by centrifugation as above, and the pellets suspended in two cell volumes Buffer A. The cells were homogenized in a Dounce homogenizer by ten strokes with pestle B. Nuclei were collected by centrifugation as above. The supernatant was collected, labeled as the crude S-100 fraction, and frozen at –80°C.

Nuclei were further centrifuged for 20 minutes in an SS-34 rotor at 14,500RPM and the supernatant discarded. Nuclei were suspended in an equal packed nuclei volume of Buffer B and lysed by ten strokes in a Dounce homogenizer with pestle B. The homogenate was gently stirred for thirty minutes and then centrifuged for thirty minutes in an SS-34 rotor at 14,500RPM to collect the nuclear membranes. The supernatant was dialyzed against 30 packed cell volumes of Buffer C for five hours, and frozen at –80°C before use. Protein yield was assessed by mixing 200 µL of Bradford reagent (Bio-Rad) with a suitable dilution of nuclear extract and measuring the absorbance after five minutes at 595 nm. The IgG standard included in the kit by the manufacturer was used to generate a standard curve.

Column Chromatography of Nuclear Extracts on DNA-Cellulose Columns

The height of the columns including the glass wool stoppers was measured and the column volumes calculated. To avoid overloading the columns, no more than 500μg of extracts was applied to the columns. The columns were allowed to flow until all the protein had entered the columns. The flow then was shut off and the columns allowed to incubate for 15 minutes (190). The flow through was eluted from the columns with 20mM HEPES, pH7.9, 100mM KCl. Step gradient elutions with increasing salt concentrations were performed in five fractions, one half column volume per fraction, for each salt concentration used. Salt concentrations of 300mM, 600mM, 900mM and 1.2M were buffered in 20mM HEPES were used to elute protein from the columns. Protein concentrations in the eluates were determined by reaction with Bradford reagent according to the manufacturers instructions. Proteins binding to the columns were visualized by silver stained (Bio-Rad) SDS-PAGE (210) in 12% gels against mid-range markers (Promega).

Results

9p21, The Cyclin-dependent Kinase Inhibitor Genes

During the course of this dissertation, four cosmids from human chromosome band 9p21 were sequenced. Cosmid c66 includes the p16INK4a gene; cosmid c86 includes the coding sequence for p15INK4b. The other two cosmids, c20 and c110, include no known gene(s).

C66

Cosmid c66 has a total insert size of 34,678bp with a G+C content of 42.2%. This cosmid was entirely sequenced using pUC18 as the sequencing vector. Four hundred subclones were isolated, of which approximately 250 were of sufficient concentration to use in Taq DNA polymerase sequencing reactions with ABI DYE-Prism Terminators.

The p16INK4a gene is entirely within c66

The coding sequence of the entire p16INK4a gene is included within the cosmid c66 insert as indicated in the schematic sequence reproduction shown in Figure 3. The output from the alignment program Crossmatch is shown in Table 1.

Table 1. Crossmatch alignment of c66 with the three exons of p16INK4a

Score	% Sub.	% Del.	% Ins.	Query	Begin	End	(Rem.)	subject	Begin	End	(Rem.)
291	1.76	0	0	C66.fa	5820	6159	28519	p16exon1	1	340	0
550	1.2	0	0	C66.fa	9400	9984	24694	p16exon2	1	585	0
386	2.15	0.24	0	C66.fa	12248	12665	22013	p16exon3	4	422	0

The first column is the score of the alignment, a higher score indicates a more identical and longer alignment between the two sequences. The next three columns give the discrepancies expressed as the percent substitutions, deletions, and insertions respectively. The fifth column is the name of the query sequence and the sixth and seventh columns show the positions within the query sequence where the homology with the subject or library sequence begins or ends, respectively. The number in the eighth column indicates the number of bases to the end of the query file beyond the end of the homologous sequence, or the (rem)ainder. The ninth column is the name of the subject sequence. The last three columns are analogous to the similarly labeled columns discussed above but pertain to the subject sequence. If the homology were to the opposite strand of the subject sequence, a "C" would appear after the name of the subject sequence indicating a match to the complementary strand of the subject sequence.

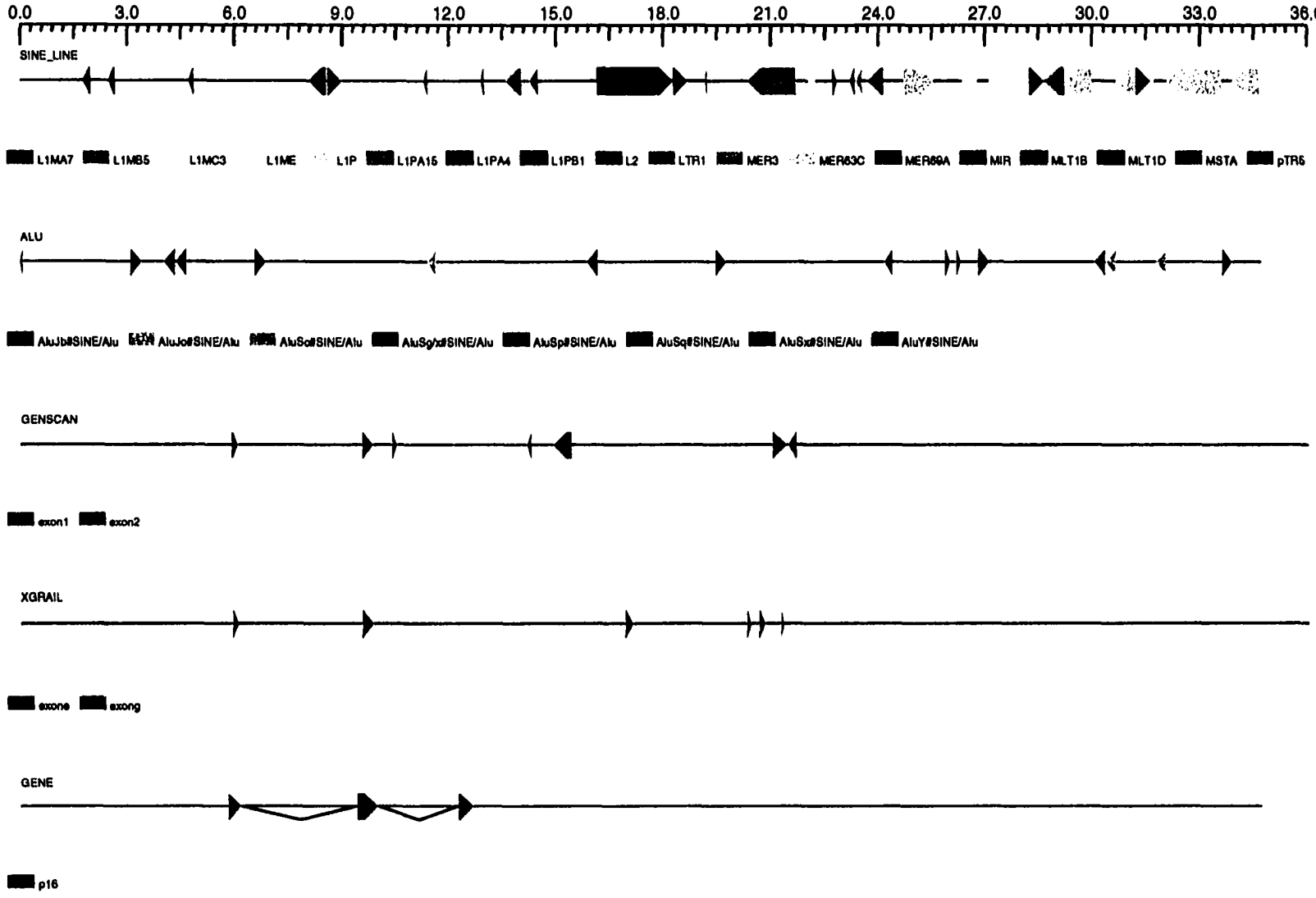


Figure 3. Sequence Feature Map of c66. Legend of Features: SINE_LINE - all LINE and medium frequency repeat elements; ALU - all repeat elements from the Alu families; GENSCAN and GRAIL - predicted exons from the respective exon prediction programs; GENE - alignment with known genes.

Inspection of Table 1 reveals that the p16INK4a gene is in the same orientation as c66 because the numbers in the "Begin" column of the subject sequence are lower than the numbers in the "End" column. These exon sequences from GenBank include flanking intron sequence. There are no insertions or deletions in c66 when compared to the first two exons of p16INK4a. On the other hand, there is a single deletion in c66 relative to p16exon3 and homology of c66 to exon 3 was found beginning at position 4 of the cDNA sequence. Thus, there are two deletions totaling four nucleotides in c66 relative to the published sequence. The beginning of the open reading frame in the p16exon3 sequence is at position 273 and the single base deletion in c66 is at position 30 of the p16exon3 sequence. Therefore, both mutations are outside the p16INK4a open reading frame and have no effect on the coding region. Note that the exon sequence used in this analysis includes sequence flanking the true exons.

The sequence alignment in Table 1 was performed with unedited c66 sequence. The substitution rate seen in the second column of Table 1 is due to a well documented sequencing artifact where the first C following a G is dropped leaving a gap in the gel reading. This artifact is only seen in reactions performed with Taq polymerase and the new mutant TaqFS™ enzyme and others containing the Tabor-Richardson mutation (211) have remedied this problem. The result of the dropped bases is that background signal is often incorrectly interpreted leading to incorrect base calls. These types of sequencing anomalies can be remedied by "double stranding". Therefore, cosmid c66 is double stranded over 99% of it's length and visual inspection of the gel readings shows that both the sequence determined here and the published exon sequence agree in all but one case, where there is a C to T transition at position 9726. This mutation changes codon 100 of

the p16INK4a coding sequence from alanine to valine, a presumably harmless polymorphism. However, it has been observed that a tryptophan to glycine mutation occurs at codon 101 in melanoma patients (12) and thus the protein does not bind CDK4 with the same efficiency as wild-type (11).

Crossmatch was used to look for the sequence of the first $p19^{ARF}$ exon of c66. On the DNA level, the $p19^{ARF}$ homology to p16INK4a extends from 229 to 447 of $p19^{ARF}$. Homology to c66 starts at 230 and ends at 452 leading to the conclusion that the alternative first exon of $p19^{ARF}$ is not found within c66.

Careful comparison of the DNA genomic DNA sequence and the known protein sequence allows the precise determination of the 5' and 3' splice sites. The first codon begins at c66 position 5924 and the first exon ends at c66 position 6073. The second exon begins at c66 position 9550 and ends at 9856. The last exon begins at c66 position 12517 and the stop codon starts at 12528. The consensus mRNA splice sites are conserved as a GT is the first two bases of both introns and an AG is the last two bases of the introns.

Both GRAIL and GENSCAN were used to analyze c66 for potential exons (Figure 3). Table 2 shows the position and quality of predicted exons from GRAIL and the genes predicted by GENSCAN from the c66 sequence.

Table 2. GRAIL Predicted Exons and GeneScan Genes

GRAIL				GeneScan					
Begin	End	Qual.	Str.	Begin	End	Str.	Begin	End	Str.
5924	6073	gd	f	5924	6073	f	Gene 1		
9550	9856	ex	f	9550	9856	f	14899	15413	r
16902	17121	ex	f	10392	10519	f	14163	14277	r
20300	20409	gd	f				Gene 2		
20649	20813	ex	f				21453	21674	r
21246	21321	gd	f				21421	21014	r

The most notable finding is that both programs accurately predict the first two exons of the p16INK4a gene but neither predicts the third exon probably because this exon only encodes four amino acids. In addition, GENSCAN predicts two genes on the reverse strand of c66 but BLASTP using the predicted amino acid sequence of these genes does not reveal any significant homology to any known gene or EST and these predicted genes are probably not real.

Transcription regulation of p16INK4a

Six thousand bases, including 2819 nt upstream of the beginning of the p16INK4a cDNA sequence were searched for the p16INK4a promoter. Careful examination of the output from NNPP reveals that because this program relies heavily on TBP binding sites to predict promoters, two nearby plus strand promoters are predicted. The predicted promoter would begin transcription at position 3344 and the second predicted promoter would begin transcription at position 4382. TATA sequences are found ~30 nt upstream from both start sites. Likewise, TSSG predicts two promoters with one would transcription start at position 5507 and the other at position 7807. The second predicted promoter can be ruled out because it is beyond position 5820, the beginning of the known cDNA sequence, even though a CpG island is detected by GRAIL between positions 5819 and 6208 of c66. A second CpG island is found spanning positions 9555 to 9843.

TBP or TFIID sites are predicted by TESS at positions 3834, 4843, and 5224, whereas MatInspector predicts one TBP site at position 4806. TESS predicts an octamer (Oct) site 40 nt upstream of the position 3834 TBP site. The predicted TBP site at position 4843 is predicted to have a Sp1 site immediately upstream and adjoining the TBP site. The TFIID site predicted at position 5224 is predicted to have a Sp1 site 45 nt

upstream, a glucocorticoid receptor site 100 nt upstream and a myogenin site 110 nt upstream. A Sp1 site is predicted at position 5796. However, no TBP or TFIID site is predicted downstream from this site. Nonetheless, the location of this predicted binding site for a general transcription factor is attractive considering that the beginning of the published cDNA sequence is position 5820. A glucocorticoid receptor site (GR) is predicted 13 nt upstream, a nuclear factor-1 (Nf-1) site is predicted 46 nt upstream, and a nuclear factor-kappaB (NF-κB) site is predicted 194 nt upstream. Transcription factors predicted by MatInspector to bind upstream of the MatInspector predicted TBP site at position 4806 are NF-1 at position 4746, hepatic NF-1 (Hnf-1) at position 4729, GATA1 at position 4710, MyoD and E47 at position 4627, and Ikaros-2 (Ik-2) at position 4619. Table 3 tabulates the results from FastM predicting protein interactions within the p16INK4a promoter. Protein interactions with TBP are predicted in the same general regions predicted by NNPP to be promoters, approximately positions 3840 and 4840. Interestingly, TESS predicts a TBP site for both regions and a TFIID binding site in the region around position 4840. MatInspector predicts a TBP site at position 4806 and predicts binding sites for E47, MyoD, Ik2, CAAT, GATA1, and Nf-1 upstream of the TBP site. Using the same Transfac database of transcription factors, FastM predicts a TBP site at the same location and finds potential binding site interactions with MyoD, Ik-2, E47, CAAT, and GATA1. Furthermore, the strength of the protein interactions in this region is predicted by FastM to be stronger than interactions in the other two regions. FastM predicts interactions of a similar strength with TBP bound to a site at position

3583, but none of the other promoter prediction programs places a promoter in this

Table 3A. FastM Predicted Protein Interactions with TBP in the p16INK4a Promoter.

TBP Position	Upstream Protein	Position	FastM Score (out of 2.0)	TBP Position	Upstream Protein	Position	FastM Score (out of 2.0)
3482	MyoD	3234	1.70	3583	GC, Sp1	3294	1.80
					CAAT	3410	1.78
					GATA1	3523	1.80
					Average Score		1.79
TBP Position	Upstream Protein	Position	FastM Score (out of 2.0)	4291	E47	4248	1.67
3830	GATA1	3703	1.67		MyoD	4249	1.72
	CAAT	3769	1.66		Average Score		1.70
	MyoD	3788	1.66				
	Average Score		1.66				
4806	MyoD	4584	1.83	5073	CAAT	5012	1.69
	IK-2	4619	1.89				
	E47	4626	1.82				
	CAAT	4691	1.79				
	GATA1	4710	1.86				
	Hnf3b	4729	1.82				
	Average Score		1.84				

Proteins with no FastM predicted interaction with TBP: Nf-1, GR, NF-kB, octamer.

Table 3B. FastM Predicted Protein Interactions between Other Transcription Factors in the p16INK4a Promoter.

Upstream Protein	Position	Downstream Protein	Position	FastM Score
GC, Sp1	4402	NF-Kb	4313	1.68
E47	5566	MyoD	5567	1.86
	5301		5302	1.78
	4626		4627	1.76
	4249		4250	1.75
	4583		4584	1.75
MyoD	5302	E47	5566	1.87
	4584		4626	1.79

region. Given that p16INK4a probably has a role in cell entrance into G0, interactions with several different cell-type specific gene regulators is to be expected. Therefore, the p16INK4a promoter is predicted to be located at position 4806 of c66. Neither GRAIL nor GENSCAN predict a plus strand promoter within this cosmid. Recently, a multi-

protein complex that includes GATA1 and E47 has been shown to form during erythroid differentiation in promoters with a consensus sequence CAGGTGN(9)GATA (212). However, a search using FindPatterns from the GCG package of the c66 cosmid sequence for this motif did not show that this sequence is in any of the predicted p16INK4a promoter regions.

GRAIL shows two polyadenylation sites downstream of p16INK4a, one at position 15546 and the second at 15823. GENSCAN places a polyadenylation site within an intron at position 11148. Computer analysis of the 6055 nt downstream of the last exon of p16INK4a was performed by PolyAH in order to predict the location of the polyadenylation site. This program predicts two strong polyadenylation signals within six base pairs of each another at positions 12973 and 12979 of c66. Given that the polyadenylation signal, AAUAAA, is six base pairs in length, two signals tandemly repeated typically are a very strong indicator of the position of the biologically active polyadenylation signal. Summarizing the data presented, a polyadenylated transcript for p16INK4a of at least 8170 nt is expected with the predicted promoter representing a cell-type specific mechanism for its transcription.

Analysis of the repeat elements within c66

From the RepeatMasker output (Table 4, also Figure 3), 55.42% of c66 is repetitive DNA. Of that, over half, 9765 nt are LINE sequence, or 28.12% of the cosmid. Only 5.2% of the total LINE DNA is found within the predicted p16INK4a gene. This is comprised of two pieces of sequence; one is a L1M, the other a Lb2. Furthermore, over 93% of the LINE DNA in c66 is 3' to p16INK4a. Approximately one third of the total LINE DNA is from the younger M family. From the remainder, all but 163 nt of L2

Table 4. RepeatMasker Output for Cosmid c66.

SW Score	% sub.	% del.	% ins.	Query	begin	end	length	(rem.)	repeat subfamily	repeat family	begin	end	(rem.)
LINE DNA													
323	22.2	14.8	0.0	C66.fa	31620	31727	108	2951 C	L1	LINE/L1	3502	3379	2644
1501	22.7	5.7	0.2	C66.fa	8594	8998	405	25680 +	L1MA7	LINE/L1	5852	6278	11
1899	19.9	5.4	1.3	C66.fa	13586	14048	463	20630 C	L1MB5	LINE/L1	6176	5695	0
1734	23.4	1.3	0.7	C66.fa	26336	26788	453	7890 C	L1MC3	LINE/L1	7740	7285	0
3739	19.9	2.2	9.0	C66.fa	27095	28247	1153	6431 C	L1MC3	LINE/L1	7291	6217	449
240	26.3	6.7	6.7	C66.fa	22030	22238	209	12440 +	L1ME	LINE/L1	5532	5740	423
3724	20.1	1.5	1.0	C66.fa	29197	29996	800	4682 C	L1P	LINE/L1	4871	4068	1275
2976	19	0.2	0.5	C66.fa	30634	31213	580	3465 C	L1P	LINE/L1	4080	3503	2066
5612	24.4	1.7	1.2	C66.fa	32051	33604	1554	1074 C	L1P	LINE/L1	3388	1828	2758
2375	21.7	4.6	0.3	C66.fa	33884	34678	795	0 C	L1P5	LINE/L1	1837	1073	4309
2494	14.4	0.6	0.4	C66.fa	28703	29195	493	5483 C	L1PA15	LINE/L1	6157	5664	0
1204	3.3	0.0	0.7	C66.fa	2480	2629	150	32049 C	L1PA4	LINE/L1	6146	5998	0
5099	12.3	2.1	4.0	C66.fa	16155	18239	2085	16439 +	L1PB1	LINE/L1	3692	5776	379
5099	12.3	2.1	4.0	C66.fa	18286	18639	354	16039 +	L1PB1	LINE/L1	5802	6155	0
272	31.2	0.0	4.6	C66.fa	11261	11369	109	23309 C	L2	LINE/L2	2737	2634	13
278	22.2	0.0	0.0	C66.fa	19154	19207	54	15471 C	L2	LINE/L2	2745	2692	5
							9765						
							28.16%	LINE DNA					
Medium Frequency Repetitive DNA													
307	26.8	0.8	6.3	C66.fa	23194	23320	127	11358 C	MER69A	DNA/AcHobo	123	4	54
540	20.7	17.2	0.0	C66.fa	23344	23532	189	11146 C	MER3	DNA/MER1_type	204	38	5
1960	22.2	6.7	6.7	C66.fa	24707	25574	868	9104 +	MER63C	DNA/MER1_type?	70	937	1
566	28.7	3.3	0.6	C66.fa	1762	1942	181	32736 C	MIR	SINE/MIR	216	31	46
292	19.0	5.8	6.6	C66.fa	4719	4855	137	29823 C	MIR	SINE/MIR	259	124	3
204	29.3	1.7	0.0	C66.fa	12908	12965	58	21713 C	MIR	SINE/MIR	121	63	141
222	30.3	9.2	0.0	C66.fa	22707	22815	109	11863 +	MIR	SINE/MIR	117	235	27
							1669						
							4.81%	Medium Frequency Repeat DNA					

LTR DNA													
1561	17.6	4.8	6.5	C66.fa	23722	24118	397	10560 C	MLT1B	LTR/MaLR	390	1	0
1664	20.7	14.2	0.7	C66.fa	8086	8530	445	26148 C	MLT1D	LTR/MaLR	505	1	0
2160	13.2	7.2	1.2	C66.fa	28248	28649	402	6029 +	MSTA	LTR/MaLR	1	426	0
2243	13.7	6.7	0.5	C66.fa	31216	31616	401	3062 +	MSTA	LTR/MaLR	1	426	0
1152	16.4	1.6	1.6	C66.fa	14232	14475	244	20203 C	LTR1	LTR/Retroviral	244	1	541
3358	16.8	2.1	3.8	C66.fa	20345	21688	1344	12990 C	pTR5	LTR/Retroviral	1439	743	0
							3233						
							9.32% LTR DNA						
Alu DNA													
841	18.2	0.0	0.0	C66.fa	25862	26004	143	8674 +	AluJb	SINE/Alu	1	143	159
1228	18.5	8.3	0.4	C66.fa	11370	11634	265	23044 C	AluJo	SINE/Alu	286	1	16
1735	16.1	1.0	0.7	C66.fa	30335	30633	299	4045 C	AluJo	SINE/Alu	301	2	1
1740	13.9	0.3	3.9	C66.fa	31740	32048	309	2630 C	AluSc	SINE/Alu	298	1	1
1567	8.2	0.0	0.9	C66.fa	24145	24363	219	10315 C	AluSg/x	SINE/Alu	302	86	0
849	10.4	0.0	0.0	C66.fa	26167	26291	125	8387 +	AluSg/x	SINE/Alu	176	300	0
2025	9.6	0.7	1.3	C66.fa	4014	4316	303	30362 C	AluSp	SINE/Alu	303	3	0
1967	13.5	0.0	0.0	C66.fa	19439	19741	303	14937 +	AluSp	SINE/Alu	1	303	0
2019	8.2	3.8	0.0	C66.fa	30031	30321	291	4357 C	AluSp	SINE/Alu	302	1	1
2030	11.6	0.0	0.0	C66.fa	26789	27081	293	7597 +	AluSq	SINE/Alu	2	294	9
1969	11.2	0.0	2.9	C66.fa	3087	3397	311	31281 +	AluSx	SINE/Alu	1	302	0
2138	8.8	0.0	0.3	C66.fa	4357	4653	297	30025 C	AluSx	SINE/Alu	297	2	5
1944	10.8	0.3	1.3	C66.fa	6560	6864	305	27814 +	AluSx	SINE/Alu	1	302	0
1956	8.7	5.2	0.0	C66.fa	15853	16139	287	18539 C	AluSx	SINE/Alu	302	1	0
600	1.4	0.0	0.0	C66.fa	1	73	73	34605 C	AluY	SINE/Alu	74	2	226
1688	6.2	0.0	8.3	C66.fa	33608	33883	276	795 +	AluY	SINE/Alu	41	293	8
							4099						
							11.82% Alu DNA						
Simple Repetitive DNA													
231	18.9	0.0	0.0	C66.fa	22836	22872	37	11806 +	(CA)n	Simple_repeat	2	38	82
210	22.7	0.0	0.0	C66.fa	25575	25618	44	9060 C	(CA)n	Simple_repeat	44	1	76
207	0.0	0.0	0.0	C66.fa	18240	18285	46	16393 +	(TA)n	Simple_repeat	1	46	74
211	3.9	0.0	0.0	C66.fa	9157	9182	26	25496 +	POLY_A	Simple_repeat	1	26	94
209	15.2	0.0	0.0	C66.fa	29997	30029	33	4649 C	POLY_A	Simple_repeat	33	1	87
							186						
							0.54% Simple Repeat DNA						

Low Complexity DNA

32	2.6	0.0	0.0	C66.fa	2442	2479	38	32199 +	AT_rich	Low_complexity	1	38	262
21	7.7	0.0	0.0	C66.fa	3484	3522	39	31156 +	AT_rich	Low_complexity	1	39	261
24	7.1	0.0	0.0	C66.fa	12719	12760	42	21918 C	AT_rich	Low_complexity	42	1	258
21	0.0	0.0	0.0	C66.fa	12973	12993	21	21685 C	AT_rich	Low_complexity	63	43	237
46	4.7	0.0	0.0	C66.fa	14088	14151	64	20527 +	AT_rich	Low_complexity	1	64	236
23	3.5	0.0	0.0	C66.fa	19827	19855	29	14823 C	AT_rich	Low_complexity	29	1	271
28	2.9	0.0	0.0	C66.fa	22557	22590	34	12088 +	AT_rich	Low_complexity	1	34	266
							267						

0.77% Low Complexity DNA
19219 Total Bases of Repetitive DNA
55.42% Repetitive DNA

See Legend for Figure 1.

DNA are from the older P family. Within the two L1 families, little or no subfamily preference is detected. There are two Alu insertions within L1 DNA; one is the virtually perfect insertion of an AluSq, position 26789, into an L1MC3 that starts at 28247, is interrupted by the Alu between 27081 and 26789, and continues to position 26336. Inspection of the L1 reveals that there is the expected duplication of L1 sequence on either end of the Alu, but otherwise, the L1 consensus is uninterrupted on either side of the Alu insertion. The duplicated L1 sequence is from 7285 to 7291 of the consensus L1MC3 sequence. The second Alu insertion involves the insertion of an AluJo into a L1P. This was followed by the insertion of an AluSp in tandem to the AluJo. In the process, 13 nt of L1P DNA are directly repeated outside the tandemly inserted Alus. The RepeatMasker output shows the duplication to be between L1P position 4068 and 4080. A L1PB1 is found between 16155 and 18639. This element has a 47 nt insertion that results in the deletion of 26 nt of L1 DNA. The insertion is not a known repeat element.

Sixteen Alu elements are found in c66. These 16 elements contain 11.82% of the total sequence in c66 and twelve can be classified as full-length. Of the other four, one is a left monomer, one is a right monomer, and one includes the final 219 nt of the consensus. The fourth is 73 nt in length and constitutes an end of the cosmid. The right monomer and the 219 nt fragment are both of the AluSg family but are oppositely oriented and separated by almost 2kB making it unlikely that they are from a common insertion event followed by a later recombination. Of the Alus considered to full-length, two are AluJo, one is AluSc, three are AluSp, one is AluSq, four are AluSx, and one is from the AluY subfamily. The preponderance of Sp and Sx subfamily Alus is in agreement with the data indicating that these Alu subfamilies were active during the time

of greatest Alu expansion within the human genome. A pair of Alus is found upstream of the predicted p16 promoter. An AluSp is found extending from 4014 to 4316 and an AluSx starts at 4357 and extends to 4653. Both are in the reverse orientation to p16INK4a. Interestingly, a TBP site and the binding sites of two proteins predicted to interact with TBP by FastM, E47 and MyoD, are within the AluSp. The AluSx between 6560 and 6864 is in the first intron and the AluJo between 11370 and 11634 is within the second intron of p16INK4a. Except for these two elements, the remaining Alu repeats are outside the predicted p16INK4a gene; however, the distribution bias of the LINE repeats for the region 3' of p16INK4a is not seen in the distribution of the Alus. In addition to the two Alus immediately upstream of the promoter, another full-length AluSx element and a partial AluY (discussed above) at the beginning of the cosmid are found 5' of p16INK4a.

Medium frequency repetitive elements including MER and MIR sequences make up 4.81% of c66. Of that total, half are found in one MER63C element. LTRs constitute 9.32% of c66. Similar to the medium frequency repeats, a single retroviral pTR5 element constitutes over 40% of the total LTR sequence.

The extent of simple repeats, or low-complexity DNA, was determined with RepeatMasker (Table 4) and Sputnik (Table 5). As expected, both programs find dinucleotide repeats. The difference between the libraries the two programs use is seen in the other types of simple repeats. RepeatMasker is able to find the mononucleotide repeats and regions of low-complexity (Table 4); whereas Sputnik finds simple tandem repeats with longer repeat units (Table 5). None of the A/T-rich regions or poly-A tracts

are associated with Alu repeats. A poly-A tract is found within the first intron of

Table 5. Sputnik Output from Cosmid c66.

Begin	End	Length	(Repeat Unit)xN
3495	3508	14	(TAAAC)x2
4002	4026	25	(TCTT)x6
6843	6864	22	(GAAA)x5
9093	9105	13	(AACA)x3
9582	9593	12	(GCT)x4
11517	11532	16	(GTTT)x4
11583	11594	12	(CCT)x4
14652	14668	17	(TTG)x5
16861	16879	19	(AAATC)x3
18240	18285	46	(TA)x23
21407	21419	13	(GGCT)x3
21965	21979	15	(TTTC)x3
22843	22856	14	(CACG)x3
25595	25605	11	(TG)x5
27373	27387	15	(ATTT)x3
30330	30343	14	(TTTTC)x2
31408	31419	12	(CT)x6
31653	31666	14	(TTGT)x3
31723	31736	14	(TTTTC)x2

p16INK4a. The 3' UTR of the p16INK4a gene includes two A/T-rich regions and a MIR element. The first A/T-rich region extends from 12719 to 12760 within the predicted 3' UTR but is not in the published cDNA sequence. The MIR is in the reverse orientation starting at position 12965 and extending to 12908 which places it within the 3'UTR predicted by placing the polyadenylation signal at 12973. The 5' end of the MIR is closely associated with the second A/T-rich region, which starts at position 12973 and is position of the predicted polyadenylation signal. Two observations lead to the possible conclusion that the polyadenylation signal predicted by POLYAH is not the biologically active site. The published cDNA sequence ends at 12665 and there is an A/T-rich region upstream and closer to the end of the published sequence.

C86

Cosmid c86 has a total insert size of 44175 and a G/C content of 40.1%. As with c66, this cosmid was sequenced using a shotgun strategy in pUC18. Approximately 400 subclones were isolated and sequence data was collected from forward and reverse universal primer reactions. An additional 40 reactions using custom primers on subclones or PCR product were performed to close gaps and double strand the sequence of the cosmid. A graphical representation of the sequence features in c86 is presented in Figure 4.

The p15INK4b gene is within c86

Table 6 shows the Crossmatch alignment of the p15INK4b cDNA sequence with C86.

Table 6. Crossmatch alignment of c86 with the p15INK4b cDNA sequence.

Score	% Sub.	% Del.	% Ins.	Query	Begin	End	(Rem.)		Subject	Begin	End	(Rem.)
331	1.94	0	0	C86.fa	15957	16316	27859	C	p15INK4b	837	478	0
403	1.03	0.083	1.24	C86.fa	18862	19345	24830	C	p15INK4b	482	1	355

A cursory inspection of Table 6 reveals that the entire p15INK4b cDNA sequence is found within c86, but in the reverse orientation to the cosmid sequence as demonstrated by the "C" before the subject column. Also, note that the numbers in the "Begin" column are larger than the numbers in the "End" column of the subject sequence. Therefore, exon 1 is on the second line of the output and the second exon is on the first line. The overlap between the beginning of the second exon and the end of the first is an artifact due to similarity of the gene sequence at the splice junction.

The substitutions reflected in the "% Sub" column are the result of the same sequencing artifacts as those discussed above in regards to c66. These two projects were

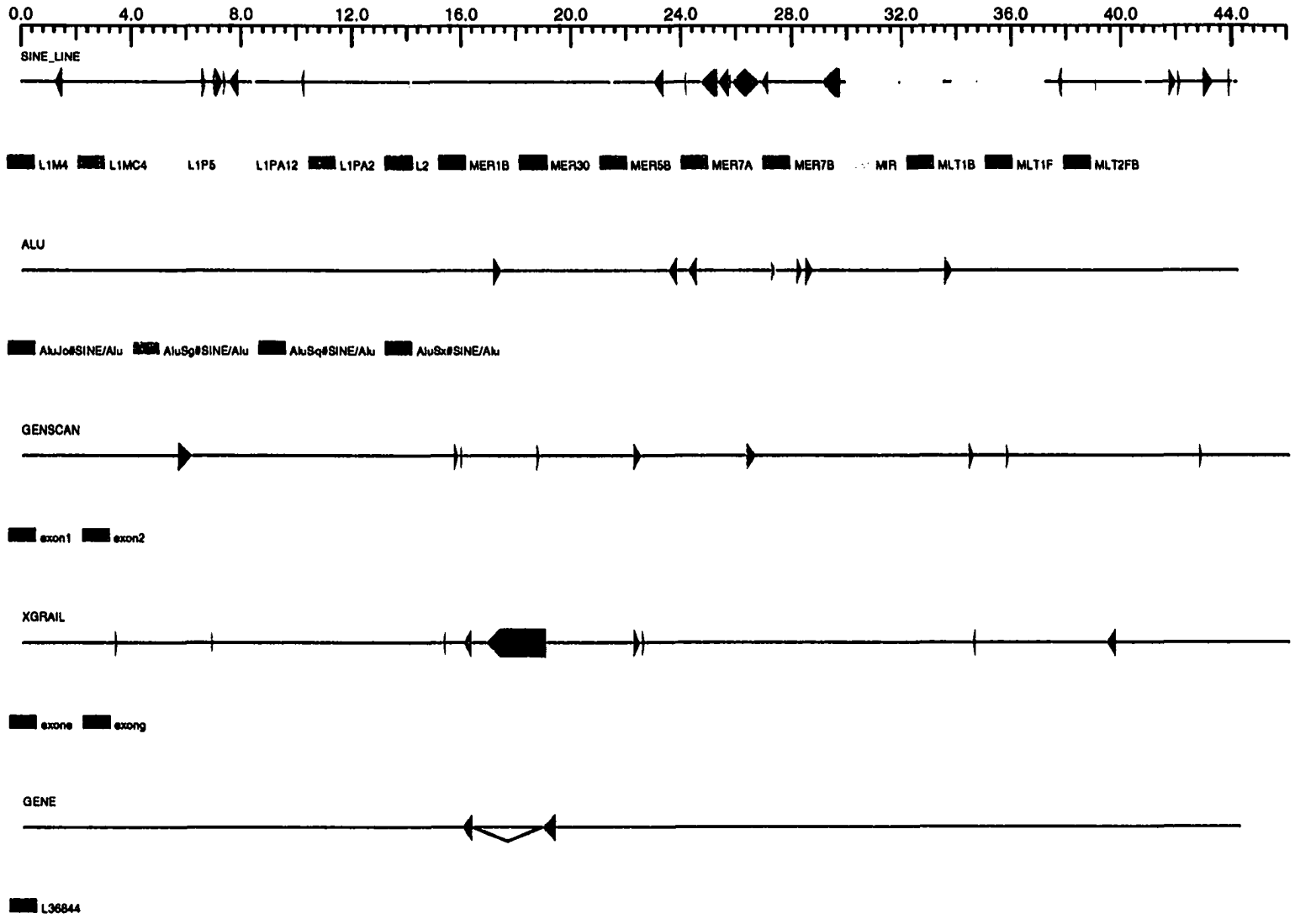


Figure 4. Sequence Feature Map from c86. See legend to Figure 3.

done at the same time using the same techniques and reagents. Detailed analysis of the deletions and insertions seen in the Crossmatch output reveals that these discrepancies between c86 and the published p15INK4b coding sequence are real. Cosmid c86 has three insertions relative to p15INK4b at c86positions 18961, 18957, and 18956. The result of these insertions is a change in the amino acid sequence from ATPA to ASAAA between codons 19 and 22 of the published p15INK4b protein sequence, a gain of a single amino acid. The effects of these mutations can only be speculated on here, but threonine and serine both have hydroxyl groups on the β-carbon so the primary effect, if any, is probably through the added residue. After alignment of the two sequences, base 18900 of c86 is an insertion relative to the p15INK4b gene but a base is deleted six bases downstream. The reading frame therefore is quickly restored as a result of these two mutations. The amino acid sequence is changed from RHSWE to RQLLE where R is amino acid 29. In the 5' UTR of the p15INK4b gene, there are four discrepancies between the two sequences. Base 19220 of c86 is an insertion and there are three bases inserted in the p15INK4b sequence after c86 position 19243.

The GRAIL predicted exons and GENSCAN predicted genes are summarized in Table 7 and shown in Figure 4.

Table7. GRAIL Exon and GeneScan Gene Predictions

GRAIL				GeneScan: Gene 1			GeneScan: Gene 2		
Begin	End	Qual.	Strand	Begin	End	Strand	Begin	End	Strand
3346	3416	gd.	f	5666	6173	f	25751	25758	f
6843	6898	gd.	f	15680	15844	f	26296	26646	f
22194	22443	ex.	f	15933	15988	f	34362	34545	f
22500	22604	ex.	f	18697	18818	f	35692	35814	f
15246	15332	gd.	r	22194	22481	f	42737	42850	f
16052	16313	ex.	r						
16864	19019	ex.	r						
34490	34590	gd.	r						
39361	39693	gd.	r						

GRAIL accurately predicts two exons of excellent quality that are the p15INK4b exons and correctly determines the p15INK4b reading frame. Therefore, BLASTP with the predicted protein sequence from these exons does find the p15INK4b gene product. GRAIL predicts the first exon of p15INK4b to begin at 19019 and end at 16864. The first methionine of p15INK4b starts at position 19019, but ends at 18864, a discrepancy of 2000 nt at the 3' end. A more detailed analysis of the splice junction reveals that the correct donor site is at position 18864 and the acceptor site is position 16313. GRAIL accurately predicts the correct open reading frame of the second exon starting it at 16313 and ending it at 16052. The consensus 5'and 3' splice sites are conserved in p15INK4b.

From the data presented in Table 7, it is evident that GENSCAN was unable to correctly predict the p15INK4b gene. BLASTP of the predicted protein sequence from GENSCAN and the remaining GRAIL predicted exons shows that the first GENSCAN predicted gene has homology to an ubiquitin-ribosomal protein fusion gene. The homology also appears in the GRAIL predicted exons on the forward strand between 22194 and 22443, and 22500 and 22604. A closer analysis of the overlap between the fusion gene and the predicted gene shows that the identity is on the order of 90% from the initiating methionine at position 294 of the predicted protein to codon 378. However, codons 379 and 382 are stop codons leading to the conclusion that this probably is a pseudo-gene.

Transcription regulation of p15INK4b

Two CpG islands are found by GRAIL, one between 15978 and 16340, and the second spanning 18745 to 18385. Seven thousand bases of c86 were analyzed to predict potential to act as a promoter for the p15INK4b gene. This region was chosen such that

there are 5.555 nt upstream of the first position in the published p15INK4b cDNA sequence, which starts at position 19345. Using the promoter.dat database, TSSG predicts a single minus strand promoter at position 19437. The strongest promoter predicted by TSSW is at 19347, only 2 nt upstream of the beginning of the p15INK4b cDNA sequence, but this predicted promoter does not have a nearby TATA site. However, TSSW does predict another promoter at 21164, labels it as having a TATA site, but gives it a very low score. As noted above, since NNPP appears to only return promoters with TATA sites in them it predicts two high scoring promoters within 2500 nt of the start of the published cDNA sequence. The predicted transcription start site for the predicted promoter closest to the published sequence is position 19935; the second is predicted to start transcription at 21164, the same position as one of the TSSW promoters discussed above.

Based on the predicted promoter positions, MatInspector and TESS were used to determine upstream transcription factors that have the potential to bind within the predicted promoters. Interactions between GC-box proteins and the upstream factors were examined using FastM. GC-box proteins were examined due to the possibility that the p15INK4b promoter is a TATA-less promoter. This conclusion was based on the results from the promoter prediction programs. A tabulation of the results from FastM is presented in Table 8. These results are catalogued by GC-box protein binding site and then sorted by position. The scores then are averaged for each GC-box binding site. When compared to the p16INK4a promoter, a much larger number of proteins were predicted to bind at least once in a p15INK4b predicted promoter by MatInspector or TESS, 32 in all. Two predicted promotes, the TSSG predicted promoter at position

Table 8. FastM Predicted Protein Interactions with GC-box Proteins in Cosmid c86.

gc box @ 19359	pos'n	score	gc box @ 19451	pos'n	score	gc box @ 19641	pos'n	score	gc box @22421	pos'n	score	gc box @23247	pos'n	score	gcbox @ 23777	pos'n	score
th1e47	19388	1.71	Ap4	19461	1.84	USF	19802	1.64	USF	22427	1.70	Ap1	23270	1.77	lyf1	23780	1.68
Ap4	19408	1.83	crebp1	19497	1.78	GATA	19864	1.70	E47	22442	1.75	Ik2	23296	1.72	USF	23798	1.73
Ap2	19418	1.78	ATF	19498	1.82	Oct1_06	19871	1.72	c-myb	22494	1.81	sry_02	23312	1.75	mycmax	23800	1.65
NF-kB_01	19428	1.71	creb-2	19499	1.86	cp2	19892	1.62	Ap1	22497	1.67	hfh2	23343	1.73	n-myc	23800	1.64
Lyf-1	19431	1.72	USF	19500	1.83	Nkx2.5	19909	1.61	cebp	22556	1.73	hnf3b	23343	1.70	E47	23817	1.65
Ik-2	19432	1.78	n-myc	19502	1.87	c-myb	19918	1.67	STAT1	22560	1.65	c-myb	23377	1.85	sry_02	23826	1.73
Average Score		1.76	th1e47	19549	1.82	hnf3b	19926	1.62	Nkx2.5	22604	1.68	Irf	23403	1.74	Ik2	23849	1.72
			c-ets1p54	19598	1.84	NF-kB_01	19935	1.67	hnf3b	22605	1.72	Oct1_06	23409	1.75	Irf1	23878	1.68
			hnf3b	19617	1.82	NF-kB_01	19936	1.67	Ap2	22626	1.65	Nkx2.5	23427	1.69	cebpb	23892	1.76
			s8	19656	1.80	c-ets1p54	19937	1.64	c-ets1p54	22635	1.75	Brn2	23433	1.77	GATA	23902	1.78
			Ik2	19657	1.87	Average Score		1.66	th1e47	22647	1.65	s8	23436	1.85	Gfi-1	23926	1.75
			Gfi-1	19676	1.83				Nkx2.5	22667	1.68	Gfi-1	23453	1.75	cp2	23928	1.75
			cebp	19689	1.84				Ik2	22677	1.74	Oct1_06	23459	1.80	hfh2	23960	1.64
			Irf1	19710	1.80				hfh2	22698	1.72	GATA	23505	1.78	Oct1_06	23981	1.71
			sry_02	19735	1.78				Average Score		1.71	Average Score		1.76	Brn2	23987	1.73
			Ap1	19752	1.79										Ap1	24022	1.77
			Brn-2	19758	1.81										Oct1_06	24022	1.74
			Average Score		1.82										s8	24056	1.68
															Nkx2.5	24063	1.63
															Average Score		1.71

The position and score returned from FastM are presented for each GC-box protein binding site predicted to be active by FastM. The interacting proteins are further sorted by score for each GC-box protein binding site in the 5555nt upstream of the published p15INK4b cDNA sequence. The score is an indication of the predicted strength of the interaction between a bound GC-box protein and the upstream transcription factor.

19437 and the TATA-less TSSW predicted promoter at position 19347, had corresponding GC-box protein binding sites. Furthermore, these two predicted promoters had the highest scores in their respective groups; in fact, TSSG only predicted the one promoter. In addition, these two promoters are closest to the 5' end of the published cDNA sequence making them the most attractive choices for the biologically active p15INK4b promoter.

None of the NNPP or the other three TSSW predicted promoters had correlated GC-box protein binding sites. Conversely, four other GC-box protein binding sites were analyzed for their interactions with upstream transcription factors, but when the position of these GC-box protein binding sites was compared with the position of the predicted promoters, no correlation was found. Therefore, it is proposed that one of the promoters discussed above, at either position 19347 or 19437, is the biologically active p15INK4b promoter. Although E47 and GATA-1 are not predicted by FastM to bind in the same p15INK4b promoter, a search for the binding site of the multi-protein complex including these two proteins was performed. However, no binding site for this complex was found in any of the p15INK4b predicted promoters. GRAIL predicts a plus strand promoter starting at 22223, which is not germane to this discussion of the p15INK4b gene.

A minus strand polyadenylation site is place by GRAIL at position 12981. Prospective polyadenylation sites were examined in 6720 nt of cosmid c86 sequence using POLYAH. Of the 6720 nt examined, 5756 nt were downstream of the 3' end of the published p15INK4b cDNA sequence. Six polyadenylation sites were predicted, but three of them are in the intron sequence. Of the remaining three, the highest scoring site is at position 12994 of c86, 2963 nt from the 3' end of the published cDNA sequence.

The predicted site closest to the end of the published cDNA sequence is at position 14656, 1301 nt from the end of the published cDNA sequence. GENSCAN only finds plus strand genes within this cosmid so no minus strand polyadenylation sites are found by this program.

Repetitive DNA within c86

Table 9 is the output from RepeatMasker showing the position of the repetitive DNA within c86 which also is graphically represented in Figure 4. In comparison to c66 which was over 55% repetitive DNA, c86 is relatively low in total repetitive DNA at only 39.96%. Cosmid c86 is made up of a lower percentage of every classification of interspersed repeat than c66, especially Alu repeats. Cosmid c66 is over 11% Alu whereas c86 is less than 5%. Conversely, c86 consists of a higher percentage of simple and simple tandem repeats than c66, but the total amount of these classes of repeats is so small as to be insignificant in comparison to the interspersed repeats.

Less than 10% of the L1 sequence in c86 was inserted after the mammalian radiation as seen by the total number of nucleotides of L1M vs. L1P L1 sequence. No LINE fragments are found between c86 positions 19437 and 12994, the predicted promoter and polyadenylation sites, respectively. Furthermore, only 103 nt, or 1% of the total LINE sequence, are found 3' of p15INK4b in c86. It is notable that 15 of 18 LINE fragments found within this cosmid lie in the direction of p15INK4b transcription and all 15 are 5' to the p15INK4b gene.

Examination of the alignment of the LINE repeats with c86 reveals several rearrangements within the L1P5 that is found from 31932 to 37220 of c86. An AluSx is inserted between 33514 and 33822. A deletion results in the loss of ~135 nt between 602

Table 9. RepeatMasker Output for Cosmid c86.

Score	% Sub.	% Del.	% Ins.	Query	Begin	End	Length	Rem.	Or.	Repeat Family	Repeat Class	Begin	End	Rem.
LINE DNA														
1081	24.0	3.4	3.1	C86	25359	25745	387	18430	C	L1M4	LINE/L1	4832	4445	1314
209	30.5	3.8	0.0	C86	24106	24210	105	19965	C	L1MC4	LINE/L1	7160	7052	817
650	31.2	1.8	1.1	C86	26920	27201	282	16974	C	L1MC4	LINE/L1	6989	6706	988
235	23.6	1.8	0.0	C86	39050	39104	55	5071	+	L1MC4	LINE/L1	6885	6940	1037
6952	20.1	2.5	0.1	C86	31932	33514	1583	10661	C	L1P5	LINE/L1	2941	1372	3205
2421	22.6	1.2	1.2	C86	33822	34410	589	9765	C	L1P5	LINE/L1	1327	739	4818
766	29.4	9.5	0.6	C86	34409	34724	316	9451	C	L1P5	LINE/L1	602	259	5543
805	21.8	9.3	3.5	C86	34755	35199	445	8976	C	L1P5	LINE/L1	185	-318	5960
7933	18.9	0.8	2.4	C86	35198	37220	2023	6955	C	L1P5	LINE/L1	608	-1383	5586
5735	9.9	1.4	1.2	C86	29946	31911	1966	12264	C	L1PA12	LINE/L1	6153	4099	12
5395	3.7	0.1	0.1	C86	24687	25355	669	18820	C	L1PA2	LINE/L1	6144	5476	2
877	4.5	0.0	0.0	C86	25716	25826	111	18349	C	L1PA2	LINE/L1	6144	6034	2
1831	3.0	0.0	0.0	C86	25826	26301	476	17874	C	L1PA2	LINE/L1	5490	5015	656
4749	2.8	0.0	0.0	C86	26302	26880	579	17295	+	L1PA2	LINE/L1	4428	5006	1140
222	24.3	8.7	1.9	C86	6575	6677	103	37498	+	L2	LINE/L2	2539	2648	102
486	32.5	11.7	1.5	C86	29092	29740	649	14435	C	L2	LINE/L2	2410	1696	340
392	27.4	2.6	0.0	C86	37713	37825	113	6350	C	L2	LINE/L2	2743	2628	7
177	26.7	2.2	0.0	C86	43869	43913	45	262	C	L2	LINE/L2	2732	2687	18
							10496 Total LINE DNA							
							23.76% LINE DNA							
LTR Interspersed Repeats														
1630	16.5	10.6	5.2	C86	23003	23371	369	20804	C	MLT1B	LTR/MaLR	389	1	1
267	30	12.7	3.0	C86	41740	42006	267	2169	+	MLT1F	LTR/MaLR	1	293	248
287	29.3	6.0	2.6	C86	42059	42174	116	2001	+	MLT1F	LTR/MaLR	304	423	118
266	13	4.3	0.0	C86	7001	7046	46	37129	C	MLT2FB	LTR/Retroviral	414	367	0
1621	16.8	9.4	1.8	C86	7575	7914	340	36261	C	MLT2FB	LTR/Retroviral	366	1	48
							1138 Total LTR DNA							
							2.58% LTR DNA							

Alu Interspersed Repeats

1451	20.8	0.3	1.3	C86	23514	23816	303	20359	C	AluJo	SINE/Alu	301	2	1
1612	15.8	1.6	2.6	C86	24229	24531	303	19644	C	AluJo	SINE/Alu	300	1	2
1351	10.3	0.0	0.5	C86	27212	27405	194	16770	+	AluSg	SINE/Alu	1	193	107
1559	8.5	0.0	0.0	C86	28120	28330	211	15845	+	AluSq	SINE/Alu	12	222	81
1915	11.6	1.3	1.0	C86	17144	17444	301	26731	+	AluSx	SINE/Alu	1	302	0
1913	10.5	2.0	0.3	C86	28443	28737	295	15438	+	AluSx	SINE/Alu	1	300	2
1595	15.8	1.0	3.0	C86	33518	33814	297	10361	+	AluSx	SINE/Alu	1	291	11

1904 Total Alu DNA

4.31% Alu DNA

Medium Frequency Interspersed Repeats

1866	14.3	1.3	1.3	C86	42995	43301	307	874	+	MER1B	DNA/MER1_type	31	337	0
1419	11	5.5	0.5	C86	1256	1474	219	42701	C	MER30	DNA/MER1_type	230	1	0
306	24.3	5.4	5.4	C86	10166	10276	111	33899	C	MER5B	DNA/MER1_type	114	4	64
1180	14.1	5.0	4.2	C86	7087	7327	241	36848	+	MER7A	DNA/MER2_type	3	245	90
563	16.7	2.8	0.0	C86	7355	7462	108	36713	+	MER7B	DNA/MER2_type	1094	1204	1
317	33.3	5.7	0.0	C86	8387	8527	141	35648	C	MIR	SINE/MIR	206	58	56
339	21.7	19.6	0.0	C86	14086	14223	138	29952	+	MIR	SINE/MIR	84	248	14
291	31.1	7.6	0.0	C86	21429	21560	132	22615	C	MIR	SINE/MIR	237	96	25
233	25.9	22.8	0.0	C86	40719	40880	162	3295	+	MIR	SINE/MIR	64	262	0
301	29.2	0.9	1.9	C86	43975	44080	106	95	C	MIR	SINE/MIR	124	20	138

1665 Total Medium Frequency Repititive DNA

3.77% Medium Frequency Repetitive DNA

Simple Repeats

287	3.0	0.0	0.0	C86	3861	3893	33	40282	C	(CA)n	Simple_repeat	34	2	86
288	0.0	0.0	0.0	C86	15178	15209	32	28966	C	(CA)n	Simple_repeat	32	1	88
360	0.0	0.0	0.0	C86	672	711	40	43464	+	(GA)n	Simple_repeat	1	40	80
204	12.5	0.0	0.0	C86	7520	7583	64	36592	+	(GA)n	Simple_repeat	1	64	56
227	25.6	1.3	5.1	C86	42362	42439	78	1736	C	(GAAAA)n	Simple_repeat	76	2	44
436	21.4	0.8	3.4	C86	9317	9433	117	34742	+	(TA)n	Simple_repeat	1	114	6
517	26.7	0.9	0.0	C86	9496	9611	116	34564	+	(TA)n	Simple_repeat	1	117	3
235	29.2	0.8	3.3	C86	8922	9098	177	35077	+	(TAAA)n	Simple_repeat	3	95	25

657 Total Simple Repetitive DNA

1.49% Simple Repeat DNA

Low Complexity DNA

24	8.3	0.0	0.0	C86	2343	2390	48	41785	+	AT_rich	Low_complexity	82	129	171
33	9.9	0.0	0.0	C86	2753	2833	81	41342	+	AT_rich	Low_complexity	1	81	219
23	3.5	0.0	0.0	C86	10795	10823	29	33352	C	AT_rich	Low_complexity	54	26	246
25	0.0	0.0	0.0	C86	12970	12994	25	31181	C	AT_rich	Low_complexity	25	1	275
23	7.3	0.0	0.0	C86	17014	17054	41	27121	C	AT_rich	Low_complexity	41	1	259
29	2.9	0.0	0.0	C86	17856	17890	35	26285	+	AT_rich	Low_complexity	1	35	265
21	3.7	0.0	0.0	C86	27565	27591	27	16584	+	AT_rich	Low_complexity	1	27	273
22	0.0	0.0	0.0	C86	37243	37264	22	6911	+	AT_rich	Low_complexity	70	91	209
27	10.1	0.0	0.0	C86	37294	37362	69	6813	C	AT_rich	Low_complexity	69	1	231
25	3.2	0.0	0.0	C86	38280	38310	31	5865	C	AT_rich	Low_complexity	100	70	200
22	12.3	0.0	0.0	C86	18801	18857	57	25318	+	GC_rich	Low_complexity	1	57	243

465 Total Low Complexity DNA
1.05% Low Complexity DNA
16325 Total Repetitive DNA
36.96% Repetitive DNA

See Legend for Table 1.

and 739 of the consensus L1P5 sequence. The deletion is at position 34410 of c86. Alignment of the L1P5 with c86 is lost between 34724 and 34775 of c86, roughly a 50 nt span. The missing L1P5 fragment in the alignment gap is from 185 to 259 of the L1P5 consensus, a gap of 73 nt. The fact that there are alignment gaps in both sequences and that they are of unequal lengths implies that a more complex mutation has taken place at this position and the mutation is not a simple insertion. Finally, the 5' end of an L1P5 repeat is found upstream of the beginning of the larger repeat spanning c86 from 35199 to 37220. The insertion ends at position 608 of the consensus L1P5. Two of the L1P5 fragments in this region have similar 3' ends. The last fragment discussed ends at 608 of the L1P5 consensus and the L1P5 fragment that spans c86 from 34410 to 34724 has an end at 602 of the L1P5 consensus. Therefore, it is tempting to speculate that the apparent duplication of the 5' end of the L1P5 repeat and the complex alignment gap between 34724 and 34775 are related.

A deletion similar to that discussed above results in the loss of 43 nt from L1PA2 at c86 position 25826. The extreme 3' end of L1PA2 is found twice within c86, both fragments terminate at base 6144 of the consensus L1PA2. The first fragment is 669 nt in length and spans positions 24687 to 25355 of c86. The second fragment is shorter than the first, only 111 nt, and spans 25716 to 25826 of c86. It is interesting to note that they lie relatively close to one another and in the same orientation within c86. Another possibly complex rearrangement within a LINE repeat occurs within the L1MC4 found in two fragments from 24106 to 27201 of c86. The gap in the L1MC4 consensus is from 6089 to 7052, a loss of 953 nt. The gap in c86 between L1MC4 fragments is from 24210 to 26920, a length of 2710 nt. A single AluJo of 303 nt is found from 24229 to 24531,

but no other retropositional event is evident within the L1MC4 gap. Furthermore, it is difficult to imagine a mechanism whereby the insertion of a single Alu can lead to the complex series of events that must have lead up to this rearrangement.

Seven Alu repeats are found by RepeatMasker within c86 and all but two of them can be considered full-length. None of the Alu repeats are specific to humans; two are of the oldest AluJo subfamily. Both of the AluJo repeats and all three of the AluSx repeats are full-length; the fragments are of the AluSp and AluSq subfamilies. Both fragments include the left monomer and the 5' portion of the right monomer. A single AluSx is found between the predicted promoter and polyadenylation signal of p15INK4b. It lies between 17144 and 17444 of c86, within the intron of p15INK4b, which is between 16316 and 18862. Similar to the LINE repeats, the remaining Alu repeats all are 5' to the p15INK4b gene.

A MIR fragment is found within the predicted p15INK4b 3'UTR between positions 14086 and 14223. No splice sites, polyadenylation signals, or other functional sites are associated with this repeat element. Of the remaining four MIRs, three are 5' of the p15INK4b gene in c86; the fourth is 3' of the gene. The MERS show a distinct preference for the region 3' of the p15INK4b gene. Of the five MERs, four are found beyond the 3' end of p15INK4b and the fifth is at the extreme end of the cosmid sequence 5' to the p15INK4b gene.

There are 23 simple repeats or tandem repeats detected by RepeatMasker (Table 9) or Sputnik (Table 10). They are notably A/T-rich, 13 are 100% A/T. Of the remaining 10, only two are less than 50% A/T. They are a CCCCA repeat from 4580 to 4593 and a CG repeat from 4679 to 4689. The distribution of simple and tandem repeats is even

Table 10. Sputnik Output for Cosmid c86

Begin	End	Length	(Repeat Unit)xN	Begin	End	Length	(Repeat Unit)xN
672	711	40	(GA)x20	17435	17455	21	(AAAT)x5
1424	1435	12	(TTC)x4	23504	23518	15	(TC)x7
3861	3893	33	(TG)x16	24222	24242	21	(ATTT)x5
4087	4100	14	(TCC)x4	24675	24693	19	(ATT)x6
4580	4593	14	(CCCCA)x2	25687	25705	19	(ATTT)x4
4679	4689	11	(CG)x5	25710	25722	13	(ATT)x4
7036	7047	12	(TG)x6	27571	27583	13	(ATTT)x3
7542	7567	26	(GA)x13	31327	31344	18	(ATTT)x4
9796	9812	17	(TTG)x5	42712	42724	13	(TC)x6
15178	15209	32	(TG)x16				

throughout the cosmid although no repeats are found within the cDNA coding sequence. The low-complexity DNA is similarly biased in favor of A/T-rich regions. Only one of the twelve low-complexity regions detected by RepeatMasker (Table 9) is G/C-rich. Interestingly, the G/C-rich low-complexity region is within the p15INK4b intron and ends seven nucleotides from the splice acceptor site. Low-complexity regions are distributed relatively evenly throughout the cosmid although a cluster of three A/T-rich regions is found in a 1.2kB span from 37200 to 38400.

Cosmids c20 and c110

These two cosmids were originally started as individual projects that were thought to have some overlap. Four hundred pUC18 subclones were isolated and sequenced with forward and reverse universal primers for each project. After assembling a database with this data, the degree of overlap was determined with Crossmatch. It was found that c20 was completely contained within c110 such that they shared one end and c110 extended approximately 2500 nt beyond the other end of c20. The databases were combined and the project closed and double stranded with 50 additional reactions. The total insert size of c110 is 42352 nt and the cosmid insert has a G/C content of 39.66%. A graphical representation of the sequence features found in c110 are presented in Figure 5.

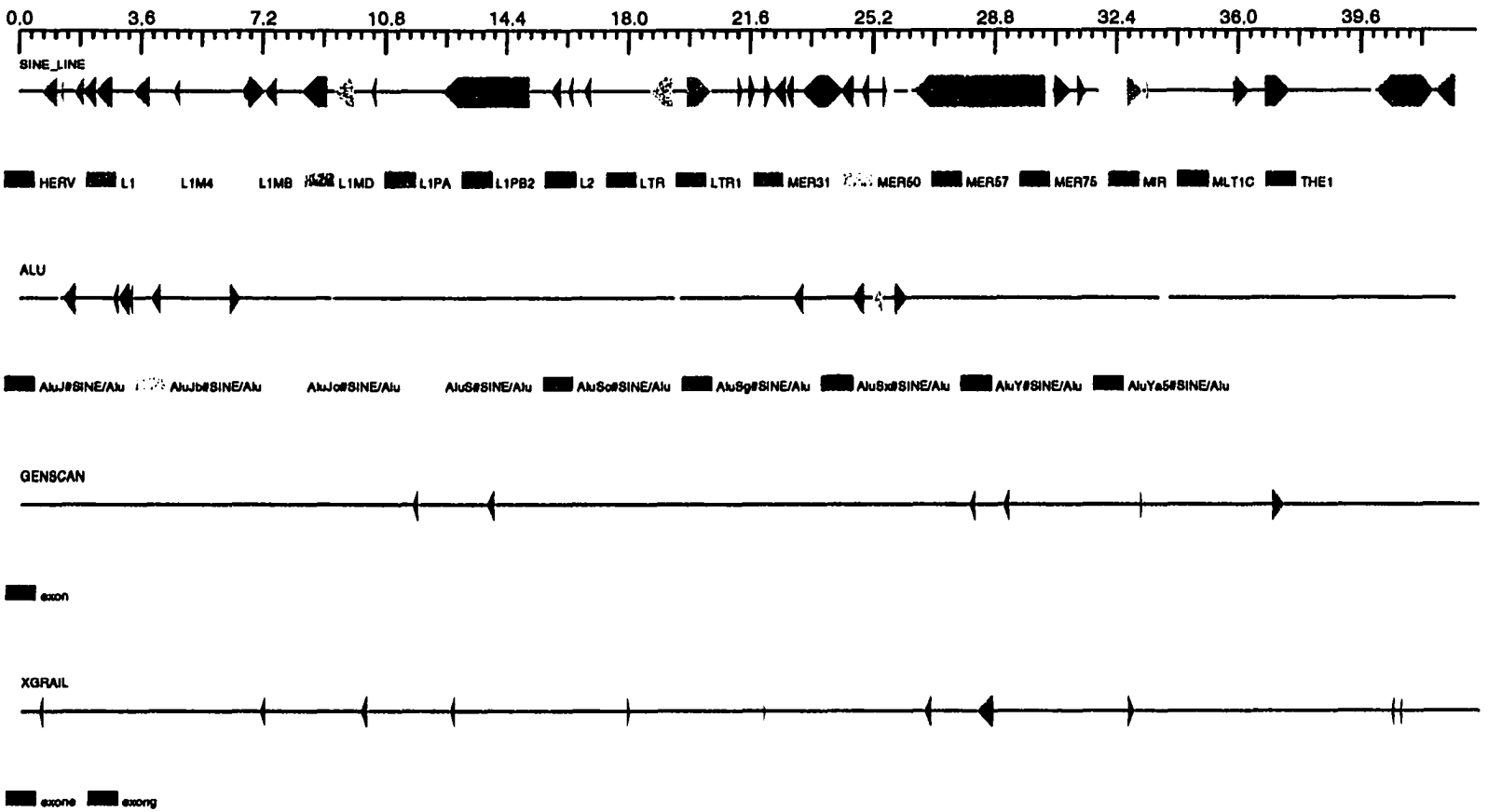


Figure 5. Sequence Feature Map of cl10. See legend to Figure 3.

The exon prediction program GRAIL and gene prediction program GENSCAN were used to predict open reading frames within c110 (Table 11 and Figure 5). The GRAIL predicted exons did not have significant homology to any known protein in SwissProt. The reverse strand gene predicted by GENSCAN has strong identity to L1 encoded reverse transcriptase within the first 41 amino acids. The same predicted gene has weaker homology to retrovirally encoded reverse transcriptases from man, ape, baboon, and mouse between amino acids 118 and 188. The forward strand predicted gene did not have significant homology to any known protein in the databases.

Table 11. GRAIL Predicted Exons and GeneScan Predicted Genes from cosmids c20 and c110.

GRAIL				GeneScan, Gene 1.			GeneScan, Gene 2.		
Begin	End	Qual.	Strand	Begin	End	Strand	Begin	End	Strand
17870	17950	gd	f	11543	11701	r	33021	33076	f
21899	21935	gd	f	13749	13978	r	36931	37273	f
32652	32832	ex	f	27981	28148	r			
539	659	ex	r	28973	29153	r			
7032	7187	gd	r						
10003	10193	ex	r						
12615	12752	gd	r						
26626	26827	ex	r						
28176	28654	gd	r						
40382	40479	gd	r						
40662	40727	ex	r						

Repetitive DNA within c20 and c110

The presence of repetitive DNA was determined by RepeatMasker and Sputnik. The output from RepeatMasker is shown in Table 12 and graphically represented in Figure 5. Cosmid c110 is over 64% repetitive DNA and over 63% of this cosmid is made up of interspersed repeats. Eighteen fragments of LINE DNA account for 25.6% of c110. Analysis of the alignments with the consensus shows that the three L1MB6 fragments probably arose from a common insertional event. In the first gap in the L1MB6, an AluY

Table 12. RepeatMasker Output from Cosmid c110.

Score	% Sub.	% Del.	% Ins.	Query	Begin	End	Length	Rem.	Or.	Repeat Family	Repeat Class	Begin	End	Rem.
LINE DNA														
2890	18	6.2	2.1	c110	19709	20452	744	21900	+	L1	LINE/L1	3045	3818	2328
782	30.8	9.3	2.6	c110	31875	32656	782	9696	+	L1M4	LINE/L1	1697	2155	3882
263	23.5	9.3	8	c110	39866	40027	162	2325	C	L1M4	LINE/L1	4603	4440	1543
1010	19.2	6.2	0	c110	25587	25810	224	16542	+	L1MB6	LINE/L1	5591	5828	347
580	19.8	0	0	c110	26194	26319	126	16033	+	L1MB6	LINE/L1	5821	5946	229
772	25.1	4.3	0.9	c110	30281	30511	231	11841	+	L1MB6	LINE/L1	5935	6173	2
2142	20.1	3.2	3.8	c110	32658	33188	531	9164	+	L1MD1	LINE/L1	5580	6107	117
392	24.2	6.5	0	c110	33241	33364	124	8988	+	L1MD1	LINE/L1	6093	6224	0
6477	7.8	0.3	0.2	c110	26407	30280	3874	12072	C	L1PA2	LINE/L1	6144	2266	2
3960	8.9	0.2	0	c110	41805	42352	548	0	C	L1PA2	LINE/L1	6144	5596	2
3735	20.6	1.6	2.4	c110	40041	41376	1336	976	C	L1PA7	LINE/L1	6106	4773	39
1002	12.5	7.1	0	c110	22646	22829	184	19523	C	L1PB2	LINE/L1	6155	5959	1
2673	12.3	1.6	0.9	c110	23133	23701	569	18651	C	L1PB2	LINE/L1	5973	5401	182
371	24.8	10.2	6.4	c110	21550	21706	157	20646	+	L2	LINE/L2	458	620	2130
285	33	6.4	0.6	c110	22010	22182	173	20170	+	L2	LINE/L2	1226	1408	1342
1550	26.1	5.2	2.5	c110	23733	24292	560	18060	+	L2	LINE/L2	1490	2064	686
319	22.5	1.4	0	c110	25480	25550	71	16802	+	L2	LINE/L2	2096	2167	583
401	29.2	13.5	4	c110	30575	31019	445	11333	+	L2	LINE/L2	2262	2748	2
							10841 Total LINE DNA							
							25.60% LINE DNA							
LTR DNA														
1685	16.9	2	0	c110	9267	9881	615	32471	C	MER50	LTR	710	1	24
3934	16.6	2.3	0.6	c110	18612	19310	699	23042	C	MER50	LTR	711	1	23
766	26.4	2.4	2.8	c110	24293	24580	288	17772	C	MLT1C	LTR/MaLR	465	179	1
682	17.9	10.5	0.6	c110	24876	25037	162	17315	C	MLT1C	LTR/MaLR	185	8	281
2167	7.5	2.6	0.3	c110	1908	2253	346	40099	C	THE1A	LTR/MaLR	354	1	0
2066	10.6	1.4	0.3	c110	22274	22631	358	19721	C	THE1B	LTR/MaLR	362	1	2
1797	11.1	3.4	5.1	c110	41377	41728	352	624	+	THE1C	LTR/MaLR	1	346	25
2245	18.9	2.9	0.9	c110	646	1095	450	41257	C	MER31-internal	LTR/MER4	4405	3947	531
449	16.5	3.3	0	c110	1214	1304	91	41048	C	MER31-internal	LTR/MER4	3813	3720	1123
1278	17.5	2.5	1.4	c110	1622	1907	286	40445	C	MER31-internal	LTR/MER4	3728	3440	1208
1056	19.8	11.1	3.9	c110	2254	2612	359	39740	C	MER31-internal	LTR/MER4	3444	3060	1492

975	22.7	5.9	8.1	c110	3345	3850	506	38502	C	MER31-internal	LTR/MER4	2884	2390	2052
514	27.6	6.1	0	c110	4523	4750	228	37602	C	MER31-internal	LTR/MER4	1980	1739	2956
639	23.6	4.5	4.1	c110	6560	6805	246	35547	C	MER31-internal	LTR/MER4	1567	1321	3369
746	21	10.3	2	c110	7241	7611	371	34741	C	MER31-internal	LTR/MER4	1322	888	3614
643	21.5	13.4	7.9	c110	2260	2714	455	39638	C	MER57_internal	LTR/MER4-group	5285	4806	2252
2944	11.6	0.5	0.2	c110	6808	7240	433	35112	+	MER57A	LTR/MER4-group	1	434	0
228	25.5	2	0	c110	7666	7716	51	34636	C	MER67C	LTR/MER4-group	251	200	459
1391	35.1	5.1	5.5	c110	12500	15080	2581	27272	C	HERV9	LTR/Retroviral	5679	3110	2720
362	35.3	1.4	4.1	c110	15709	16000	292	26352	C	HERV9	LTR/Retroviral	2483	2200	5916
294	38.4	2	0	c110	16220	16370	151	25982	C	HERV9	LTR/Retroviral	1957	1804	6442
371	33.5	1.4	3.7	c110	16692	16906	215	25446	C	HERV9	LTR/Retroviral	1516	1307	6883
366	34.9	0	0	c110	10384	10535	152	31817	C	HERVH	LTR/Retroviral	7284	7133	429
390	11.5	0	2.6	c110	35819	35896	78	6456	+	LTR1	LTR/Retroviral	152	227	558
2238	11.1	4.9	4.4	c110	35897	36303	407	6049	+	LTR1	LTR/Retroviral	377	785	0
3762	12.7	8.1	5.2	c110	36752	37514	763	4838	+	LTR1	LTR/Retroviral	1	785	0
204	26.7	8.3	0	c110	17725	17784	60	24568	+	LTR2	LTR/Retroviral	310	374	75
1159	23.7	4.4	11.6	c110	8333	9076	744	33276	C	LTR8	LTR/Retroviral	691	1	0
							11739 Total LTR DNA							
							27.72% LTR DNA							
Alu DNA														
796	20	3	0.6	c110	2747	2911	165	39441	C	AluJ	SINE/Alu	302	134	0
717	17.5	0.8	0	c110	3219	3344	126	39008	C	AluJ	SINE/Alu	127	1	175
493	19.4	1.1	0	c110	9174	9266	93	33086	C	AluJb	SINE/Alu	281	188	21
1421	21.9	1.7	1.7	c110	25153	25453	301	16899	C	AluJb	SINE/Alu	302	2	0
1044	16.8	0	4.9	c110	19311	19513	203	22839	C	AluJo	SINE/Alu	193	1	109
1593	17	0.3	3.5	c110	33613	33924	312	8428	C	AluJo	SINE/Alu	302	1	0
692	8.1	0	1	c110	1104	1202	99	41150	C	AluS	SINE/Alu	99	2	204
1927	9.6	4.3	0	c110	3852	4131	280	38221	C	AluSc	SINE/Alu	292	1	7
2235	8	0	0	c110	6204	6503	300	35849	+	AluSg	SINE/Alu	1	300	0
1718	14	0	6	c110	1307	1620	314	40732	C	AluSx	SINE/Alu	295	1	7
1935	13.4	0	0.3	c110	2912	3210	299	39142	C	AluSx	SINE/Alu	299	2	3
1920	12.8	0.7	0	c110	24584	24873	290	17479	C	AluSx	SINE/Alu	293	2	9
1964	10.6	0	3.2	c110	25811	26121	311	16231	+	AluY	SINE/Alu	1	301	0
2558	0.7	0.3	0	c110	22833	23132	300	19220	C	AluYa5	SINE/Alu	301	1	0
							3393 Total Alu DNA							
							8.01% Alu DNA							

Medium Frequency Repetitive DNA														
723	24.5	2	4.1	c110	31238	31482	245	10870	+	MIR	SINE/MIR	23	262	0
523	27.6	1.8	0	c110	35509	35678	170	6674	C	MER86	Unknown	174	2	32
843	26.1	5.5	3.9	c110	227	536	310	41816	C	MER21B	Unknown/MER21_g	653	339	142
229	8.3	2.8	0	c110	7925	7960	36	34392	+	MER34	Unknown/MER21_g	483	519	27
1056	8.3	0.6	0.6	c110	21210	21365	156	20987	+	MER75	DNA/T2_type	359	514	0
							917 Total Medium Repetitive DNA							
							2.17% Medium Repetitive DNA							
Simple Tandem Repetitive DNA														
261	11.9	5.1	3.4	c110	34680	34738	59	7614	C	(GA)n	Simple_repeat	60	1	60
210	13.3	0	0	c110	2717	2746	30	39606	C	(GAAA)n	Simple_repeat	33	4	87
499	15.1	0	1.2	c110	25064	25149	86	17203	C	(GGAA)n	Simple_repeat	88	4	32
241	9.7	0	0	c110	26122	26193	72	16159	+	(GGAA)n	Simple_repeat	1	72	48
384	12	8	1.3	c110	41729	41803	75	549	C	(GGAA)n	Simple_repeat	80	1	40
							322 Total Simple Repetitive DNA							
							0.76% Simple Repetitive DNA							
Low Complexity DNA														
26	0	0	0	c110	97	122	26	42230	C	AT_rich	Low_complexity	26	1	274
							26 Total Low Complexity DNA							
							0.06% Low Complexity DNA							
							27238 Total Repetitive DNA							
							64.31% Repetitive DNA							

See :Legend for Table 1.

in the same orientation and a 72 nt GGAA simple tandem repeat are found. The presence of the Alu element is easily explained by an insertional event, but the presence of the tandem repeat is more puzzling. An element of this length could arise due to an expansion similar to that seen in Huntington's disease and Fragile X syndrome. However, these two diseases are the result of trinucleotide repeat expansions and this element is made up of a tetranucleotide repeat. Sputnik shows that at least 10 copies of this poly-purine element are perfectly repeated within this region. Furthermore, Sputnik determines that the region upstream of the GGAA repeat contains 10 perfectly repeated copies of an AAAG repeat (Table 13). The second gap in the L1MB6 is the result of the insertion of a L1PA2 in the opposite orientation. Therefore, a primate L1 is inserted in a mammalian L1 in agreement with the age of the respective elements.

Table 13. Simple Tandem Repeats found with Sputnik for Cosmid c110

Begin	End	Length	(Repeat Unit)xN	Begin	End	Length	(Repeat Unit)xN
1475	1496	22	(ATT)x7	25081	25105	25	(CCTT)x6
2722	2740	19	(TTTC)x4	25127	25150	24	(CCCT)x6
2745	2760	16	(TTTC)x4	26106	26146	41	(AAAG)x10
4202	4212	11	(TG)x5	26147	26189	43	(AAGG)x10
6698	6710	13	(TGAT)x4	35686	35696	11	(TG)x5
7639	7654	16	(TTC)x5	41729	41758	30	(TTCC)x7*
19777	19789	13	(GAAA)x3	41789	41803	15	(CCTT)x5

* Denotes imperfect tandem repeat.

The two L1PB2 fragments also probably arose as the result of a single insertional event and they were later separated by insertion in the same orientation of an AluYa5. In this case, there appears to have been some duplication of L1 sequence at the Alu boundaries. The two L1MD1 fragments are separated by 53 nt of unknown origin. As with the L1PB2, some duplication of L1 sequence appears to have taken place in the insertional event that separated the two fragments.

If all the L2 fragments are pieced together, a single virtually full-length L2 element is seen within cosmid c110. The first ~450 nt are missing as is ~600 nt between 620 and 1226 of the L2 consensus. The L2 element spans 10kB of c110 and judging from the greater percent divergence from the consensus relative to other interspersed repetitive elements, may be one of the oldest repeat elements detected in this cosmid. The 600 nt gap in the L2 consensus has no detectable repeat elements within it and spans a 304 nt region of c110. However, the remaining breaks in the L2 element are easily seen to be the result of insertion of other repeat elements resulting in the four-fold expansion of the genome size with respect to the original L2. The inserted elements include the L1PB2 and L1MB6 regions discussed above as well as a third complicated region that extends from 24293 to 25480. Here, a MLT1C has been split by an AluSx. Both elements are on the opposite strand relative to the L2. Upstream of the MLT1C is an 86 nt GGAA simple tandem repeat which is further preceded by a AluJb in the same orientation as the MLT1C and the AluSx. Again, the origin of a GGAA repeat of this length can only be speculated on, but Sputnik reveals that this repeat is perfectly repeated at least six times (Table 13).

Cosmid c110 is 27.72% LTR DNA and similar to the LINE DNA discussed above, there are two very complicated LTR elements that have been fragmented by later insertion of other repeat DNA. The first of these is a MER31 that has expanded from an original size of 3.5kB to 7.0kB due to insertional events. The MER31 is in eight fragments and lacks the first 887 nt and the last 531 nt although the position of MER31 within c110 makes it possible that the last 531 nt are outside c110. Starting from the 5' end of the MER31, the first gap is the result of insertion of a full-length MER57A in the

opposite orientation to the MER31. The next two gaps in the MER31 are the due to insertion of an AluSg and an AluSc. The AluSg is inserted opposite to the MER31 and AluSc inserted the same orientation. The fourth gap in the MER31 resulted from the insertion of an AluJ, which was later interrupted by the insertion of an AluSx in the central poly-A tract of the AluJ. At the other end of the MER31 gap is a MER57 and separating the AluJ from the MER57 is a 30 nt GAAA tandem repeat. Both Alus and the MER57 are inserted in the same orientation as the MER31. The remaining three gaps contain a THE1A, a full-length AluSx, and an AluS fragment respectively, all in the same orientation as the MER31. The MER31 is lacking about 150 nt relative to the consensus around the AluS fragment.

The second notable LTR region is a HERV9 that is found in four fragments by Crossmatch. None of the fragments is separated by another repetitive element. However, comparison of the size of the gaps in the consensus HERV9 and the length of the c110 sequence between fragments reveals that the genomic gaps and the HERV9 gaps are similar in length. Two LTR1s are found in three fragments. One of these is a full-length element extending from 36752 to 37514 of c110. The other two fragments probably arose from a single element that either had or subsequently underwent a deletion of 150 nt followed by religation.

There are 14 Alu elements in c110, three of which are fragments. None of the Alu elements found within this cosmid are from the monomer families, five are from an AluJ subfamily, seven are from an AluS subfamily, and two are of AluY origin. The significance of many of the Alu elements has been discussed in relation to other repeat classes. However, one other feature worth noting is a minus strand AluJo fragment that

ends at 19311. A nearly full-length MER50 truncates the Alu and begins at 19310. The remaining 109 nt from the AluJo are not found downstream of the MER50 and may have been lost in the same event that led to the loss of the last 23 nt of MER50 sequence.

The low complexity and simple tandem repeat DNA detected by Crossmatch totals less than one percent of c110. Likewise, Sputnik detects 299 nt of simple repetitive DNA, the significance of which has been discussed above. The A/T-rich character of c110 is reflected in the A/T-rich character of the simple repeats. Of the 14 simple repeats detected by Sputnik, 13 are 50% A/T or more and account for 92% of the total simple repetitive DNA found.

9p22, The af9 Gene Region

Four cosmids making up three contigs on the genomic level have been sequenced in the af9 gene region. From centromere to telomere, the cosmids are 34a5, 92f5, c48, and 213a7. A graphical representation of the sequence features found in the three af9 gene region contigs is presented in Figures 6a, b, and c. The Crossmatch output of the alignment of the af9 cDNA sequence with the three 9p22 contigs is shown in Table 14 and represented in Figures 6a, b, and c.

Table 14. Crossmatch Alignment of the af9 cDNA within the three 9p22 contigs

Score	% Sub.	% Del.	% Ins.	Query	Begin	End	Rem.	Or.	Subject	Begin	End	Rem.
201	0	0	0	34a5	13478	13684	27006		af9	1	207	3169
177	0.55	0	0	34a5	15096	15277	25413		af9	207	388	2988
678	0.14	0	0	ctg2	63631	62926	3609	C	af9	615	1320	2056
73	0	0	0	ctg2	14950	14873	52290	C	af9	1321	1398	1978
127	0	0	0	ctg2	12816	12685	54424	C	af9	1395	1526	1850
100	0	0	0	ctg2	10054	9952	57186	C	af9	1524	1626	1750
67	0	0	0	ctg2	4091	4017	63149	C	af9	1625	1699	1677
72	0	0	0	ctg2	2808	2735	64432	C	af9	1697	1770	1606
1557	0	0.06	0	213a7	38038	39644	2572	C	af9	3376	1769	0

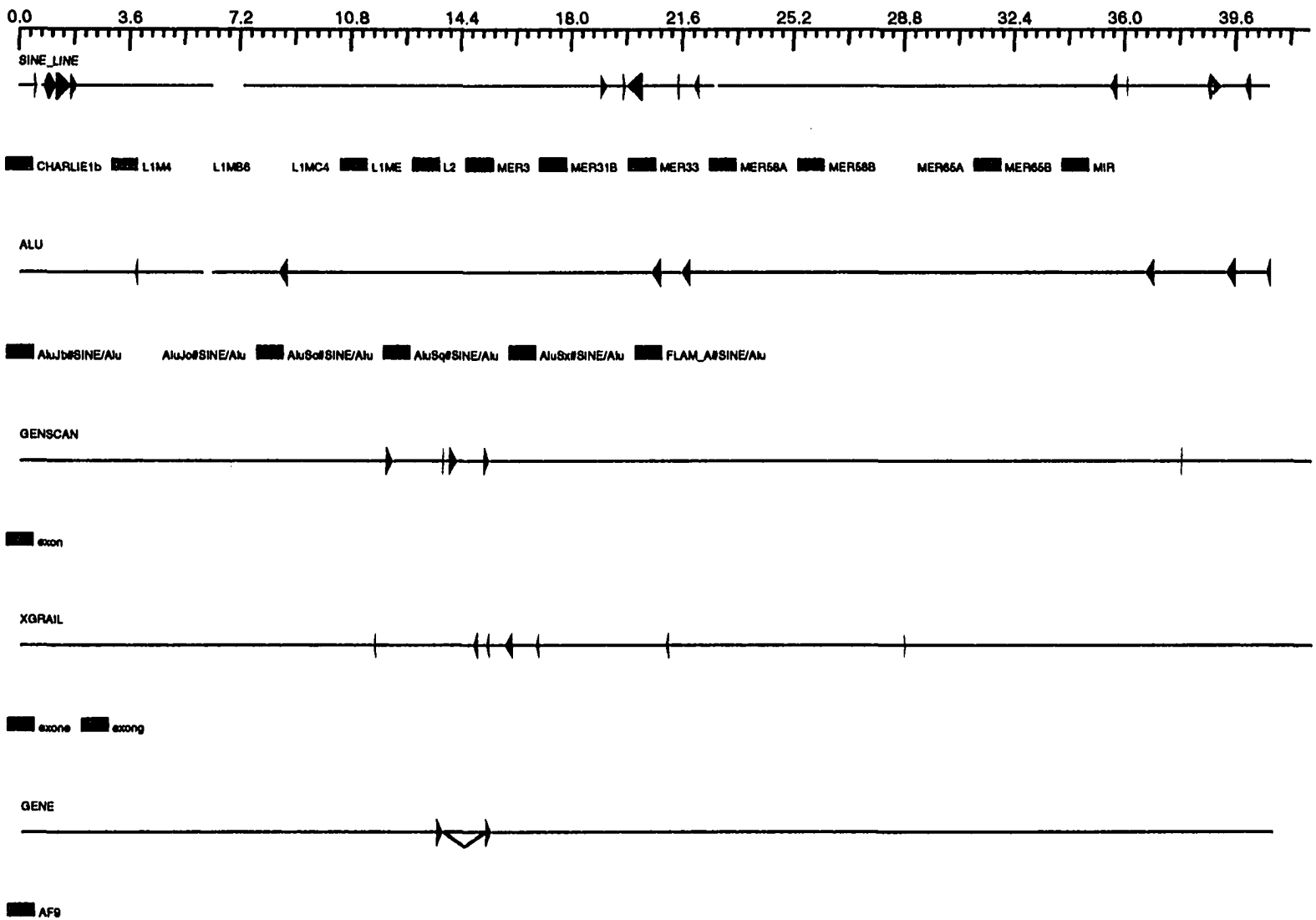


Figure 6a. Sequence Feature Map of 34a5 from af9 Gene Region See legend to Figure 3.

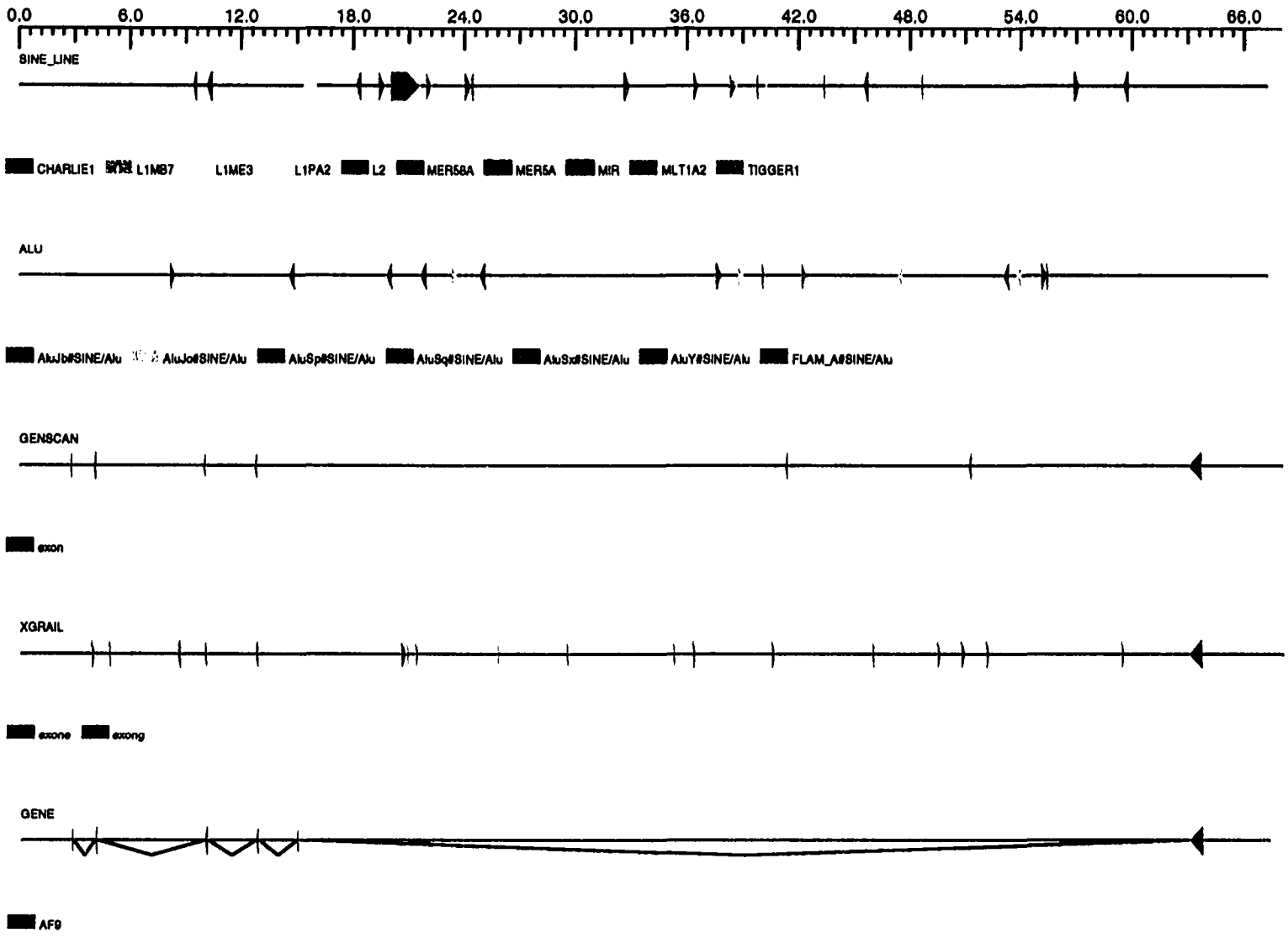


Figure 6b. Sequence Feature Map of Contig 2 from the af9 Gene Region. See legend to Figure 3.

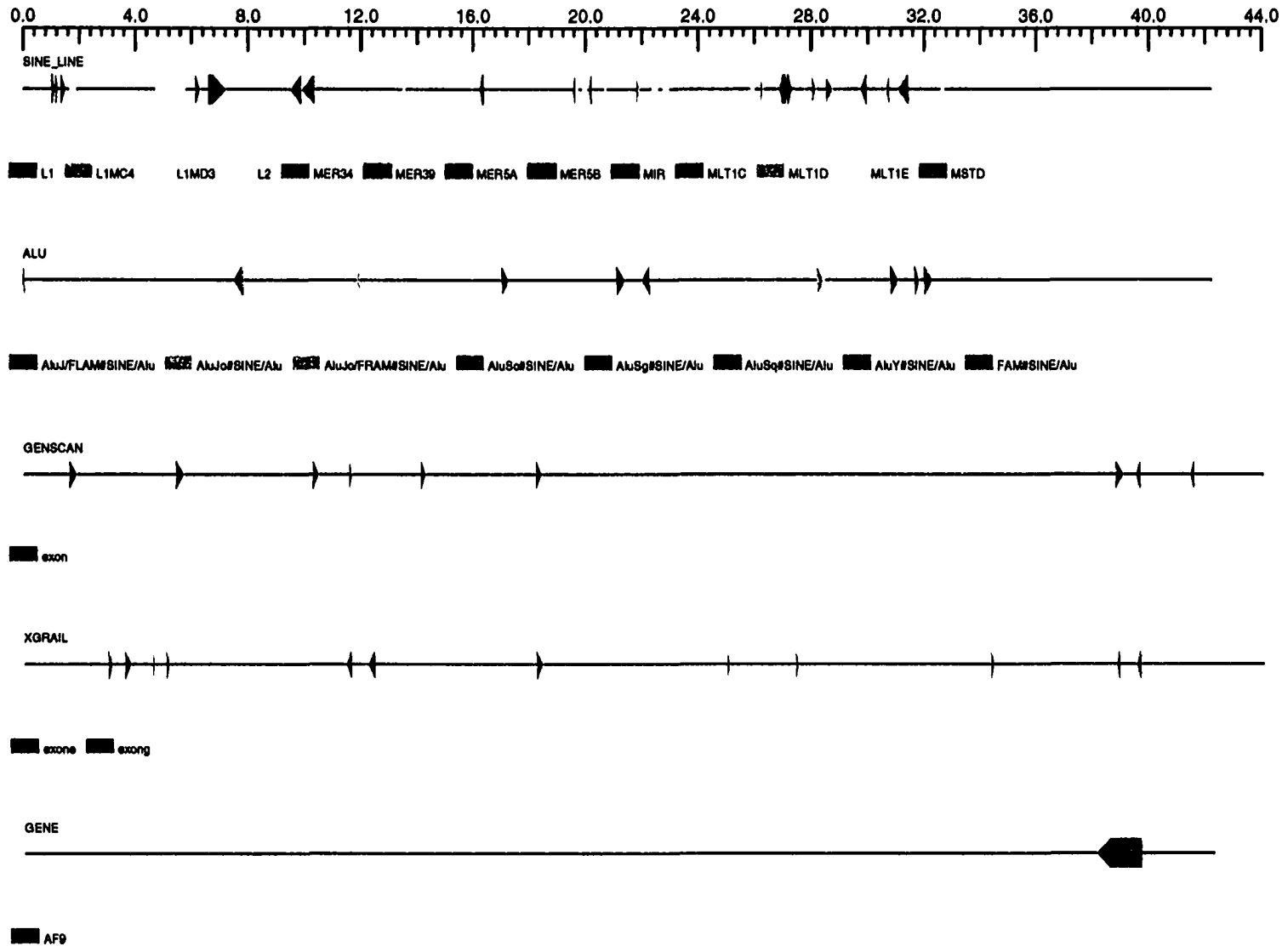


Figure 6c. Sequence Feature Map of Cosmid 213a7 from the af9 Gene Region. See legend to Figure 3.

The actual observed position of the splice junctions is shown in Table 15. These were determined by finding the consensus splice site sequence positions in proximity to the ends of the Crossmatch alignments and confirmed by the protein sequence.

Table 15. Splice Junctions within the af9 cDNA Sequence and the Three Genomic DNA Contigs.

Exon	af9 cDNA Positions		Genomic Clone Positions		
Number	Begin	End	Begin	End	Contig
1	1	207	13478	13684	1
2	208	388	15097	15277	1
3					
4	615	1320	63631	62926	2
5	1321	1396	14950	14875	2
6	1397	1526	12818	12685	2
7	1527	1626	10051	9952	2
8	1627	1698	4089	4018	2
9	1699	1770	2806	2715	2
10	1771	3376	39646	38038	3

The differences between the Crossmatch alignments and the actual splice junctions arise because Crossmatch often extends alignments over the splice sites by a few bases. 92f5 and c48 overlap but there are small gaps between this central contig and the two flanking contigs, each of which is a single cosmid. The size of the gap between the first and second contigs is unknown and several attempts to PCR across this region with genomic DNA have failed to amplify this gap even though the gap between the second and third contigs is approximately 1.5kB as determined by PCR off the same genomic DNA. Two exons from the af9 gene are found within 34a5, six more exons are within the central contig, and a single large exon is in 213a7. Approximately 220 nt of cDNA sequence and the immediately flanking genomic sequence are missing due to the gap between 34a5 and the central contig. Thus, it is most likely that the missing exon(s) is contained within this unsequenced gap region.

The genomic organization of af9 is interesting in that the exons are not evenly distributed. The first two exons are within 2kB of each other but are separated by at least 25kB from the third exon. The first two exons are similar in size, roughly 200 nt each. In comparison, the fourth exon is quite large, 700 nt. As noted above, another large intron follows the fourth exon. The next five exons are within 12.2kB and are similar in size; the largest of these five exons is ~130 nt and none of the other four are larger than 100 nt. The last exon is large, but not separated from the preceding exons by more than 6kB.

GRAIL and GENSCAN were used to determine the presence of additional genes within the af9 gene region (Table 16, Figures 6a, b, and c). The second exon of af9 is accurately predicted by GENSCAN but not by GRAIL and the first exon is not predicted by either program. The initiating methionine is at position 196 of the cDNA sequence so only the first four codons of the protein are in the first exon which is probably why the exon prediction programs were unable to find it. The large fourth exon is accurately predicted by both programs as are the sixth, and seventh exons. The eighth and ninth exons are predicted by GENSCAN but not GRAIL. The open reading frame of the tenth exon, which ends at 39511 in contig 3, also is predicted by both programs. The exons not predicted by either program are all under 100 nt in length although GENSCAN is able to predict the eighth and ninth exons, which are also less than 100 nt.

Transcription regulation of the af9 gene

GRAIL predicts one minus strand promoter within 34a5. Three CpG islands are noted by GRAIL. They are closely spaced in relation to one another spanning positions 13229 to 13572, 13850 to 15102, and 15311 to 15518. GENSCAN predicts a plus strand

Table 16. GRAIL Predicted Exons and GeneScan Predicted Genes within 9p22.

GRAIL Predicted Exons within 34a5. GeneScan Predicted Gene within 34a5.

Begin	End	Qual.	Strand	Begin	End	Strand
28718	28767	gd.	f	6895	6900	f
11453	11518	gd.	r	11891	12130	f
14737	14890	ex.	r	13754	13794	f
15156	15262	ex.	r	13962	14247	f
15734	16010	gd.	r	15097	15277	f
16760	16870	gd.	r	37787	37809	f
20987	21094	ex.	r			

GRAIL Predicted Exons within Contig 2. GeneScan Predicted Gene within Contig 2.

3849	3976	ex	f	2735	2806	r
20442	20689	gd	f	4018	4089	r
20752	20807	gd	f	9952	10051	r
21203	21334	gd	f	12685	12814	r
25656	25698	gd	f	41253	41347	r
35187	35217	ex	f	51108	51232	r
36234	36298	ex	f	62926	63630	r
40498	40610	ex	f			
49362	49487	gd	f			
50667	50865	ex	f			
52026	52157	ex	f			
59292	59383	gd	f			
4787	4826	ex	r			
8470	8605	ex	r			
9950	10050	ex	r			
12684	12814	ex	r			
29422	29497	ex	r			
45812	45895	gd	r			
62928	63631	ex	r			

GRAIL Predicted Exons within 213a7. GeneScan Predicted Genes within 213a7.

3007	3122	ex	f	1624	1886	f
3580	3798	gd	f	5431	5703	f
4572	4622	gd	f	10274	10471	f
5057	5151	gd	f	11565	11649	f
18221	18415	ex	f	14114	14266	f
24956	25018	gd	f	18221	18415	f
27386	27463	gd	f			
34312	34399	gd	f	38774	39058	f
11403	11604	gd	r			
12163	12437	gd	r	39511	39642	r
38782	38890	gd	r	41420	41556	r
39511	39642	ex	r			

promoter spanning positions 4993 to 5032. The promoter prediction programs NNPP, TSSW, and TSSG were used to analyze 1380 nt of contig 1. Of the 1380 nt, 1058 nt are upstream of the published af9 cDNA sequence. The strongest promoter predicted by

NNPP is at position 13522, which is within the published cDNA sequence and, therefore, can not be the true promoter. TSSG and TSSW each predict a TATA-less promoter at 13370 of contig 1 and a TATA-box containing promoter at 13736. Using the same reasoning as for the NNPP promoter, the TATA-box promoter predicted by TSSG and TSSW must be ruled out.

The positions of transcription factor binding sites were predicted using MatInspector and TESS. Oddly, MatInspector does not predict any general transcription factor binding sites in the promoter region predicted by TSSG and TSSW. However, TESS predicts an overlapping site for TBP, TFIID, and Sp1 centered around 13355. Furthermore, several other transcription factors are predicted to bind upstream of this site. Possible interactions between TBP or Sp1 and the upstream factors were investigated using FastM (Table 17).

Interactions of TBP with upstream binding factors were not predicted 5' of the af9 cDNA sequence position by FastM. Possible interactions with other bound factors were found for Sp1 bound to five different sites (Table 17). The site closest to the TSSW and TSSG predicted promoters would bind Sp1 at position 13283, 87 bases upstream of the predicted promoters. Oddly, the scores of the interactions between transcription factors in this region are the lowest of any Sp1 site in the af9 promoter. The highest scored interactions are for Sp1 bound to two sites further upstream at positions 12798 and 12818. In comparison to p15INK4b and p16INK4a, relatively few proteins are predicted to bind in the af9 promoter. This may be a reflection of a more specialized role for this protein within the cell. In any case, a clear picture of af9 regulation does not emerge from the analysis of the proteins predicted to bind and interact within the af9 promoter.

Table 17. FastM Predicted Protein Interactions in the af9 Promoter

Sp1 Position	Upstream Protein	Upstream Position	FastM Score
12798	Mzf1	12588	1.91
	Elk1	12711	1.86
	Ik1	12588	1.80
	Nf1	12790	1.80
	Hfh2	12758	1.79
	Oct_1	12762	1.79
	Average Score		1.83

Sp1 Position	Upstream Protein	Upstream Position	FastM Score
12818	Mzf1	12802	1.89
	GC-box	12798	1.82
	Nf1	12814	1.74
	Average Score		1.82

Sp1 Position	Upstream Protein	Upstream Position	FastM Score
12841	GC-box	12818	1.69
	Mzf1	12834	1.67
	Average Score		1.68

Sp1 Position	Upstream Protein	Upstream Position	FastM Score
13156	Mzf1	12854	1.77
	Oct_1	12892	1.72
	Nf1	13096	1.70
	Hfh2	12948	1.67
	Elk1	12998	1.66
	Average Score		1.70

Sp1 Position	Upstream Protein	Upstream Position	FastM Score
13283	Nf1	13177	1.69
	Elk1	13222	1.69
	GC-box	13156	1.65
	Hfh2	13231	1.63
	Mzf1	13275	1.62
	Average Score		1.66

Judging from the location of the first CpG island and the promoters predicted by TSSW and TSSG, a transcription start site downstream of the Sp1 site at 13283 is likely.

GRAIL predicts three minus strand polyadenylation sites within 213a7 at positions 30077, 23257, and 17848. GENSCAN places a polyadenylation site at 39381. A 4980 nt contig of 213a7 was used to predict polyadenylation sites downstream of af9 with POLYAH. The contig used included 4411 nt of sequence 3' to the af9 ORF. The first two sites at 39380 and 38721 of 213a7 are within the published cDNA sequence and the first is predicted by GENSCAN. The third site has a low score. The next three sites are at 36994, 36778, and 35528 of 213a7. Since the mechanism of polyadenylation

involves addition of rATP 3' of the polyadenylation site, the first two sites are more likely to be biologically active sites.

Repetitive DNA within the af9 gene region

In contrast to the two 9p21 gene regions already discussed, 9p22 is low in total repetitive DNA, 18.5% overall (Table 18). Contig 1 has seven LINE fragments within it, six of which are in the second intron of af9. The remaining LINE element encompasses 1021 nt, almost half of the total LINE DNA in contig 1, and lies 5' or centromeric to af9. There are eight Alu elements within contig 1, including a full-length FLAM, only one of which can be considered less than full-length. The Alu and LINE DNA is clustered in the second intron of af9 and almost exclusively oriented opposite to the direction of af9 transcription. In contrast to the interspersed repeats, the simple repetitive elements are clustered in the central part of this region. Approximately 75% of the simple repeat elements are found in a window between positions 10000 and 24000 of contig 1 (Table 18, 19). A 37 nt CAG repeat and a 24 nt AGCG repeat are found in the 5'UTR of af9 (Table 18, 19).

Contig 2 is relatively lower in LINE DNA in agreement with the intragenic position of this contig. Only six fragments are found (Table 18). Two L1PA2 fragments are in conjunction but in opposite orientations. A trivial hypothesis is that both are from a common insertional event followed by excision, inversion and reinsertion of one piece of the L1PA2. The LINE DNA is all in the large intron between the fourth and fifth af9 exons. Fifteen Alu elements are in contig 2 and all can be considered full-length. Seven of these elements are evolutionarily old, either FLAM or AluJ in origin. Of the

Table 18. RepeatMasker Output from the af9 Gene Region

RepeatMasker Output from Cosmid 34a5 (9p22 contig 1)

Score	% Sub.	% Del.	% Ins.	Query	Begin	End	Length	Rem.	Or.	Repeat Family	Repeat Class	Begin	End	Rem.
LINE DNA														
1338	19.1	3.3	1.2	34a5	38770	39168	399	1522	+	L1M4	LINE/L1	4271	4677	1469
654	14.1	2.3	0.8	34a5	22612	22739	128	17951	C	L1MB6	LINE/L1	6175	6046	0
1686	24.6	2.8	10.4	34a5	6300	7320	1021	33370	C	L1MC4	LINE/L1	7882	6876	95
339	31.2	6.2	2.2	34a5	18928	19199	272	21491	+	L1ME	LINE/L1	5520	5802	366
397	33.2	1.4	0.5	34a5	21937	22144	208	18546	C	L1ME	LINE/L1	5745	5536	419
204	10.2	0	12.2	34a5	21422	21470	49	19220	C	L2	LINE/L2	2750	2708	0
270	28	10.8	0	34a5	35529	35734	206	4956	C	L2	LINE/L2	2708	2536	2
							2283							
							5.61% LINE DNA							
LTR DNA														
2404	14.6	8.6	0.2	34a5	1233	1676	444	39014	+	MER31B	LTR/MER4-group	1	481	4
277	30.3	3	3.5	34a5	551	748	198	39942	C	MER65A	LTR/MER4-group	445	249	0
445	22	3.8	5.7	34a5	802	960	159	39730	C	MER65B	LTR/MER4-group	156	1	375
							801							
							1.97% LTR DNA							
Alu DNA														
1436	18.7	0	5.9	34a5	20545	20865	321	19825	C	AluJb	SINE/Alu	302	1	0
722	17.1	0	2.9	34a5	40548	40687	140	3	C	AluJb	SINE/Alu	301	166	1
1512	17.7	0.3	3	34a5	5970	6274	305	34416	+	AluJo	SINE/Alu	6	302	0
2099	9.9	0.3	0	34a5	8408	8700	293	31990	C	AluSc	SINE/Alu	295	2	4
2090	10	0	0.3	34a5	36614	36912	299	3778	C	AluSc	SINE/Alu	298	1	1
2118	10.6	0.3	0	34a5	21512	21813	302	18877	C	AluSq	SINE/Alu	303	1	0
1899	11.4	0.3	2.4	34a5	39232	39529	298	1161	C	AluSx	SINE/Alu	292	1	10
692	14.2	0	0	34a5	3742	3847	106	36843	C	FLAM_A	SINE/Alu	130	25	2
							2064							
							5.07% Alu DNA							
Medium Frequency Repeat DNA														
1584	22.5	6.3	4.1	34a5	19789	20295	507	20395	C	CHARLIE1b	DNA/MER1_type	518	1	0
534	18.9	8.5	0.7	34a5	39912	40064	153	626	C	MER3	DNA/MER1_type	204	40	5
231	19.7	0	1.6	34a5	491	550	60	40140	+	MER33	DNA/MER1_type	25	83	241

720	17.1	7.5	0.4	34a5	962	1189	228	39501	+	MER33	DNA/MER1_type	81	324	0
232	16.1	1.8	5.4	34a5	1199	1232	34	39458	+	MER58A	DNA/MER1_type	4	35	189
988	20.8	3.7	0.4	34a5	1670	1914	245	38776	+	MER58B	DNA/MER1_type	16	268	70
230	16.8	15.8	7.9	34a5	19683	19783	101	20907	+	MIR	SINE/MIR	82	190	72
210	19.2	5.8	0	34a5	36060	36111	52	4579	C	MIR	SINE/MIR	141	87	121
211	27.6	4.1	4.1	34a5	38672	38769	98	1921	C	MIR	SINE/MIR	238	141	24
							1478							

3.63% Medium Frequency Repetitive DNA

Low Complexity DNA

23	7.3	0	0	34a5	8810	8850	41	31840	C	AT_rich	Low_complexity	41	1	259
22	0	0	0	34a5	22317	22338	22	18352	C	AT_rich	Low_complexity	91	70	219
21	3.7	0	0	34a5	24384	24410	27	16280	+	AT_rich	Low_complexity	92	118	182
25	7	0	0	34a5	28652	28694	43	11996	C	AT_rich	Low_complexity	69	27	231
26	0	0	0	34a5	30525	30550	26	10140	C	AT_rich	Low_complexity	26	1	274
32	2.6	0	0	34a5	40487	40524	38	166	C	AT_rich	Low_complexity	38	1	262
30	6.7	0	0	34a5	14875	14919	45	25771	C	GC_rich	Low_complexity	163	119	137
33	14.4	0	0	34a5	15335	15452	118	25238	C	GC_rich	Low_complexity	118	1	182
							360							

0.88% Low Complexity DNA

Simple Tandem Repeat DNA

292	13.3	0	0	34a5	13870	13914	45	26776	+	(CA)n	Simple_repeat	1	45	75
407	10	0	1.7	34a5	33112	33171	60	7519	C	(CA)n	Simple_repeat	60	2	60
209	16.2	0	0	34a5	13451	13487	37	27203	C	(CAG)n	Simple_repeat	37	1	83
395	25.4	2.7	1.8	34a5	11959	12068	110	28622	+	(GAAA)n	Simple_repeat	4	114	6
209	25.7	0	5.7	34a5	5667	5736	70	34954	+	(GAAAA)n	Simple_repeat	3	68	52
213	20	2.2	0	34a5	10867	10911	45	29779	C	(GAAAA)n	Simple_repeat	47	2	73
264	10.5	0	0	34a5	21474	21511	38	19179	C	(GGGA)n	Simple_repeat	38	1	82
208	30.1	4.8	1.9	34a5	23361	23463	103	17227	C	(TAAAA)n	Simple_repeat	107	2	13
206	21.4	2.4	0	34a5	10380	10421	42	30269	+	(TGG)n	Simple_repeat	1	43	77
345	7	0	0	34a5	15634	15676	43	25014	C	CA)n	Simple_repeat	44	2	76
233	6.5	0	0	34a5	35781	35811	31	4879	C	POLY_A	Simple_repeat	31	1	89
196	25.5	0	0	34a5	13819	13865	47	26825	C	POLY_G	Simple_repeat	47	1	73
							671							

1.65%

7657 Total Repetitive DNA from 34a5

18.82% Repetitive DNA from 34a5

RepeatMasker Output from 9p22 Central Contig (contig 2)

Score	% Sub.	% Del.	% Ins.	Query	Begin	End	Length	Rem.	Or.	Repeat Family	Repeat Class	Begin	End	Rem.	
LINE DNA															
1891	20.9	1.1	2.7	9p22	38251	38694	444	28546	+	L1MB7	LINE/L1	5676	6112	63	
379	15.4	3.2	6.5	9p22	40178	40300	123	26940	C	L1ME3	LINE/L1	6157	6039	7	
4740	11.3	1.1	0.1	9p22	15266	15988	723	51252	C	L1PA2	LINE/L1	6145	5416	1	
510	5.3	0	0	9p22	15985	16060	76	51180	+	L1PA2	LINE/L1	5348	5423	740	
246	31.2	0	3.7	9p22	39670	39778	109	27462	C	L2	LINE/L2	2749	2645	1	
223	26.2	8.3	1.2	9p22	48558	48641	84	18599	C	L2	LINE/L2	2734	2645	16	
							1559								
							2.32% LINE DNA								
LTR DNA															
1440	19.7	5.2	4.1	9p22	24036	24380	345	42860	+	MLT1A2	LTR/MaLR	26	374	0	
							345								
							0.51% LTR DNA								
Alu DNA															
1681	17.3	0	1	9p22	53025	53325	301	13915	C	AluJb	SINE/Alu	298	1	4	*
1724	16.8	0	0.3	9p22	23307	23597	291	43643	+	AluJo	SINE/Alu	2	291	11	*
1458	22.6	0.3	1.6	9p22	38698	39003	306	28237	+	AluJo	SINE/Alu	1	302	0	*
1161	18.6	10.2	0	9p22	47308	47571	264	19669	C	AluJo	SINE/Alu	302	12	0	*
1558	19.8	0.7	0.3	9p22	53713	54010	298	13230	C	AluJo	SINE/Alu	299	1	3	*
1917	10.6	0.4	0	9p22	42100	42373	274	24867	+	AluSp	SINE/Alu	22	296	7	*
2082	7.2	4.1	0	9p22	8085	8375	291	58865	+	AluSq	SINE/Alu	1	303	0	*
1980	12.4	0	0	9p22	19723	20013	291	47227	C	AluSq	SINE/Alu	296	6	7	*
2012	12.9	0	0	9p22	55035	55336	302	11904	+	AluSx	SINE/Alu	1	302	0	*
1715	10.2	0.7	4	9p22	14546	14820	275	52420	C	AluY	SINE/Alu	299	34	2	
1837	9.1	3.6	5.5	9p22	21606	21912	307	45328	C	AluY	SINE/Alu	301	1	0	*
2273	6.3	0	0.3	9p22	24811	25110	300	42130	C	AluY	SINE/Alu	301	3	0	*
2103	9.2	0	0	9p22	37512	37804	293	29436	+	AluY	SINE/Alu	1	293	8	*
576	22.6	0	0	9p22	40037	40151	115	27089	+	FLAM_A	SINE/Alu	1	115	17	
618	15	0	2.6	9p22	55351	55463	113	11777	+	FLAM_A	SINE/Alu	21	130	2	
							4021								
							5.98% Alu DNA								

Medium Frequency Repeat DNA

210	24.1	5.6	0	9p22	43334	43387	54	23853	C	CHARLIE1	DNA/MER1_type	2444	2388	295
977	22	1.3	4.7	9p22	36299	36530	232	30710	+	MER58A	DNA/MER1_type	1	224	0
536	21.6	2.5	7.5	9p22	18172	18370	199	48870	C	MER5A	DNA/MER1_type	189	1	0
1593	13.4	1.7	0.6	9p22	19371	19720	350	47520	+	TIGGER1	DNA/MER2_type	108	461	1957
7748	10.6	2.1	0.3	9p22	20017	21605	1589	45635	+	TIGGER1	DNA/MER2_type	459	2076	342
1427	19.5	9.3	0.3	9p22	21917	22239	323	45001	+	TIGGER1	DNA/MER2_type	2066	2417	1
300	26.3	9	0.8	9p22	9408	9540	133	57700	C	MIR	SINE/MIR	262	119	0
464	26.1	18.6	0.5	9p22	10180	10394	215	56846	C	MIR	SINE/MIR	255	2	7
207	15.6	0	4.4	9p22	24403	24447	45	42793	C	MIR	SINE/MIR	141	99	121
501	28.7	7.3	3.2	9p22	32590	32836	247	34404	+	MIR	SINE/MIR	6	262	0
275	26.6	17.5	2.8	9p22	45533	45709	177	21531	C	MIR	SINE/MIR	249	47	13
622	25.1	3.1	2.6	9p22	56875	57069	195	10171	+	MIR	SINE/MIR	49	244	18
494	25.4	7.3	1.7	9p22	59559	59735	177	7505	C	MIR	SINE/MIR	218	32	44
							3936							

5.85% Medium Frequency Repetitive DNA

Low Complexity DNA

21	3.7	0	0	9p22	1911	1937	27	65303	C	AT_rich	Low_complexity	27	1	273
21	0	0	0	9p22	11917	11937	21	55303	C	AT_rich	Low_complexity	21	1	279
26	6.8	0	0	9p22	30832	30875	44	36365	+	AT_rich	Low_complexity	1	44	256
22	3.6	0	0	9p22	32849	32876	28	34364	C	AT_rich	Low_complexity	53	26	247
25	0	0	0	9p22	34259	34283	25	32957	+	AT_rich	Low_complexity	45	69	231
22	8.7	0	0	9p22	34691	34736	46	32504	C	AT_rich	Low_complexity	71	26	229
23	3.5	0	0	9p22	41100	41128	29	26112	C	AT_rich	Low_complexity	29	1	271
26	0	0	0	9p22	54169	54194	26	13046	+	AT_rich	Low_complexity	1	26	274
24	0	0	0	9p22	66017	66040	24	1200	+	AT_rich	Low_complexity	1	24	276
							270							

0.40% Low Complexity DNA

Simple Tandem Repeat DNA

313	11.1	0	0	9p22	7190	7234	45	60006	+	(CA)n	Simple_repeat	1	45	75
813	2.5	0	0	9p22	63488	63607	120	3633	C	(CAG)n	Simple_repeat	120	1	0
222	25.5	0	0	9p22	7140	7186	47	60054	C	(GA)n	Simple_repeat	47	1	73
241	2.8	5.6	0	9p22	66382	66417	36	823	C	(TA)n	Simple_repeat	39	2	81
398	10.5	4	1.3	9p22	66546	66621	76	619	C	(TA)n	Simple_repeat	79	2	41
216	30.5	0	0	9p22	25705	25763	59	41477	C	(TAA)n	Simple_repeat	59	1	61
217	17.9	0	0	9p22	4292	4330	39	62910	+	POLY_A	Simple_repeat	1	39	81
							422							

0.63% Simple Repetitive DNA

10553 Total Repetitive DNA from 9p22 Central Contig

15.69% Repetitive DNA from 9p22 Central Contig

RepeatMasker Output for Cosmid 213a7 (9p22 contig 3)

Score	% Sub.	% Del.	% Ins.	Query	Begin	End	Length	Rem.	Or.	Repeat Family	Repeat Class	Begin	End	Rem.
LINE DNA														
206	36.1	2.4	1.2	213a7	6147	6315	169	35901	+	L1	LINE/L1	3991	4161	1985
285	26.8	5.5	0.8	213a7	6613	7254	642	34962	+	L1	LINE/L1	4633	5350	796
316	21	5.9	7.6	213a7	26208	26326	119	15890	+	L1MC4	LINE/L1	7861	7977	0
347	27.4	0.6	5.7	213a7	25845	26019	175	16197	+	L1MD3	LINE/L1	7348	7513	227
532	29.8	5.2	3.1	213a7	1612	1900	289	40316	C	L2	LINE/L2	2605	2311	145
1116	30.5	10	1.9	213a7	4685	5791	1107	36425	C	L2	LINE/L2	2749	1605	1
253	28.4	12.7	1.5	213a7	13450	13583	134	28633	C	L2	LINE/L2	2708	2560	2
388	22.6	7	2.6	213a7	19632	19746	115	22470	+	L2	LINE/L2	2122	2241	509
257	27.1	18.6	0.9	213a7	19870	20079	210	22137	+	L2	LINE/L2	2457	2703	47
207	27.8	15.7	0	213a7	20654	20761	108	21455	+	L2	LINE/L2	2626	2750	0
292	33.7	7	2.6	213a7	22713	22982	270	19234	+	L2	LINE/L2	2129	2410	340
236	24	0	0	213a7	27731	27780	50	14436	+	L2	LINE/L2	2695	2744	6
420	24.2	8.5	3	213a7	32569	32733	165	9483	C	L2	LINE/L2	2745	2572	5
							3553							

8.42% LINE DNA

LTR DNA													
1587	20.6	4.9	6.4	213a7	9893	10358	466	31857	C	MLT1C	LTR/MaLR	459	1	7
330	24	1.6	5.6	213a7	21795	21932	138	20284	+	MLT1D	LTR/MaLR	14	160	381
670	21	9.6	0	213a7	28025	28191	167	14025	+	MLT1D	LTR/MaLR	7	189	316
1049	18.6	1.5	1.5	213a7	28514	28787	274	13429	+	MLT1D	LTR/MaLR	232	505	0
671	23.3	12	4.4	213a7	22316	22590	275	19626	+	MLT1E	LTR/MaLR	265	560	8
898	27.2	5.3	6	213a7	9496	9892	397	32324	C	MSTD	LTR/MaLR	394	1	0
							1717							
							4.07% LTR DNA							
Alu DNA														
315	5	0	0	213a7	1	40	40	42176	C	AluJ/FLAM	SINE/Alu	62	23	240
1734	15.7	1.4	0.3	213a7	28199	28485	287	13731	+	AluJo	SINE/Alu	7	296	6
635	20.9	0	0	213a7	11815	11943	129	30273	C	AluJo/FRAM	SINE/Alu	293	165	9
2119	9.8	0	0	213a7	21958	22252	295	19964	C	AluSc	SINE/Alu	295	1	4
984	21.4	11.8	0.8	213a7	16997	17258	262	24958	+	AluSg	SINE/Alu	1	291	9
1804	13	0.3	4.8	213a7	21079	21394	316	20822	+	AluSq	SINE/Alu	1	302	1
2029	9.1	0.3	0.3	213a7	30779	31065	287	11151	+	AluSq	SINE/Alu	2	288	15
2144	9.3	0	0.3	213a7	7503	7804	302	34412	C	AluY	SINE/Alu	301	1	0
2162	7.9	0	1	213a7	31980	32283	304	9933	+	AluY	SINE/Alu	1	301	0
771	23.8	0.6	0	213a7	31655	31818	164	10398	+	FAM	SINE/Alu	11	175	0
							2386							
							5.65% Alu DNA							
Medium Frequency Repeat DNA														
300	15.6	18.2	1.3	213a7	1027	1103	77	41113	+	MER5A	DNA/MER1_type	4	93	96
180	16	6	4	213a7	1162	1211	50	41005	+	MER5A	DNA/MER1_type	138	188	1
263	24.8	1.7	7.7	213a7	26881	26997	117	15219	C	MER5A	DNA/MER1_type	164	55	25
314	22.8	0	1.1	213a7	16262	16353	92	25863	C	MER5B	DNA/MER1_type	162	72	16
214	29.8	18.3	0	213a7	27054	27157	104	15059	+	MER5B	DNA/MER1_type	9	131	47
346	24.2	10.5	1.6	213a7	27165	27288	124	14928	+	MER5B	DNA/MER1_type	13	147	31
380	33.9	3.3	0.6	213a7	1335	1517	183	40699	+	MIR	SINE/MIR	52	239	23
207	22.2	5.6	0	213a7	19549	19620	72	22596	C	MIR	SINE/MIR	256	181	6
309	18.5	9.9	0	213a7	20120	20200	81	22016	C	MIR	SINE/MIR	160	72	102
422	27.4	8	2	213a7	29733	29933	201	12283	C	MIR	SINE/MIR	261	49	1
311	26.5	0.8	0.8	213a7	30636	30752	117	11464	C	MER34	Unknown/MER21_	534	418	12
1168	21.8	6.3	0.7	213a7	31066	31426	361	10790	C	MER39	Unknown/MER21_	412	13	281
							1579							
							3.74% Medium Frequency Repetitive DNA							

Low Complexity DNA														
22	3.6	0	0	213a7	25425	25452	28	16764	C	AT_rich	Low_complexity	28	1	272
22	0	0	0	213a7	34994	35015	22	7201	C	AT_rich	Low_complexity	47	26	253
25	0	0	0	213a7	37445	37469	25	4747	C	AT_rich	Low_complexity	25	1	275
							75							
							0.18%	Low Complexity DNA						
Simple Tandem Repeat DNA														
255	3.2	0	0	213a7	33192	33254	63	8962	+	(CA)*n*	Simple_repeat	1	63	57
202	18.4	0	0	213a7	41887	41924	38	292	+	(CA)n	Simple_repeat	1	38	82
322	9.3	0	0	213a7	35894	35936	43	6280	C	(TA)n	Simple_repeat	43	1	77
261	0	0	0	213a7	36254	36282	29	5934	+	POLY_A	Simple_repeat	1	29	91
251	20	0	0	213a7	39421	39475	55	2741	+	POLY_A	Simple_repeat	1	55	65
259	11.4	0	0	213a7	40144	40178	35	2038	+	POLY_A	Simple_repeat	1	35	85
							263							
							0.62%	Simple Repetitive DNA						
							9573	Total Repetitive DNA from 213a7						
							22.68%	Repetitive DNA from 213a7						
							27783	Total Repetitive DNA from af9 Gene Region						
							18.50%	Repetitive DNA from af9 Gene Region						

See Legend from Table 1.

Table 19. Sputnik Output of Simple Tandem Repeats from af9 Gene Region

Contig 1 (34a5).				Contig 2.				Contig 3 (213a7).			
Begin	End	Length	(Repeat Unit)xN	Begin	End	Length	(Repeat Unit)xN	Begin	End	Length	(Repeat Unit)xN
4179	4192	14	(AAAAT)x2	7192	7225	34	(CA)x17	6153	6164	12	(AT)x6
10874	10893	20	(TTTTC)x4	13574	13584	11	(CA)x5	9188	9207	20	(TTTTG)x4
13451	13471	21	(GCT)x7	14536	14554	19	(TTTTG)x3	14932	14945	14	(AAAAG)x4
13497	13520	24	(AGCG)x6	18564	18574	11	(CA)x5	21388	21402	15	(AAAC)x3
13887	13906	20	(AC)x10	21604	21618	15	(TTTC)x3	26302	26312	11	(AT)x5
13935	13947	13	(CCTC)x3	24806	24819	14	(TTTTC)x2	32274	32290	17	(AAAAT)x3
14010	14208	19	(CA)x9	28919	28929	11	(CA)x5	33195	33254	60	(AC)x30
14066	14078	13	(GCC)x4	36002	36012	11	(TA)x5	35905	35930	26	(AT)x13 *
14470	14481	12	(GT)x6	44491	44502	12	(ATT)x4	40064	40075	12	(AC)x6
15641	15676	36	(GT)x18	48105	48119	15	(AAAC)x3			187	
20028	20042	15	(CTTTT)x3	54117	54130	14	(GGAA)x3				
20727	20741	15	(TG)x7	55827	55839	13	(TTTC)x3				
21421	21434	14	(CATT)x3	56969	56984	16	(AAT)x5				
21472	21484	13	(TGCC)x3	63485	63607	123	(CTG)x4 *				
21485	21506	22	(CCCT)x5	64988	64998	11	(TA)x5				
21508	21522	15	(TTCC)x3	66387	66417	31	(TA)x14 *				
23227	23239	13	(TA)x6	66569	66586	18	(TA)x8 *				
23450	23463	14	(TTTTG)x2	66601	66613	13	(TA)x6				
30529	31543	15	(ATAA)x3			392					
33129	33169	41	(TG)x20								
40543	40559	17	(TTTTA)x3								
		386									

* Denotes imperfect tandem repeat.

remaining eight elements, four are AluY subfamily members and therefore are inserted since the human divergence from the ape. Two Alu elements are found tandemly repeated, a FLAM-A starts at 55351 and an AluSx ends at 55336. The Tigger1 element starting at 19371 is at least as old as the AluSq family as seen by the fact that it is interrupted by two Alus, an AluSq and an AluY. There are no Alu elements between positions 24000 and 37000 of contig 2. Likewise, there is no LINE DNA between 16060 and 38000. There also is a paucity of medium frequency interspersed elements in the region between 24000 and 37000, two MIRs and a MER58A are found within this span. The interspersed repeats are evenly distributed as to orientation within contig 2. Fifteen are oriented telomere to centromere, 18 are oriented centromere to telomere.

The amount of low complexity and simple tandem repeat DNA is considerably lower in contig 2. Contig 1 is 2.53% whereas contig 2 is only 1% (Table 18, 19). By far the largest single block of tandemly repeated DNA in this gene region is within the fourth exon and encodes the poly-serine tract of af9. The distribution of simple tandem repetitive and low complexity DNA is uneven. The centromeric half of contig 2 contains 75% of these classes of DNA.

Contig 3 is higher in LINE DNA than contig 1 or 2 with 8.42%. All LINE elements are telomeric to the af9 gene. A greater proportion the LINE DNA is L2 DNA, 9 of 13 fragments and 69% of the total LINE DNA in contig 3. Over 4% of contig 3 is LTR DNA, double the ratio of contig 1 and eight times the ratio of LTR DNA in contig 2. Alu DNA is found in contig 3 at the same ratio as seen in the other two af9 contigs, but similar to contig 2, a greater than expected proportion of the Alu elements are of the older FLAM, FRAM, FAM, or AluJ subfamilies. Two of the elements from the older

subfamilies are fragments. The Alu elements are exclusively telomeric of the af9 gene. The medium frequency repeat elements are clustered in a central portion of the contig except for three outliers at the extreme telomeric end of 213a7.

The simple tandem repeat and low complexity DNA is at a much lower level in contig 3 than in either contig 1 or 2 and is positionally skewed to the centromeric end of the contig, closer to the af9 gene. The bias of the two classes of simple repeat for the centromeric end of the af9 gene region is very evident and remarkable.

Breakpoints within the af9 Gene Region

Ten (9;11) translocation breakpoints have been cloned and sequenced such that the chromosome 9 sequence at the breakpoint is known. Crossmatch was used to place each breakpoint within the af9 gene region (Table 20). The stringency was reduced to a minimum match of five and a minimum score of 10 due to the short length of the patient 1 and monomac 6 breakpoint sequences.

Several sequence elements have been implicated in recombinations. Translin, which has been implicated in lymphoid leukemia translocations (213), has two consensus binding sites, ATGCA and GGCC(A/T)CCT. As discussed in the introduction, the immunoglobulin recombination sites have been implicated in recombinogenic processes as well. The immunoglobulin heptamer and nonamer consensus sites are CAC(A/T)GTG and GGTTTTTGT, respectively. A 26 nt long sequence feature commonly found within Alu repeats has been found in close proximity to recombination in many diseases (214). This last element includes a sequence similar to *chi*, the prokaryotic recombination site. The consensus site for *chi* is GCTGGTGG. The therapy related nature of 11q23 recombinations has lead to previous investigations of topoisomerase II cut sites within the

Table 20. Breakpoint Positions within the af9 Gene Region.

Score	% Sub.	% Del.	% Ins.	Query	Begin	End	Rem.	Or.	Subject	Begin	End	Rem.	Reference
34	0	0	0	Patient 3 (3173)	1	36	0	C	contig2	16525	16490*	50715	Atlas et al. Unpublished Results
24	3.7	0	0	Patient 5 (6230)	1	27	0	C	contig2	20378	20352*	46862	Atlas et al. Unpublished Results
22	0	0	0	Patient 6 (1930)	1	23	0	C	contig2	28737	28715*	38503	Atlas et al. Unpublished Results
24	0	0	0	Patient 10 (6784)	1	26	1	C	contig2	22594	22569*	44646	Atlas et al. Unpublished Results
97	0	2.65	0	Gill 2	1	113	48	C	contig2	27274	27159*	39966	Gill, Personal Communication
10	0	0	0	Gill 2	116	126	35		contig2	22840	22850	44390	
202	0	0.46	0.91	Gill 1	1	219	0	C	contig2	15042	14825*	52198	Gill, Personal Communication
19	0	0	0	Patient 1	1	20	0	C	contig2	15095	15076*	52145	Super et al. Genes Chromo. and Cancer 20: 185-95
15	0	0	0	monomac 6	1	19	0	C	contig2	26983	26965*	40257	Super et al. Genes Chromo. and Cancer 20: 185-95
83	1.04	1.04	0	Negrini bkpt	1	96	0	C	contig2	3922	3826 *	63318	Negrini et.al. Cancer Res. 53: 4489-92
10	7.69	0	0	S79541	6	18	208	C	contig2	3402	34390	32838	Negrini et.al. Cancer Genetics and
147	0.62	0.62	0.62	S79541	65	226	0	C	contig2	24717	24556*	42523	Cytogenetics 83: 65-70

* Denotes Breakpoint Position

11q23 breakpoint region (58). Topoisomerase II sites will be discussed separate of the other recombinogenic DNA elements.

In order to look for recombinogenic features at the breakpoints. FindPatterns from the GCG package was used to examine the three af9 gene region contigs for translin sites, the immunoglobulin heptamer and nonamer sequences, and *chi*. A single mismatch was allowed in this search. To search for the recombination site from Rudiger et al. (214) (the "Rudiger element"), Crossmatch was used at a reduced stringency of a minimum match of five consecutive bases and a minimum score of 15.

A comparison of the recombination sites proposed by Rudiger et al. (214) and the positions of Alu sequences shows that these sites were found within Alu sequences. Under the stringency discussed above, 12 sites were found in contig 2. It should be noted that the AluY starting at 14820 does not have a Rudiger element within or associated with it as determined by Crossmatch.

Breakpoints were found in introns four, five, and eight of af9 (Table 20, Figure 6b). Introns 5 and 8 have one breakpoint each. The remaining eight breakpoints are relatively evenly distributed between 15000 and 29000 of contig 2. Two breakpoints, one from patient Gill 1 (215) at position 14825 and the other from Patient 1 (216) at position 15076, are very close together in a region between two interspersed repeats, the L1PA2 which ends at position 15266 and an AluY beginning at 14820. Interestingly this is the Alu element lacking a Rudiger element. Two of the five nucleotide translin sites are found in proximity to the Patient 1 (216) breakpoint, one at 15064 and one at 15140. The breakpoint found in Patient 6 (217) is the most centromeric breakpoint within contig 2 at position 28715. A translin site, gTGCA, where the lower case letter denotes a

mismatch from the consensus, was found at 28751. The breakpoint from patient Gill 2 (215) is at 27159 of contig 2. A translin site consisting of tTGCA is found nearby at 27136. The next most centromeric breakpoint is from the monomac 6 cell line (216) and is at position 26983. At 27031, a translin site of ATcCA is found. Breakpoint S79541 (65) is at 24717, close to a translin site, ATcCA, at 24734. The breakpoint found in Patient 10 (217) is at 22594, between two translin sites, one a perfect match at 22553 and the other, ATGgA at 22571. The breakpoint of Patient 5 (217) is located at 20352, a translin site consisting of ATGCt begins at 20375. The last breakpoint in intron 4 is from Patient 3 (217) and is located at 16525, which has a translin site, tTGCA starting at 16534. None of the other sites investigated is found within 50 bases of any of the breakpoints. The Negrini breakpoint (74) in intron 8 is not associated with any of the repetitive elements or recombinogenic sites investigated here.

The high probability of cleavage by topoisomerase II *in vivo* has previously been determined to rely on at least a 70% match to the consensus binding site on the first strand and a 40% match to the consensus on the opposite strand (68). The first strand is defined as the strand with the highest concordance with the RNYNNCNNGYNGKTNYNY consensus sequence. FindPatterns from the GCG package was used to search for possible topoisomerase II sites by allowing a mismatch of three. Since there are ten diagnostic positions in the topoisomerase II consensus binding site, a mismatch of three finds all sites with a 70% match on at least one strand. The sites in close proximity to breakpoints were further examined to determine their concordance to the consensus on the opposite strand. The topoisomerase II sites meeting the criteria of Ebert et al. (68) are shown in Table 21. The search for elements that match the consensus

Table 21. Topoisomerase II Cut Sites in Proximity to 9p22 Breakpoints.

Breakpoint	Breakpoint Position	Topo II Cut Site Position	Topo II Site, 5' End Position	Contig 2 Topo II Binding Site Sequence	Topo II Site Strand	Plus Strand Score	Minus Strand Score
Negrini bkpt.	3825	3821	3814	AAGGCCAAGCTTAATACT	minus	50%	70%
Gill 1	14824	14815	14806	AGCCTCCCAAAGTGCCAT	plus	70%	50%
		14814	14807	GCCTCCCAAAGTGCCATT	minus	50%	70%
		14822	14815	AAGTGCCATTAAGGTGTT	minus	50%	70%
		14825	14816	AGTGCCATTAAGGTGTTT	plus	80%	40%
Patient 1	15075	15058	15049	ACCGTGATGCATTTCATT	plus	70%	50%
		15082	15075	TTACTAACACAGGAGAGT	minus	40%	70%
		15097	15088	AGAGTCATGTTGTCTTTT	plus	80%	50%
Patient 3 (3173)	16489	16477	16470	AAACCACTGAAATATAGT	minus	40%	70%
		16500	16491	AGTATCATGTTTTCTTGT	plus	80%	40%
Patient 5 (6230)	20351	20349	20342	AGGAAGCTGCAGACAAAA	minus	40%	70%
		20357	20350	GCAGACAAAAAGTTTGAT	minus	60%	70%
Patient 10 (6784)	22568	22553	22546	GCAAAATCATTTTGTAAT	minus	40%	70%
S79541	24555	24551	24542	AATTTAAAGCAGTGTTCT	plus	80%	50%
Monomac 6	26964	26964	26955	GGCATTAAGAGTTTATTT	plus	70%	30%
		26944	26937	ATTTACTATCATGGCAGT	minus	50%	70%
Gill 2	27158	None					
Patient 6 (1930)	28714	28706	28699	AATTACTGGCTAGTAAGA	minus	70%	70%
		28708	28699	AATTACTGGCTAGTAAGA	plus	70%	70%
		28711	28703	ACTGGCTAGTAAGACTTT	plus	80%	50%

at least 70% found 2372 sites on the top strand of contig 2 and 2936 sites on the bottom strand. Twenty sites were further investigated as possible topoisomerase II sites. Sixteen sites meet the criteria of a 40% match on the opposite strand. A plus strand site that has a 30% match on the opposite strand is included in Table 21 because the topoisomerase II cleavage site is in exact agreement with the breakpoint position of the monomac 6 cell line (216). No topoisomerase II sites are found in relation to the breakpoint of Gill 2 (215) discussed above (Table 21, Figure 6b). Two breakpoints are found at the position of the opposite strand cleavage site of topoisomerase II (Table 21, Figure 6b), the Negrini breakpoint (74) and S79541 (65). The breakpoints of patients 1, 5, and 6 (217), and Gill 1 (215) are all within two nucleotides of topoisomerase II cleavage sites on one strand or the other (Table 21, Figure 6b). Patients 3 and 10 (217) have topoisomerase II cleavage sites a greater distance from the breakpoint (Table 21, Figure 6b).

3q26, The MDS1 Gene Region

The P1 clone P856 has a total insert size of 62415 and a G/C content of 36% and the sequence features found within this clone are presented in Figure 7. The MDS1B cDNA is partially within the P1 clone and is on the minus strand. As seen in Table 22, the alignment of P856 with the MDS1B cDNA begins at position 340 and ends at 685 of the cDNA.

Table 22. Alignment of P856 with the MDS1 cDNA.

Score	% Sub.	% Del.	% Ins.	Query	Begin	End	Rem.	Or.	Subject	Begin	End	Rem.
341	0.29	0	0	P856	60668	61013	3202	C	mds1	685	340	592

The GenBank entry (U43292) states that the first codon is at position 318 and the stop codon is at position 817 of the published sequence. Therefore, a single exon of the MDS1B gene is within P856. The intron following this exon must be quite large since

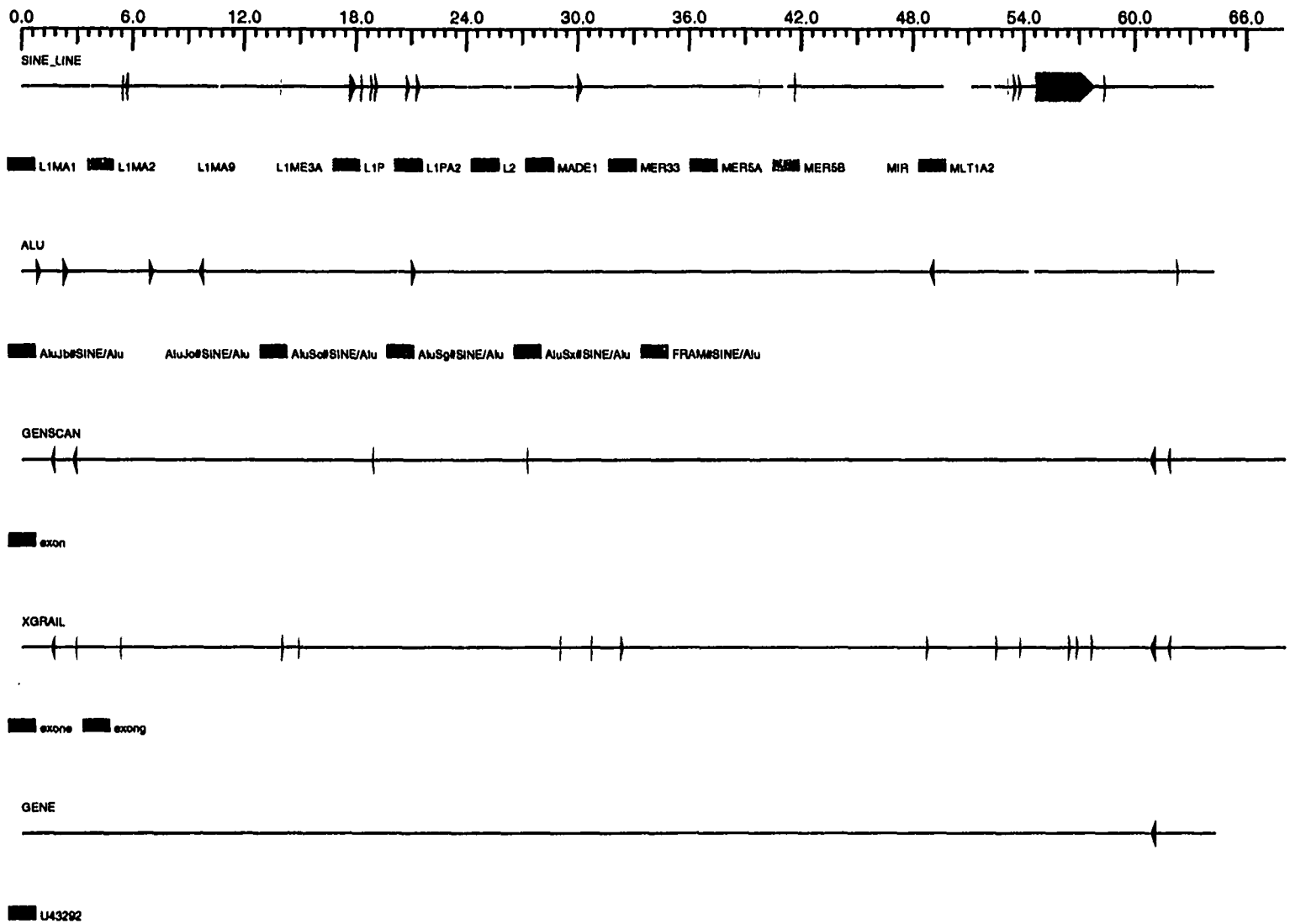


Figure 7. Sequence Feature Map from p856. See legend to Figure 3.

the exon sequence ends at position 60668. A low stringency search at a minimum match of five and a minimum score of 10 was performed to be sure that no small exons were missed. No additional exons were found with this search.

GRAIL and GENSCAN (Table 23) both predict an exon to begin and end approximately at the positions of the Crossmatch alignment. Both programs predict an open reading frame from 61008 to 60671 despite a transition that results in the "G" that would be the 3' splice site being an "A" in the p856 sequence. A BLASTP of the GENSCAN predicted gene against the protein database shows that the predicted gene has high homology to MDS1B, and identity to MDS1A over a four amino acid shorter region

Table 23. GRAIL Predicted Exons and GeneScan Predicted Genes within P856.

GRAIL Predicted Exons.				GRAIL. cont.				GeneScan Predicted Genes		
Begin	End	Qual.	Str.	Begin	End	Qual.	Str.	Begin	End	Str.
13960	14105	ex	f	56296	56464	gd	f	1538	1760	r
14857	14943	ex	f	56711	56856	ex	f	2694	2943	r
28991	29048	ex	f	57510	57650	ex	f	18797	18897	r
30671	30730	ex	f	1538	1760	gd	r	27171	27254	r
32218	32434	gd	f	2848	2942	gd	r	60671	61008	r
48625	48797	gd	f	5239	5332	gd	r	61627	61822	r
52361	52513	gd	f	60671	61008	ex	r			
53668	53769	gd	f	61627	61822	gd	r			

of overlap. The four missing amino acids include the only two amino acids not identical in the overlap of the predicted gene with MDS1B.

Two GRAIL predicted exons have homology to a reverse transcriptase gene. The two exons are both forward strand exons (Figure 7). Both are within an L1PA2 element (Table 24) where the first predicted exon begins at position 56296 while the second begins at 56711. The first is given a good rating by GRAIL and the second is given an excellent rating. These observations are consistent with the presence of an open reading frame homologous to reverse transcriptase in L1 elements (183).

As discussed above, the reverse strand exon predicted by GRAIL to begin at 60671 and end at 61008 has high homology to the MDS1 genes. The lack of alignment of any more of the MDS1 cDNA, and the lack of other predicted exons with homology to MDS1 leads to the conclusion that only a single exon of MDS1 is within this P1 clone. Since that is the case, an analysis of the promoter and polyadenylation site structure of this gene is not possible from the data presented here.

Repetitive DNA within p856

The repetitive DNA is presented in Table 24 as the output from RepeatMasker (see also Figure 7). The total repetitive DNA as determined using RepeatMasker is 18.42%. Of that, a little over half is found in 12 LINE DNA fragments. The two L1MA1 fragments that begin at 20700 and 21242 probably arose as from a common insertional event as seen by the adjoining alignment of the two fragments with the consensus L1 sequence (Table 24, Figure 7). Separation of these two fragments probably came about from insertion of the AluSc that begins at 20942.

There are eight Alu repeats in P856, four are from the older FRAM or AluJ families. None are from the younger AluY subfamily. Combining the LINE and Alu repeats, 70% are inserted in the direction opposite to MDS transcription. Only two Alu repeats are found between 10000 and 50000 of the P1 clone, but half the LINE repeats are found in this region. If the region were extended to 53000, two more L1s would be included in this total (Table 24, Figure 7).

Thirteen medium frequency repeats are found in P856, or 2.74% of the clone. The orientation bias observed for the LINE and Alu repeats does not extend to this class, 7 of 13 elements are oriented in the same direction as MDS1B transcription (Table 24,

Table 24. RepeatMasker Output from P856.

Score	% Sub.	% Del.	% Ins.	Query	Begin	End	Length	Rem.	Or.	Repeat Family	Repeat Class	Begin	End	Rem.
LINE DNA														
1217	11.6	8.6	1.3	P856	20700	20932	233	43283	+	L1MA1	LINE/L1	5815	6064	240
1535	8.9	2.1	0.4	P856	21242	21477	236	42738	+	L1MA1	LINE/L1	6065	6304	0
565	11.1	9.1	0.0	P856	53056	53154	99	11061	C	L1MA2	LINE/L1	6308	6201	0
4961	15.1	3.8	3.6	P856	49636	51172	1537	13043	+	L1MA9	LINE/L1	4771	6307	1
344	22.3	0.0	8.5	P856	26397	26526	130	37689	+	L1ME3A	LINE/L1	5707	5825	338
305	7.5	0.0	0.0	P856	39713	39752	40	24463	C	L1P	LINE/L1	133	94	6013
5754	12.8	0.1	0.0	P856	54644	57803	3160	6412	+	L1PA2	LINE/L1	2975	6141	5
182	22	2.0	2.0	P856	5412	5461	50	58754	C	L2	LINE/L2	2675	2626	35
187	26	4.1	7.3	P856	5596	5718	123	58497	C	L2	LINE/L2	2491	2373	219
519	32	3.0	4.6	P856	17637	17964	328	46251	+	L2	LINE/L2	2369	2691	19
189	33.3	0.0	0.0	P856	18250	18318	69	45897	+	L2	LINE/L2	2613	2681	69
190	25.6	0.0	0.0	P856	58324	58366	43	5849	+	L2	LINE/L2	2699	2741	9
							6048 Total LINE DNA							
							9.42% LINE DNA							
LTR DNA														
699	22.4	0.0	4.9	P856	53426	53608	183	10607	+	MLT1A2	LTR/MaLR	1	174	200
865	22.5	0.5	1.1	P856	53685	53871	187	10344	+	MLT1A2	LTR/MaLR	189	374	0
							370 Total LTR DNA							
							0.58% Percent LTR DNA							
Alu DNA														
1528	19.4	1.0	0.3	P856	9509	9802	294	54413	C	AluJb	SINE/Alu	296	1	6
1841	15.4	0.0	1.0	P856	48828	49132	305	15083	C	AluJb	SINE/Alu	302	1	0
1265	20.9	4.3	4.3	P856	54235	54536	302	9679	+	AluJo	SINE/Alu	1	302	0
2106	9.7	0.0	0.3	P856	20942	21241	300	42974	+	AluSc	SINE/Alu	1	299	0
1975	11.2	0.0	0.3	P856	739	1033	295	63182	+	AluSg	SINE/Alu	1	294	6
1973	12.3	0.0	0.7	P856	6836	7136	301	57079	+	AluSg	SINE/Alu	1	299	1
1932	13.2	0.0	0.7	P856	2195	2498	304	61717	+	AluSx	SINE/Alu	1	302	0
836	18.1	3.5	1.8	P856	62185	62355	171	1860	+	FRAM	SINE/Alu	7	166	0
							2272 Total Alu DNA							
							3.54% Alu DNA							

Medium Frequency Repetitive DNA

274	4.4	2.2	0.0	P856	41621	41665	45	22550	C	MADE1	DNA/Mariner	80	35	0
948	19.9	8.7	0.7	P856	29961	30247	287	33968	+	MER33	DNA/MER1_type	7	316	8
568	16.4	4.3	0.0	P856	18756	18895	140	45320	+	MER5A	DNA/MER1_type	44	189	0
383	22.1	14.5	6.4	P856	18992	19163	172	45052	+	MER5A	DNA/MER1_type	4	189	0
362	26.1	1.5	0.0	P856	13884	14013	130	50202	C	MER5B	DNA/MER1_type	152	21	26
211	15.4	2.6	0.0	P856	427	465	39	63750	C	MIR	SINE/MIR	133	94	129
308	22.7	4.5	0.0	P856	3665	3752	88	60463	C	MIR	SINE/MIR	205	114	57
377	30.6	3.8	0.6	P856	10562	10721	160	53494	C	MIR	SINE/MIR	238	74	24
235	24.7	14.0	2.1	P856	23996	24088	93	40127	+	MIR	SINE/MIR	46	149	113
213	29.3	0.0	0.0	P856	26011	26068	58	38147	+	MIR	SINE/MIR	48	105	157
311	25.6	0.0	1.2	P856	29766	29847	82	34368	+	MIR	SINE/MIR	64	144	118
797	22.1	2.0	6.3	P856	40988	41241	254	22974	C	MIR	SINE/MIR	244	2	18
475	28.2	7.0	2.8	P856	52190	52402	213	11813	C	MIR	SINE/MIR	253	32	9

1761 Total Medium Frequency Repetitive DNA

2.74% Medium Frequency Repetitive DNA

Simple Repetitive DNA

288	0.0	0.0	0.0	P856	18083	18114	32	46101	C	(CA)n	Simple_repeat	33	2	87
233	19.2	0.0	0.8	P856	57824	57972	149	6243	+	(CAAG)n	Simple_repeat	1	120	0
342	4.2	0.0	0.8	P856	57958	58076	119	6139	+	(GAAA)n	Simple_repeat	3	120	0
338	22.6	1.1	3.2	P856	37257	37349	93	26866	+	(GAAAA)n	Simple_repeat	3	93	27
290	14.9	0.0	0.9	P856	13297	13410	114	50805	C	(GGAA)n	Simple_repeat	115	3	5
232	24.2	1.5	0.0	P856	10350	10415	66	53800	+	(TA)n	Simple_repeat	1	67	53
231	25.8	0.0	3.2	P856	26756	26848	93	37367	+	(TAAA)n	Simple_repeat	1	90	30
198	22.1	1.3	6.5	P856	60195	60271	77	3944	+	(TAAAA)n	Simple_repeat	2	74	46
332	15.7	0.0	0.0	P856	53351	53401	51	10814	C	(TGAA)n	Simple_repeat	53	3	67
404	15.9	0.0	0.0	P856	39765	39833	69	24382	+	POLY_A	Simple_repeat	1	69	51

863 Total Simple Repetitive DNA

1.34% Simple Repetitive DNA

Low Complexity DNA

24	0.0	0.0	0.0	P856	6809	6832	24	57383	C	AT_rich	Low_complexity	24	1	276
26	0.0	0.0	0.0	P856	9968	9993	26	54222	C	AT_rich	Low_complexity	54	29	246
28	0.0	0.0	0.0	P856	11475	11502	28	52713	C	AT_rich	Low_complexity	28	1	272
22	3.6	0.0	0.0	P856	12846	12873	28	51342	C	AT_rich	Low_complexity	106	79	194
26	8.0	0.0	0.0	P856	13097	13146	50	51069	+	AT_rich	Low_complexity	29	78	222
22	0.0	0.0	0.0	P856	13521	13542	22	50673	C	AT_rich	Low_complexity	100	79	200
30	11.5	0.0	0.0	P856	21677	21772	96	42443	+	AT_rich	Low_complexity	1	96	204
21	0.0	0.0	0.0	P856	21971	21991	21	42224	C	AT_rich	Low_complexity	184	164	116
21	0.0	0.0	0.0	P856	26183	26203	21	38012	+	AT_rich	Low_complexity	1	21	279
28	6.5	0.0	0.0	P856	29452	29497	46	34718	C	AT_rich	Low_complexity	142	97	158
29	8.5	0.0	0.0	P856	30593	30651	59	33564	C	AT_rich	Low_complexity	59	1	241
21	9.8	0.0	0.0	P856	32820	32870	51	31345	+	AT_rich	Low_complexity	87	137	163
27	0.0	0.0	0.0	P856	33138	33164	27	31051	C	AT_rich	Low_complexity	86	60	214
22	0.0	0.0	0.0	P856	47905	47926	22	16289	+	AT_rich	Low_complexity	1	22	278
26	5.3	0.0	0.0	P856	63667	63704	38	511	C	AT_rich	Low_complexity	38	1	262

559 Total Low Complexity DNA

0.87% Low Complexity DNA

11873 Total Repetitive DNA

18.5% Repetitive DNA

See Legend for Table 1.

Figure 7). The A/T rich character of P856 is reflected in the simple tandem repeats and low complexity DNA. The simple tandem repeats are all at least 50% A/T and all the low complexity regions are classified as AT rich. The two classes together account for over 2% of P856 by RepeatMasker analysis. Sputnik find five more simple tandem repeats, one of which is 16 nt in length and less than 50% A/T. One large region of tandemly repeated DNA, >250 nt, is found from 57822 to 58077 (Table 25). Both RepeatMasker and Sputnik find this region though with slightly different boundaries, and both programs break it into two pieces. Interestingly, each piece is classified as being a different class of repeat. RepeatMasker classifies the first piece as a CAAG repeat

Table 25. Sputnik Output from P856.

Begin	End	Length	(Repeat Unit)xN
13327	13377	51	(CCTT)x12*
13378	13393	16	(CCCT)x4
18083	18114	32	(GT)x16
41707	41718	12	(CAA)x4
50999	51029	31	(AC)x15
53612	53641	30	(TC)x15
57822	57867	46	(AAGG)x11*
57870	58077	208	(AAAG)x52*

* Denotes imperfect tandem repeat.

whereas Sputnik classifies it as an AAGG repeat. RepeatMasker overlaps the two pieces, but Sputnik places a small gap between them. When the RepeatMasker output is corrected for the 14 nt duplication in the overlap, the actual percentage of simple tandem repeat DNA in P856 is 1.32%.

9q34, A Gene Hunt

In an effort to find the tuberous sclerosis (TSC) gene on chromosome 9, cosmid B1 was sequenced. The total insert size of this clone is 35004 nt and the G/C content is 53.7%. The sequence features of cosmid B1 are presented in graphical form in Figure 8.

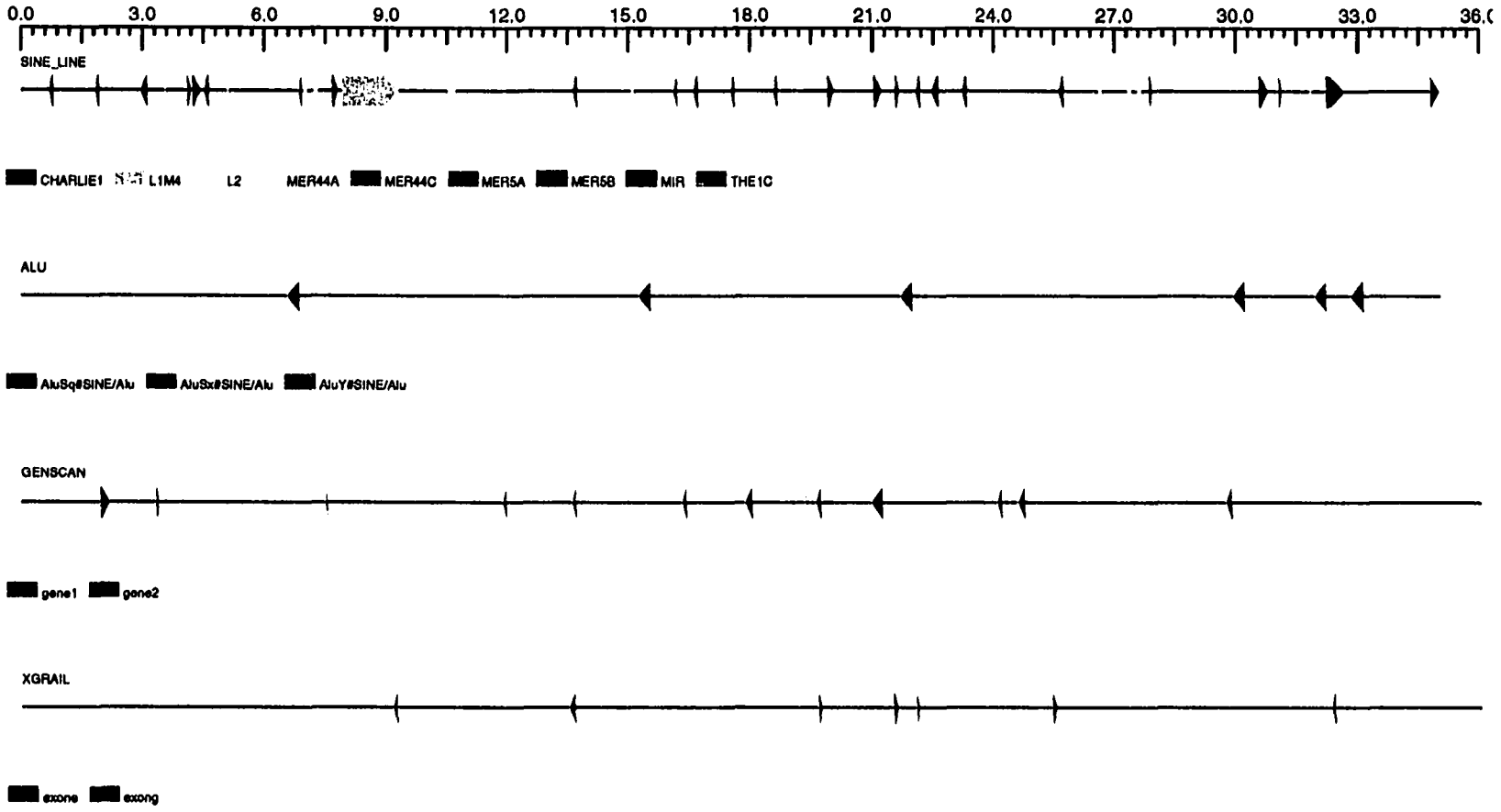


Figure 8. Sequence Feature Map of Bl. See legend to Figure 3.

Oddly, of all the projects presented as part of this dissertation research, this cosmid, which as well as can be determined at this time encodes no genes, has the highest G/C content. BLASTN of the sequence from B1 did not reveal significant homology to any known gene. BLASTP of the predicted exons from GRAIL or the predicted genes from GENSCAN also did not reveal homology to any known protein. The positions of the exons predicted by GRAIL and the GENSCAN predicted genes are in Table 26. These programs are unable to predict any identical exons, although one reverse strand exon predicted by GRAIL to be from position 13489 to 13629 overlaps an exon predicted by GENSCAN to span B1 from 13577 to 13668.

Table 26. GRAIL Predicted Exons and GeneScan Predicted Genes from Cosmid B1.

GRAIL				GeneScan, Gene 1			GeneScan, Gene2		
Begin	End	Qual.	Str.	Begin	End	Str.	Begin	End	Str.
19638	19725	gd	f	1953	2183	f	29705	29840	r
21494	21606	ex	f	3327	3392	f	24566	24728	r
22032	22112	gd	f				24055	24154	r
25407	25515	gd	f				20931	21211	r
9140	9223	ex	r				19587	19694	r
13489	13629	gd	r				19328	19346	r
32307	32368	ex	r				17833	18010	r
							16282	16380	r
							13577	13668	r
							11858	11954	r
							7470	7510	r

A single CpG island is shown by GRAIL in B1. This specialized region spans positions 20398 to 21052. GRAIL positions a minus strand promoter within B1, but it is not associated with the CpG island. The CAAT box for this promoter is at position 13497. A TATA box with the sequence GGATAGG is at position 13415 and the transcription start site is at 13385. The plus strand gene predicted by GENSCAN does not have a promoter within B1, but the polyadenylation site for this gene is at 3564. Likewise, the predicted minus strand gene does not start within B1. The polyadenylation site predicted for this gene starts at 5975.

Repetitive DNA within B1

The RepeatMasker output showing the repetitive DNA within B1 is in Table 27 and graphically represented in Figure 8. RepeatMasker shows 24.18% of B1 to be repetitive DNA. Of that, approximately one fourth is LINE DNA. In contrast to the other projects discussed above, only one of the ten LINE fragments is from a L1, the other nine are L2 DNA. However, the single L1 fragment accounts for 40% of the LINE DNA. Of the nine L2 fragments, seven are on the forward strand of B1. An analysis of these fragments and their alignment with the L2 consensus does not reveal that any two or more of these fragments could have arisen from a common insertional event.

There are six Alu repeats in B1, all full-length or nearly so, and all in the reverse orientation or minus strand. There are no Alus from the monomer or AluJ subfamilies in B1 and four of the Alus are from the AluSx subfamily. The remaining two are from the AluSq and human specific AluY subfamilies. The AluSq, one of the AluSx and the AluY repeats are clustered between 29891 and 33091 such that this 3200 nt region is over 28% Alu repeat DNA (Table 27, Figure 8).

The medium frequency repeats are the most common class of repeats in B1, both from the standpoint of number of elements and percentage. Two MERs are found in close association with Alu repeats (Table 27, Figure 8). The first of these is a reverse strand MER5A fragment that starts at B1 position 6834 and ends at 6899. This MER is tandemly associated with a reverse strand AluSx, which starts at 6831 and ends at 6528. The second case is a plus strand MER44A that starts at 31797 and ends at 31865. This MER fragment is closely associated with a second reverse strand AluSx, starting at 31886

Table 27. RepeatMasker Output from Cosmid B1.

Score	% Sub.	% Del.	% Ins.	Query	Begin	End	Length	Rem.	Or.	Repeat Family	Repeat Class	Begin	End	Rem.
LINE DNA														
2667	25.6	4.2	2.1	CosB1	7903	9307	1405	25697	+	L1M4	LINE/L1	315	1806	4607
180	24.4	2.4	0.0	CosB1	3536	3576	41	31428	+	L2	LINE/L2	2696	2737	13
175	23.4	4.3	0.0	CosB1	5074	5120	47	29884	C	L2	LINE/L2	2692	2644	18
248	24.1	5.6	1.9	CosB1	6947	7054	108	27950	+	L2	LINE/L2	1503	1614	1136
189	30.9	11.0	2.2	CosB1	7188	7323	136	27681	+	L2	LINE/L2	2597	2744	6
285	30.0	11.0	1.3	CosB1	10489	10715	227	24289	+	L2	LINE/L2	2282	2530	180
277	20.6	5.4	2.2	CosB1	15054	15145	92	19859	+	L2	LINE/L2	2428	2522	188
302	32.1	0.9	0.9	CosB1	26584	26695	112	8309	+	L2	LINE/L2	2596	2707	43
260	31.2	1.6	0.8	CosB1	27310	27434	125	7570	+	L2	LINE/L2	2358	2483	227
355	28.8	2.7	0.0	CosB1	27558	27668	111	7336	C	L2	LINE/L2	2736	2623	14
							2404 Total LINE DNA							
							6.87% LINE DNA							
LTR DNA														
1065	13.1	0.5	3.6	CosB1	34782	35003	222	1	+	THE1C	LTR/MaLR	1	215	156
							222 Total LTR DNA							
							0.63% LTR DNA							
Alu DNA														
1901	11.1	0.3	2.0	CosB1	29891	30188	298	4816	C	AluSq	SINE/Alu	293	1	10
1964	11.8	0.3	1.0	CosB1	6528	6831	304	28173	C	AluSx	SINE/Alu	302	1	0
1960	11.0	0.3	2.9	CosB1	15212	15521	310	19483	C	AluSx	SINE/Alu	302	1	0
1986	11.3	1.0	1.0	CosB1	21641	21942	302	13062	C	AluSx	SINE/Alu	302	1	0
2140	9.9	0.0	0.3	CosB1	31886	32188	303	2816	C	AluSx	SINE/Alu	302	1	0
2179	7.4	0.0	2.9	CosB1	32782	33091	310	1913	C	AluY	SINE/Alu	301	1	0
							1827 Total Alu DNA							
							5.22% Alu DNA							

Medium Frequency Repetitive DNA

285	25.4	1.4	0.0	CosB1	31053	31123	71	3881	+	CHARLIE1	DNA/MER1_typ	2666	2737	2
373	25.8	11.7	0.6	CosB1	2939	3101	163	31903	C	MER5A	DNA/MER1_typ	181	1	8
468	17.0	2.0	3.0	CosB1	4096	4195	100	30809	+	MER5A	DNA/MER1_typ	31	129	60
273	21.2	4.5	0.0	CosB1	6834	6899	66	28105	C	MER5A	DNA/MER1_typ	74	6	115
361	20.6	4.1	5.2	CosB1	16090	16186	97	18818	C	MER5A	DNA/MER1_typ	98	3	91
431	28.3	3.0	0.6	CosB1	4239	4404	166	30600	+	MER5B	DNA/MER1_typ	9	178	0
260	23.6	0.0	0.0	CosB1	31797	31865	69	3139	+	MER44A	DNA/MER2_typ	41	110	618
1918	24.5	3.4	0.6	CosB1	32189	32662	474	2342	+	MER44C	DNA/MER2_typ	234	720	8
220	27.3	6.1	0.0	CosB1	714	779	66	34225	C	MIR	SINE/MIR	192	123	70
213	22.2	1.9	0.0	CosB1	1854	1907	54	33097	C	MIR	SINE/MIR	162	108	100
191	28.2	14.1	0.0	CosB1	4540	4617	78	30387	C	MIR	SINE/MIR	154	66	108
216	26.7	6.9	5.9	CosB1	7679	7779	101	27225	+	MIR	SINE/MIR	67	168	94
293	20.6	1.5	1.5	CosB1	13652	13719	68	21285	C	MIR	SINE/MIR	144	77	118
297	28.9	0.0	0.0	CosB1	16613	16688	76	18316	C	MIR	SINE/MIR	154	79	108
182	27.7	0.0	0.0	CosB1	17545	17591	47	17413	C	MIR	SINE/MIR	96	50	166
277	17.2	3.5	0.0	CosB1	18592	18649	58	16355	C	MIR	SINE/MIR	141	82	121
400	28.6	4.8	0.0	CosB1	19914	20039	126	14965	+	MIR	SINE/MIR	65	196	66
309	27.1	13.9	1.3	CosB1	21048	21198	151	13806	+	MIR	SINE/MIR	93	262	0
209	26.7	5.0	0.0	CosB1	21570	21629	60	13375	+	MIR	SINE/MIR	189	251	11
229	19.1	13.2	0.0	CosB1	22079	22146	68	12858	C	MIR	SINE/MIR	147	71	115
383	30.9	1.3	2.0	CosB1	22462	22610	149	12394	C	MIR	SINE/MIR	261	114	1
294	25.3	12.7	0.0	CosB1	23226	23304	79	11700	C	MIR	SINE/MIR	162	74	100
237	25.8	3.4	5.6	CosB1	25632	25720	89	9284	C	MIR	SINE/MIR	175	89	87
187	25.6	0.0	0.0	CosB1	27871	27913	43	7091	+	MIR	SINE/MIR	103	145	117
760	23.8	2.6	2.1	CosB1	30583	30771	189	4233	+	MIR	SINE/MIR	24	213	49

2708 Total Medium Frequency Repetitive DNA

7.74% Medium Frequency Repetitive DNA

Low Complexity DNA

21	7.7	0.0	0.0	CosB1	30982	31020	39	3984	C	AT_rich	Low_complexity	39	1	260
41	16	0.0	0.0	CosB1	20621	20826	206	14178	C	GC_rich	Low_complexity	206	1	94

245 Total Low Complexity DNA

0.70% Low Complexity DNA

Simple Repetitive DNA

348	23.6	1.1	0.0 CosB1	10799	10887	89	24117	C	(CA)n	Simple_repeat	90	1	30
509	25.2	0.8	0.0 CosB1	11060	11178	119	23826	C	(CA)n	Simple_repeat	120	1	0
581	20.2	0.0	3.2 CosB1	11295	11418	124	23586	C	(CA)n	Simple_repeat	120	1	0
597	20.2	0.8	0.0 CosB1	11430	11548	119	23456	C	(CA)n	Simple_repeat	120	1	0
196	13.6	0.0	4.0 CosB1	11553	11677	125	23327	C	(CA)n	Simple_repeat	120	1	0
292	27.8	1.6	7.1 CosB1	10889	11014	126	23990	+	(CATG)n	Simple_repeat	2	120	0
242	34.3	5.6	0.0 CosB1	11182	11289	108	23715	C	(CATG)n	Simple_repeat	114	1	6
197	7.7	0.0	0.0 CosB1	15865	15890	26	19114	+	(CCAA)n	Simple_repeat	2	27	93
230	17.8	2.2	0.0 CosB1	22750	22794	45	12210	+	(GA)n	Simple_repeat	1	46	74
580	17.3	5.5	0.0 CosB1	1528	1665	138	33339	C	(GGAA)n	Simple_repeat	119	3	1
296	7.7	0.0	0.0 CosB1	9977	10015	39	24989	+	(GGGA)n	Simple_repeat	2	40	80

1058 Total Simple Repetitive DNA

3.02% Simple Repetitive DNA

8464 Total Repetitive DNA

24.2% Repetitive DNA

See Legend for Table 1.

and ending at 32188. Since the Alus are full-length and the MERs are not, it is likely that the Alus inserted into the MERs and not vice versa. A similar association is seen between a plus strand MIR that spans positions 21570 to 21629 and a reverse strand AluSx that starts at 21641 and ends at 21942. In this case, the MIR lacks 11 nt from ending at the consensus position (Table 27). Since this is the length of the gap between the two repeats it is more difficult to make any determination concerning the order of insertion. However, if it is assumed that the 11 nt are MIR sequence that were not recognized by the alignment program due to mutation, and that the two repeats would undergo similar mutation rates, then the Alu was inserted after the MIR. The orientation of the medium frequency repeats within B1 is roughly even, 11 on the plus strand and 13 on the minus.

A much larger proportion of B1 is simple tandem repetitive DNA than any of the other projects presented herein. Over 3% of B1 is made up of this class of repeat, largely due to a single region starting at position 10799 and ending at 11677. RepeatMasker breaks this region up into five regions of CA and two regions of CATG repeats. Probably due to a more stringent scoring matrix, Sputnik (Table 28) only shows five GT (CA) repeats in this region and they are of a smaller size than those shown by RepeatMasker (Table 27). Neither of the CATG repeats is shown by Sputnik. The increased sensitivity of Sputnik is shown in the region from 1440 to 1668 (Table 28). This region is shown by Sputnik to be a complicated region of penta- and tetranucleotide repeats. The region starts with the purine-rich TTCCC and TCCCC repeats. There is a shift to a repeating unit that is half pyrimidine and half purine, CCTT. The region ends with a pyrimidine-rich repeating unit of TTTTC.

Table 28. Sputnik Output from Cosmid B1.

Begin	End	Length	(Repeat Unit)xN	Begin	End	Length	(Repeat Unit)xN
636	647	12	(TAA)x3	11589	11601	13	(TG)x6
1440	1454	15	(TTCCC)x3	11642	11678	37	(TG)x18
1515	1558	44	(TCCCC)x8*	13471	13483	13	(TG)x6
1560	1595	36	(CCTTT)x7*	15211	15228	18	(CTTTT)x3
1603	1634	32	(CCTT)x7*	15871	15887	17	(CCAA)x6
1641	1653	13	(CCTT)x3	16140	16159	20	(TGC)x6
1654	1668	15	(TTTTC)x3	20511	20523	13	(GGGC)x3
2403	2416	14	(CAGG)x3	22750	22761	12	(GA)x6
9986	10008	23	(AGGG)x5	26243	26255	13	(AGG)x4
11348	11358	11	(TG)x5	28934	28947	14	(GTTTT)x2
11376	11387	12	(TG)x6	29390	29403	14	(AGGG)x3
11452	11462	11	(TG)x5				

The higher G/C content of the overall cosmid is reflected in the higher G/C content of the simple repeats. Eighteen of the 23 tandem repeat units found by Sputnik are 50% G/C or more (Table 28).

Analysis of Alu and L1 Repeat Divergence

An analysis of the divergence of the Alu and L1 repeats from the sequence data presented above was performed. Alu and L1 repeats are each thought to arise from very small numbers of progenitor elements. Therefore, an analysis of the percent divergence from the consensus for each subfamily can be a useful indicator of the relative ages of each subfamily, assuming that each subfamily has a similar mutation rate. The RepeatMasker output from each project was pooled and then grouped by repeat subfamily. The percent substitution, percent deletion, and percent insertion of each element was added together, converted to the fractional form and multiplied by the length of each repeat element. The sum of these numbers was multiplied by the sum of the elements in each subfamily to arrive at a weighted average percent divergence for each subfamily.

The results of this analysis of the Alu repeats is found in Table 29. Four Alu repeat elements were omitted from this analysis because RepeatMasker was unable to

unambiguously determine the subfamily of these elements. Surprisingly, the monomer subfamilies, FAM, FLAM, and FRAM, have a similar average divergence to the oldest dimer families, AluJ, AluJo and AluJb. All the AluS subfamilies have similar divergence values from their respective consensus sequences except the AluSg, which has a 15.32% percent divergence. However, a single AluSg element from 213a7 has a significantly higher substitution rate than the remaining four elements in this subfamily. Exclusion of this single element lowers the divergence of the group to 10.83%, a figure that is similar to the much younger AluY subfamily, which has a 10.44% divergence from the consensus. The sample size is too small to differentiate between the AluSc, AluSg, AluSp, AluSq, and AluSx subfamilies, but the fact that they are similar in age (151,154) is well demonstrated by the similarity in the divergence of each subfamily from its respective consensus.

The L1 repeat subfamilies were subjected to the same analysis (Table 30). In an effort to maintain significant sample sizes, the L1 repeats were grouped according to a broader classification than the individual subfamilies such that L1MA1 and L1MA2 subfamilies were grouped into a L1MA family. With the exception of the L1MA and L1MB subfamilies, the remaining L1 families originated before the primate radiation and have an average divergence approaching or exceeding 30%. L1MB, L1P and L1P5 have an intermediate level of divergence around 25%. The L1PB family is only represented by four fragments from two elements and the figures calculated for this family probably are not significant. The L1PA family is represented by a large and apparently homogeneous group of elements that only have an average divergence from the

Table 29. Analysis of the Alu Repeats from all Projects.

Score	% Sub.	% Del.	% Ins.	Weight	Query	Begin	End	Length	Rem.	Or.	Repeat Family	Repeat Class	Begin	End	Rem.
717	17.5	0.8	0.0	23.06	c110	3219	3344	126	39008	C	AluJ	SINE/Alu	127	1	175
796	20.0	3.0	0.6	38.94	c110	2747	2911	165	39441	C	AluJ	SINE/Alu	302	134	0
Percent Divergence, Family				**21.31**		Length, Family		291							
722	17.1	0.0	2.9	28.00	34a5	40548	40687	140	3	C	AluJb	SINE/Alu	301	166	1
1436	18.7	0.0	5.9	78.97	34a5	20545	20865	321	19825	C	AluJb	SINE/Alu	302	1	0
1681	17.3	0.0	1.0	55.08	9p22	53025	53325	301	13915	C	AluJb	SINE/Alu	298	1	4
493	19.4	1.1	0.0	19.07	c110	9174	9266	93	33086	C	AluJb	SINE/Alu	281	188	21
1421	21.9	1.7	1.7	76.15	c110	25153	25453	301	16899	C	AluJb	SINE/Alu	302	2	0
841	18.2	0.0	0.0	26.03	c66	25862	26004	143	8674	+	AluJb	SINE/Alu	1	143	159
1528	19.4	1.0	0.3	60.86	P856	9509	9802	294	54413	C	AluJb	SINE/Alu	296	1	6
1841	15.4	0.0	1.0	50.02	P856	48828	49132	305	15083	C	AluJb	SINE/Alu	302	1	0
Percent Divergence, Family				**20.77**		Length, Family		1898							
1734	15.7	1.4	0.3	49.94	213a7	28199	28485	287	13731	+	AluJo	SINE/Alu	7	296	6
1512	17.7	0.3	3.0	64.05	34a5	5970	6274	305	34416	+	AluJo	SINE/Alu	6	302	0
1161	18.6	10.2	0.0	76.03	9p22	47308	47571	264	19669	C	AluJo	SINE/Alu	302	12	0
1458	22.6	0.3	1.6	74.97	9p22	38698	39003	306	28237	+	AluJo	SINE/Alu	1	302	0
1558	19.8	0.7	0.3	61.98	9p22	53713	54010	298	13230	C	AluJo	SINE/Alu	299	1	3
1724	16.8	0.0	0.3	49.76	9p22	23307	23597	291	43643	+	AluJo	SINE/Alu	2	291	11
1044	16.8	0.0	4.9	44.05	c110	19311	19513	203	22839	C	AluJo	SINE/Alu	193	1	109
1593	17.0	0.3	3.5	64.90	c110	33613	33924	312	8428	C	AluJo	SINE/Alu	302	1	0
1228	18.5	8.3	0.4	72.08	c66	11370	11634	265	23044	C	AluJo	SINE/Alu	286	1	16
1735	16.1	1.0	0.7	53.22	c66	30335	30633	299	4045	C	AluJo	SINE/Alu	301	2	1
1451	20.8	0.3	1.3	67.87	C86	23514	23816	303	20359	C	AluJo	SINE/Alu	301	2	1
1612	15.8	1.6	2.6	60.60	C86	24229	24531	303	19644	C	AluJo	SINE/Alu	300	1	2
1265	20.9	4.3	4.3	89.09	P856	54235	54536	302	9679	+	AluJo	SINE/Alu	1	302	0
Percent Divergence, Family				**22.17**		Length, Family		3738							

692	8.1	0.0	1.0	9.01	c110	1104	1202	99	41150	C	AluS	SINE/Alu	99	2	204
Percent Divergence, Family				**9.10**		Length, Family		99							
2119	9.8	0.0	0.0	28.91	213a7	21958	22252	295	19964	C	AluSc	SINE/Alu	295	1	4
2090	10.0	0.0	0.3	30.80	34a5	36614	36912	299	3778	C	AluSc	SINE/Alu	298	1	1
2099	9.9	0.3	0.0	29.89	34a5	8408	8700	293	31990	C	AluSc	SINE/Alu	295	2	4
1927	9.6	4.3	0.0	38.92	c110	3852	4131	280	38221	C	AluSc	SINE/Alu	292	1	7
1740	13.9	0.3	3.9	55.93	c66	31740	32048	309	2630	C	AluSc	SINE/Alu	298	1	1
2106	9.7	0.0	0.3	30.00	P856	20942	21241	300	42974	+	AluSc	SINE/Alu	1	299	0
Percent Divergence, Family				**12.07**		Length, Family		1776							
984	21.4	11.8	0.8	89.08	213a7	16997	17258	262	24958	+	AluSg	SINE/Alu	1	291	9
2235	8.0	0.0	0.0	24.00	c110	6204	6503	300	35849	+	AluSg	SINE/Alu	1	300	0
1351	10.3	0.0	0.5	20.95	C86	27212	27405	194	16770	+	AluSg	SINE/Alu	1	193	107
1973	12.3	0.0	0.7	39.13	P856	6836	7136	301	57079	+	AluSg	SINE/Alu	1	299	1
1975	11.2	0.0	0.3	33.93	P856	739	1033	295	63182	+	AluSg	SINE/Alu	1	294	6
Percent Divergence, Family				**15.32**		Length, Family		1352							
1917	10.6	0.4	0.0	30.14	9p22	42100	42373	274	24867	+	AluSp	SINE/Alu	22	296	7
1967	13.5	0.0	0.0	40.91	c66	19439	19741	303	14937	+	AluSp	SINE/Alu	1	303	0
2019	8.2	3.8	0.0	34.92	c66	30031	30321	291	4357	C	AluSp	SINE/Alu	302	1	1
2025	9.6	0.7	1.3	35.15	c66	4014	4316	303	30362	C	AluSp	SINE/Alu	303	3	0
Percent Divergence, Family				**12.05**		Length, Family		1171							

1804	13.0	0.3	4.8	57.20	213a7	21079	21394	316	20822	+	AluSq	SINE/Alu	1	302	1
2029	9.1	0.3	0.3	27.84	213a7	30779	31065	287	11151	+	AluSq	SINE/Alu	2	288	15
2118	10.6	0.3	0.0	32.92	34a5	21512	21813	302	18877	C	AluSq	SINE/Alu	303	1	0
1980	12.4	0.0	0.0	36.08	9p22	19723	20013	291	47227	C	AluSq	SINE/Alu	296	6	7
2082	7.2	4.1	0.0	32.88	9p22	8085	8375	291	58865	+	AluSq	SINE/Alu	1	303	0
2030	11.6	0.0	0.0	33.99	c66	26789	27081	293	7597	+	AluSq	SINE/Alu	2	294	9
1559	8.5	0.0	0.0	17.94	C86	28120	28330	211	15845	+	AluSq	SINE/Alu	12	222	81
1901	11.1	0.3	2	39.93	CosB1	29891	30188	298	4816	C	AluSq	SINE/Alu	293	1	10
Percent Divergence, Family				**12.18**		Length, Family		2289							
1899	11.4	0.3	2.4	42.02	34a5	39232	39529	298	1161	C	AluSx	SINE/Alu	292	1	10
2012	12.9	0.0	0.0	38.96	9p22	55035	55336	302	11904	+	AluSx	SINE/Alu	1	302	0
1718	14.0	0.0	6.0	62.80	c110	1307	1620	314	40732	C	AluSx	SINE/Alu	295	1	7
1920	12.8	0.7	0.0	39.15	c110	24584	24873	290	17479	C	AluSx	SINE/Alu	293	2	9
1935	13.4	0.0	0.3	40.96	c110	2912	3210	299	39142	C	AluSx	SINE/Alu	299	2	3
1944	10.8	0.3	1.3	37.82	c66	6560	6864	305	27814	+	AluSx	SINE/Alu	1	302	0
1956	8.7	5.2	0.0	39.89	c66	15853	16139	287	18539	C	AluSx	SINE/Alu	302	1	0
1969	11.2	0.0	2.9	43.85	c66	3087	3397	311	31281	+	AluSx	SINE/Alu	1	302	0
2138	8.8	0.0	0.3	27.03	c66	4357	4653	297	30025	C	AluSx	SINE/Alu	297	2	5
1595	15.8	1.0	3.0	58.81	C86	33518	33814	297	10361	+	AluSx	SINE/Alu	1	291	11
1913	10.5	2.0	0.3	37.76	C86	28443	28737	295	15438	+	AluSx	SINE/Alu	1	300	2
1915	11.6	1.3	1.0	41.84	C86	17144	17444	301	26731	+	AluSx	SINE/Alu	1	302	0
1960	11.0	0.3	2.9	44.02	CosB1	15212	15521	310	19483	C	AluSx	SINE/Alu	302	1	0
1964	11.8	0.3	1.0	39.82	CosB1	6528	6831	304	28173	C	AluSx	SINE/Alu	302	1	0
1986	11.3	1.0	1.0	40.17	CosB1	21641	21942	302	13062	C	AluSx	SINE/Alu	302	1	0
2140	9.9	0.0	0.3	30.91	CosB1	31886	32188	303	2816	C	AluSx	SINE/Alu	302	1	0
1932	13.2	0.0	0.7	42.26	P856	2195	2498	304	61717	+	AluSx	SINE/Alu	1	302	0
Percent Divergence, Family				**13.83**		Length, Family		5119							

2144	9.3	0.0	0.3	28.99	213a7	7503	7804	302	34412	C	AluY	SINE/Alu	301	1	0
2162	7.9	0.0	1.0	27.06	213a7	31980	32283	304	9933	+	AluY	SINE/Alu	1	301	0
1715	10.2	0.7	4.0	40.98	9p22	14546	14820	275	52420	C	AluY	SINE/Alu	299	34	2
1837	9.1	3.6	5.5	55.87	9p22	21606	21912	307	45328	C	AluY	SINE/Alu	301	1	0
2103	9.2	0.0	0.0	26.96	9p22	37512	37804	293	29436	+	AluY	SINE/Alu	1	293	8
2273	6.3	0.0	0.3	19.80	9p22	24811	25110	300	42130	C	AluY	SINE/Alu	301	3	0
1964	10.6	0.0	3.2	42.92	c110	25811	26121	311	16231	+	AluY	SINE/Alu	1	301	0
600	1.4	0.0	0.0	1.02	c66	1	73	73	34605	C	AluY	SINE/Alu	74	2	226
1688	6.2	0.0	8.3	40.02	c66	33608	33883	276	795	+	AluY	SINE/Alu	41	293	8
2179	7.4	0.0	2.9	31.93	CosB1	32782	33091	310	1913	C	AluY	SINE/Alu	301	1	0
2558	0.7	0.3	0.0	3.00	c110	22833	23132	300	19220	C	AluYa5	SINE/Alu	301	1	0
Percent Divergence, Family				**10.44**		Length, Family		3051							
771	23.8	0.6	0.0	40.02	213a7	31655	31818	164	10398	+	FAM	SINE/Alu	11	175	0
692	14.2	0.0	0.0	15.05	34a5	3742	3847	106	36843	C	FLAM_A	SINE/Alu	130	25	2
576	22.6	0.0	0.0	25.99	9p22	40037	40151	115	27089	+	FLAM_A	SINE/Alu	1	115	17
618	15.0	0.0	2.6	19.89	9p22	55351	55463	113	11777	+	FLAM_A	SINE/Alu	21	130	2
836	18.1	3.5	1.8	40.01	P856	62185	62355	171	1860	+	FRAM	SINE/Alu	7	166	0
Percent Divergence, Family				**21.07**		Length, Family		669							
315	5.0	0.0	0.0		213a7	1	40	40	42176	C	AluJ/FLAM	SINE/Alu	62	23	240
635	20.9	0.0	0.0		213a7	11815	11943	129	30273	C	AluJo/FRAM	SINE/Alu	293	165	9
849	10.4	0.0	0.0		c66	26167	26291	125	8387	+	AluSg/x	SINE/Alu	176	300	0
1567	8.2	0.0	0.9		c66	24145	24363	219	10315	C	AluSg/x	SINE/Alu	302	86	0

Table 30. Analysis of the L1 Repeats from all Projects.

Score	% Sub.	% Del.	% Ins.	Weight	Query	Begin	End	Length	Rem.	Or.	Repeat Family	Repeat Class	Begin	End	Rem.
206	36.1	2.4	1.2	67.09	213a7	6147	6315	169	35901	+	L1	LINE/L1	3991	4161	1985
285	26.8	5.5	0.8	212.50	213a7	6613	7254	642	34962	+	L1	LINE/L1	4633	5350	796
2890	18.0	6.2	2.1	195.67	c110	19709	20452	744	21900	+	L1	LINE/L1	3045	3818	2328
323	22.2	14.8	0.0	39.96	c66	31620	31727	108	2951	C	L1	LINE/L1	3502	3379	2644
Percent Divergence, Family				**30.98**		Length, Family		1663							
1338	19.1	3.3	1.2	94.16	34a5	38770	39168	399	1522	+	L1M4	LINE/L1	4271	4677	1469
263	23.5	9.3	8.0	66.10	c110	39866	40027	162	2325	C	L1M4	LINE/L1	4603	4440	1543
782	30.8	9.3	2.6	333.91	c110	31875	32656	782	9696	+	L1M4	LINE/L1	1697	2155	3882
1081	24	3.4	3.1	118.04	C86	25359	25745	387	18430	C	L1M4	LINE/L1	4832	4445	1314
2667	25.6	4.2	2.1	448.20	CosB1	7903	9307	1405	25697	+	L1M4	LINE/L1	315	1806	4607
Percent Divergence, Family				**33.82**		Length, Family		3135							
1217	11.6	8.6	1.3	50.10	P856	20700	20932	233	43283	+	L1MA1	LINE/L1	5815	6064	240
1535	8.9	2.1	0.4	26.90	P856	21242	21477	236	42738	+	L1MA1	LINE/L1	6065	6304	0
565	11.1	9.1	0.0	20.00	P856	53056	53154	99	11061	C	L1MA2	LINE/L1	6308	6201	0
1501	22.7	5.7	0.2	115.83	c66	8594	8998	405	25680	+	L1MA7	LINE/L1	5852	6278	11
4961	15.1	3.8	3.6	345.83	P856	49636	51172	1537	13043	+	L1MA9	LINE/L1	4771	6307	1
Percent Divergence, Family				**22.26**		Length, Family		2510							
1899	19.9	5.4	1.3	123.16	c66	13586	14048	463	20630	C	L1MB5	LINE/L1	6176	5695	0
654	14.1	2.3	0.8	22.02	34a5	22612	22739	128	17951	C	L1MB6	LINE/L1	6175	6046	0
580	19.8	0.0	0.0	24.95	c110	26194	26319	126	16033	+	L1MB6	LINE/L1	5821	5946	229
772	25.1	4.3	0.9	69.99	c110	30281	30511	231	11841	+	L1MB6	LINE/L1	5935	6173	2
1010	19.2	6.2	0.0	56.90	c110	25587	25810	224	16542	+	L1MB6	LINE/L1	5591	5828	347
1891	20.9	1.1	2.7	109.67	9p22	38251	38694	444	28546	+	L1MB7	LINE/L1	5676	6112	63
Percent Divergence, Family				**25.17**		Length, Family		1616							

1734	23.4	1.3	0.7	115.06	c66	26336	26788	453	7890	C	L1MC3	LINE/L1	7740	7285	0
3739	19.9	2.2	9.0	358.58	c66	27095	28247	1153	6431	C	L1MC3	LINE/L1	7291	6217	449
316	21.0	5.9	7.6	41.06	213a7	26208	26326	119	15890	+	L1MC4	LINE/L1	7861	7977	0
1686	24.6	2.8	10.4	385.94	34a5	6300	7320	1021	33370	C	L1MC4	LINE/L1	7882	6876	95
209	30.5	3.8	0.0	36.02	C86	24106	24210	105	19965	C	L1MC4	LINE/L1	7160	7052	817
235	23.6	1.8	0.0	13.97	C86	39050	39104	55	5071	+	L1MC4	LINE/L1	6885	6940	1037
650	31.2	1.8	1.1	96.16	C86	26920	27201	282	16974	C	L1MC4	LINE/L1	6989	6706	988
Percent Divergence, Family				**32.84**		Length, Family		3188							
392	24.2	6.5	0.0	38.07	c110	33241	33364	124	8988	+	L1MD1	LINE/L1	6093	6224	0
2142	20.1	3.2	3.8	143.90	c110	32658	33188	531	9164	+	L1MD1	LINE/L1	5580	6107	117
347	27.4	0.6	5.7	58.98	213a7	25845	26019	175	16197	+	L1MD3	LINE/L1	7348	7513	227
Percent Divergence, Family				**29.03**		Length, Family		830							
339	31.2	6.2	2.2	107.71	34a5	18928	19199	272	21491	+	L1ME	LINE/L1	5520	5802	366
397	33.2	1.4	0.5	73.01	34a5	21937	22144	208	18546	C	L1ME	LINE/L1	5745	5536	419
240	26.3	6.7	6.7	82.97	c66	22030	22238	209	12440	+	L1ME	LINE/L1	5532	5740	423
379	15.4	3.2	6.5	30.87	9p22	40178	40300	123	26940	C	L1ME3	LINE/L1	6157	6039	7
344	22.3	0.0	8.5	40.04	P856	26397	26526	130	37689	+	L1ME3A	LINE/L1	5707	5825	338
Percent Divergence, Family				**35.52**		Length, Family		942							
2976	19	0.2	0.5	114.26	c66	30634	31213	580	3465	C	L1P	LINE/L1	4080	3503	2066
3724	20.1	1.5	1.0	180.80	c66	29197	29996	800	4682	C	L1P	LINE/L1	4871	4068	1275
5612	24.4	1.7	1.2	424.24	c66	32051	33604	1554	1074	C	L1P	LINE/L1	3388	1828	2758
305	7.5	0.0	0.0	3.00	P856	39713	39752	40	24463	C	L1P	LINE/L1	133	94	6013
Percent Divergence, Family				**24.29**		Length, Family		2974							

2375	21.7	4.6	0.3	211.47	c66	33884	34678	795	0	C	L1P5	LINE/L1	1837	1073	4309
766	29.4	9.5	0.6	124.82	C86	34409	34724	316	9451	C	L1P5	LINE/L1	602	259	5543
805	21.8	9.3	3.5	153.97	C86	34755	35199	445	8976	C	L1P5	LINE/L1	185	-318	5960
2421	22.6	1.2	1.2	147.25	C86	33822	34410	589	9765	C	L1P5	LINE/L1	1327	739	4818
6952	20.1	2.5	0.1	359.34	C86	31932	33514	1583	10661	C	L1P5	LINE/L1	2941	1372	3205
7933	18.9	0.8	2.4	447.08	C86	35198	37220	2023	6955	C	L1P5	LINE/L1	608	-1383	5586
Percent Divergence, Family				**25.11**		Length, Family		5751							
5735	9.9	1.4	1.2	245.75	C86	29946	31911	1966	12264	C	L1PA12	LINE/L1	6153	4099	12
2494	14.4	0.6	0.4	75.92	c66	28703	29195	493	5483	C	L1PA15	LINE/L1	6157	5664	0
510	5.3	0.0	0.0	4.03	9p22	15985	16060	76	51180	+	L1PA2	LINE/L1	5348	5423	740
4740	11.3	1.1	0.1	90.38	9p22	15266	15988	723	51252	C	L1PA2	LINE/L1	6145	5416	1
3960	8.9	0.2	0.0	49.87	c110	41805	42352	548	0	C	L1PA2	LINE/L1	6144	5596	2
6477	7.8	0.3	0.2	321.54	c110	26407	30280	3874	12072	C	L1PA2	LINE/L1	6144	2266	2
877	4.5	0.0	0.0	5.00	C86	25716	25826	111	18349	C	L1PA2	LINE/L1	6144	6034	2
1831	3.0	0.0	0.0	14.28	C86	25826	26301	476	17874	C	L1PA2	LINE/L1	5490	5015	656
4749	2.8	0.0	0.0	16.21	C86	26302	26880	579	17295	+	L1PA2	LINE/L1	4428	5006	1140
5395	3.7	0.1	0.1	26.09	C86	24687	25355	669	18820	C	L1PA2	LINE/L1	6144	5476	2
5754	12.8	0.1	0.0	407.64	P856	54644	57803	3160	6412	+	L1PA2	LINE/L1	2975	6141	5
1204	3.3	0.0	0.7	6.00	c66	2480	2629	150	32049	C	L1PA4	LINE/L1	6146	5998	0
3735	20.6	1.6	2.4	328.66	c110	40041	41376	1336	976	C	L1PA7	LINE/L1	6106	4773	39
Percent Divergence, Family				**11.24**		Length, Family		14161							
5099	12.3	2.1	4.0	383.64	c66	16155	18239	2085	16439	+	L1PB1	LINE/L1	3692	5776	379
5099	12.3	2.1	4.0	65.14	c66	18286	18639	354	16039	+	L1PB1	LINE/L1	5802	6155	0
1002	12.5	7.1	0.0	36.06	c110	22646	22829	184	19523	C	L1PB2	LINE/L1	6155	5959	1
2673	12.3	1.6	0.9	84.21	c110	23133	23701	569	18651	C	L1PB2	LINE/L1	5973	5401	182
Percent Divergence, Family				**17.83**		Length, Family		3192							

consensus of 11.24%. This figure would indicate that this family is of approximately the same age as the AluY subfamily, which has an average divergence of 10.44%.

Nuclear Extract Binding to Alu Enriched DNA

Cosmid B1was mapped to determine the location of *Bam*HI and *Bgl*II restriction enzyme cut sites and their relation to the Alu sequences within the cosmid. Three full-length Alu elements were found grouped within the same 5896 *Bgl*II fragment, which extends from 27737 to 33633 of B1 (Table 27, Figure 8). This fragment and a 4469 *Bam*HI fragment spanning 19069 to 23538 of B1 containing no full length Alu sequence were cloned into M13mp18. Single-stranded DNA was isolated and bound to cellulose to produce the DNA-cellulose columns. M13mp18 DNA was isolated as a non-human DNA control. The hypothesis here was that the single-stranded Alu DNA will renature to form double-stranded regions with the remaining regions being single-stranded. The double-stranded regions are further hypothesized to serve as binding sites for Alu-specific DNA binding proteins.

As detailed in Materials and Methods, two liters of cell culture typically yielded 10μg of single-stranded DNA from the supernatant. After binding the DNA to the cellulose, the absorbance readings from the cellulose wash steps showed that greater than 85% of the initial input DNA was bound to the cellulose. The column resin bed volume was 1mL and five 0.5mL fractions were collected for each elution buffer salt concentration, except as noted. The peak fraction from each salt concentration was trichloroacetic acid precipitated and assayed for the presence of protein by SDS-PAGE. Four columns were used in these experiments: M8, the Alu-enriched DNA; M17, a column with human DNA without Alu sequence; M1C, M13mp18 ssDNA; and ND,

cellulose with no DNA bound. A double-stranded DNA control column (CT) was made from calf-thymus DNA-cellulose (2μg of DNA per gram of cellulose; Pharmacia).

Single stranded DNA binding protein (SSBP, USB) and Taq DNA polymerase were used as controls for the presence of single- and double-stranded DNA, respectively. Ten micrograms of SSBP was bound to columns M1C, NC, CT, and M8. Similar amounts of total protein were recovered from the four columns. As expected, SSBP bound strongly to column M1C as indicated by the small amount of protein eluted in the flow through fractions (Figure 9). Most of the binding activity of SSBP eluted from column M1C in a single peak in the 0.3M NaCl fractions. In contrast, SSBP did not bind columns ND (Figure 10) or CT (Figure 11) with the same affinity as M1C as seen by the amount of SSBP that eluted from these columns in the flow-through fractions. A SSBP binding activity eluted from these two columns following the addition of 0.6M NaCl. A consistent but low level of SSBP binding activity did elute from column M8 at all salt concentrations (Figure 12). These results indicate that, as expected, the M1C column is predominantly single stranded DNA. SSBP could bind to cellulose non-specifically as shown by the elution of a binding activity in 0.6M salt from columns ND, CT, and to a lesser extent, M8. Column M8 has a double-stranded DNA component as demonstrated by the higher percentage of the total SSBP that eluted from the column in the flow through fraction when compared to column M1C.

These conclusions were confirmed by the elution profile of partially purified Taq DNA polymerase from the M1C, ND, and M8 columns (Figures 13, 14, and 15, respectively). In contrast to the binding of SSBP, a binding activity was observed that elutes from the M1C and M8 columns at 0.3M NaCl that was not observed on the ND

column. These observations indicated that Taq DNA polymerase had a similar affinity for double-stranded and single-stranded DNA under these conditions. The DNA on column M8 was expected to be mostly single-stranded. A profile similar to that observed from column M8 would be seen if Taq DNA polymerase had a higher affinity for single-stranded DNA than double-stranded DNA and Taq DNA polymerase was not added to the column in a sufficient amount to saturate the single-stranded DNA.

Nuclear extracts from buffy coats were bound to the ssDNA columns and protein eluted from the columns in single fractions totaling three column volumes of each buffer. Comparison of the elution profiles revealed that each column had similar protein binding characteristics and that there was a high non-specific binding activity to the cellulose as noted above. Sixty percent of the total protein eluted from column ND was in the flow-through non-bound fraction (Figure 16). A similar elution profile also was observed from column M17 (Figure 17) since the amount of protein eluting from each column was highest in the flow-through fractions and decreased throughout the remaining salt concentrations. In contrast, in columns M1C (Figure 18) and M8 (Figure 19) over 50% of the total protein eluted in the 0.4M salt fraction. However, the proteins that elute from each column at each salt concentration appear to be the same set of proteins (Figure 20) leading to the conclusion that on renatured Alu DNA, buffy coats do not express a protein that binds specifically to Alu DNA that could be measured by the SDS-PAGE assay. Finally, in the case of the double-stranded DNA-cellulose columns, when nuclear extracts were bound and eluted, the results were inconsistent and thus were not reproducible. One speculation for this lack of consistent results with the dsDNA-cellulose columns is that

the nuclear extracts might contain proteases or double strand specific nucleases but this hypothesis was not tested further.

Figure 9. SSBP Binding to M1C ssDNA Column

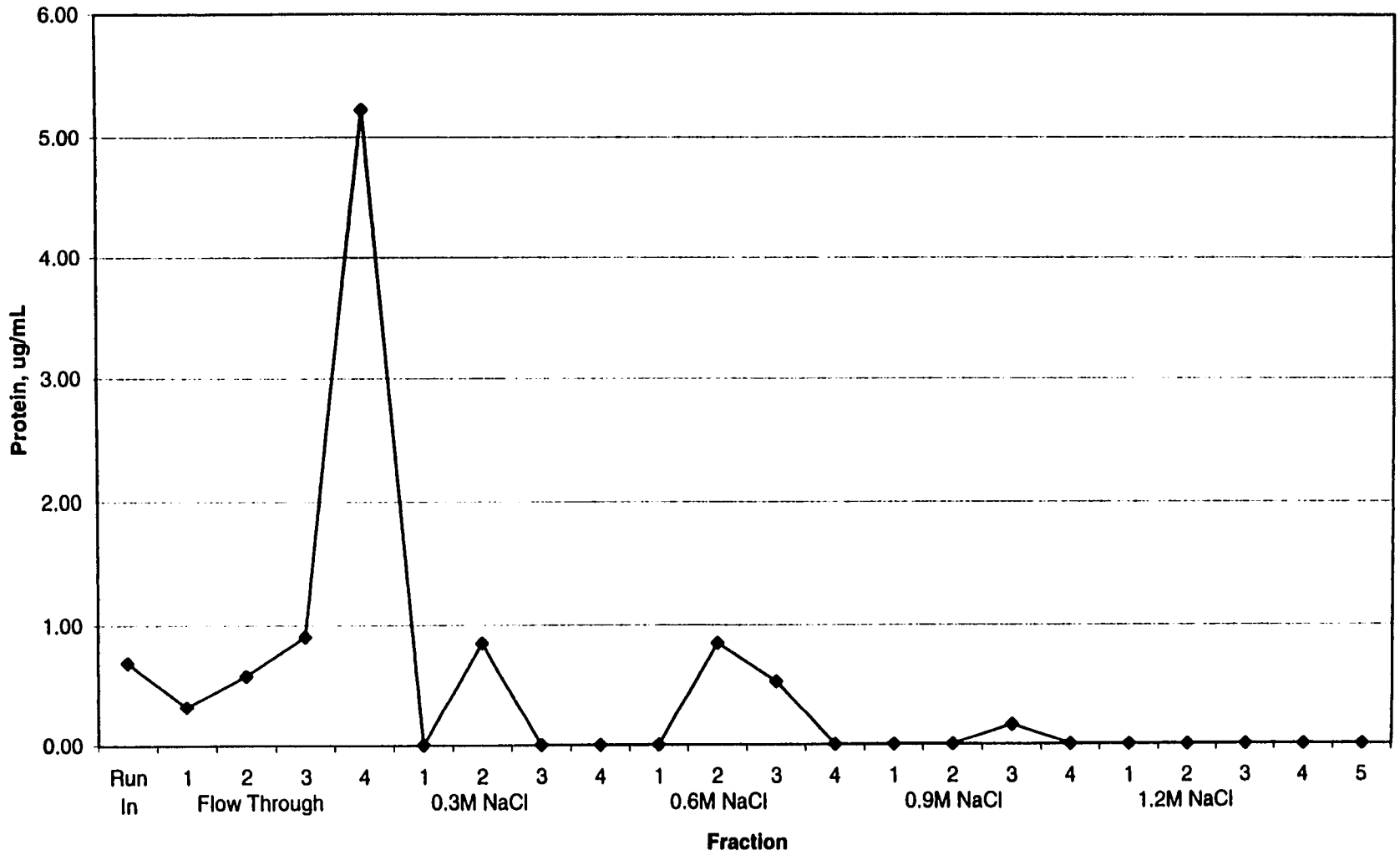


Figure 10. SSBP Binding to ND ssDNA Column

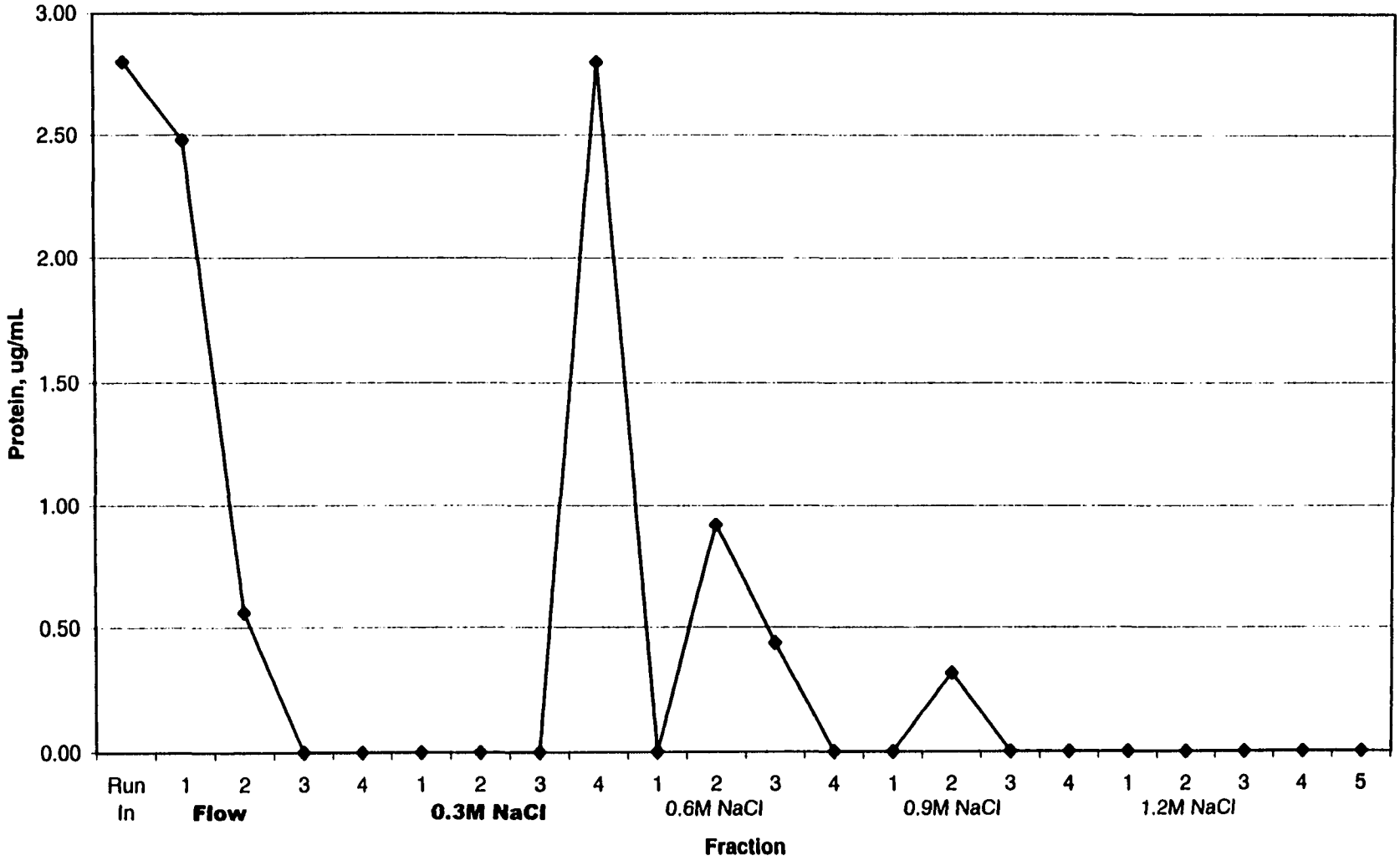


Figure 11. SSBP Binding to CT ssDNA Column

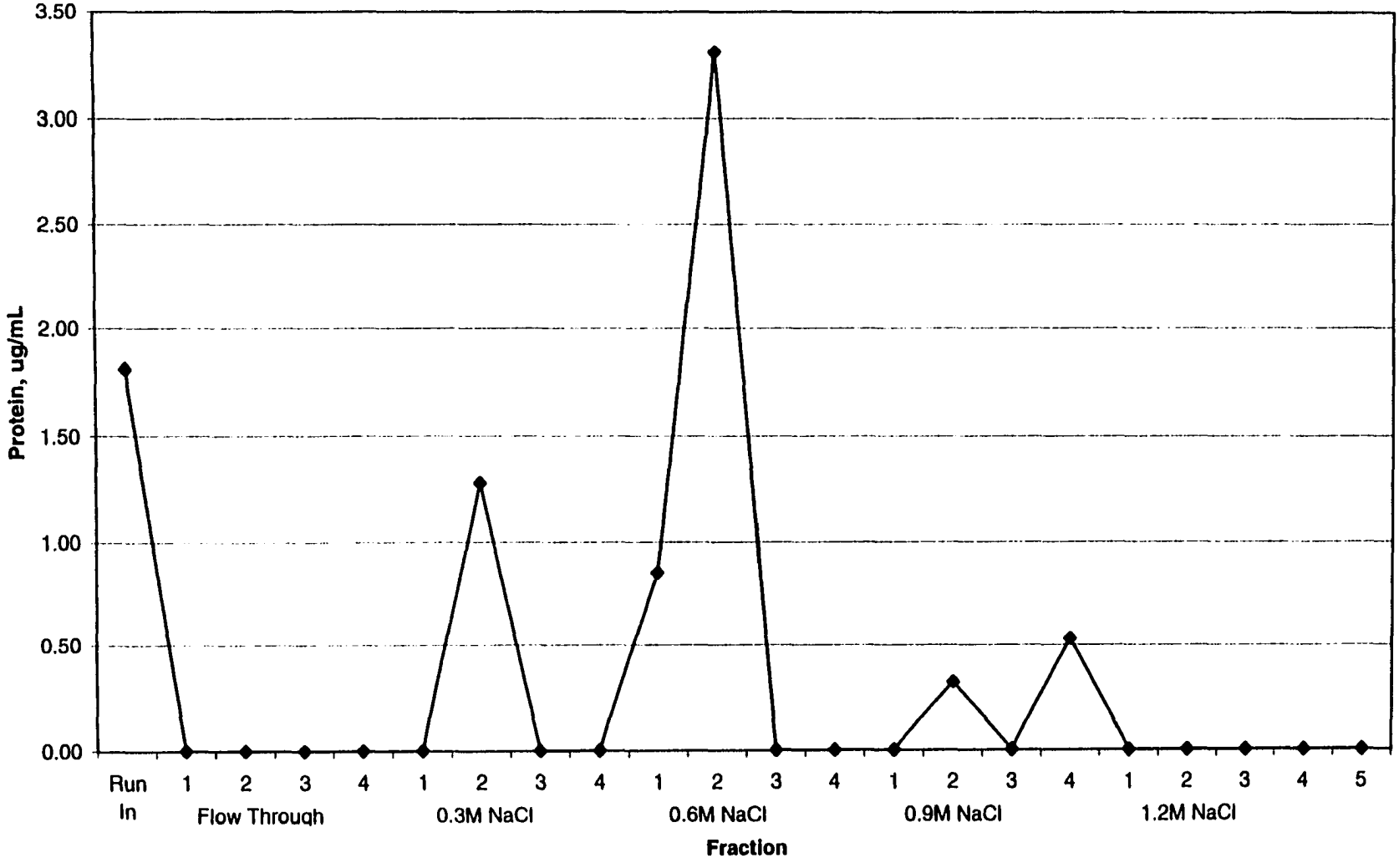


Figure 12. SSBP Binding to M8 ssDNA Column

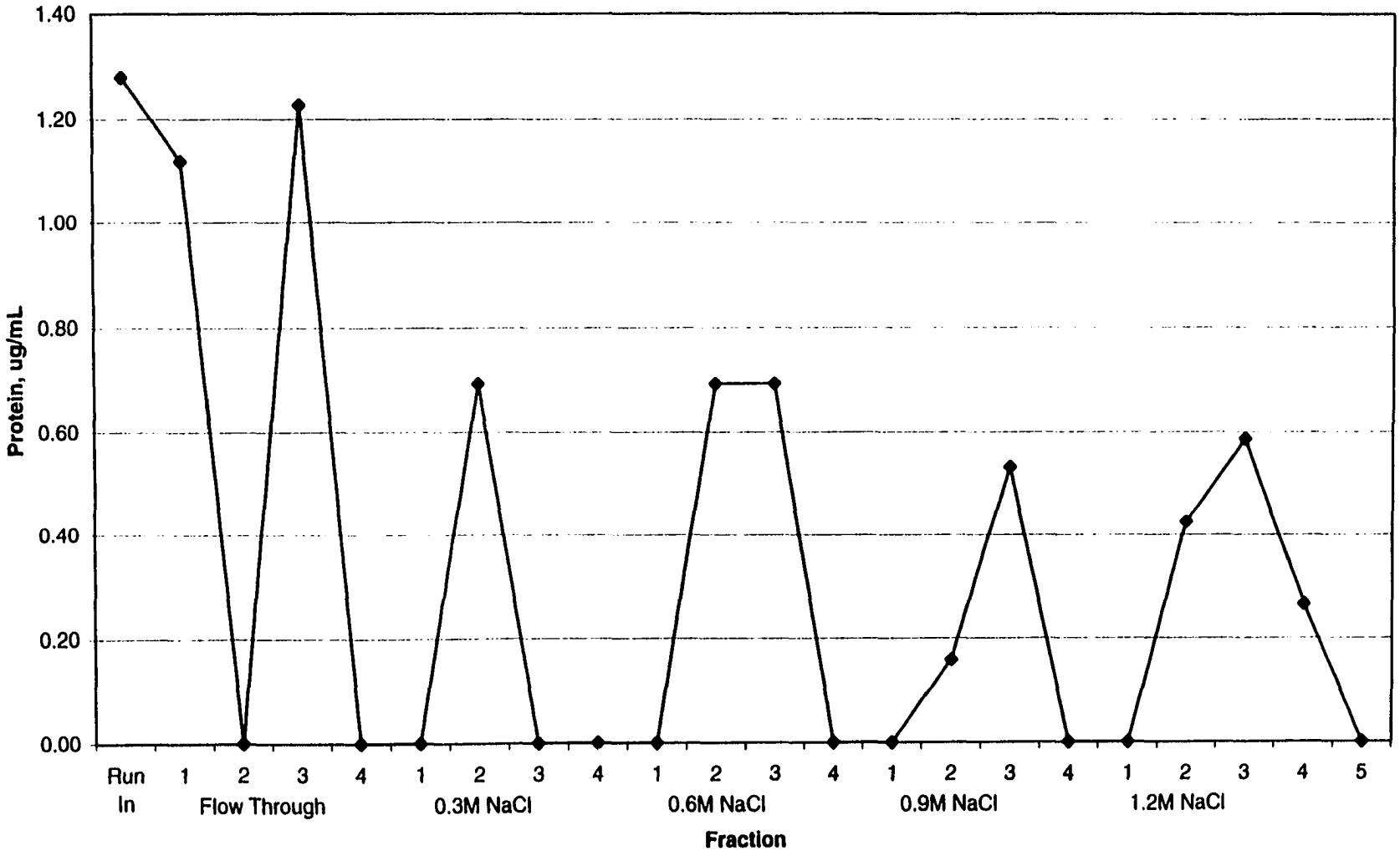


Figure 13. Taq DNA Polymerase Binding to M1C ssDNA Column

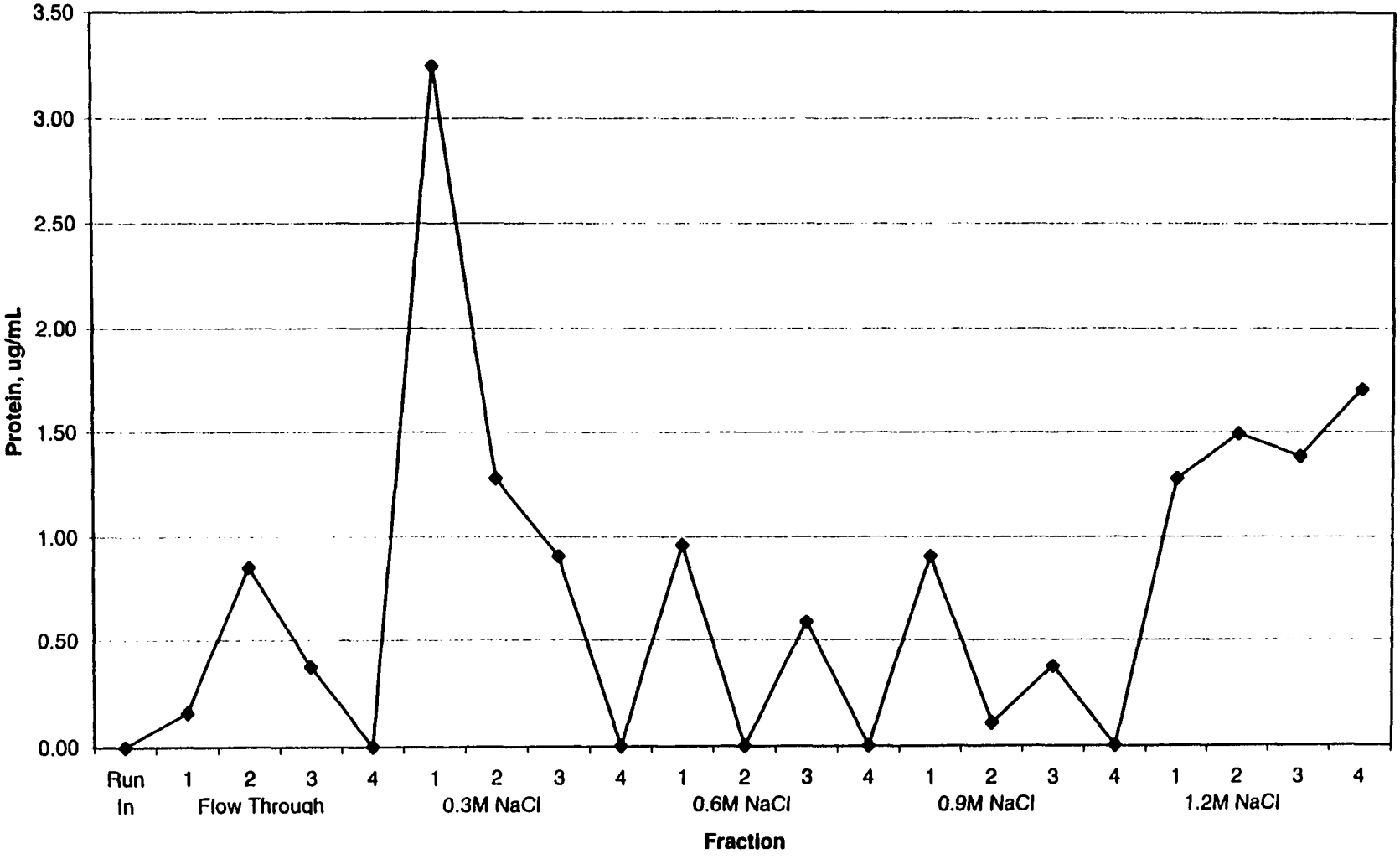


Figure 14. Taq DNA Polymerase Binding to ND ssDNA Column

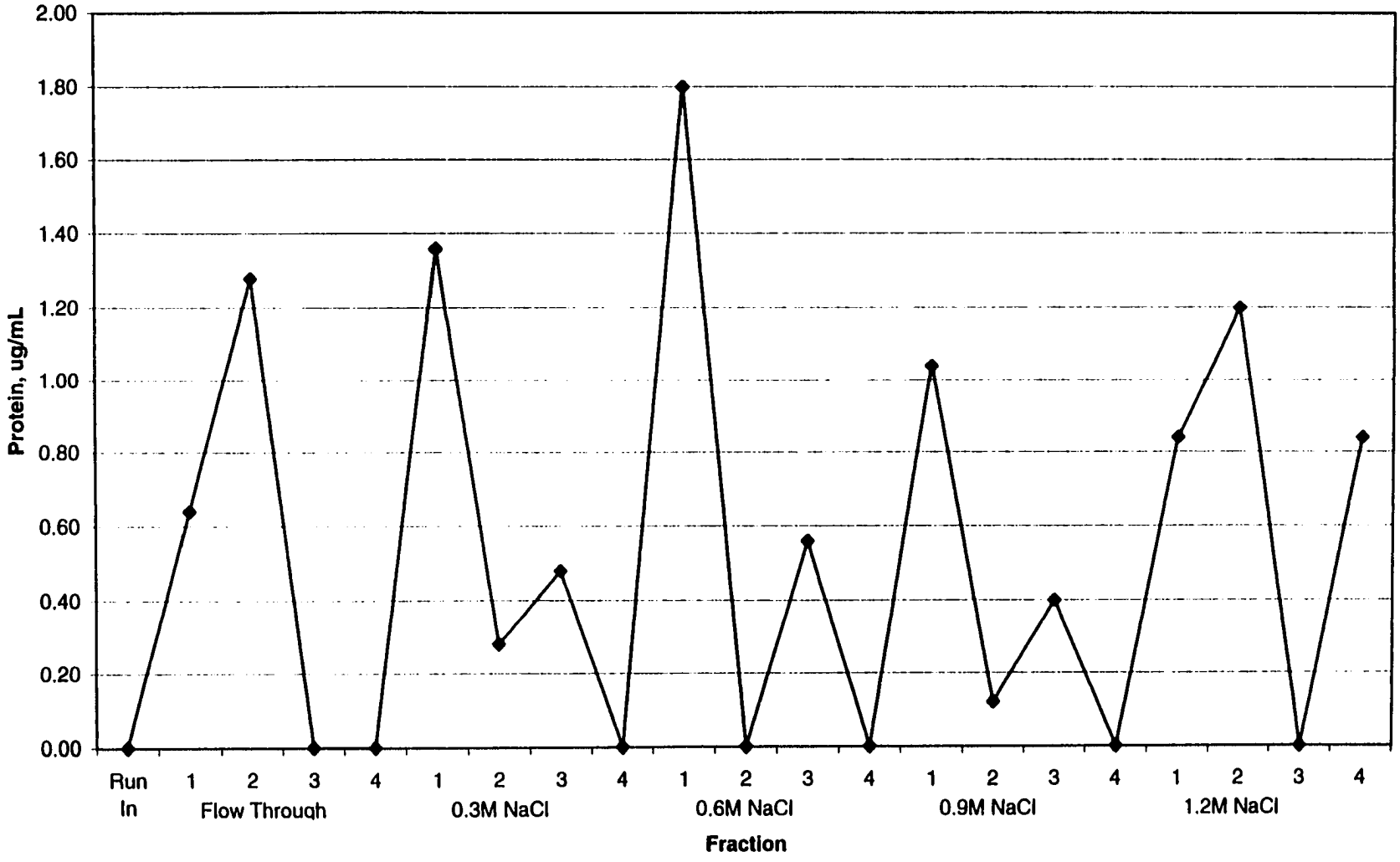


Figure 15. Taq DNA Polymerase Binding to M8 ssDNA Column

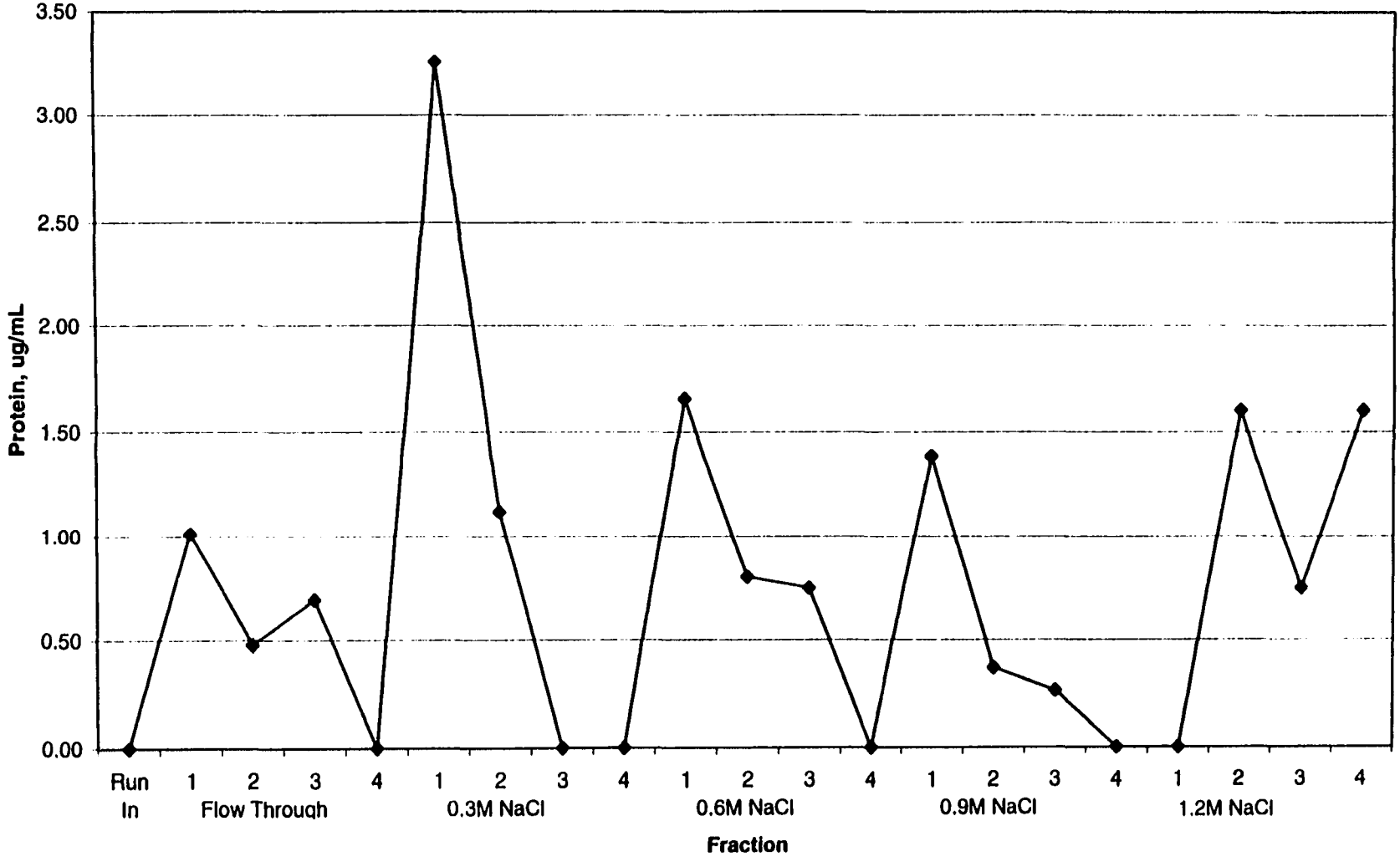


Figure 16. Nuclear Extract Binding to ND ssDNA Column

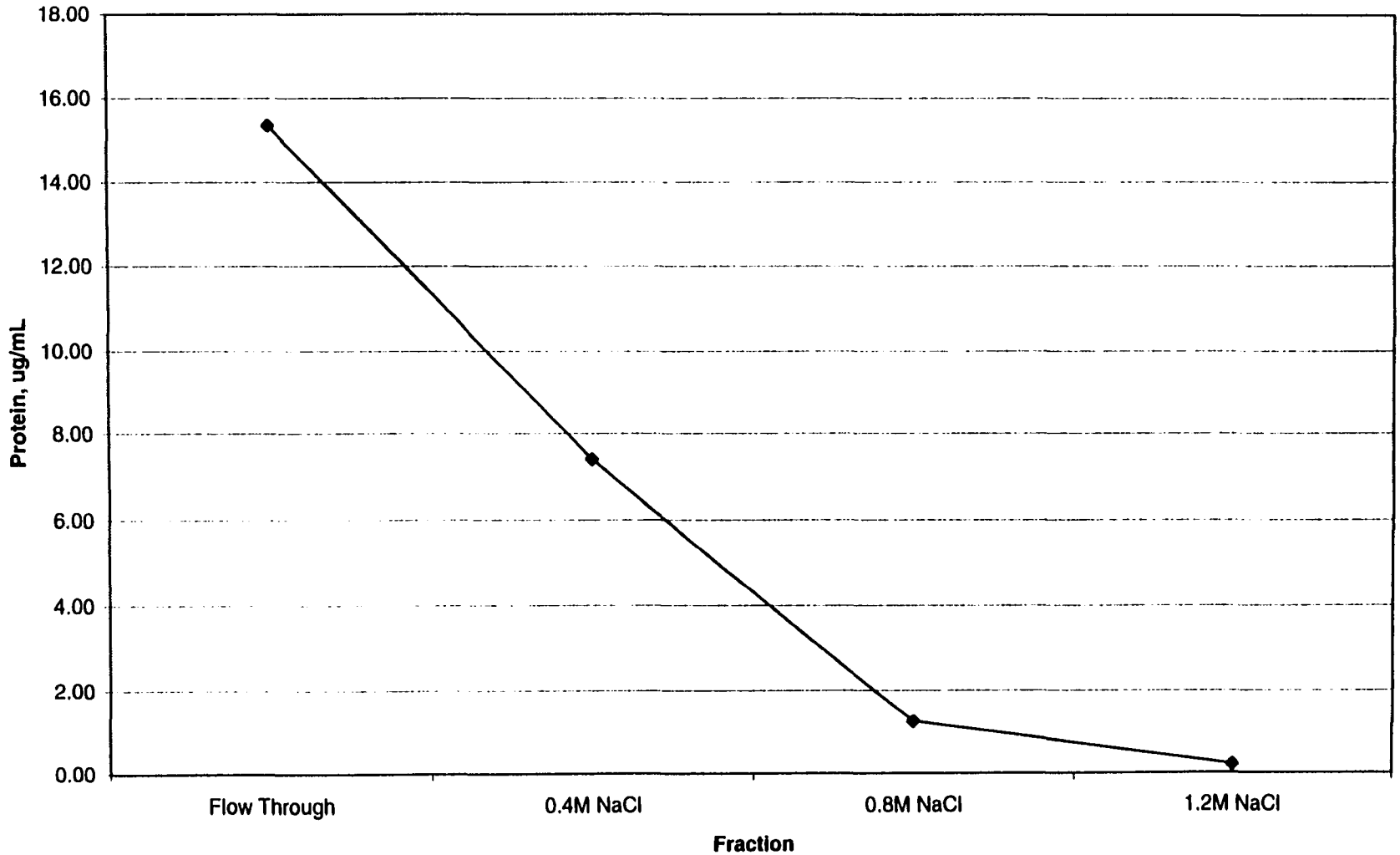


Figure 17. Nuclear Extract Binding to M17 ssDNA Column

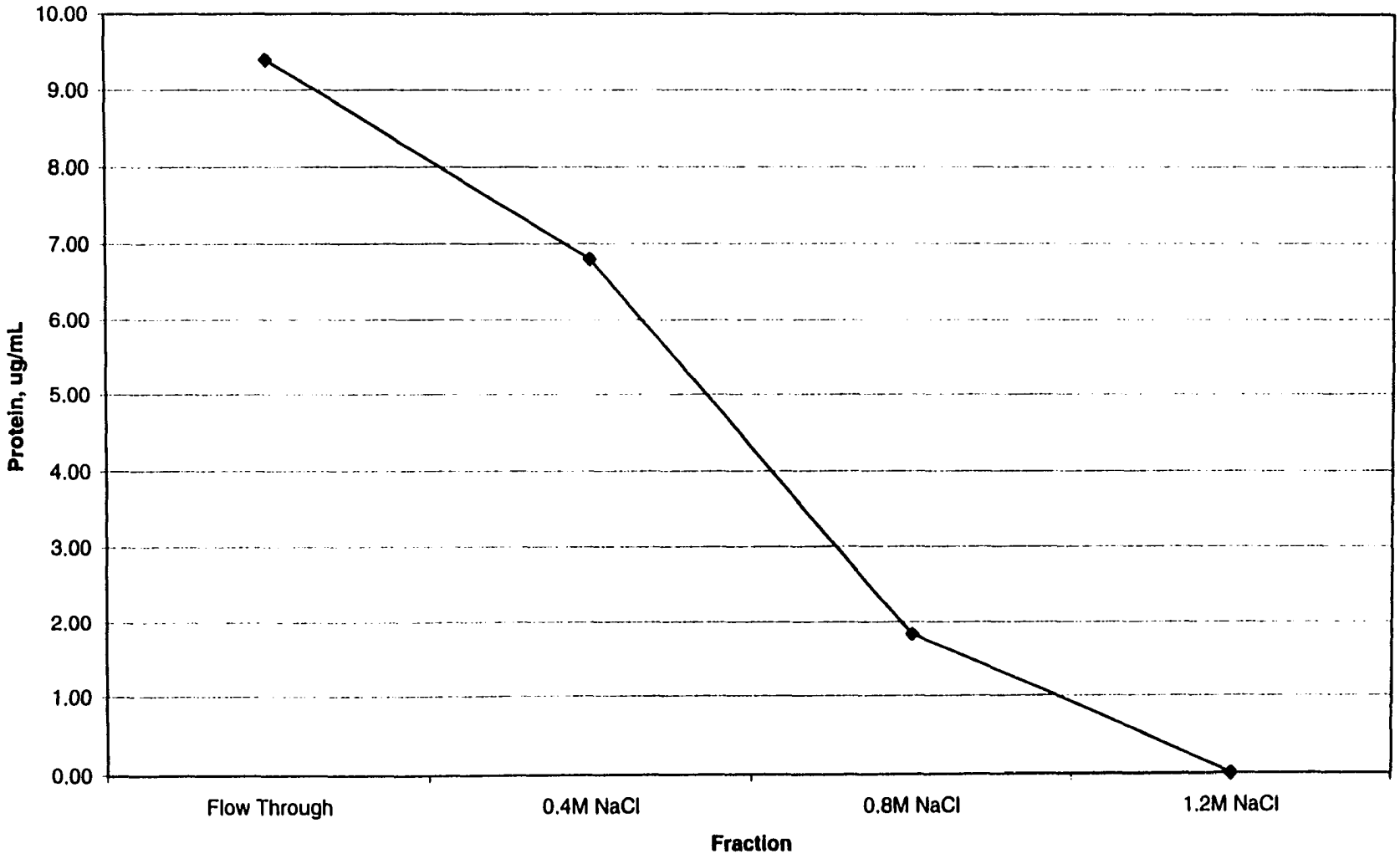


Figure 18. Nuclear Extract Binding to M1C ssDNA Column

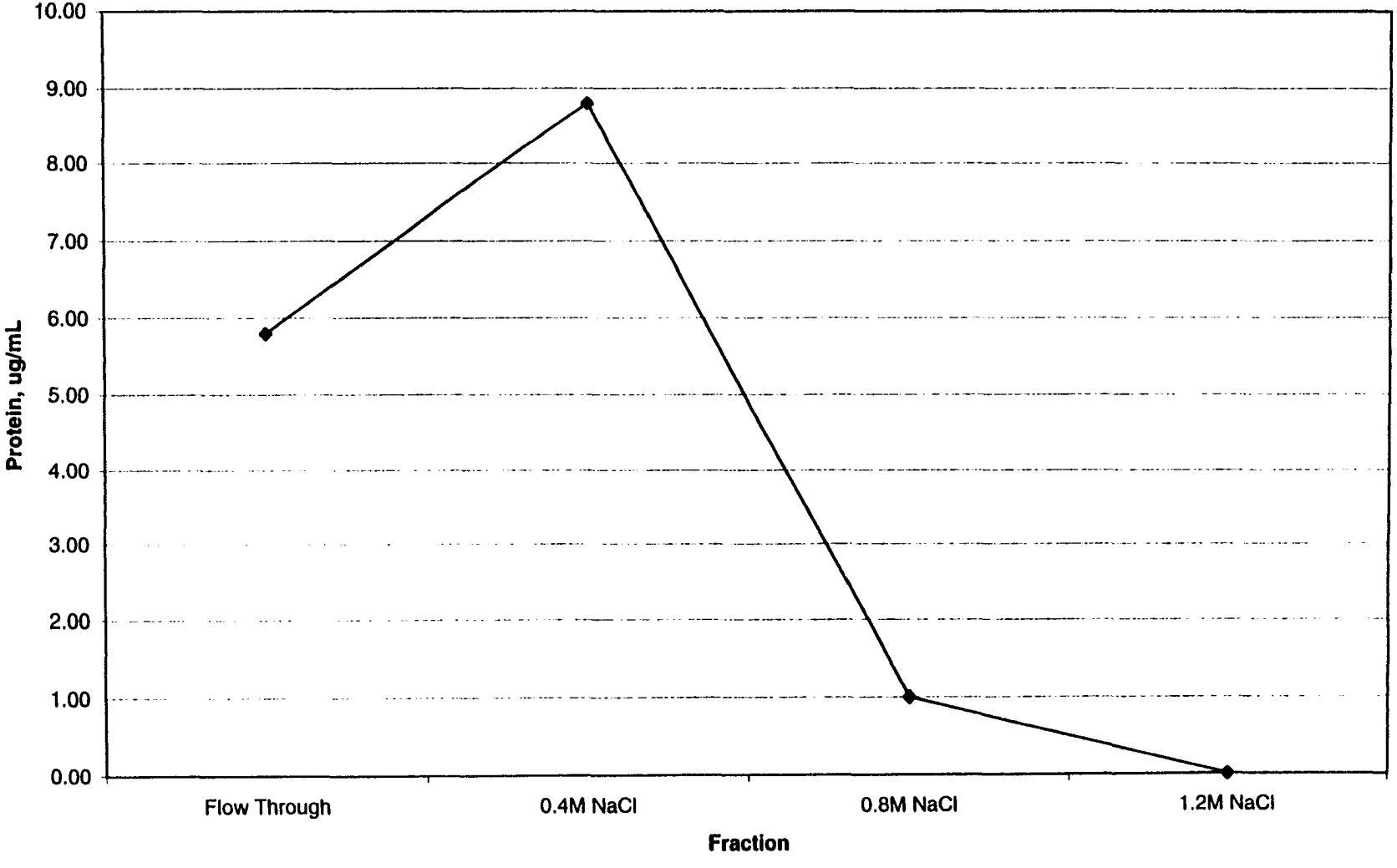


Figure 19. Nuclear Extract Binding to M8 ssDNA Column

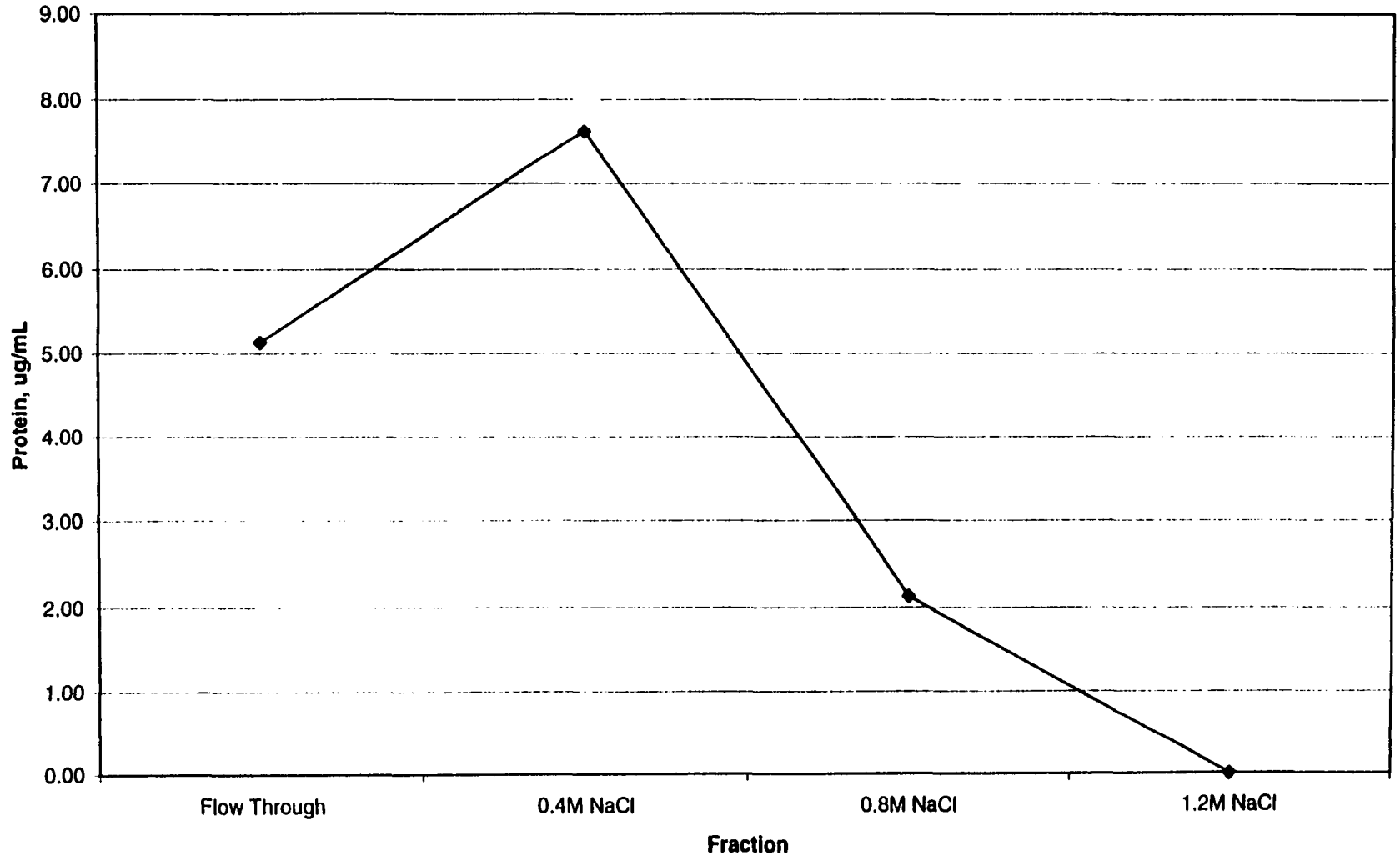
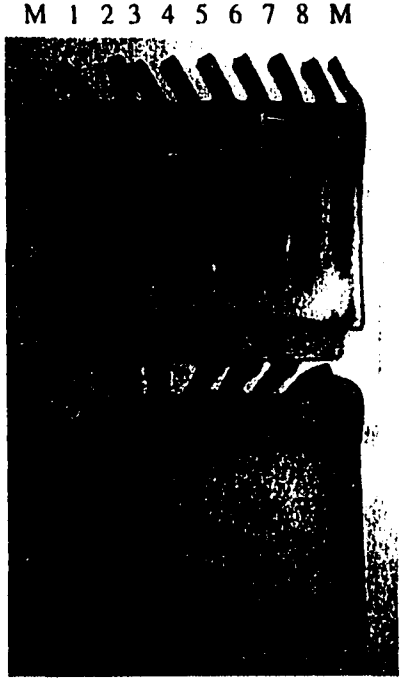


Figure 20. Assay of Proteins Bound to ssDNA Columns



M Mid-Range Protein Markers

Lane 1 Column ND, Flow Through
Lane 2 Column M1C, Flow Through
Lane 3 Column M8, Flow Through
Lane 4 Column M17, Flow Through

Lane 5 Column ND, 0.4M NaCl
Lane 6 Column M1C, 0.4M NaCl
Lane 7 Column M8, 0.4M NaCl
Lane 8 Column M17, 0.4M NaCl

Lane 9 Column ND, 0.8M NaCl
Lane 10 Column M1C, 0.8 NaCl
Lane 11 Column M8, 0.8M NaCl
Lane 12 Column M17, 0.8M NaCl

Lane 13 Column ND, 1.2M NaCl
Lane 14 Column M1C, 1.2M NaCl
Lane 15 Column M8, 1.2M NaCl
Lane 16 Column M17, 1.2M NaCl

Discussion

In dissertation research presented, 333,061 nt of sequence from three genomic regions, 9p21, 9p22, and 3q26, were determined. These three genomic regions all have been implicated in many types of human cancers. The genomic sequences and their biological relevance now will be discussed from a structure-function perspective for each region.

9p21

The genomic sequence of two tumor suppressor genes from 9p21, p16INK4a and p15INK4b, has been elucidated and the regulatory regions of these two genes analyzed. A third cosmid, c110, from this genomic region has been sequenced and analyzed for the presence of a new gene(s). A third gene, $p19^{ARF}$, is known to utilize the second exon from p16INK4a in a different reading frame and the first exon from $p19^{ARF}$ is upstream from the first exon of p16INK4a (42). Further sequencing work in this chromosome band should be targeted towards closing the gap between c86 and c66 in order to find the first exon of $p19^{ARF}$ and determine the promoter features important in the regulation of this gene and how these features differ from those of p16INK4a.

9p21, the p16INK4a gene

The p16INK4a gene is positioned in c66 such that the first exon is less than 6000 bases from the end of the cosmid. The first exon of $p19^{ARF}$ was not found in c66 or the next cosmid telomeric to c66, c86 (24), and must lie in the gap between the two cosmids. Analysis of c66 with two gene prediction programs, GRAIL (201,218) and GENSCAN yielded several putative exons other than p16INK4a exons, but a search of the databases

with the predicted protein sequence from these exons did not find significant homology to other known genes.

Proteins predicted to bind in the p16INK4a promoter by TESS (192) and MatInspector (191) were further analyzed for predicted protein interactions by FastM (199). Cellular differentiation relies on the cell cycle machinery being halted so that differentiation can begin. Given the number of different cell types in which p16INK4a deletions have been linked to cancer, it is surprising that so few tissue specific transcription factors are predicted to bind and interact with the general transcription apparatus (Table 3). One possible explanation for these findings is that $p19^{ARF}$ plays a much larger role in cell cycle control than previously thought. Evidence in favor of this hypothesis is found in the recent finding that deletion of the first exon of $p19^{ARF}$ mimics the phenotype of the previously characterized deletions of 9p21 in mice with a $p16INK4a^{+/+}$ background (219). These effects were independent of pRb (see also (220)) and were dependent on p53. Tissue specific expression of p16INK4a was observed in four-week-old mice in lung, liver and kidney and in 15-month-old mice in lung, spleen and testis (221).

MyoD binding was predicted upstream of four of the six TBP binding sites (Table 3A) and an interaction with the co-activator E47 was predicted in two of these promoters, TBP binding site positions 4291 and 4806 of c66 (Table 3A and 3B). The predicted TBP binding site at position 4806 has two putative MyoD/E47 heterodimer binding sites (Table 3B), E47 binding sites 4626 and 4583. MyoD is expressed at the earliest stages of myogenesis and continues through late myotube differentiation (222) and ectopic expression of MyoD in fibroblast, melanoma, liver, neuroblastoma and adipocyte cell

lines results in the expression of myogenic proteins (222,223). E47 is ubiquitously expressed (224) and binds DNA as a helix-loop-helix mediated homo- or heterodimer with other factors including MyoD and E12. E12 is thought to be a splice variant of E47 (224). E47 is known to bind immunoglobulin heavy and light chain enhancers and the creatine kinase enhancer in myocytes and B-cells (225) and upregulates the insulin control element in pancreatic beta cells (226). B-cells overexpressing E47 increased DJ rearrangements by 150-times (227).

Three of the transcription factors predicted to bind in the p16INK4a promoter by TESS or MatInspector play a direct and cell-type restricted role in blood cell differentiation (Table 3A), NF-κB, Ik-2 and GATA. NF-κB is important in B- and T-cell maturation and is induced by IL-1, among other cytokines (228). The Ikaros (Ik) family of proteins has five members and expression of the family is restricted to lymphocytes (229). Ik-2, the Ik gene family member predicted to regulate p16INK4a, is expressed in the fetal liver, maturing thymus, and postnatal spleen, all sites of early or embryonic lymphocyte maturation (230). Interestingly, Ik-2 binds to the enhancers of both IL-1 and NF-κB. Although NF-κB is not predicted by FastM to interact with TBP in the P16INK4a promoter, a regulatory progression can be envisioned where Ik-2 initially up-regulates p16INK4a, IL-1, and NF-κB expression. Ik-2 induced expression of p16INK4a then is "handed off" to NF-κB at a later maturation stage through either a direct Ik-2 mediated increase in intracellular NF-κB or an extracellular effect through IL-1 induced expression of NF-κB. The third circulatory system transcription factor predicted to regulate p16INK4a, GATA, is expressed in erythroid cells, mast cells, and megakaryocytes (231). GATA binds to globin gene promoters and locus control regions

and can exert an effect as much as 100kB away from the binding site. Recently, a complex of five proteins has been isolated from mouse hematopoietic cells that includes GATA and E47 and was required for transactivation from a binding site that included the E47 and GATA consensus binding sites (212). The *in vivo* function of this complex is unknown at this time, but *in vivo* transactivation of a reporter construct was observed when the proteins from the complex were co-expressed. FindPatterns from the GCG package at low stringency was used to search c66 for the consensus binding site of this complex, but no sites were found upstream of p16INK4a (data not shown).

The liver specific transcription factor, hepatic nuclear factor-3b (Hnf3b) (232), is predicted by TESS to bind upstream of the TBP site at position 4806 and is predicted by FastM to interact with TBP in this predicted promoter. Hnf3b has been shown to be an early regulator of hepatocyte differentiation (232). Furthermore, p16INK4a expression has been observed in the liver of four-week old mice (221).

Taken together, p16INK4a mediated entrance or maintenance of muscle, lymphoid and erythroid, and liver precursors into a differentiation mode can be predicted from the various activating transcription factors predicted to bind and interact in the p16INK4a promoter. Furthermore, p16INK4a expression patterns in mice substantiate a possible role for p16INK4a in maturation of hematopoietic cells and maintenance of differentiated liver tissue in a G0 state. Further biochemical evidence will be needed to confirm these findings.

9p21, the p15INK4b gene

The gene prediction programs GRAIL and GENSCAN were used to search for putative or unknown genes in c86. Similar results to those for c66 were observed for c86,

no regions of homology to known genes other than p15INK4b were discovered by this analysis. An analysis of the promoter region upstream of p15INK4b was performed, but in contrast to the results discussed above for the p16INK4a promoter, a much larger number of transcription factors were predicted to bind and interact in the p15INK4b promoter (Table 8). These predicted interactions are in agreement with the observation that p15INK4b is expressed in virtually all tissues in mice older than four weeks of age (221). Because NNPP predicted that the p15INK4b promoter was TATA-less, FastM interaction predictions were performed to GC-box proteins.

Ap1, Ap2, and Ap4 are predicted to interact with GC-box proteins at various positions upstream of p15INK4b (Table 8). Ap1 is a heterodimer of the proto-oncogenes c-fos and c-jun (233). Activation through any of the three factors is induced by phorbol esters implying activation via PKC (234,235) and cAMP mediated activation has been observed in promoters binding either Ap2 (236) or Ap4 (237) implicating PKA regulation of promoters binding either Ap2 or Ap4. Binding sites for two or all three factors overlap in some promoters (237,238). Expression of Ap2 may be limited as it is not expressed in the HepG2 cell line (236) Competition for Ap1 or Ap2 binding sites was not observed in liver extracts but was seen with HeLa extracts (238). An analysis of the expression of Ap2 in mouse embryos showed expression in 13.5 days post-coitus embryos in the hindbrain and spinal cord, face, body, and limbs (239). Binding site overlap between Ap1 and GATA (240) or Oct1 (241) has been observed as well. A synergistic activation of the SV40 promoter is observed in response to Ap1 and Ap4 binding to adjacent sites (242) but binding of Ap1 to a negative regulatory site also has been observed (241).

The cAMP response element binding protein (CREB) is predicted to bind and interact with GC-box proteins in the p15INK5b promoter. CREB mediated activation can be in response to phosphorylation by PKA and subsequent heterodimer formation with CREB-binding protein (243) and complex formation of CREB with ATF1 (244) or ATF2 (245,246) has also been observed. The CREB/ATF2 complex is mediated by a third protein in some cases (245,246). CREB directly interacts with TAF110, a TFIID subunit (247).

Oct1 is ubiquitously expressed and has been isolated from B- and HeLa cells (248,249). The protein is known to activate transcription of immunoglobulin genes (248,250,251) and the histone H2b gene (249). A brain specific member of the Oct family (252), Brn2, is predicted to bind in the promoter of p15INK4b. Oddly, this transcription factor also is expressed in SCLC (253).

Upstream Stimulatory Factor (USF) is another ubiquitously expressed protein predicted to interact with other transcription factors in the p15INK4b promoter. The protein is localized to the nucleus in a cell cycle independent manner (254). Two forms of USF are observed, a 43Kd and a 44kD form that has been implicated as a regulatory subunit (254). USF binds to and stimulates transcription from both the alcohol dehydrogenase (255) and the adenovirus major late promoters (256). In the case of the adenovirus major late promoter (256), activation was decreased in highly purified USF fractions implying that additional factors may be necessary for full activation in this promoter.

The myc/max heterodimer has interesting structure in that dimer formation depends on the presence of a basic helix-loop-helix (bHLH) and a leucine zipper in each

subunit (257,258). Furthermore, dimers are not observed between other proteins that have either a bHLH or a leucine zipper (257). Transformation by Ha-Ras requires formation of the heterodimer (259). Transactivation is not observed in myc homodimers (260) and depends on phosphorylation at a specific serine residue by either MAP kinase or cdc2 (261). This last finding implies that myc/max is regulated extracellularly and in response to cell cycle controls.

The Stat family of transcription factors are activated in response to interferon-α (IFNα) -β or -γ stimulated phosphorylation of both serine and tyrosine residues (262-264). Epidermal growth factor (EGF) and IL-6 also stimulate complex formation and cell treatment with EGF induces phosphorylation and nuclear localization (265). A member of a second family of interferon responsive transactivating proteins is predicted to interact with GC-box proteins in the p15INK4b promoter, Irf1. Irf1 expression is induced up to 20-fold by IFNγ (266) and is responsive to IFNα, TNFα, IL-1, viral infection, and IL-6 (266,267).

A variety of cell-type specific transcription factors is predicted to bind in the p15INK4b promoter. *Ets* isoforms have differing effects when complexed with Ap1. *Ets*-p54 can act as an inhibitor in the proper context (268), but can be activating in other promoters (269,270). *Ets*-p54 is expressed in lymphocytes from embryonic thymus and the Bursa of Fabricus as well as from cell lines of B-, T-, and nonB-nonT origin (271) implicating it in lymphocyte maturation or differentiation. Two members of the forkhead family have predicted binding sites and interactions within the p15INK4b promoter. Hnf3b is a liver specific transcription factor thought to be an early regulator of hepatocyte differentiation (232) while the hfh2 specific binding site is bound strongly bound by

nuclear extracts from kidney (272). Lyf1 is expressed in only B- and T-cells and activates the terminal deoxyribonucleotidyl transferase (tDt) gene important in immunoglobulin gene rearrangements (273). Lyf1 is thought to play an early role in lymphocyte maturation because tDt expression is an early step. Gfi1 is a repressor protein as demonstrated by down-regulation of the Gfi1 gene in myelomonocytes induced to differentiate (274).

S8 is expressed in embryonic mice in the craniofacial mesenchyme, heart and somites, but is not expressed in any nervous system tissues or precursors (275), or hematopoietic tissues (276). Nkx2.5 is embryonically expressed in myocardial precursors and very closely related tissues of mice (277). Nkx2.5 protein is first detected at the late primitive streak stage of development and continues into adulthood. E47, GATA, Ik2, and NF-κB regulation are discussed above.

Because transcription factor binding sites specific for PKA, PKC, and cytokine regulated transcription factors as well as cell-type specific factors from tissues including heart, liver, and hematopoietic cells are predicted (Table 8), regulation of p15INK4b is probably quite complex. Cell-type specific factors are thought to play a role in activation by several of the ubiquitously expressed proteins discussed. It is noteworthy that two embryonically expressed transcription factors are both expressed specifically in heart tissue. This finding is inconsistent with the observation that p15INK4b is not readily detected in mouse embryos (221). An interesting finding of this analysis is that SRY, the testis determining factor, has predicted interactions with GC-box proteins in the p15INK4b promoter at three GC-box protein sites. The implications of this interaction are unknown but could be related to spermatocyte differentiation. Tissue specific

expression of p15INK4b is observed in mice in hematopoietic tissues and brain consistent with predicted tissue-specific transcription factor interactions in the p15INK4b promoter and confirming a putative role for p15INK4b in the differentiation or maintenance of these tissues. In any event, further studies of the regulation of this gene based on the above analysis now would be useful.

9p22

Four cosmids from the af9 gene region have been sequenced. All but presumably one exon from the af9 gene have been found. Analysis of this region with GRAIL and GENSCAN did not reveal the presence of any other known genes. Analysis of the promoter region of af9 revealed a limited set of predicted protein interactions. Similar to the p15INK4b promoter, the af9 promoter was predicted to be TATA-less. FastM interactions were performed using the Sp1 protein, which had several binding sites predicted by TESS and MatInspector. Two ubiquitously expressed proteins are predicted to bind and interact with Sp1, Oct1, which has been discussed above, and Nf1. Nf1 was originally isolated from HeLa cells as a cellular factor necessary for adenovirus genome replication (278). It was later found to act as a transcriptional activator of the α-collegenase (279,280) in response to TGFβ (280). Nf1 is expressed as a cell-type specific family of transcripts ranging in size from 0.6kB to 4.5kB including a testis specific transcript of 0.8kB and a liver specific transcript of 0.6kB (281). Expression also is observed in thyroid, kidney, spleen, brain, ovary and heart tissues (281) and differentiating heart (282). Activation of the human growth hormone gene by Nf1 (283) has been observed.

The Ikaros family member predicted to bind in the af9 promoter is Ik1 which is expressed in a similar manner to the Ik2 gene discussed above except that it is present in the neonatal mouse brain and binds to a slightly different sequence element (229). Elk1 is a member of the ets family that has been reported to be expressed in the mouse lung and testes (284), and myeloid and lymphoid cells (285). Elk1 is a potent transactivator when bound to its target sequence in the Moloney murine sarcoma virus LTR, E74, or polyoma virus enhancer (285). Regulation of c-fos has been shown to be mediated by formation of a ternary complex that includes Elk1 and serum response factor (286). The myeloid cell specific transcription factor Mzf1 is the last protein not already discussed that is predicted to bind in the af9 promoter. This protein is not expressed in lymphoid cells, monocytes, macrophages, lung, brain, melanoma, muscle, pancreas or endothelial cells (287). The highest levels of expression were in HL-60 cells induced with retinoic acid leading to the conclusion that Mzf1 may function in hematopoiesis. The Hfh2 protein from kidney is predicted to bind in the af9 promoter as well.

Since the cellular function of the af9 gene is unknown at this time, there may be more to gain from a careful analysis of predicted protein interactions in the af9 promoter. The restricted nature of the cell types expressing proteins like Ik1, Mzf1, and Hfh2 offer a starting point. The opportunity now exists to use the sequence data presented here and study the expression of the af9 gene *in vitro* armed with antibodies to the proteins predicted to bind and interact within the af9 promoter. Northern blots have revealed af9 expression at high levels in the thymus and peripheral blood cells and low levels in the spleen (57) consistent with a role for Ik-1, Mzf-1, and Elk-1 in the regulation of af9. Regulation upstream of the transcription factors remains a question although Mzf1

activation is in response to retinoic acid in at least one promoter and Ikaros gene regulation is uncharacterized at this time.

A detailed investigation of the sequence surrounding the known breakpoints now reveals that seven of the 10 breakpoints are within two nucleotides of topoisomerase II cleavage sites (Table 21). However, two of these seven breakpoints, the Negrini breakpoint (74) and S79541 (65) are *de novo* rearrangements. In both cases, the breakpoint exactly corresponds to the enzymatic cleavage site on the opposite strand. In the case of the Negrini breakpoint (74), the authors believed immunoglobulin recombinase to play a role in this translocation. The monomac 6 cell line (61) breakpoint corresponds to a 70/30 topoisomerase cut site, below the accepted criteria for topoisomerase II cleavage (68). Patient Gill 2 (215) is a therapy related translocation but a topoisomerase II site is not found close enough to the breakpoint of this patient to be significant. Patients 3 and 10 (217) have therapy-related translocations, but although a topoisomerase II site is close to these breakpoints, the cleavage sites do not correspond to the breakpoints. The remaining four patients, patients 1, 5, and 6 (217), and Gill 1 (215), have therapy-related translocations and each has a topoisomerase II cleavage site within two nucleotides of their breakpoint. A mechanism can be envisioned for these cases that begins with topoisomerase II cleavage followed by an exonuclease mediated loss or terminal deoxynucleotidyl transferase mediated addition of a small number of bases followed by ligation of the modified end to form the translocation. The data presented here are in contrast to the data from breakpoints in 11q23 where an association was made to SARs but not to topoisomerase II sites (58). During the analysis of 9p22, a very large number of topoisomerase II sites were found that matched one strand at least 70%. A

seven nucleotide sequence would be expected to randomly occur every 16000 bases. If the topoisomerase II site obeyed random chance, four sites would be expected on each strand of contig 2. The actual number is 650 times greater than that, each strand of contig 2 averages 2654 topoisomerase II sites of 70% or greater. An analysis of all the 70% or greater topoisomerase II sites to determine how many meet the additional criteria of a 40% match of the opposite strand was not done. If the ratio of the 20 sites that were analyzed were to hold, 75% of the sites would be considered true topoisomerase II cleavage sites. One conclusion that could be drawn from this is that a large SAR may be in this region of 9p22. Additional work will be needed to determine the location of SARs in 9p22 and clone more breakpoints, but until such time as the sequence surrounding more t(9;11) breakpoints is known, the role of topoisomerase II in therapy related leukemia is very much in question.

The five nucleotide Translin site is the only other site that was found in proximity to any of the breakpoints. None of the breakpoints is closer than 12 nt of a Translin site. In addition, the chances of finding a five nucleotide site with one mismatch by random chance would be the same as finding any four base site, one every 256 bases. When this is added to the observation that only one exact match to the Translin site was seen close to a breakpoint, the logical conclusion is that the Translin protein probably does not play much of a role in the currently known translocations in 9p21.

Nine of the ten known breakpoints are within a 15kB region upstream of af9 exon 5 (Figure 6b). This observation could lead to the hypothesis that there is a region of genomic instability in intron 4 of the af9 gene. However, there is currently no way to test

this hypothesis since the occurrence of translocations without a disease phenotype is unknown.

Future sequencing work in this chromosome band should include closure of the two gaps to determine the remaining genomic features of the af9 gene.

3q26

A single genomic DNA clone from this region has been sequenced and one of the three exons from the MDS1 gene was found in this clone. GRAIL and GENSCAN did not predict any exons with homology to presently known genes other than MDS1 and reverse transcriptase from an L1 repeat.

During the course of this work, not only was the sequence of the coding regions and the intron-exon structure of the genes in over a third of a megabase of genomic DNA determined, but through the use of all available "promoter finding" programs, the regions that regulate the complex expression patterns for the sequenced tumor-related genes have been investigated. Specifically, as shown in Tables 3, 8, and 17, the expression of these critical genes can be dissected by investigation of the 5' flanking regions and the results obtained by this computer analysis are generally consistent with the observed expression patterns of these genes. An important distinction needs to be made. As sequence data continues to be generated by labs around the world from many different organisms, new genes are continuously being discovered and it is important to bear in mind that today's predicted exon could be tomorrow's gene. Another important point is that the protein binding sites and interactions are predicted by a computer using a matrix to determine the likelihood of binding or interaction. This is a starting point for further work, but there is

no biochemical evidence to support protein binding sites or sites for activation of p16INK4a, p15INK4b or af9 in the above sequence data.

Nuclear Extract Binding to Alu Enriched DNA

A preliminary set of experiments has been performed to determine if there is specific binding of a unique nuclear protein to Alu DNA. The results from buffy coat nuclei were negative. Binding to the ssDNA columns was observed, but the same set of proteins bound to all columns. A subset of nuclear proteins already is known to bind Alu DNA including histones (176) and RNA PolIII. It is possible that if a unique Alu-DNA binding factor exists, binding of this factor is difficult to detect due to the high background of other proteins binding to these columns. Fractionation of the nuclear extracts on an ion-exchange or size exclusion column could decrease the background to a level that allowed detection of the unique protein. Alternatively, histones and RNA PolIII could be immunoprecipitated from the nuclear extracts to remove these ubiquitous proteins and decrease the background.

Buffy coat nuclei were used in these experiments because they are readily available. Other cell types and tissues might be used to cover a wider range of protein expression patterns from terminally differentiated cells. A second problem with buffy coats is that they are predominantly monocytes and as such are high in proteinases and nucleases (288). A modified procedure that makes more liberal use of the proteinase inhibitor PMSF was found (288) after the nuclear extracts had been prepared. The experiments within this dissertation were not repeated with nuclear extracts prepared using this modified protocol, but preliminary evidence of protein degradation was observed and is consistent with the variable yields between different preparations of the

nuclear extracts. Specifically, when cells that had been thawed on ice and re-frozen were used, the yield per mL was about one fourth that obtained when fresh cells were used. In conclusion, these preliminary experiments show that the basic strategy of allowing single stranded DNA to anneal to itself thereby enriching the DNA for the sequence of interest is sound and might be useful for others to pursue.

References Cited

1. Weinberg, R. A. (1995) *Cell* **81,** 323-330
2. Reed, S. I., Bailly, E., Dulic, V., Hengst, L., Resnitzky, D., and Slingerland, J. (1994) *J. Cell Sci. Suppl.* **18,** 69-73
3. Peters, G. (1994) *J. Cell Sci. Suppl.* **18,** 89-96
4. Sherr, C. H., and Roberts, J. M. (1995) *Genes Dev.* **9,** 1149-1163
5. Hannon, G. J., and Beach, D. (1994) *Nature* **371**(6494), 257-61
6. Sandhu, C., Garbe, J., Bhattacharya, N., Daksis, J., Pan, C. H., Yaswen, P., Koh, J., Slingerland, J. M., and Stampfer, M. R. (1997) *Mol. Cell. Biol.* **17**(5), 2458-67
7. Reynisdottir, I., Polyak, K., Iavarone, A., and Massague, J. (1995) *Genes Dev.* **9**(15), 1831-45
8. Reynisdottir, I., and Massague, J. (1997) *Genes Dev.* **11**(4), 492-503
9. Serrano, M., Lee, H., Chin, L., Cordon-Cardo, C., Beach, D., and DePinho, R. A. (1996) *Cell* **85**(1), 27-37
10. Alcorta, D. A., Xiong, Y., Phelps, D., Hannon, G., Beach, D., and Barrett, J. C. (1996) *Proc. Natl. Acad. Sci. U S A* **93**(24), 13742-7
11. Koh, J., Enders, G. H., Dynlacht, B. D., and Harlow, E. (1995) *Nature* **375**(6531), 506-10
12. Yang, R., Gombart, A. F., Serrano, M., and Koeffler, H. P. (1995) *Cancer Res.* **55**(12), 2503-6
13. Tsuzuki, T., Tsunoda, S., Sakaki, T., Konishi, N., Hiasa, Y., and Nakamura, M. (1996) *Cancer* **78**(2), 287-93
14. Hangaishi, A., Ogawa, S., Imamura, N., Miyawaki, S., Miura, Y., Uike, N., Shimazaki, C., Emi, N., Takeyama, K., Hirosawa, S., Kamada, N., Kobayashi, Y., Takemoto, Y., Kitani, T., Toyama, K., Ohtake, S., Yazaki, Y., Ueda, R., and Hirai, H. (1996) *Blood* **87**(12), 4949-58
15. Diaz, M. O., Ziemin, S., Beau, M. M. L., Pitha, P., Smith, S. D., Chilcote, R. R., and Rowley, J. D. (1988) *Proc. Natl. Acad. Sci. U S A.* **85,** 5259-5263
16. Diaz, M. O., Rubin, C. M., Harden, A., Ziemin, S., A.Larson, R., LeBeau, M. M., and Rowley, J. D. (1990) *N. Eng. J. Med.* **322**(2), 77-82
17. Fountain, J. W., Karayiorgou, M., Ernstoff, M. S., Kirkwood, J. M., Vlock, D. R., Titus-Ernstoff, L., Bouchard, B., Vijatasaradhi, S., Houghton, A. N., Lahti, J., Kidd, V. J., Housman, D. E., and Dracopoli, N. C. (1992) *Proc. Natl. Acad. Sci. U S A.* **89,** 10557-10561
18. Olopade, O. I., Jenkins, R. B., Ransom, D. T., Malik, K., Pomykala, H., Nobori, T., Cowan, J. M., Rowley, J. D., and Diaz, M. O. (1992) *Cancer Res.* **52,** 2523-2529
19. Olopade, O. I., Buchhagen, D. L., Malik, K., Sherman, J., Nobori, T., Bader, S., Nau, M. M., Gazdar, A. F., Minna, J. D., and Diaz, M. O. (1993) *Cancer Res.* **53,** 2410-2415
20. Cheng, J. Q., Jhanwar, S. C., Lu, Y. Y., and Testa, J. R. (1993) *Cancer Res.* **53,** 4761-4763
21. James, C. D., He, J., Collins, V. P., Allalunis-Turner, M. H., and III, R. S. D. (1993) *Cancer Res.* **53,** 3674-3676
22. Merlo, A., Gabrielson, E., Askin, F., and Sidransky, D. (1994) *Cancer Res.* **54,** 640-642

23. Riet, P. v. d., Nawroz, H., Hruban, R. H., Corio, R., Tokino, K., Koch, W., and Sidransky, D. (1994) *Cancer Res.* **54**, 1156-1158
24. Kamb, A., Gruis, N. A., Weaver-Feldhaus, J., Liu, Q., Harshman, K., Tavtigian, S. V., Stockert, E., Day, R. S., 3rd, Johnson, B. E., and Skolnick, M. H. (1994) *Science* **264**(5157), 436-40
25. Serrano, M., Hannon, G. J., and Beach, D. (1993) *Nature* **366**(6456), 704-7
26. Nobori, T., Miura, K., Wu, D. J., Lois, A., Takabayashi, K., and Carson, D. A. (1994) *Nature* **368**(6473), 753-6
27. Flores, J. F., Walker, G. J., Glendening, J. M., Haluska, F. G., Castresana, J. S., Rubio, M. P., Pastorfide, G. C., Boyer, L. A., Kao, W. H., Bulyk, M. L., Barnhill, R. L., Hayward, N. K., Housman, D. E., and Fountain, J. W. (1996) *Cancer Res.* **56**(21), 5023-32
28. Stranks, G., Height, S. E., Mitchell, P., Jadayel, D., Yuille, M. A., De Lord, C., Clutterbuck, R. D., Treleaven, J. G., Powles, R. L., Nacheva, E., and et al. (1995) *Blood* **85**(4), 893-901
29. Koduru, P. R., Zariwala, M., Soni, M., Gong, J. Z., Xiong, Y., and Broome, J. D. (1995) *Blood* **86**(8), 2900-5
30. Southgate, J., Proffitt, J., Roberts, P., Smith, B., and Selby, P. (1995) *Br. J. Cancer* **72**(5), 1214-8
31. Williamson, M. P., Elder, P. A., Shaw, M. E., Devlin, J., and Knowles, M. A. (1995) *Hum. Mol. Genet.* **4**(9), 1569-77
32. Geradts, J., and Wilson, P. A. (1996) *Am. J. Pathol.* **149**(1), 15-20
33. Herman, J. G., Civin, C. I., Issa, J. P., Collector, M. I., Sharkis, S. J., and Baylin, S. B. (1997) *Cancer Res.* **57**(5), 837-41
34. Nakamaki, T., Bartram, C., Seriu, T., Kahan, J., Fukuchi, K., Tsuruoka, N., Janssen, J. W., Miller, C. W., and Koeffler, H. P. (1997) *Leuk. Res.* **21**(3), 235-40
35. Farrell, W. E., Simpson, D. J., Bicknell, J. E., Talbot, A. J., Bates, A. S., and Clayton, R. N. (1997) *Cancer Res.* **57**(13), 2703-9
36. Gonzalez-Zulueta, M., Bender, C. M., Yang, A. S., Nguyen, T., Beart, R. W., Van Tornout, J. M., and Jones, P. A. (1995) *Cancer Res.* **55**(20), 4531-5
37. Herman, J. G., Jen, J., Merlo, A., and Baylin, S. B. (1996) *Cancer Res.* **56**(4), 722-7
38. Batova, A., Diccianni, M. B., Yu, J. C., Nobori, T., Link, M. P., Pullen, J., and Yu, A. L. (1997) *Cancer Res.* **57**(5), 832-6
39. Ng, M. H., Chung, Y. F., Lo, K. W., Wickham, N. W., Lee, J. C., and Huang, D. P. (1997) *Blood* **89**(7), 2500-6
40. Malumbres, M., Perez de Castro, I., Santos, J., Melendez, B., Mangues, R., Serrano, M., Pellicer, A., and Fernandez-Piqueras, J. (1997) *Oncogene* **14**(11), 1361-70
41. Herman, J. G., Graff, J. R., Myohanen, S., Nelkin, B. D., and Baylin, S. B. (1996) *Proc. Natl. Acad. Sci. U S A* **93**(18), 9821-6
42. Quelle, D. E., Zindy, F., Ashmun, R. A., and Sherr, C. J. (1995) *Cell* **83**(6), 993-1000
43. Stone, S., Jiang, P., Dayananth, P., Tavtigian, S. V., Katcher, H., Parry, D., Peters, G., and Kamb, A. (1995) *Cancer Res.* **55**, 2988-2994

44. Mao, L., Merlo, A., Bedi, G., Shapiro, G. I., Edwards, C. D., Rollins, B. J., and Sidransky, D. (1995) *Cancer Res.* **55**, 2995-2997
45. Yunis, J. J., Bloomfield, C. D., and Ensrud, K. (1981) *N. Eng. J. Med.* **305**(3), 135-9
46. Bennett, J. M., Catovsky, D., Daniel, M.-T., Flandrin, G., Galton, D. A. G., Gralnick, H. R., and Sultan, C. (1976) *Br. J. Haematol.* **33**, 451-458
47. Kaneko, Y., Maseki, N., Takasaki, N., Sakurai, M., Hayashi, Y., Nakazawa, S., Mori, T., Sakurai, M., Takeda, T., and al, T. S. e. (1986) *Blood* **67**, 484-491
48. Pui, C. H., Raimondi, S. C., Murphy, S. B., Ribeiro, R. C., Kalwinsky, D. K., Dahl, G. V., Crist, W. M., and Williams, D. L. (1987) *Blood* **69**(5), 1289-93
49. Super, H. J., McCabe, N. R., Thirman, M. J., Larson, R. A., Le Beau, M. M., Pedersen-Bjergaard, J., Philip, P., Diaz, M. O., and Rowley, J. D. (1993) *Blood* **82**(12), 3705-11
50. Pedersen-Bjergaard, J., and Philip, P. (1991) *Blood* **78**(4), 1147-8
51. Rowley, J. D., Diaz, M. O., Espinosa III, R., Patel, Y. D., van Melle, E., Ziemen, S., Taillon-Miller, P., Lichter, P., Evans, G. A., Kersey, J. H., Ward, D. D., Domer, P. H., and LeBeau, M. M. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 9358-9362
52. Cimino, G., Moir, D. T., Canaani, O., Williams, K., Crist, W. M., Katzav, S., Cannizzaro, L., Lange, B., Nowell, P. C., Croce, C. M., and Canaani, E. (1991) *Cancer Res.* **51**(24), 6712-4
53. Ziemin-van der Poel, S., McCabe, N. R., Gill, H. J., Espinosa, R., III, Patel, Y., Harden, A., Rubinelli, P., Smith, S. D., LeBeau, M. M., Rowley, J. D., and et al. (1991) *Proc. Natl. Acad. Sci. U S A* **88**(23), 10735-9
54. Tkachuk, D. C., Kohler, S., and Cleary, M. L. (1992) *Cell* **71**(4), 691-700
55. Yamamoto, K., Seto, M., Komatsu, H., Iida, S., Akao, Y., Kojima, S., Kodera, Y., Nakazawa, S., Ariyoshi, Y., Takahashi, T., and et al. (1993) *Oncogene* **8**(10), 2617-25
56. McCabe, N. R., Burnett, R. C., Gill, H. J., Thirman, M. J., Mbangkollo, D., Kipiniak, M., van Melle, E., Ziemin-van der Poel, S., Rowley, J. D., and Diaz, M. O. (1992) *Proc. Natl. Acad. Sci. U S A* **89**(24), 11794-8
57. Iida, S., Seto, M., Yamamoto, K., Komatsu, H., Tojo, A., Asano, S., Kamada, N., Ariyoshi, Y., Takahashi, T., and Ueda, R. (1993) *Oncogene* **8**(11), 3085-92
58. Broeker, P. L., Super, H. G., Thirman, M. J., Pomykala, H., Yonebayashi, Y., Tanabe, S., Zeleznik-Le, N., and Rowley, J. D. (1996) *Blood* **87**(5), 1912-22
59. Biondi, A., Rambaldi, A., Rossi, V., Elia, L., Caslini, C., Basso, G., Battista, R., Barbui, T., Mandelli, F., Masera, G., and et al. (1993) *Blood* **82**(10), 2943-7
60. Yamamoto, K., Seto, M., Iida, S., Komatsu, H., Kamada, N., Kojima, S., Kodera, Y., Nakazawa, S., Saito, H., Takahashi, T., and et al. (1994) *Blood* **83**(10), 2912-21
61. Super, H. J., Martinez-Climent, J., and Rowley, J. D. (1995) *Blood* **85**(3), 855-6
62. Marschalek, R., Greil, J., Lochner, K., Nilson, I., Siegler, G., Zweckbronner, I., Beck, J. D., and Fey, G. H. (1995) *Br. J. Haematol.* **90**(2), 308-20
63. Nam, D. K., Honoki, K., Yu, M., and Yunis, J. J. (1996) *Gene* **178**(1-2), 169-75

64. Ida, K., Taki, T., Bessho, F., Kobayashi, M., Taira, F., Hanada, R., Yamamoto, K., Okimoto, Y., Seto, M., Ueda, R., and Hayashi, Y. (1997) *Med. Pediatr. Oncol.* **28**(5), 325-32
65. Negrini, M., Cuneo, A., Nakamura, T., Baffa, R., Sabbioni, S., Alder, H., Castoldi, G., and Croce, C. M. (1995) *Cancer Genet. Cytogenet.* **83**(1), 65-70
66. Gu, Y., Alder, H., Nakamura, T., Schichman, S. A., Prasad, R., Canaani, O., Saito, H., Croce, C. M., and Canaani, E. (1994) *Cancer Res.* **54**(9), 2326-30
67. Spitzner, J. R., and Muller, M. T. (1988) *Nucleic Acids Res.* **16**(12), 5533-56
68. Ebert, S. N., Shtrom, S. S., and Muller, M. T. (1990) *J. Virol.* **64**(9), 4059-66
69. Laemmli, U. K., Kas, E., Poljak, L., and Adachi, Y. (1992) *Curr. Op. Genet. Dev.* **2**, 275-285
70. Joh, T., Kagami, Y., Yamamoto, K., Segawa, T., Takizawa, J., Takahashi, T., Ueda, R., and Seto, M. (1996) *Oncogene* **13**(9), 1945-53
71. Lavau, C., Szilvassy, S. J., Slany, R., and Cleary, M. L. (1997) *EMBO J.* **16**(14), 4226-4237
72. Corral, J., Lavenir, I., Impey, H., Warren, A. J., Forster, A., Larson, T. A., Bell, S., McKenzie, A. N., King, G., and Rabbitts, T. H. (1996) *Cell* **85**(6), 853-61
73. Gu, Y., Cimino, G., Alder, H., Nakamura, T., Prasad, R., Canaani, O., Moir, D. T., Jones, C., Nowell, P. C., Croce, C. M., and et al. (1992) *Proc. Natl. Acad. Sci. U S A* **89**(21), 10464-8
74. Negrini, M., Felix, C. A., Martin, C., Lange, B. J., Nakamura, T., Canaani, E., and Croce, C. M. (1993) *Cancer Res.* **53**(19), 4489-92
75. Desiderio, S. V., Yancopoulos, G. D., Paskind, M., Thomas, E., Boss, M. A., Landau, N., Alt, F. W., and Baltimore, D. (1984) *Nature* **311**(5988), 752-755
76. Hiom, b., and Gellert, M. (1997) *Cell* **88**(1), 65-72
77. Steen, S. B., Gomelsky, L., Speidel, S. L., and Roth, D. B. (1997) *EMBO J.* **16**(10), 2656-264
78. Ingham, P. W. (1983) *Nature* **306**, 591-593
79. Reeves, R., and Nissen, M. S. (1990) *J. Biol. Chem.* **265**(15), 8573-8582
80. Zeleznik-Le, N. J., Harden, A. M., and Rowley, J. D. (1994) *Proc. Natl. Acad. Sci. U S A* **91**(22), 10610-4
81. Le Coniat, M., Romana, S. P., Bernard, O., Flexor, M. A., Hillion, J., Cherif, D., and Berger, R. (1993) *C. R. Acad. Sci. III* **316**(7), 692-7
82. Harris, B., Davis, E., Beau, M. L., Bitter, M., Kaminer, L., Morgan, E., and Rowley, J. (1988) *Cancer Genet. Cytogenet.* **30**, 171-175
83. Nakamura, T., Alder, H., Gu, Y., Prasad, R., Canaani, O., Kamada, N., Gale, R. P., Lange, B., Crist, W. M., Nowell, P. C., and et al. (1993) *Proc. Natl. Acad. Sci. U S A* **90**(10), 4631-5
84. Rubnitz, J. E., Morrissey, J., Savage, P. A., and Cleary, M. L. (1994) *Blood* **84**(6), 1747-52
85. Verma, R. S., da Costa, M., Gogineni, S. K., and Palermo, G. (1993) *Am. J. Hematol.* **42**(4), 398-9
86. Secker-Walker, L. M., Mehta, A., and Bain, B. (1995) *Br. J. Haematol.* **91**(2), 490-501

87. Morishita, K., Parganas, E., William, C. L., Whittaker, M. H., Drabkin, H., Oval, J., Taetle, R., Valentine, M. B., and Ihle, J. N. (1992) *Proc. Natl. Acad. Sci. U S A* **89**(9), 3937-41
88. Akahoshi, M., Oshimi, K., Mizoguchi, H., Okada, M., Enomoto, Y., and Watanabe, Y. (1987) *Cancer* **60**(11), 2654-61
89. Bitter, M. A., Neilly, M. E., Le Beau, M. M., Pearson, M. G., and Rowley, J. D. (1985) *Blood* **66**(6), 1362-70
90. Bernstein, R., Bagg, A., Pinto, M., Lewis, D., and Mendelow, B. (1986) *Blood* **68**(3), 652-7
91. Rubin, C. M., Larson, R. A., Bitter, M. A., Carrino, J. J., Le Beau, M. M., Diaz, M. O., and Rowley, J. D. (1987) *Blood* **70**(5), 1338-42
92. Horsman, D. E., Gascoyne, R. D., and Barnett, M. J. (1995) *Leuk. Lymphoma* **16**(5-6), 369-77
93. Miyazaki, H., Ino, T., Sobue, R., Kojima, H., Wakita, M., Nomura, T., Tuzuki, M., and Hirano, M. (1994) *Cancer Genet. Cytogenet.* **74**(2), 84-6
94. Pedersen-Bjergaard, J., Pedersen, M., Roulston, D., and Philip, P. (1995) *Blood* **86**(9), 3542-52
95. Nucifora, G., Begy, C. R., Kobayashi, H., Roulston, D., Claxton, D., Pedersen-Bjergaard, J., Parganas, E., Ihle, J. N., and Rowley, J. D. (1994) *Proc. Natl. Acad. Sci. U S A* **91**(9), 4004-8
96. Fears, S., Mathieu, C., Zeleznik-Le, N., Huang, S., Rowley, J. D., and Nucifora, G. (1996) *Proc. Natl. Acad. Sci. U S A* **93**(4), 1642-7
97. Soderholm, J., Kobayashi, H., Mathieu, C., Rowley, J. D., and Nucifora, G. (1997) *Leukemia* **11**, 352-358
98. Morishita, K., Parganas, E., Douglass, E. C., and Ihle, J. N. (1990) *Oncogene* **5**(7), 963-971
99. Buyse, I. M., Shao, G., and Huang, S. (1995) *Proc. Natl. Acad. Sci. U S A* **92**, 4467-4471
100. Matsugi, T., Morishita, K., and Ihle, J. N. (1990) *Mol. Cell. Biol.* **10**(3), 1259-1264
101. Morishita, K., Parker, D. S., Mucenski, M. L., Jenkins, N. A., Copeland, N. G., and Ihle, J. N. (1988) *Cell* **54**(6), 831-840
102. Perkins, A. S., Fishel, R., Jenkins, N. A., and Copeland, N. G. (1991) *Mol. Cell. Biol.* **11**(5), 2665-2674
103. Delwel, R., Funabiki, T., Kreider, B. L., Morishita, K., and Ihle, J. N. (1993) *Mol. Cell. Biol.* **13**(7), 4291-4300
104. Funabiki, T., Kreider, B. L., and Ihle, J. N. (1994) *Oncogene* **9**(6), 1575-1581
105. Morishita, K., Suzukawa, K., Taki, T., Ihle, J. N., and Yokata, J. (1995) *Oncogene* **10**, 1961-1967
106. Kreider, B. L., Orkin, S. H., and Ihle, J. N. (1993) *Proc. Natl. Acad. Sci. U S A* **90**, 6454-6458
107. Morishita, K., Parganas, E., Matsugi, T., and Ihle, J. N. (1992) *Mol. Cell. Biol.* **12**(1), 183-189
108. Romeo, P.-H., Prandini, M.-H., Joulin, V., Mignotte, V., Prenant, M., Vainchenker, W., Marguerie, G., and Uzan, G. (1990) *Nature* **344**(6265), 447-449

109. Suzukawa, K., Parganas, E., Gajjar, A., Abe, T., Takahashi, S., Tani, K., Asano, S., Asou, H., Kamada, N., Yodota, J., Morishita, K., and Ihle, J. N. (1994) *Blood* **84**(8), 2681-2688
110. Sacchi, N., Nisson, P. E., Watkins, P. C., Faustinella, F., Wijsman, J., and Hagemeijer, A. (1994) *Genes Chromo. Cancer* **11**(4), 226-36
111. Johansson, B., Fioretos, T., Garwicz, S., Heim, S., and Mitelman, F. (1996) *Br. J. Haematol.* **92**(2), 429-31
112. Nucifora, G., Begy, C. R., Erickson, P., Drabkin, H. A., and Rowley, J. D. (1993) *Proc. Natl. Acad. Sci. U S A* **90**(16), 7784-8
113. Speck, N. A., and Stacy, T. (1995) *Critical Reviews in Eukaryotic Gene Expression* **5**(3&4), 337-364
114. Nusslein-Volhard, C., and Wieschaus, E. (1980) *Nature* **287**(5785), 795-801
115. Mitani, K., Ogawa, S., Tanaka, T., Miyoshi, H., Kurokawa, M., Mano, H., Yazaki, Y., Ohki, M., and Hirai, H. (1994) *EMBO J.* **13**(3), 504-510
116. Kurokawa, M., Ogawa, S., Tanaka, T., Mitani, K., Yazaki, Y., Witte, O. N., and Hirai, H. (1995) *Oncogene* **11**(5), 833-840
117. Zent, C. S., Mathieu, C., Claxton, D. F., Zhang, D. E., Tenen, D. G., Rowley, J. D., and Nucifora, G. (1996) *Proc. Natl. Acad. Sci. U S A* **93**(3), 1044-8
118. Peeters, P., Wlodarska, I., Baens, M., Criel, A., Selleslag, D., Hagemeijer, A., Van den Berghe, H., and Marynen, P. (1997) *Cancer Res.* **57**(4), 564-9
119. Raynaud, S. D., Baens, M., Grosgeorge, J., Rodgers, K., Reid, C. D., Dainton, M., Dyer, M., Fuzibet, J. G., Gratecos, N., Taillan, B., Ayraud, N., and Marynen, P. (1996) *Blood* **88**(2), 682-9
120. Jousset, C., Carron, C., Boureux, A., Quang, C. T., Oury, C., Dusanter-Fourt, I., Charon, M., Levin, J., Bernard, O., and Ghysdael, J. (1997) *EMBO J.* **16**(1), 69-82
121. Kwiatkowski, D. J., and Short, M. P. (1994) *Arch. Dermatol.* **130**(3), 348-54
122. Fryer, A. E., Connor, J. M., Povey, S., Yates, J. R. W., Chalmers, A., Fraser, I., Yates, A. D., and Osborne, J. P. (1987) *Lancet* **1**, 659-661
123. Haines, J. L., Short, M. P., Kwiatkowski, D. J., Jewell, A., Andermann, E., Bejjani, B., Yang, C. H., Gusella, J. F., and Amos, J. A. (1991) *Am. J. Hum. Genet.* **49**(4), 764-72
124. Northrup, H., Kwiatkowski, D. J., Roach, E. S., Dobyns, W. B., Lewis, R. A., Herman, G. E., Jr., E. R., Daiger, S. P., and Blanton, S. H. (1992) *Am. J. Hum. Genet.* **51**, 709-720
125. Sepp, T., Yates, J. R., and Green, A. J. (1996) *J. Med. Genet.* **33**(11), 962-4
126. Carbonara, C., Longa, L., Grosso, E., Borrone, C., Garre, M. G., Brisigotti, M., and Migone, N. (1994) *Hum. Mol. Genet.* **3**(10), 1829-32
127. Green, A. J., Johnson, P. H., and Yates, J. R. (1994) *Hum. Mol. Genet.* **3**(10), 1833-4
128. Green, A. J., Sepp, T., and Yates, J. R. (1996) *Hum. Genet.* **97**(2), 240-3
129. Consortium, T. E. C. T. S. (1993) *Cell* **75**(7), 1305-1313
130. Zhou, C. Y., Wu, K. Y., Leversha, M. A., Furlong, R. A., Ferguson-Smith, M. A., and Affara, N. A. (1995) *Genomics* **25**(1), 304-8
131. van Slegtenhorst, M., de Hoogt, R., Hermans, C., Nellist, M., Janssen, B., Verhoef, S., Lindhout, D., van den Ouweland, A., Halley, D., Young, J., Burley, M., Jeremiah, S., Woodward, K., Nahmias, J., Fox, M., Ekong, R., Osborne, J.,

Wolfe, J., Povey, S., Snell, R. G., Cheadle, J. P., Jones, A. C., Tachataki, M., Ravine, D., Kwiatkowski, D. J., and et al. (1997) *Science* **277**(5327), 805-8
132. Kwiatkowski, D. J. (1992) Personal Communication.
133. Ottolenghi, S., and Giglioni, B. (1982) *Nature* **300**, 770-771
134. Calabretta, B., Robberson, D. L., Barrera-Saldana, H. A., Lambrou, T. P., and Saunders, G. F. (1982) *Nature* **296**(5854), 219-25
135. Lehrman, M. A., Schneider, W. J., Sudhof, T. C., Brown, M. S., Goldstein, J. L., and Russell, D. W. (1985) *Science* **227**(4683), 140-6
136. Lehrman, M. A., Goldstein, J. L., Russell, D. W., and Brown, M. S. (1987) *Cell* **48**, 827-835
137. Nicholls, R. D., Fischel-Ghodsian, N., and Higgs, D. R. (1987) *Cell* **49**, 369-378
138. Markert, M. L., Hutton, J. J., Wiginton, D. A., States, J. C., and Kaufman, R. E. (1988) *J. Clin. Invest.* **81**, 1323-1327
139. Schmid, C. W., and Jelinek, W. R. (1982) *Science* **216**(4550), 1065-70
140. Matera, A. G., Hellmann, U., and Schmid, C. W. (1990) *Mol. Cell. Biol.* **10**(10), 5424-32
141. Hwu, H. R., Roberts, J. W., Davidson, E. H., and Britten, R. J. (1986) *Proc. Natl. Acad. Sci. U S A* **83**(11), 3875-9
142. Ullu, E., and Tschudi, C. (1984) *Nature* **312**(5990), 171-2
143. Daniels, G. R., and Deininger, P. L. (1985) *Nucleic Acids Res.* **13**(24), 8939-54
144. Quentin, Y. (1992) *Nucleic Acids Res.* **20**(13), 3397-401
145. Britten, R. J., Baron, W. F., Stout, D. B., and Davidson, E. H. (1988) *Proc. Natl. Acad. Sci. U S A* **85**(13), 4770-4
146. Quentin, Y. (1992) *Nucleic Acids Res.* **20**(3), 487-93
147. Labuda, D., and Striker, G. (1989) *Nucleic Acids Res.* **17**(7), 2477-91
148. Shen, M. R., Batzer, M. A., and Deininger, P. L. (1991) *J. Mol. Evol.* **33**(4), 311-20
149. Quentin, Y. (1994) *Nucleic Acids Res.* **22**(12), 2222-7
150. Jurka, J., and Milosavljevic, A. (1991) *J. Mol. Evol.* **32**, 105-121
151. Jurka, J., and Smith, T. (1988) *Proc. Natl. Acad. Sci. U S A* **85**(13), 4775-8
152. Batzer, M. A., Deininger, P. L., Hellmann-Blumberg, U., Jurka, J., Labuda, D., Rubin, C. M., Schmid, C. W., Zietkiewicz, E., and Zuckerkandl, E. (1996) *J. Mol. Evol.* **42**(1), 3-6
153. Jurka, J. (1995) in *The Impact of Short Interspersed Elements (SINEs) on the Host Genome* (Maraia, R. J., ed), pp. 25-42, R. G. Landes Company, Austin
154. Kapitonov, V., and Jurka, J. (1996) *J. Mol. Evol.* **42**(1), 59-65
155. Geiduschek, E. P., and Tocchini-Valentini, G. P. (1988) (Richardson, C. C., Boyer, P. D., Dawid, I. B., and Meister, A., eds) Vol. 57, pp. 873-914, Annual Reviews Inc., Palo Alto, CA 94303-0897
156. Sinnett, D., Richer, C., Deragon, J. M., and Labuda, D. (1992) *J. Mol. Biol.* **226**(3), 689-706
157. Liu, W. M., Maraia, R. J., Rubin, C. M., and Schmid, C. W. (1994) *Nucleic Acids Res.* **22**(6), 1087-95
158. Hellmann-Blumberg, U., Hintz, M. F., Gatewood, J. M., and Schmid, C. W. (1993) *Mol. Cell. Biol.* **13**(8), 4523-30

159. Chu, W. M., Liu, W. M., and Schmid, C. W. (1995) *Nucleic Acids Res.* **23,** 1750-1757
160. Chang, D. Y., and Maraia, R. J. (1993) *J. Biol. Chem.* **268**(9), 6423-8
161. Chang, D. Y., Nelson, B., Bilyeu, T., Hsu, K., Darlington, G. J., and Maraia, R. J. (1994) *Mol. Cell. Biol.* **14**(6), 3949-59
162. Hsu, K., Chang, D. Y., and Maraia, R. J. (1995) *J. Biol. Chem.* **270**(17), 10179-86
163. Panning, B., and Smiley, J. R. (1995) in *The Impact of Short Interspersed Elements (SINEs) on the Host Genome* (Maraia, R. J., ed), pp. 143-161, R. G. Landes Company, Austin
164. Liu, W.-M., Chu, W.-M., Choudary, P. V., and Schmid, C. W. (1995) *Nucleic Acids Res.* **23**(10), 1758-1765
165. Maraia, R. J., and Sarrowa, J. (1995) in *The Impact of Short Interspersed Elements (SINEs) on the Host Genome* (Maraia, R. J., ed), pp. 163-196, R. G. Landes Company, Austin
166. Chang, D. Y., Hsu, K., and Maraia, R. J. (1996) *Nucleic Acids Res.* **24**(21), 4165-70
167. Sarrowa, J., Chang, D. Y., and Maraia, R. J. (1997) *Mol. Cell. Biol.* **17**(3), 1144-51
168. Chang, D. Y., Sasaki-Tozawa, N., Green, L. K., and Maraia, R. J. (1995) *Mol. Cell. Biol.* **15**(4), 2109-16
169. Bovia, F., Wolff, N., Ryser, S., and Strub, K. (1997) *Nucleic Acids Res.* **25**(2), 318-26
170. Britten, R. J. (1994) *Proc. Natl. Acad. Sci. U S A* **91**(13), 5992-6
171. Sinnett, D., Richer, C., Deragon, J. M., and Labuda, D. (1991) *J. Biol. Chem.* **266**(14), 8675-8
172. Vidal, F., and Cuzin, F. (1995) in *The Impact of Short Interspersed Elements (SINEs) on the Host Genome* (Maraia, R. J., ed), pp. 125-131, R. G. Landes Company, Austin
173. Wallace, M. R., Andersen, L. B., Saulino, A. M., Gregory, P. E., Glover, T. W., and Collins, F. S. (1991) *Nature* **353**(6347), 864-866
174. Ellis, N. A., Goodfellow, P. J., Pym, B., Smith, M., Palmer, M., Frischauf, A. M., and Goodfellow, P. N. (1989) *Nature* **337**(6202), 81-4
175. Zucman-Rossi, J., Batzer, M. A., Stoneking, M., Delattre, O., and Thomas, G. (1997) *Hum. Genet.* **99**(3), 357-63
176. Englander, E. W., Wolffe, A. P., and Howard, B. H. (1993) *J. Biol. Chem.* **268**(26), 19565-73
177. Lehrman, M. A., Russell, D. W., Goldstein, J. L., and Brown, M. S. (1986) *Proc. Natl. Acad. Sci. U S A* **83**(11), 3679-83
178. Kigawa, K., Kihara, K., Miyake, Y., Tajima, S., Funahashi, T., Yamamura, T., and Yamamoto, A. (1993) *J. Biochem. (Tokyo)* **113**(3), 372-6
179. Marshall, B., Isidro, G., and Boavida, M. G. (1996) *Gene* **174**(1), 175-9
180. Li, L., and Bray, P. F. (1993) *Am. J. Hum. Genet.* **53,** 140-149
181. Puget, N., Torchard, D., Serova-Sinilnikova, O. M., Lynch, H. T., Feunteun, J., Lenoir, G. M., and Mazoyer, S. (1997) *Cancer Res.* **57**(5), 828-31
182. Adams, J. W., Kaufman, R. E., Kretschmer, P. J., Harrison, M., and Nienhuis, A. W. (1980) *Nucleic Acids Res.* **8**(24), 6113-28

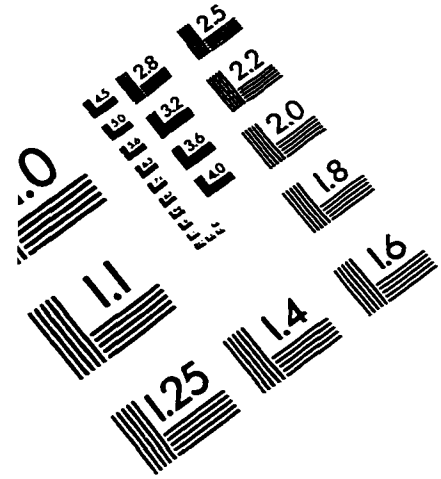
183. Hattori, M., Kuhara, S., Takenaka, O., and Sakaki, Y. (1986) *Nature* **321**(6070), 625-8
184. Smit, A. F., Toth, G., Riggs, A. D., and Jurka, J. (1995) *J. Mol. Biol.* **246**(3), 401-417
185. Jurka, J. (1990) *Nucleic Acids Res.* **18**(1), 137-41
186. Jurka, J., Zietkiewicz, E., and Labuda, D. (1995) *Nucleic Acids Res.* **23**(1), 170-5
187. Smit, A. F., and Riggs, A. D. (1995) *Nucleic Acids Res.* **23**(1), 98-102
188. Okada, N., and Ohshima, K. (1995) in *The Impact of Short Interspersed Elements (SINEs) on the Host Genome* (Maraia, R. J., ed), pp. 61-79, R. G. Landes Company, Austin
189. Paulson, K. E., Deka, N., Schmid, C. W., Misra, R., Schindler, C. W., Rush, M. G., Kadyk, L., and Leinwand, L. (1985) *Nature* **316**(6026), 359-61
190. Litman, R. M. (1968) *J. Biol. Chem.* **243**(23), 6222-6233
191. Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995) *Nucleic Acids Res.* **23**(23), 4878-84
192. Schug, J., and Overton, G. C. (1997) in *Technical Report CBIL-TR-1997-1001-v0.0, of the Computationai Biology and Informatics Laboratory*, School of Medicine, University of Pennsylvania, URL: http://agave.humgen.upenn.edu/tess/index.html
193. Solovyev, V., and Salamov, A. (1997) *Ismb* **5**, 294-302
194. Wingender, E., Dietze, P., Karas, H., and Knuppel, R. (1996) *Nucleic Acids Res.* **24**(1), 238-41
195. Reese, M. G. (1994), Ph.D. Dissertation, German Cancer Res. Center, Heidelberg
196. Reese, M. G., and Eeckman, F. H. (1995) in *The Seventh International Genome Sequencing and Analysis Conference,*, Hilton Head Island, South Carolina
197. Reese, M. G., Harris, N. L., and Eeckman, F. H. (1996) in *Proceedings of the Pacific Symposium on Biocomputing* (Klein, L. H. a. T., ed), Kona, Hawaii,
198. Prestridge, D. S. (1995) *J. Mol. Biol.* **249**(5), 923-32
199. Frech, K., Danescu-Mayer, J., and Werner, T. (1997) *J. Mol. Biol.* **270**(5), 674-87
200. Salamov, A. A., and Solovyev, V. V. (1997) *Comput. Appl. Biosci.* **13**(1), 23-8
201. Xu, Y., Mural, R., Shah, M., and Uberbacher, E. (1994) *Genet. Eng. (N Y)* **16**, 241-53
202. Burge, C., and Karlin, S. (1997) *J. Mol. Biol.* **268**(1), 78-94
203. Burge, C. (1997), Stanford University, Stanford, CA
204. Smits, A. , http://www.genome.washington.edu, Seattle WA
205. Abajian, C. (1994), University of Washington
Dept. of Molecular Biotechnology FJ-20
Fluke Hall, Mason Road
Seattle WA 98195
206. Roe, B., Crabtree, J., and Kahn, A. (eds) (1996) *DNA Isolation and Sequencing Essential Techniques*. Essential Techniques Series. Edited by Rickwood, D., John Wiley and Sons in association with Bios Scientific Publishers, Ltd., Oxford, UK
207. Roe, B. A. (ed) (1997) *Protocols for recombinant DNA isolation, cloning, and sequencing*, Norman, OK
208. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) *J. Mol. Biol.* **215**, 403-410

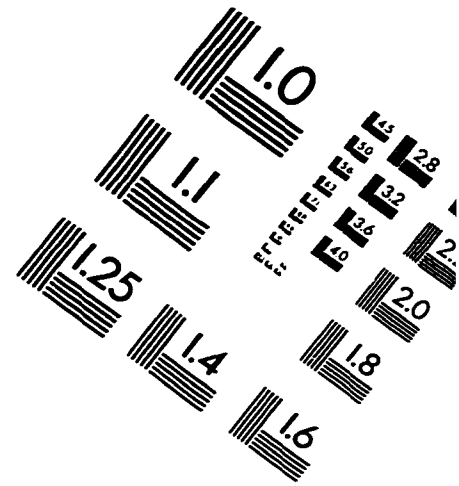
209. Dignam, J. D., Lebovitz, R. M., and Roeder, R. G. (1983) *Nucleic Acids Res.* **11**(5), 1475-1489
210. Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) *Molecular cloning: a laboratory manual*, 2 Ed., 3. 3 vols., Cold Springs Harbor Laboratory Press, New York City
211. Tabor, S., and Richardson, C. C. (1995) *Proc. Natl. Acad. Sci. U S A* **92**(14), 6339-43
212. Wadman, I. A., Osada, H., Grutz, G. G., Agulnick, A. D., Westphal, H., forster, A., and Rabbitts, T. H. (1997) *EMBO J.* **16**(11), 3145-3157
213. Aoki, K., Suzuki, K., Sugano, T., Tasaka, T., Nakahara, K., Kuge, O., Omori, A., and Kasai, M. (1995) *Nature Genet.* **10**, 167-174
214. Rudiger, N. S., Gregersen, N., and Kielland-Brandt, M. C. (1995) *Nucleic Acids Res.* **23**(2), 256-60
215. Gill-Super, H., and Zeleznik-Le, N. (1996) Personal Communication.
216. Super, H. G., Strissel, P. L., Sobulo, O. M., Burian, D., Reshmi, S. C., Roe, B., Zeleznik-Le, N. J., Diaz, M. O., and Rowley, J. D. (1997) *Genes Chromo. Cancer* **20**(2), 185-95
217. Atlas, M., Head, D., Behm, F., Schmidt, E., Gill-Super, H., Zeleznik-Le, N., Roe, B., Burian, D., Rowley, J., and Domer, P. (1997) Unpublished results.
218. Uberbacher, E. C., Xu, Y., and Mural, R. J. (1996) *Methods Enzymol* **266**, 259-81
219. Kamijo, T., Zindy, F., Roussel, M. F., Quelle, D. E., Downing, J. R., Ashmun, R. A., Grosveld, G., and Sherr, C. J. (1997) *Cell* **91**(5), 649-659
220. Arap, W., Knudsen, E., Sewell, D. A., Sidransky, D., Wang, J. Y., Huang, H. J., and Cavenee, W. K. (1997) *Oncogene* **15**(17), 2013-20
221. Zindy, F., Quelle, D. E., Roussel, M. F., and Sherr, C. J. (1997) *Oncogene* **15**(2), 203-11
222. Tapscott, S. J., Davis, R. L., Thayer, M. J., Cheng, P. F., Weintraub, H., and Lassar, A. B. (1988) *Science* **242**(4877), 405-11
223. Weintraub, H., Tapscott, S. J., Davis, R. L., Thayer, M. J., Adam, M. A., Lassar, A. B., and Miller, A. D. (1989) *Proc. Natl. Acad. Sci. U S A* **86**(14), 5434-8
224. Murre, C., McCaw, P. S., and Baltimore, D. (1989) *Cell* **56**(5), 777-83
225. Murre, C., Voronova, A., and Baltimore, D. (1991) *Mol. Cell. Biol.* **11**(2), 1156-60
226. Cordle, S. R., Henderson, E., Masuoka, H., Weil, P. A., and Stein, R. (1991) *Mol. Cell. Biol.* **11**(3), 1734-8
227. Schlissel, M., Voronova, A., and Baltimore, D. (1991) *Genes Dev.* **5**(8), 1367-76
228. Grilli, M., Chiu, J. J., and Lenardo, M. J. (1993) *Int. Rev. Cytol.* **143**, 1-62
229. Molnar, A., and Georgopoulos, K. (1994) *Mol. Cell. Biol.* **14**(12), 8292-303
230. Roitt, I., Brostoff, J., and Male, D. (1989) *Immunology*, 2 Ed., Gower Medical Publishing, London
231. Orkin, S. H. (1990) *Cell* **63**(4), 665-72
232. Pani, L., Overdier, D. G., Porcella, A., Qian, X., Lai, E., and Costa, R. H. (1992) *Mol. Cell. Biol.* **12**(9), 3723-32
233. Bohmann, D., Bos, T. J., Admon, A., Nishimura, T., Vogt, P. K., and Tjian, R. (1987) *Science* **238**(4832), 1386-92
234. Lee, W., Mitchell, P., and Tjian, R. (1987) *Cell* **49**(6), 741-52

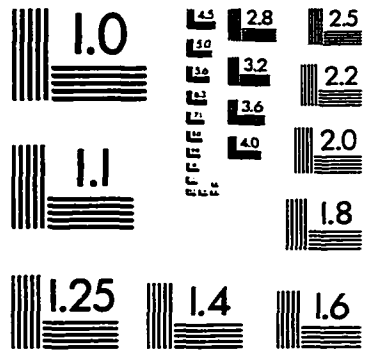
235. Angel, P., Imagawa, M., Chiu, R., Stein, B., Imbra, R. J., Rahmsdorf, H. J., Jonat, C., Herrlich, P., and Karin, M. (1987) *Cell* **49**(6), 729-39
236. Imagawa, M., Chiu, R., and Karin, M. (1987) *Cell* **51**(2), 251-60
237. Comb, M., Mermod, N., Hyman, S. E., Pearlberg, J., Ross, M. E., and Goodman, H. M. (1988) *EMBO J.* **7**(12), 3793-805
238. Lucero, M. A., Sanchez, D., Ochoa, A. R., Brunel, F., Cohen, G. N., Baralle, F. E., and Zakin, M. M. (1989) *Nucleic Acids Res.* **17**(6), 2283-300
239. Mitchell, P. J., Timmons, P. M., Hebert, J. M., Rigby, P. W., and Tjian, R. (1991) *Genes Dev.* **5**(1), 105-19
240. Jarman, A. P., Wood, W. G., Sharpe, J. A., Gourdon, G., Ayyub, H., and Higgs, D. R. (1991) *Mol. Cell. Biol.* **11**(9), 4679-89
241. Takimoto, M., Quinn, J. P., Farina, A. R., Staudt, L. M., and Levens, D. (1989) *J. Biol. Chem.* **264**(15), 8992-9
242. Mermod, N., Williams, T. J., and Tjian, R. (1988) *Nature* **332**(6164), 557-61
243. Chrivia, J. C., Kwok, R. P., Lamb, N., Hagiwara, M., Montminy, M. R., and Goodman, R. H. (1993) *Nature* **365**(6449), 855-9
244. Lin, Y. S., and Green, M. R. (1988) *Proc. Natl. Acad. Sci. U S A* **85**(10), 3396-400
245. Maguire, H. F., Hoeffler, J. P., and Siddiqui, A. (1991) *Science* **252**(5007), 842-4
246. Wagner, S., and Green, M. R. (1993) *Science* **262**(5132), 395-9
247. Ferreri, K., Gill, G., and Montminy, M. (1994) *Proc. Natl. Acad. Sci. U S A* **91**(4), 1210-3
248. Wang, J. Y., Nishiyama, K., Araki, K., Kitamura, D., and Watanabe, T. (1987) *Nucleic Acids Res.* **15**(24), 10105-16
249. Fletcher, C., Heintz, N., and Roeder, R. G. (1987) *Cell* **51**(5), 773-81
250. Currie, R. A., and Roeder, R. G. (1989) *Mol. Cell. Biol.* **9**(10), 4239-47
251. Sen, R., and Baltimore, D. (1986) *Cell* **46**(5), 705-16
252. He, X., Treacy, M. N., Simmons, D. M., Ingraham, H. A., Swanson, L. W., and Rosenfeld, M. G. (1989) *Nature* **340**(6228), 35-41
253. Schreiber, E., Himmelmann, A., Malipiero, U., Tobler, A., Stahel, R., and Fontana, A. (1992) *Cancer Res.* **52**(21), 6121-4
254. Pognonec, P., and Roeder, R. G. (1991) *Mol. Cell. Biol.* **11**(10), 5125-36
255. Potter, J. J., Cheneval, D., Dang, C. V., Resar, L. M., Mezey, E., and Yang, V. W. (1991) *J. Biol. Chem.* **266**(23), 15457-63
256. Sawadogo, M. (1988) *J. Biol. Chem.* **263**(24), 11994-2001
257. Blackwood, E. M., and Eisenman, R. N. (1991) *Science* **251**(4998), 1211-7
258. Dang, C. V., McGuire, M., Buckmire, M., and Lee, W. M. (1989) *Nature* **337**(6208), 664-6
259. Amati, B., Brooks, M. W., Levy, N., Littlewood, T. D., Evan, G. I., and Land, H. (1993) *Cell* **72**(2), 233-45
260. Dang, C. V., Barrett, J., Villa-Garcia, M., Resar, L. M., Kato, G. J., and Fearon, E. R. (1991) *Mol. Cell. Biol.* **11**(2), 954-62
261. Seth, A., Alvarez, E., Gupta, S., and Davis, R. J. (1991) *J. Biol. Chem.* **266**(35), 23521-4
262. Wen, Z., Zhong, Z., and Darnell, J. E., Jr. (1995) *Cell* **82**(2), 241-50
263. Pine, R., and Darnell, J. E., Jr. (1989) *Mol. Cell. Biol.* **9**(8), 3533-7

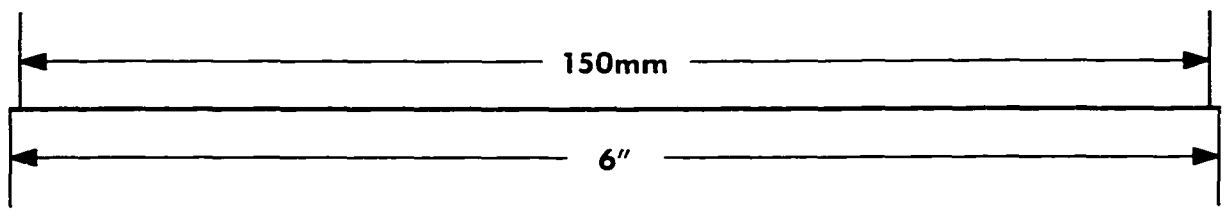
264. Khan, K. D., Shuai, K., Lindwall, G., Maher, S. E., Darnell, J. E., Jr., and Bothwell, A. L. (1993) *Proc. Natl. Acad. Sci. U S A* **90**(14), 6806-10
265. Sadowski, H. B., Shuai, K., Darnell, J. E., Jr., and Gilman, M. Z. (1993) *Science* **261**(5129), 1739-44
266. Pine, R., Decker, T., Kessler, D. S., Levy, D. E., and Darnell, J. E., Jr. (1990) *Mol. Cell. Biol.* **10**(6), 2448-57
267. Harada, H., Takahashi, E., Itoh, S., Harada, K., Hori, T. A., and Taniguchi, T. (1994) *Mol. Cell. Biol.* **14**(2), 1500-9
268. Goldberg, Y., Treier, M., Ghysdael, J., and Bohmann, D. (1994) *J. Biol. Chem.* **269**(24), 16566-73
269. Gegonne, A., Bosselut, R., Bailly, R. A., and Ghysdael, J. (1993) *EMBO J.* **12**(3), 1169-78
270. Wasylyk, B., Wasylyk, C., Flores, P., Begue, A., Leprince, D., and Stehelin, D. (1990) *Nature* **346**(6280), 191-3
271. Ghysdael, J., Gegonne, A., Pognonec, P., Dernis, D., Leprince, D., and Stehelin, D. (1986) *Proc. Natl. Acad. Sci. U S A* **83**(6), 1714-8
272. Overdier, D. G., Porcella, A., and Costa, R. H. (1994) *Mol. Cell. Biol.* **14**(4), 2755-66
273. Lo, K., Landau, N. R., and Smale, S. T. (1991) *Mol. Cell. Biol.* **11**(10), 5229-43
274. Zweidler-Mckay, P. A., Grimes, H. L., Flubacher, M. M., and Tsichlis, P. N. (1996) *Mol. Cell. Biol.* **16**(8), 4024-34
275. Opstelten, D. J., Vogels, R., Robert, B., Kalkhoven, E., Zwartkruis, F., de Laaf, L., Destree, O. H., Deschamps, J., Lawson, K. A., and Meijlink, F. (1991) *Mech. Dev.* **34**(1), 29-41
276. Kongsuwan, K., Webb, E., Housiaux, P., and Adams, J. M. (1988) *EMBO J.* **7**(7), 2131-8
277. Lints, T. J., Parsons, L. M., Hartley, L., Lyons, I., and Harvey, R. P. (1993) *Development* **119**(2), 419-31
278. Rosenfeld, P. J., and Kelly, T. J. (1986) *J. Biol. Chem.* **261**(3), 1398-408
279. Jones, K. A., Kadonaga, J. T., Rosenfeld, P. J., Kelly, T. J., and Tjian, R. (1987) *Cell* **48**(1), 79-89
280. Rossi, P., Karsenty, G., Roberts, A. B., Roche, N. S., Sporn, M. B., and de Crombrugghe, B. (1988) *Cell* **52**(3), 405-14
281. Paonessa, G., Gounari, F., Frank, R., and Cortese, R. (1988) *EMBO J.* **7**(10), 3115-23
282. Gustafson, T. A., and Kedes, L. (1989) *Mol. Cell. Biol.* **9**(8), 3269-83
283. Courtois, S. J., Lafontaine, D. A., Lemaigre, F. P., Durviaux, S. M., and Rousseau, G. G. (1990) *Nucleic Acids Res.* **18**(1), 57-64
284. Rao, V. N., Huebner, K., Isobe, M., ar-Rushdi, A., Croce, C. M., and Reddy, E. S. (1989) *Science* **244**(4900), 66-70
285. Rao, V. N., and Reddy, E. S. (1992) *Oncogene* **7**(1), 65-70
286. Hipskind, R. A., Rao, V. N., Mueller, C. G., Reddy, E. S., and Nordheim, A. (1991) *Nature* **354**(6354), 531-4
287. Hromas, R., Collins, S. J., Hickstein, D., Raskind, W., Deaven, L. L., P, O. H., Hagen, F. S., and Kaushansky, K. (1991) *J. Biol. Chem.* **266**(22), 14183-7
288. Webb, C. (1997), Personal Communication.

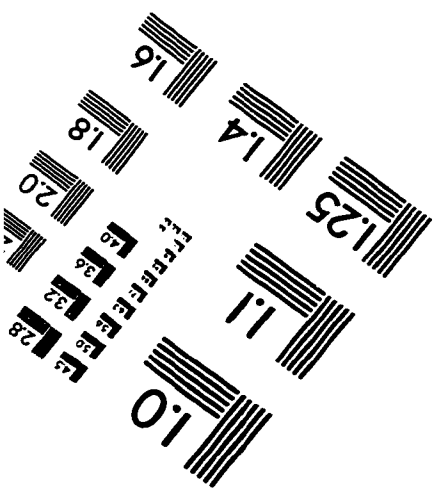
IMAGE EVALUATION
TEST TARGET (QA-3)

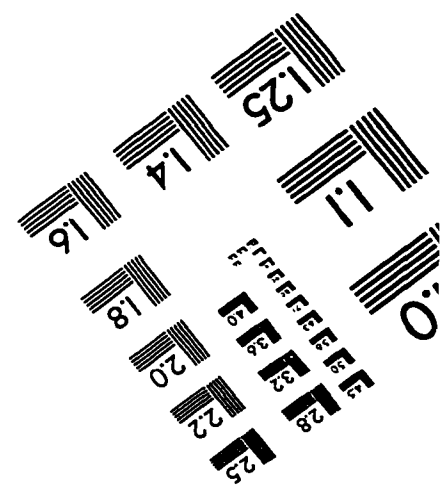












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989

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