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

SEQUENCING ANALYSIS OF REGIONS OF HUMAN CHROMOSOMES 11 AND 22: MENINGIOMA (22), CAT EYE SYNDROME (22), AND MULTIPLE ENDOCRINE NEOPLASIA, TYPE 1 (11)

A Dissertation

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

JUDY SUE CRABTREE

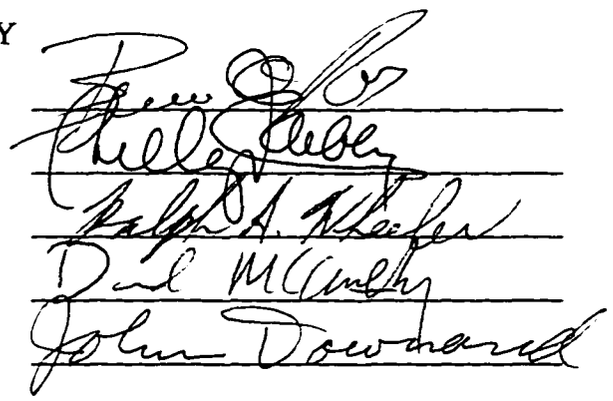
Norman, Oklahoma

1997

SEQUENCING ANALYSIS OF REGIONS OF HUMAN CHROMOSOMES 11 AND 22: MENINGIOMA (22), CAT EYE SYNDROME (22) AND MULTIPLE ENDOCRINE NEOPLASIA, TYPE 1 (11)

A Dissertation APPROVED FOR THE
DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

BY



Acknowledgements

I would like to express my sincere gratitude to my parents, Bob and Sue Lobsinger, for their unwavering support, guidance and love. Thanks for instilling in me the desire to learn, to set high standards, and to always give everything my best shot. I truly learned by your example. Mom, thanks for listening as we wandered the malls and Dad, thanks for everything you couldn't find the words to say. I would like to thank my brothers, Mike, John and Steve Lobsinger, and my cousin, Mark Lobsinger, for the intellectual banter and the not-so-intellectual hack and slay. I appreciate your support, and certainly the comic relief, more than you could ever imagine. I also wish to thank Al and Cecile Crabtree, and Cecil and Johnna Gray for the encouragement, support, catfish, and weekends of R & R. I would like to thank Cecile and my parents for critical reading of the manuscript.

A special thanks to my brother and sister-in-law, Steve and Darci Lobsinger, for their generous hospitality during the closing stages of this dissertation work, and for a friendship that will last a lifetime.

I would like to thank my major professor, Dr. Bruce Roe, for his guidance, support, inspiration, and friendship during the past ten years. Thanks for being my teacher, sounding board, therapist, role model and friend all rolled into one. I also extend my appreciation to the other members of my graduate advisory committee, Dr. David McCarthy, Dr. John Downard, Dr. Ralph Wheeler and Dr. Phil Klebba for their assistance, and the DOE and NHGRI for the financial support.

I would like to extend a big thank you to everyone in the Roe Lab, both past and present, for making it such a positive, friendly place to work and learn: Mueed Ahmad, Cathy Anadu, Dennis Burian, Linda Cantu, Feng Chen, Stephanie Chissoe, Lingzhi Chu, Sandy Clifton, Arlena Coulberson, Angela Dorman, Whitney Elkins, Jennifer Gray, Karen Hartman, Jennifer Hausner, Andrew Horning, Naiqing Hu, Emily Huang,

Axin Hua, Kala Iyer, Steve Kenton, Akbar Khan, Doris Kupfer, Hongshing Lai, Lisa Lane, Michele Lasley, Sharon Lewis, Shaoping Lin, Eda Malaj, Sean Meadows, Fares Najar, Nien Ma, Yichen Ma, Elaine Mardis, Thuan Nguyen, Huaqin Pan, Murli Rao, Adonis Reece, Qun Ren, Will Tankersley, Steve Toth, Yonathan Tilahun, Ha Vu, YingPing Wang, Zhili Wang, Heather Wright, Ziyun Yao, Xiling Yuan, Min Zhan, Guozhong Zhang, and Hua Zhu. I would like to especially thank Diana Willingham and Marsha Lobsinger for their efforts on the b18h3 project, the long nights and weekends are greatly appreciated! I would also like to thank my special giants, Ginger Coleman and Jim Martin for a great start in the right direction; Cheri Crabtree for always being there with a hug and a smile; Mary Ann Roberts for the continuous encouragement, inspiration and loads of fun; Kaylynn Hale for the humor, smiles and for always finding the silver lining; and my many friends, both old and new, for their endless support and encouragement along the way.

Last, but certainly not least, I would like to thank my husband, Chuck Crabtree, for his undying patience, eternal faith in my abilities, constant love, and continuous support during the last five years. Thanks for understanding the late nights and weekends, thanks for bringing me food when I didn't take the time to eat, and thanks for always being there with me...to celebrate in the good times and to console in the bad. I couldn't have done it without you!

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABI - Applied Biosystems Incorporated (a division of Perkin Elmer)

ALU - a human DNA repeat element with homology to 7SL RNA

BAC - bacterial artificial chromosome

BLAST - basic local alignment search tool

Bst - Bacillus stereothermophilus

CAP2 - computer assembly program, v. 2. Computer program for shotgun sequence assembly.

CCD - charge coupled device

cDNA - complementary deoxyribonucleic acid

CEC - Cat Eye chromosome

CES - Cat Eye Syndrome

CONSED - consensus editor. Graphical user interface for use with databases generated by the Phred/phrap assembly program.

Crossmatch - computer sequence comparison program

DGCR - DiGeorge Sydrome critical region

DMSO - dimethyl sulfoxide

Dotter - two dimensional dot plot sequence comparison program

bp - base pair

dc^{7}GTP - deoxy-7-deaza guanosine triphosphate

dITP - deoxyinosine triphosphate

dNTP - deoxynucleotide

ddNTP - dideoxynucleotide

DNA - deoxyribonucleic acid

E. coli - Escherichia coli

EDTA - disodium ethylenediamine tetraacetate

EMBL - European Molecular Biology Laboratory

EST - expressed sequence tag

EWS - Ewing Sarcoma

FAKII - fragment assembly kernel, v.II. Computer program for shotgun sequence assembly.

FAU - Finkel-Biskis-Reilly murine sarcoma virus [FBR-MuSV] associated ubiquitously expressed gene.

GCG - a software package from the Genetic Computer Group in Madison, WI.

HST1 - human stomach transforming factor type 1

HSTF1 - heparin secretory transforming factor type 1 (formerly HST1)

Hu EST - human EST

IGLC - immunoglobin light chain constant region

IFLV - immunoglobin light chain variable region

INT2 - human homologue of the mouse mammary tumor virus integration site 2

IPTG - isopropyl-β-D-thiogalactopyranoside

kb - kilobase

LINE (L1) - long interspersed repeat element. A human DNA repeat element

MB - myoglobin locus

Mb - megabase

MDR - meningioma deletion region

MEN-1 - multiple endocrine neoplasia, type I

MER - medium element of repeat. A human DNA repeat element

M EST - mouse EST

Met - methionine transcription start codon

MIR - mammalian-wide interspersed repeats

MLT - human repeat element

MN-1 - gene found in the meningioma deletion region 22q11

MRC - Medical Research Council

mRNA - messenger ribonucleic acid

MST - human repeat element

NCBI - National Center for Biotechnology Information

NF2 - bilater acoustic neurofibromatosis, type 2

NHGRI - National Human Genome Research Institute

NIH - National Institutes of Health

PCR - polymerase chain reaction

Phred/Phrap - Phil's (Phil Green) read editor/Phil's read assembly program. Computer programs for shotgun sequence basecalling/shotgun sequence assembly.

PMT - photomultiplier tube

PP1a - protein phosphatase 1 catalytic subunit

PRAD1 - parathyroid adenomatosis, type 1

PYGM - human muscle glycogen phosphorylase

R EST - rat EST

RFLP - restriction fragment length polymorphism

SDS - sodium dodecyl sulfate

STS - sequence tagged site

SV40 - simian virus 40

SVA - human repeat element

Taq - Thermus aquaticus

TAR1 - human repeat element

TED - trace editor. An editor for DNA sequencing trace data.

TEMED - N, N, N', N'-tetramethylethylenediamine

THE - transposon like human repeat element

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

XGAP - X windows based genome assembly program

Xgrail - X windows based gene recognition and analysis internet link. Used for sequence feature prediction.

YENB - cell growth media containing yeast extract and nutrient broth

ZFM1 - zinc finger motif

ABSTRACT

The complete nucleotide sequence of a meningioma deletion region (MDR) from human chromosome 22q11, a segment of the Cat Eye Syndrome (CES) region from human chromosome 22, and portions of the Multiple Endocrine Neoplasia, type 1 (MEN-1) from human chromosome 11 has been determined by sequencing an overlapping set of cosmids or BAC genomic clones previously mapped to these regions. The sequencing strategy entailed isolation of the genomic, target clones by a modified diatomaceous earth-based procedure, physical shearing by nebulization, and subcloning the size-selected fragments into pUC-based vectors. The sequencing templates were isolated using a modified alkaline lysis procedure that was automated on a Beckman Biomek 2000 robotic workstation as part of this research. Each template was incubated in cycle sequencing reactions with either forward or reverse universal sequencing primer, *Taq* DNA polymerase and dye labeled terminators. After automated sample electrophoresis, detection, data collection and base calling on ABI 373/377 DNA sequencers, the resulting sequence data was edited, assembled and proofread using one of the several editing and assembly programs (TED, XGAP, fakII, cap2, phred/phrap, consed) on a Sun SPARCstation. Each final contiguous sequence was analyzed for potential coding regions, repeated elements and database homology using the XGrail and powblast programs.

Additionally, sequence analysis in the CES and MEN-1 regions served as pilot projects for a novel sequencing strategy, sequence scanning. Results presented here demonstrate how this approach is used to localize the majority of coding regions after only the shotgun phase of DNA sequencing. This strategy is particularly useful when searching for genes related to disease phenotypes.

Analysis of these regions reveals several putative genes in the MEN-1 region, including the PYGM locus, a zinc finger motif gene, a GC kinase gene, the DNA

polymerase a gene and nine putative genes, including the gene for MEN-1. Sequence analysis of the CES region indicated a repeat density of 80% and no putative or known genes present. The MDR region contains the MN-1 gene, as expected, and the genomic organization of this gene is reported. The entire sequence presented here represents approximately 1% of the 57 million bases of human chromosome 22.

Chapter I

INTRODUCTION

I. Meningioma

A. Historical overview of nomenclature

Harvey Cushing, in his 1922 Cavendish Lecture[1], coined the term meningioma as an all-encompassing category for benign tumors of the central nervous system. These tumors were studied for many decades, but were called a variety of names reflecting the gross appearance, location or histological origin of the tumor. Although Bright[2] first recognized the similarity between meningioma cells and arachnoid cells, many pathologists continued to classify each variant meningioma as a distinct tumor class. However, Cushing circumvented this histology-based classification scheme and instead categorized all tumors of the meninges as meningiomas (see figure 1).

Cushing advocated the concept of one primary meningioma, composed of a basic cell, and variants of this pure form[3]. Electron microscopy since has revealed several important structural characteristics of the basic tumor cells which allowed detection of the same basic cell in a number of variants of meningiomas. However, the controversy regarding the histogenesis of certain primary tumors involving the meninges continued, because it was argued that primary tumors involving the meninges as a variant of meningiomas or as a completely unrelated neoplasm developing in and around the meninges might lack any histological relationship to meningiomas.

The search for adequate nomenclature for these tumors of the central nervous system caused great difficulty because meningiomas arise from the central layer of the meninges, i.e., the arachnoid or more specifically, the outermost cell layers of the arachnoid, the arachnoidal cap cells. These external covering cells are unique in their structure and function in that they both cover (or line) a surface, and also act as

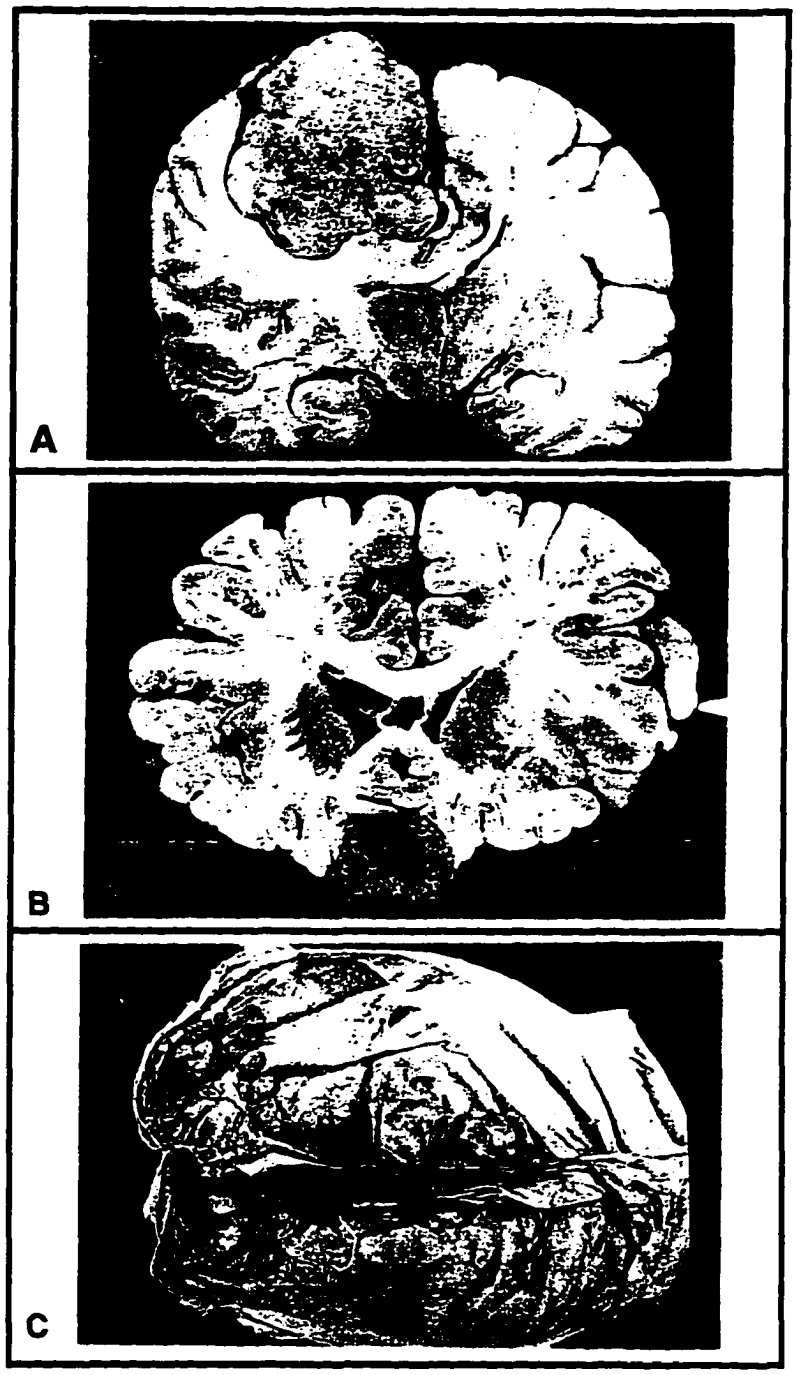


Figure 1: Meningioma present in a) convexity of the brain, b) base of brain and c) small multiple meningiomas attached to the meninges[96].

vascularized connective tissue. Similar to the synovial cells which line the joints, arachnoidal cap cells are non epithelial cells performing functions of the epithelium. Yet, in contrast, the arachnoidal cap cells do not form the inner lining of a cavity such as the mesothelial cells of the serous membranes, but rather seal it off from the outside, i.e., if one considers the subarachnoid space which contains the cerebrospinal fluid. Thus, the arachnoidal cells face the dura mater (see figure 2) and in addition to sealing off the subarachnoid space, they also served as conduits for the cerebrospinal fluid into the venous circulation of the sinuses. These conduits occur in specific regions called the arachnoid villa. While a great portion of the tumors present in the brain have an attachment to the dura, only those which specifically originated from the arachnoid villa are designated as meningiomas[4].

Another distinguishing feature of the meninges is its histological characterization. Meningiothelial cells tend to form whorls, both in normal and neoplastic cells. These whorls are an inherent characteristic of these cells that also occurs in tissue culture if provided the appropriate scaffolding[5]. No other tissue is known to form whorls to this extent or to have the dual responsibilities exhibited by meningiothelial cells. It was this histological characterization and the unique differentiation of these cells that challenged investigators such as Cushing to find an all-encompassing classification scheme for tumors of the meninges.

B. General statistics

1. Intracranial meningioma

During the past 50 years, approximately 13% to 19% of all primary brain tumors have been classified as intracranial meningiomas. In the United States, several groups reported results from large neurosurgical studies: Cushing and Eisenhardt[3] found 13.4% in a series of 2203 brain tumors; Grant[6] at the University of Pennsylvania found 17% meningiomas among 2326 brain tumors. The European figures were only slightly higher:

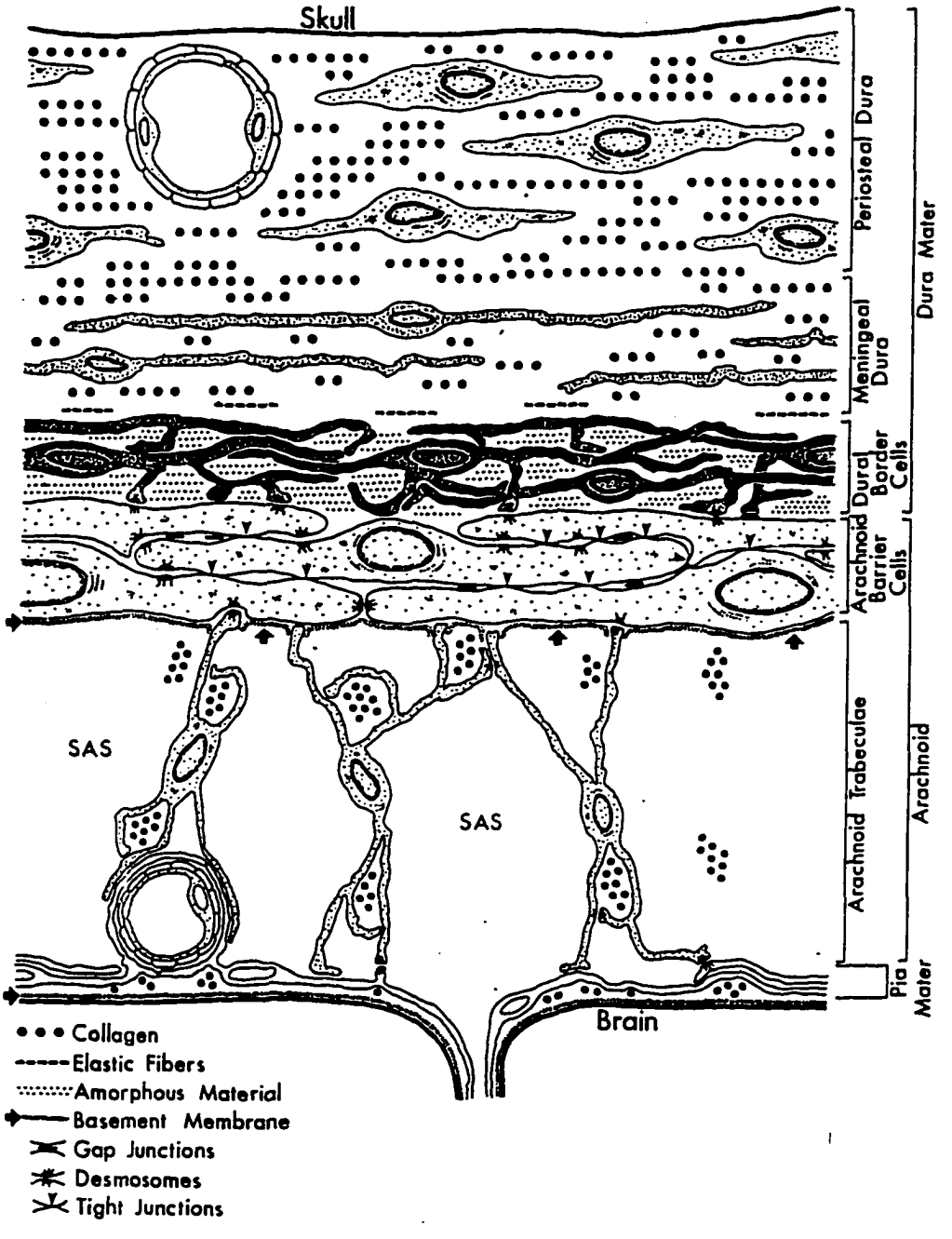


Figure 2: Arachnoidal cells in the dura mater[2].

Zulich[7] found 18% in his series of 5372 tumors and 20.6% in the Olivecrona series[8] consisting of 3909 tumors. In other studies, Zimmerman[9] and Schoenberg, et al.[10], demonstrated 27.3% of 2262 and 16.5% of 2371 tumors, respectively, were meningiomas. Studies from India reported 13.1% from Dastur[11], 13% from Balasubramaniam and Ramamurthi[12], and in Japan, 15.9% of the samples were meningiomas as reported by Katsura[13]. Overall, it was accepted that meningiomas accounted for about 15% of all intracranial tumors worldwide. However, it was noted that figures from Africa are appreciably higher. Meningiomas were found to represent 23.5% of brain tumors on the Ivory Coast[14]; Levy[15] reports 28.6% in the study of Malawi, Rhodesia and Zambia; 29.9% in Nigerian Africa by Odeku and Adeloye[16]; 30.3% in the Bantus of Transvaal by Froman and Lipshitz[17] and 38% in a small series of 13 cases in Ethiopia[18].

2. Spinal meningioma

Meningiomas of the spinal column are much less frequent than intracranial meningiomas. In the Russell and Rubenstein[19] series, 12% of meningiomas were present in the spinal canal. However, if primary tumors of the spinal column are considered independently, meningiomas make up a significant portion of common spinal tumors. It is generally accepted that the thoracic region is involved in spinal meningiomas more often than other regions, with the cervical segments occasionally affected and the lumbar region rarely affected. As in the cranium, spinal tumors usually are fixed with the dura, but not always, and occasionally are found in a variety of relationships with the spinal cord and the soft meninges surrounding the spinal cord[20].

Meningiomas, spinal or intracranial, are most commonly detected in the fourth and fifth decades of life and occur predominantly in women. For example, a recent review of intracranial meningiomas diagnosed in Manitoba, Canada[21], reported the ratio of female to male patients was 2:1 and the incidence of occurrence increased with age. The mean age of patients with benign meningiomas was 58 +/- 15 years. The only

exceptions are the African studies mentioned above which demonstrated a male predominance and in which 20% of the tumors developed during the second decade of life. No explanation of this discrepancy was reported.

3. Childhood and elderly incidence of meningioma

Incidence of childhood meningiomas is quite rare, and accounts for less than 2% of intracranial tumors. Taptas, in a review of various reports totaling 1760 meningiomas, found only 19 examples (1.1%)[22]. The report by Cushing and Eisenhardt[3] contained only five childhood cases, or 1.9% and Crouse and Berg[23] reported 13 patients with meningioma (only 2.9% of all pediatric intracranial tumors) at the University of California, San Francisco. However, meningiomas first diagnosed in childhood (i.e., under age 20), are more likely to become malignant as demonstrated by Deen, Scheithauer and Ebersold[24]. Nakamura and Backer[25] confirmed this observation, and also noted that a high incidence of malignant meningiomas occurred in infancy and childhood. Two out of seven childhood meningiomas were histologically malignant as reported by Cooper and Dahn[26] and Crouse and Berg[23] reported five in 13 meningiomas for an astonishing 38%. Childhood meningiomas do not demonstrate the female predominance seen in adults. For example, in the Cushing and Eisenhardt studies[3], of the five patients under age 20, three were male and two were female. In addition, Paillas, et al.[27], reviewed 110 cases of pediatric meningiomas and found 57 boys and 53 girls.

Meningiomas occurring in the elderly (patients in sixth through eighth decade of life) are common, with many of the neoplasms detected incidentally at necropsy. Wood, White and Kernohan[28] found that of 300 incidental brain tumors, 100 could be classified as meningiomas (33.3%) and the peak incidence was in the seventh decade of life with no sex predominance. Cooney and Solitare[29] found that of 17,000 patients from Yale-New Haven Hospital over age 60, 497 had general brain tumors and of these, 176 were classified as meningiomas (35%). Up to 80% of these meningiomas also were

subsidiary meningiomas discovered only at autopsy, suggesting that some of these tumors were slow-growing and undetected during life.

4. Familial occurrence of meningioma

Familial occurrence of meningiomas usually is found in conjunction with von Recklinghausen's disease (NF1) or bilateral acoustic neurofibromatosis-2 (NF2). However, there are documented cases of familial meningiomas with no stigmata of neurofibromatosis. For example, in 1959, Gaist and Piazza[30] described a case of brother and sister with meningioma and two cranial meningiomas, respectively. Neither patient, nor any other member of the family, demonstrated any history of neurofibromatosis. Similarly, Joynt and Perret[31,32] reported two cases of single meningiomas in a mother and daughter, Delleman[33] presented a case of multiple familial meningioma in which three members of one generation and two members in the next generation had multiple meningiomas, and Memon[34] reported a case of multiple meningiomas in a mother and son, neither of whom presented a history of neurofibromatosis. Of particular interest is a case observed by Sedzimer, et al.[35] involving twin boys, each presenting cranial and spinal meningiomas at ages eight and 13, respectively. It therefore is believed that cases of familial meningioma with no presence of neurofibromatosis, could indeed be instances of very subtle symptoms of neurofibromatosis which are easily overlooked.

C. Etiology

The etiology of meningiomas includes a variety of environmental and genetic factors including trauma, irradiation, steroid hormone receptor presence (reflecting the female bias) and the aforementioned association with neurofibromatosis.

1. Trauma related meningioma

Historically, many patients with meningiomas recalled head trauma, suggesting a statistically significant reason for these tumors. Cushing and Eisenhardt[3] reported 101 of their 313 cases (33%) were related to head trauma. They also noted that a swelling, a

scar, or a depressed skull fracture was present at the site of meningioma that in all cases also supported their theory of head trauma resulting in meningioma. Other studies by Walsh, Gye and Connelly[36], and Gardeur, et al.[37], supported the same notion by reporting examples of meningiomas that presented at or near the site of previous skull fracture. Spinal trauma also was reported to result in meningioma formation as in the case described by Von Holander, et al.[38] (in a review by Kepes[4]) of a 38 year old female with traumatic fracture of the twelfth thoracic vertebra. Nineteen years later, a meningioma developed at the same site as the fracture. In two reports, Preston-Martin, et al.[39,40] demonstrated a statistically significant number of meningiomas resulted from head trauma using case-controlled studies. These studies were sex-controlled, using only women, and the patients with meningioma demonstrated greater recall of head trauma than a control group of friends and neighbors. Despite the presented proof, some investigators are skeptical with regard to trauma in meningioma formation. Dunsmore and Roberts[41] expressed strong objection, mentioning that trauma did not offer any explanation of childhood or malignant meningiomas, nor did the notion of head trauma appear significant to neoplasms located at the base of the skull. Further controlled case studies performed by Choi, et al.[42] also failed to identify a correlation between trauma and meningiomas after interviewing 24 patients with meningioma. They concluded that the patients with meningiomas did not differ from matched controls in the frequency of previous head trauma.

2. Radiation induced meningioma

Radiation also was examined as a possible cause for the development of human meningiomas. Originally, this concept was based on the study performed by Dimant, et al.[43] that demonstrated meningioma formation in rabbits implanted with cobalt-60. These rabbits developed three tumors in the region of implant: two meningioma and a sarcoma. Since radiation is a known mutagen capable of producing both single-stranded and double-stranded breaks in DNA, these physical changes could result in major DNA

abnormalities (insertions, deletions or translocations), which were associated with neoplastic transformation of cells in other tumor cell lines. In most reports, there were two classes of meningiomas resulting from radiation: meningioma resulting from radiation occurring after treatment for a primary head or neck tumor, or meningiomas resulting from other brain irradiation (e.g. mycotic infection). In the first class of tumors, meningiomas developed after several years of latency with a greater latent period following lower doses of radiation. The average latent period between original treatment and meningioma formation was 20 years (ranging from 12 to 27 years). For example, Bogdanowitz and Sachs[44] reported meningioma formation in a patient with pituitary adenoma 25 years after primary treatment, and 17 years after primary treatment of another patient with medulloblastoma. In both instances, the treatment involved relatively high doses of radiation. Norwood, et al.[45] reported similar findings in an optic nerve glioma patient resulting in meningioma formation 25 years following primary treatment, and Waga and Handa[46], in their review, reported only a 12 year latency in a patient with craniopharyngioma. More recently, Pagni, et al.[47] reported radiation induced meningioma in a patient with no history of neurofibromatosis or familial meningioma. The primary tumor in this case was a skin carcinoma that was excised and treated with radiation to decrease chances of recurrence. Fourteen years later, a meningioma developed in the irradiated region of the skull.

In the second class of meningiomas, those developing after primary irradiation, a review by Kepes[4] described a case reported by Horanyi, et al.[48] in which a five year old girl received scalp irradiation treatment for tinea capitis (micropsoriasis) and developed a malignant meningioma two years later, an observation consistent with the notion that childhood meningiomas are malignant. Meningiomas also developed in patients following scalp irradiation as reported by Munk, et al.[49], Waterson and Shapiro[50], and Russell and Rubenstein[19]. Patients described by Feiring and Foer[51], Bogdanowitz and Sachs[44], and Watts[52] were treated for "port wine" stains of the face and in all reported

cases, meningiomas formed specifically on the treated side of the face. Other more recent reports by Soffer, et al.[53], Giaquinto, et al.[54], and Rubenstein, et al.[55], also described how low dose scalp irradiation for fungal disease or "port wine" stains pointed to the tendency for multiple meningioma and recurrence of meningioma after apparently complete surgical removal.

In a much smaller study affecting far fewer patients, the development of meningioma also has been attributed to the carcinogenic effect of Thorotrast (thorium dioxide suspension) which was used for a short time for radiological visualization. Several patients developed meningiomas after enduring Thorotrast treatment for unrelated ailments[56-58].

3. Viral agents and meningioma

Viral agents were suspected by some investigators to have a role in meningioma formation. Weiss, et al.[59] reported finding simian virus 40 (SV40)-related antigens in two of seven meningiomas studied. Building on this study, a total of 336 human intracranial tumors were examined for presence of papovavirus large T antigen. The large T antigen was found in only 36 of these tumors, with 32 classified as meningiomas[59-62]. However, it must be noted that different results were obtained when samples from the same tumor were examined by different investigators.

Hybridization studies also were used to associate viruses with human tumors by identifying unique viral DNA sequences in tumor cell DNA. Here, a labeled viral DNA probe was allowed to form double-stranded complexes with tumor cell DNA if identical sequences were present in both DNAs. A number of investigators explored this technique using human intracranial tumors[63-65]. Again, not all used control tumor tissues and DNA probes, and different investigators reported different results when examining the same cell lines.

Southern blot was another technique used to assay 30 human brain tumors in the search for DNA sequences homologous to viral DNA. Rachlin, et al.[66] assayed for

SV40 DNA and two of nine meningiomas (22%) had these homologous sequences integrated into the tumor cell DNA. These same meningiomas did not contain DNA sequences homologous to adenovirus when control probes and non-tumor tissue were examined.

It should be noted that the presence of viral DNA or proteins within a tumor cell does not conclusively prove that the tumor was caused by the particular virus. Papovavirus DNA and T antigen are commonly found in meningioma; however, the presence of this viral DNA and protein does not reveal whether the tumor was caused by viral transformation and maintenance, or whether the virus merely infected a tumor cell which was already transformed. Therefore, whether the virus acts as a mutagen or maintains a transformed state remains undefined. Of notable interest, however, the human homologue of the simian sarcoma virus transforming gene (oncogene c-sis) has been mapped to the long arm of chromosome 22, and as will be discussed in a later section, karyotypic analysis of meningiomas reveals characteristic monosomy or partial deletion of chromosome 22.

4. Steroid hormones and meningioma

The clear predominance of meningiomas in females has led to some discussion about the relationship between female sex hormones and meningiomas. There have been several reported cases of rapid growth of meningiomas during pregnancy[3,67-70]. In these cases, all the meningiomas occurred near the optic nerves resulting in increased pressure on the optic nerve and causing changes in visual acuity that was easily noticed by patients.

Numerous cases also have been reported in which patients were afflicted with both meningioma and breast carcinoma. Smith, et al.[71] reported two women with breast carcinoma who subsequently developed meningioma. Schoenberg, et al.[72] discussed the possibility of CNS neoplasms and primary cancers in other sites and found the only significant combination to be meningioma and malignant breast cancer. Others have

reported observations of meningioma associated with mammary and/or genital cancer in the same patient with similar conclusions[73,74].

In studies performed by Donnell, et al.[75], six meningiomas were examined for the presence of estrogen receptors. Of the six, four were found to contain estrogen receptors, two at a high level. Interestingly, one of the four estrogen receptor positive meningiomas was from a male patient. Tilzer also performed sex hormone receptor studies but examined meningioma tumors for the presence of estrogen and progesterone receptors. He found that 13 out of 22 meningiomas contained estrogen receptors (59%) and that all 22 meningiomas contained progesterone receptors[4]. The presence of progesterone receptors correlated directly with the observed increase in tumor growth during pregnancy[76].

5. Genomic imprinting in meningioma

Neoplastic cellular transformation resulted from alterations in primary genomic sequences by: 1) mutations that convert a proto-oncogene to an oncogene, or 2) by inactivation of a tumor suppressor gene. In several cases there were reports of loss of heterozygosity in cases of partial deletion or monosomy 22 with preference for the paternal homologue[77,78]. This genomic alteration suggests the possibility of genomic imprinting in meningiomas since similar precedence for paternal homologue bias has been established in cases of Wilm's tumor[79], rhabdomyosarcoma[80] and osteosarcoma[81]. The initial hypothesis was that genomic imprinting also occurs in meningioma. This hypothesis was supported by the observation that the parental origin could be established in nine cases of meningioma. These results indicated that unlike the tumor cases mentioned above, there was no evidence of maternal or paternal imprinting in meningioma since both parent homologues were lost randomly.

D. Chromosomal patterns in meningioma

1. Karyotype analysis

In the cases of meningioma mentioned above, there is a clear indication that alteration in the genetic material is the predominant causative factor. In cases of radiation treatment, DNA damage was a well-documented side effect. Where viral genes have been observed, they might be incorporated into the genome and instigate changes in the genetic material by insertion or deletion that disrupts a tumor suppressor or converts a proto-oncogene to an oncogene. Alternatively, the presence or absence of receptors based on the activation or repression of the transcription machinery to produce mRNA encoding the proteins or the delivery machinery of the receptors also is considered genetically linked. The underlying questions remain: Is the transformation from a normal to a meningioma cell induced by or followed by a disturbance in the genome? Does a specific external disturbance (i.e. trauma) result in cell transformation or do random events result in meningioma formation?

Volumes of data have been collected reporting the relationships of chromosomal aberration with neoplasms[82-86]. With respect to meningiomas, a distinction should be drawn between the karyotype studies of somatic cells in patients with meningiomas and the karyotype studies of the meningiomas themselves. Of significant interest is the latter group in which the studies of Zang and Singer[87] present cytogenetic evidence of brain tumor culture cells which show that eight of eight tumors analyzed lacked one G group chromosome. Four years later, the monosomy G group chromosome noted by Zang and Singer[87] was identified as chromosome 22 by fluorescence banding[88,89], an observation also confirmed by Giemsa banding[90]. Zankl[91] expanded these studies and demonstrated that the missing chromosome was not lost due to a chromosomal translocation. Other chromosomes also were observed to be missing or altered in similar karyotype analyses[87,92-94],.specifically, chromosomes 1, 6, 11, 12, 14 and in six out of 24 meningiomas in male patients, the Y chromosome also was missing. Katsuyama noted the possible involvement of oncogenes since many oncogenes were assigned to the same

chromosomes found missing or altered in meningioma karyotype studies, and alluded to a relationship between the two.

If one examines the many reports on cytogenetic studies in cases of meningioma, the most consistent finding is the total loss or partial deletion of chromosome 22[93,96-103]. The presence of an abnormal chromosome 22 was characterized further by Casalone[94] in his study of 31 cases of meningioma. Here, Casalone found 14 cases of meningioma with chromosomal abnormalities and of those 14 cases, 11 involved aberrations in chromosome 22. The remaining 17 cases had normal karyotypes, which was a finding also noted by others, including Mark[98], who reported that approximately 24% of meningiomas studied showed normal karyotype. Maltby[104] confirmed this observation, reporting that in a series of 50 meningioma, almost half had normal diploid karyotypes. The remainder of tumors in this study carried monosomy 22, with a low frequency of other chromosomal deletions and rearrangements in 1, 10, 14, 17, 19 and Y. In this regard, Maltby also pointed out that the chromosomal rearrangements seen in benign meningiomas were similar to those seen in senescent cells and suggested a common mechanism. Within the structural rearrangements seen in other chromosomes, a distinct marker chromosome resulting from translocation of satellites from short arms to other deleted chromosomes was reported and confirmed by silver staining by Zankl and Huwer[105]. Another distinct type of chromosomal marker was formed by the fusion of telomeric ends of two chromosomes in meningioma (for example, chromosomes 9 and 14)[98,104]. In cases with dicentric chromosome formation, telomeric breakdown also may be involved[106-108], reflecting the similarities in chromosome structure between tumor cells and senescent cell lines.

Other investigators reported karyotype analysis of meningioma patients with a constitutional ring chromosome originating from chromosome 22[109-111]. Arinami[111] discussed a patient with chromosomal breaks at 22p12 and 22q13.3 resulting in a ring chromosome. Essentially this ring chromosome behaved as a constitutional deletion of

22q13.3-qter in chromosome 22 and supported the theory presented by Zankl and Zang[102] that genetic information critical to cell proliferation could be located in the long arm region of chromosome 22 and deletions may predispose to meningioma formation. Ring chromosomes were of particular interest since they were uncommon in malignant tumors, but rare in benign tumors such as meningioma. Mark[101] reported 3 of 30 tumor cultures containing ring chromosomes, although the origin of the ring chromosome was different in each case. The appearance of ring chromosomes supports the hypothesis that telomeric breakdown may be involved in chromosome structural rearrangements.

However, the correlation between chromosomal damage present in meningioma and senescent cells cannot be overlooked since the rearrangements which formed dicentric chromosomes resulted from telomeric fusion. Such dicentric fusions also were reported by Benn[112], who first suggested the instability of telomeres or of the associated enzymes in the overall process of senescence. Here, one culture demonstrated a monosomy 22 and then formed telomeric fusions resulting in dicentric chromosomes. Mark, in his chapter in Chromosomes and Cancer[95], suggested that "mechanisms for chromosome behavior in benign tumors are analogous to those occurring in senescence; in the case of tumors, such changes are perhaps triggered earlier than in normal cells by the influence of oncogenic factors, but do not necessarily lead to cell death."

2. Mapping analysis

The literature reveals that there are two types of meningioma karyotypes, one with no chromosomal abnormalities and another characterized by monosomy or partial deletion of chromosome 22 with the possibility of other chromosomal rearrangements. Studies from several laboratories have focused on the latter group of meningiomas in the search for the specific region critical in meningioma formation. By utilizing polymorphic markers for several chromosomes to investigate directly the genetic constitution of primary tumor samples, these studies provided significant, new information and insight into the molecular basis of meningioma and represented a change in direction from the

earlier karyotyping studies performed on cultured tumor cells. In addition, the use of polymorphic markers circumvented the confusion resulting from chromosomal aberrations that occur in an *in vitro* environment.

In view of the striking clinical association established between meningioma and bilateral acoustic neurofibromatosis (NF2), and the discovery that acoustic neuromas also displayed a loss of chromosome 22, many researchers speculated that a common mechanism involving chromosome 22 must be at work in both tumor types. Seizinger[113,114] was the first to investigate the relationship between meningioma and NF2 using polymorphic markers. They used several different probes known to give restriction fragment length polymorphisms (RFLPs) on several different human chromosomes. In analyzing chromosome 22, four specific loci were examined with the following probes: sis (the platelet derived growth factor beta locus homologous to the sis oncogene) that is mapped to 22q12.3-13.1, the D22S1 marker that is mapped to 22q12-13, the D22S9 marker that is mapped to 22q11 and IGLC (the immunoglobin light chain constant region) that is mapped to 22q11 (see figure 3). These markers were the only ones available at the time of the study which mapped to the long arm of chromosome 22. The short arm was not investigated because all evidence to date showed loss of chromosome 22 or deletion of the long arm. The results of these studies indicated that the majority of karyotyping information from cultured tumor cells was correct and that both acoustic neuroma and meningioma demonstrated loss of alleles on chromosome 22, and narrowed the possible meningioma and NF2 loci to the region distal to marker D22S9 in band 22q11. In two meningiomas studied that showed normal karyotypes, a loss of heterozygosity on chromosome 22 in the polymorphic marker analysis was observed. These two meningiomas represented a small number of cases with microdeletions that were undetectable by cytogenetic analysis.

Similar deletion mapping studies were undertaken by Dumanski[115] using some of the same markers used by Seizinger, as well as additional markers. Here, D22S9, IGLC,

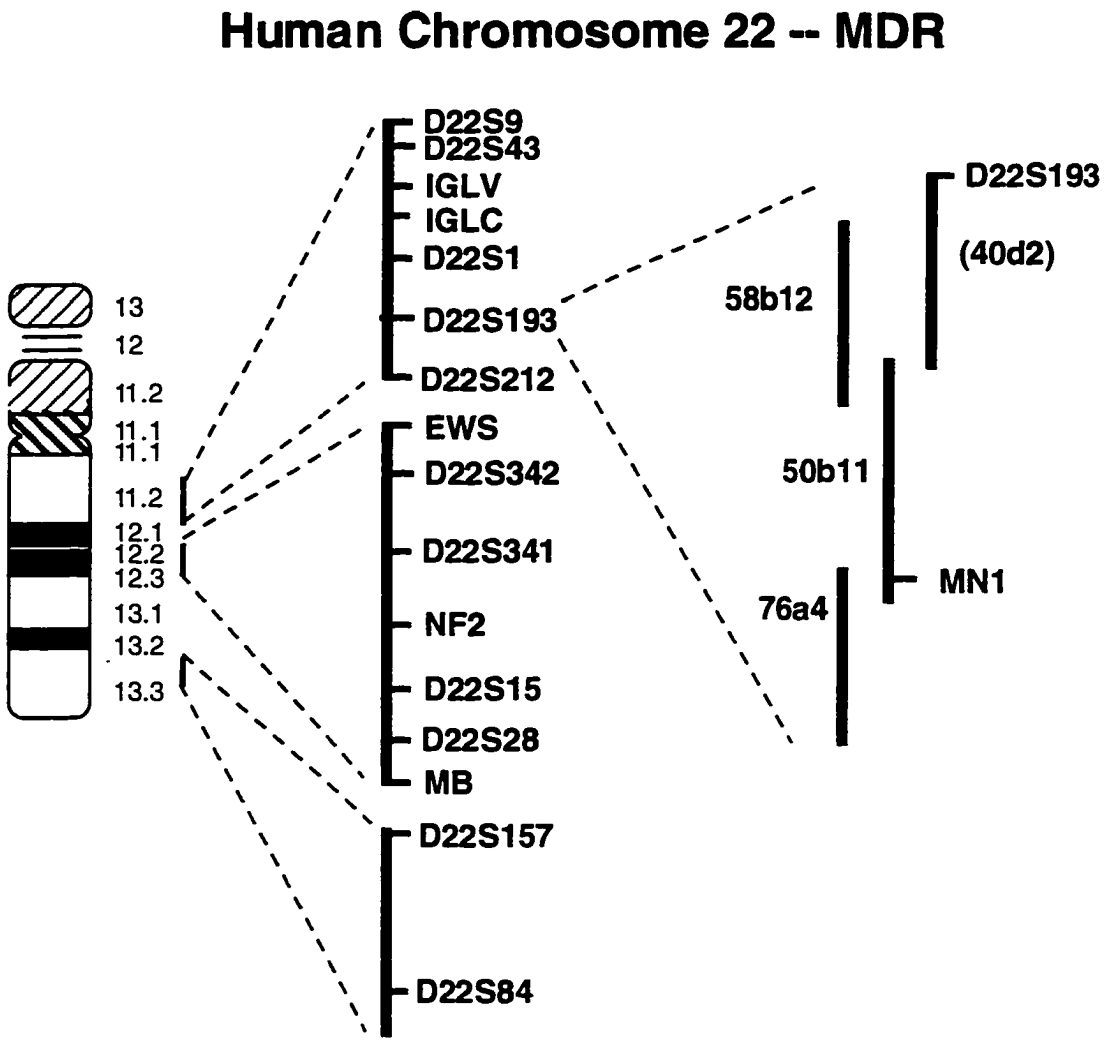


Figure 3: Meningioma deletion regions and relevant markers on human chromosome 22

D22S1, D22S10, IGLV (immunoglobin light chain variable region), MB (myoglobin locus) and sis all were used as probes to analyze 35 patients with meningiomas. Sixteen tumors retained the constitutional heterozygosity, whereas 14 (40%) demonstrated loss of heterozygosity of all loci in tumor DNA (consistent with monosomy 22), and the other five tumors showed loss of heterozygosity at specific loci in tumor cells while retention of heterozygosity at other loci. These observations are consistent with the hypothesis that a partial deletion of chromosome 22 results in meningioma. Based on this study, Dumanski concluded that a meningioma locus was located at 22q12.3-qter, considering that the only portion of chromosome 22 consistently lost in his studies was located between the MB locus and the telomere. This region corresponded to 22q12.3-qter.

Okasaki, et al.[116], and Schneider[117] reported similar experiments using many of the same markers as Seizinger and reached the same conclusions, although citing the need for more polymorphic markers on the q arm of chromosome 22. In subsequent studies, Dumanski[118] answered the call for more markers and reported 195 probes mapping to four different regions of the long arm of chromosome 22. Twenty-eight of these probes which map to 22q12-qter generate RFLPs which allowed for generation of a linkage map to narrow down the region containing potential NF2 and meningioma genes. This meningioma region thus was tentatively localized to 22q12.3-qter as mentioned above and the NF2 region now was known to be distal to the anonymous marker D22S1 which mapped to the 22q12-13 region. Because there was potential overlap between these two regions, and since meningiomas occur frequently in NF2 patients, it was conceivable that a single gene locus was involved in both afflictions. Using these new probes, tumor tissue from 81 patients was analyzed by Dumanski[119]. Here, 42 tumors (52%) showed loss of heterozygosity for all alleles on chromosome 22 (monosomy 22), and nine tumors (11%) showed loss of heterozygosity for telomeric loci but retained heterozygosity for more centromeric loci, an observation consistent with the hypothesis that a partial deletion of the q arm of chromosome 22 resulted in meningioma. This study confirmed the

previous work by Dumanski which mapped the meningioma locus to the 22q12.3-qter region but was unsuccessful in narrowing the suspected meningioma region. However, the fact that no small deletions of 22q were found suggested that the meningioma locus was probably located distal to the MB locus, but closer to MB than to the telomere of the chromosome. The hypothesis presented was that the small terminal deletions would not be sufficient to induce neoplastic transformation.

Dumanski[119] also emphasized that the distinction between chromosomal aberration with respect to sex of the patient was quite significant. The conclusion he reached was that in cases of chromosomal aberration, males had significantly smaller abnormalities than females, if the assumption was made that the cases with no apparent cytogenetic aberrations did have small abnormalities in the meningioma locus. This was based on the observation that the number of male cases with no apparent chromosomal aberrations was much higher (50% versus 34% in females), although, it was recognized that the mechanism of cellular transformation might be different in cases of no apparent abnormalities than in cases of gross deletion or monosomy 22.

Dumanski also mentioned that the hypothesis that the NF2 gene and the meningioma locus were the same was proven false by Rouleau[120,121] who localized the NF2 region centromeric to the D22S28 marker on chromosome 22. In a case in Dumanski's study, the retention of heterozygosity at D22S28 and MB suggested that the meningioma gene and NF2 genes were separate loci, a hypothesis confirmed by Rouleau. Other evidence for separate loci was presented by Pulst[122] who suggested that familial meningiomas must have a loci distinct from NF2 based on linkage analysis of a pedigree with no presence of NF2. In this respect, it was proposed that a familial meningioma locus was present distal to the NF2 locus, but no confirming studies were reported[122].

Due to the clinical finding that NF2 and meningioma were so intimately related based on linkage studies, Cogen[123] concluded that the neoplastic transformation in meningioma formation was a result of activation of a different locus than that presented

by Dumanski, and the proposed locus on chromosome 22 most likely was located proximal to the newly discovered gene for NF2, around 22q11. Cogen also hypothesized that base pair mutations and/or deletions of chromosome 22 DNA may have been present but not detectable by the techniques used by others. In support of his proposal, Cogen cited point mutations that were demonstrated in at least one tumor suppressor gene (p53) and several other neoplasms[124]. Sanson[125], in his report, concurred with Cogen and speculated that the oncogenesis of meningiomas which did not demonstrate loss of heterozygosity may have one of the two alleles of a tumor suppresser gene rendered dysfunctional, alluding to microdeletions and/or base changes.

To this end, Bello[126] suggested that in the cases of meningioma with no apparent chromosomal aberration, perhaps duplication of the retained chromosome occurred after the loss of the other copy in the pair which accounted for the retention of heterozygosity, but correlated with the accepted characteristics of monosomy 22 in meningiomas. Citing the case of retinoblastoma, Bello supported Cogen's hypothesis that in the cases of no apparent chromosomal abnormality, the mechanism may be a subtle anomaly such as microdeletion or inactivating mutations in the meningioma locus on both copies of chromosome 22, keeping both chromosomes 22 seemingly intact at the cytogenetic level.

Lekanne Deprez[127], shifted the focus to the meningioma cases resulting from translocations or other aberrations of chromosome 22, in an attempt to localize the putative tumor suppressor gene involved in meningioma[93,94,115,128-132]. These authors supported the hypothesis put forth earlier by Cogen and Bello which suggested inactivation of both alleles of a tumor suppressor gene as the mechanism of meningioma tumorigenesis. They reported a case with a stemline karyotype 45, XY, -1, 4p+, 22q-, 22q+ that had rearrangements on both copies of chromosome 22. The first allele of 22 was found as a dicentric chromosome resulting from translocation with chromosome 1 (22pter-22q11::1pter-1q11) with the reciprocal product presumably lost due to the loss of a centromere, and therefore resulted in a loss of sequences distal to 22q11. In addition,

they hypothesized that a remaining tumor suppressor allele of 22 was disrupted due to an additional but balanced translocation t(4;22)(p16;q11). This translocation breakpoint was mapped to a region between D22S1 and D22S15, also the same region where the NF2 gene was mapped, but in conflict with reports from Dumanski who reported localization of the meningioma tumor suppressor gene at 22q13.3-qter.

Rey, et al.[133] performed other mapping studies and agreed with Dumanski that a meningioma locus most likely was present distal to the MB locus at 22q12.3-qter, based on deletions found at 22q12 (distal to marker D22S32). However, the authors admitted that the relative location between the D22S32 and D22S15 loci had not yet been established, and evidence for a breakpoint distal to D22S32 might have been proximal to D22S15. If this was the case, the data presented by Rey, et al. was in agreement with that of Lekanne Deprez, but not Dumanski.

In contrast, it is conceivable that two or more meningioma loci may be present on chromosome 22 (q11 and q12.3-ter), and according to Ruttledge[134], the meningioma tumor suppressor gene located at 22q11 was the NF2 gene (assuming loss of the second allele of chromosome 22). Ruttledge also suggested that an additional gene (perhaps in the 22q12.3-qter region) was responsible for the remainder of sporadic meningiomas, based on the evidence that the NF2 gene was the primary target for loss of heterozygosity on chromosome 22 in 60% of sporadic meningiomas. Within these tumors demonstrating loss of heterozygosity for the NF2 gene, 27.6% harbored inactivating base substitutions on the remaining allele in 8 of the 17 exons of the NF2 gene studied. It was suggested that additional mutations neither detected nor screened for in the NF2 gene were the cause of tumorigenesis in the additional cases of meningioma demonstrating loss of heterozygosity for the NF2 gene. Ruttledge also claimed that an additional mechanism (an additional tumor suppressor gene region) was responsible for the other 40% of sporadic meningioma, which allowed for the notion of another meningioma locus on chromosome 22 as suggested by Dumanski.

Recent studies searching for mutations in the NF2 gene in meningioma patients were performed by Wellenreuther[135]. The entire coding region of the NF2 gene in 70 sporadic meningiomas was examined and 43 mutations in 41 patients were reported. All of the observed mutations resulted in either a truncated, an abnormally spliced transcript, or an inactivated predicted protein product. Interestingly, all mutations occurred within the first 13 exons of the NF2 gene which was the region of similarity with moesin, ezrin and radixin[141], which are members of a well conserved family of cytoskeleton associated proteins[142-147]. The most significant homology present between merlin (the NF2 gene product name: moesin, ezrin and radixin like protein) and moesin, ezrin and radixin occurred in the first 340 residues at the N terminus of the protein, which was the region encoded by the exons containing mutations in Wellenreuther's study, which suggested that the mechanism of action could be related to the function of this class of proteins.

For example, other proteins in the same family as merlin include erythrocyte protein 4.1 which functions to maintain membrane stability and overall cellular shape by connecting integral membrane proteins to the cytoskeleton[136,137]. Hereditary eliptocytosis[138] is a direct result of disruptions in this gene, specifically mutations which occur in the sequence coding the N-terminal domain of the resulting protein. This region has been found to contain binding sites required for protein:protein interactions.

Talin is another protein in the merlin family. This protein, found in regions of focal adhesion at cell-to-cell or cell-substrate contacts, binds to integral proteins in the cell membrane and is responsible for connecting the extracellular adhesion matrix to the cytoskeleton[139,140].

Because of this significant homology with cytoskeleton associated proteins, it was proposed by Bianchi[141] and others[148-151] that the NF2 gene constituted a tumor suppressor gene of more general importance in tumorigenesis, not only in NF2 and meningioma, but also in malignant mesotheliomas[148,151], ependymomas[150] and astrocytomas[150].

3. Sequencing analysis

Although there was a clear correlation between meningiomas and alterations in the NF2 gene, these alterations in the NF2 gene also were accompanied by loss of the other allele in approximately one third of all cases of sporadic meningioma[152]. Thus, there still was evidence that other non-NF2 gene-related loci present on the long arm of chromosome 22 might be involved in meningioma formation.

Lekanne Deprez reported the presence of a new meningioma locus at 22q11 based on a patient with balanced translocation (4;22). Subsequently, they cloned and characterized an expressed gene (MN1) from 22q11 which was disrupted in this meningioma patient and thus might play a role in meningioma formation. The MN1 cDNA characterized by sequence analysis has a total length of 7.5 kilobases (kb). Mapping studies subsequently demonstrated that the MN1 gene has 2 large exons, a 5' exon of 4.7 kb and a 3' exon of 2.8 kb, with the 3' exon contributing only 59 amino acids to the final predicted protein and the 5' exon contributing a varied number of amino acids (as discussed below). These exons were mapped back to the chromosome 22 and the results indicate that this gene spanned approximately 70 kb with a 60 kb intron. These studies also indicated that the 5' exon (4.7 kb) was disrupted in the balanced translocation reported in this patient and it was thought that this interruption of the MN1 gene resulted in meningioma formation.

Studies[152] involving hybridization of an MN1 cDNA probe to DNA extracted from other species, revealed hybridization signals in species as evolutionarily distant as *Xenopus laevis* and *Drosophila melanogaster*, an observation consistent with the hypothesis that the MN1 gene is evolutionarily conserved. The MN1 cDNA displayed an open reading frame starting at nucleotide 888 and ending at nucleotide 4913. The predicted protein could begin at one of two Met start sites yielding a protein of 1319 amino acids or 1290 amino acids. The Met start site which resulted in the 1290 amino acid protein had a much better Kosak consensus sequence[153,154] and all further analyses

were based on translation from this start site. The MN1 gene also contains two trinucleotide CAG (glutamine) repeats at positions 1839-1883 and at position 2526-2606, which have been analyzed for potential expansion as is seen in a number of neurological diseases including Huntington's disease, Kennedy's disease, spinocerebellar ataxia-type 1, monotomic dystrophy (DM) and dentatorubral and pallidoluysian atrophy[155,156]. These studies observed that the most 5' repeat was not polymorphic whereas the 3' repeat contained 3 alleles differing 6 or 9 nucleotides from the smallest allele. However, neither repeat showed any evidence of expansion.

The MN1 gene demonstrated several structural features which allude to the function of the protein. First, the 5' untranslated region of the MN1 gene is very GC rich and contains four ATG start codons followed by stop codons. Several growth regulatory genes have presented this feature and the hypothesis is that they inhibit translation[153,154]. Second, the MN1 gene is very rich in prolines, specifically in the region 339 to 365. In this region, 15 of the 26 amino acids are proline, and proteins with stretches of proline typically function as a transactivating domain in a transcriptional regulator. Further documentation for a putative function in transcription is the third feature; i.e., stretches of glutamines that result from the CAG repeats. According to Gerber, et al.[157] (1994), over 80% of proteins which harbor stretches of 20 or more glutamine residues are transcription factors, and these glutamine residues behave as transactivating domains of a protein. To support this notion, they also demonstrated that glutamine and proline transactivating domains both could be used in the activation of a reporter gene *in vitro* or *in vivo*.

Because of the obvious implications of this gene in meningioma, cosmid libraries were screened using the probe D22S193, and four cosmids were selected which present a minimal tiling path over this suspected meningioma gene region of chromosome 22. The genomic sequence and analysis of these four cosmids is reported in the results section of this dissertation.

Genomic sequence analysis of the meningioma locus located at 22q12 (the region described by Dumanski) revealed the presence of the gene encoding β'-adaptin. Ponnambalam, et al.[158] first reported the DNA sequence of the β'-adaptin cDNA from human fibroblasts, rat lymphocytes, rat brain[159] and bovine lymphocytes in 1990, noting the significant nucleotide and peptide homology among these species.

Adaptins are a class of proteins which play a role in clathrin coated vesicle transport in receptor mediated endocytosis[160] and transmembrane receptor transit from the Golgi apparatus to the plasma membrane (see figure 4)[161]. In clathrin coated pits and vesicles, adaptins are the major components of adaptors which link clathrin to transmembrane proteins or receptors. There are three classes of adaptins: α, β and γ, with the adaptors of the plasma membrane consisting of α and β adaptins and the adaptors of the Golgi contain β' and γ adaptins[162]. Studies have demonstrated that cytoplasmic domains of selected membrane proteins bind to adaptors[163,164], and that adaptors, in turn, bind to clathrin[165-167]. As mentioned above, significant peptide sequence homology exists between species within each of the three classes of adaptins, exhibiting the specialized conservation of the domains, although weaker homology exists between the three classes. Robinson[162] reports 25% overall homology between the peptide sequences of α and γ adaptins, and a weaker homology between β and γ adaptins, although Ponnambalam disputed this finding, reporting that no significant homology between α and β adaptins was present in either protein or nucleic acid sequence. However, Ponnambalam notes that the gross properties of the three classes of molecules are similar. One such similarity exists in a proline- and glycine-rich (sometimes alanine) hinge region which divides the protein into parts: the N terminal and the C terminal domains, with the N terminus comprising the so-called brick region, the C terminus making up the ear region, with the hinged stalk region in between (see figure 5). Keen and Beck[168] indicate that the brick domain is the region which interacts with clathrin and the ear regions bind to the varied types of molecules

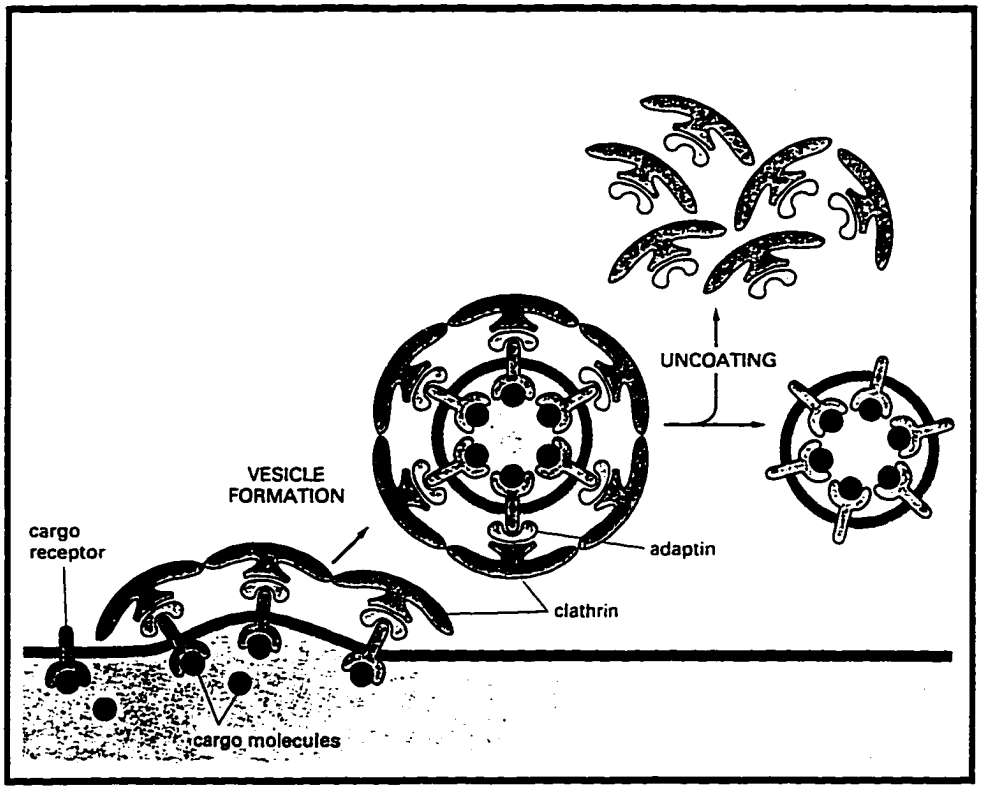


Figure 4: Receptor mediated endocytosis. The role of adaptins in clathrin coated pit assembly. Adapted from reference 253.

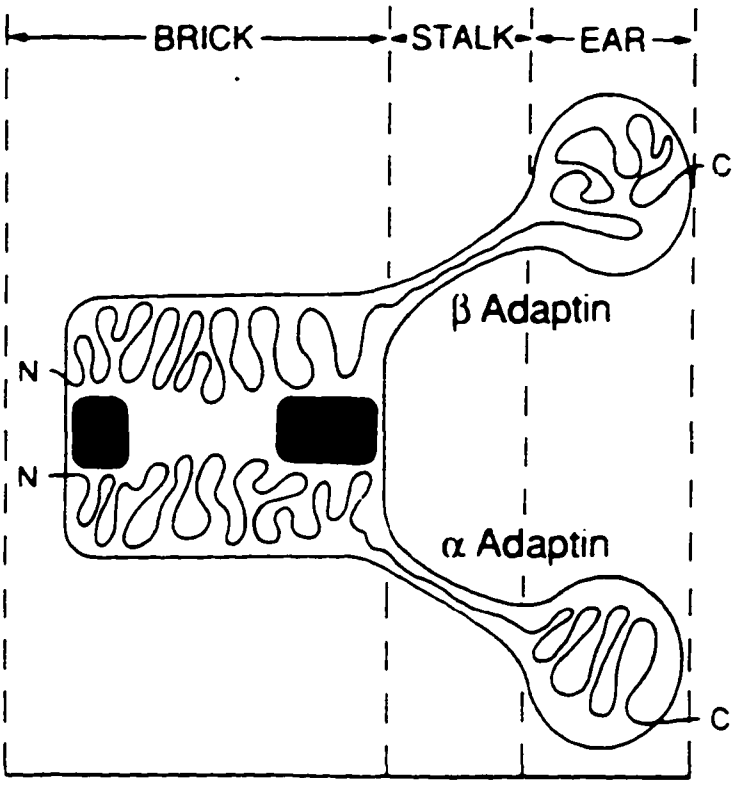


Figure 5: General structure of adaptins[158].

transported by the coated vesicles. The stalk regions vary in length, flexibility and conformation and may facilitate the ear region interactions with the transmembrane proteins or receptors[158].

Structural perturbations in any of the three regions of these proteins could result in drastic modification of the binding ability either for clathrin or for a unique set of required (or related) transmembrane proteins. Alternatively, slight structural alterations can be envisioned which decrease the specificity of binding to transmembrane proteins leaving the clathrin binding affinity intact, or decrease the binding affinity for clathrin resulting in abnormal clathrin binding or decreased receptor trapping within the vesicles. In this type of scenario, the overall presence of receptors on the cell surface is increased because the down-regulation mechanism is faulty, yielding the cell surface containing an abnormally high concentration of receptors with bound growth factors. This results in virtually unregulated cell growth such as that observed in meningioma. It must be noted that β adaptins present in the brain have pronounced diversification with respect to the rest of the β adaptin family. Others[169-171] reported brain specific forms of the α and β adaptins and hypothesize that an independent mechanism for receptor mediated intracellular transport exists in the brain. This specialized pathway present in the brain may lend credence to the hypothesis developed by Dumanski and Roe[172] that unregulated cells present in the tissue surrounding the brain (and hence, meningioma formation) may be a direct consequence of β adaptin absence or dysfunction resulting from complete or partial loss of the q arm of chromosome 22 which encodes the β adaptin gene.

II. Cat Eye Syndrome

Cat Eye Syndrome (or Schmidt-Fraccaro syndrome) is a genetic-based disease that is characterized by a variety of congenital defects[173]. These include ocular colobomata, anal atresia, preauricular dimples and/or tags, heart anomalies, urogenital malformations, and intestinal defects, which vary in severity and number from patient to patient, and in some cases, result in mild retardation (see figure 6). It is estimated, from patients observed in Northeastern Switzerland during the last 20 years, that Cat Eye Syndrome (CES) is thought to affect between 1:50,000 and 1:150,000 individuals. CES, unlike meningioma as discussed above, does not have any sex predominance, and typically affects individuals from birth.

Haab first described the association between anal atresia and colobomata of the iris in 1878[174]. Schmidt, Fraccaro and Schachenmann[173] noted the further association of these two major defects with the presence of a supernumerary chromosome. They described the presence of an abnormal chromosome in three patients and hypothesized that this may be the characteristic of a new "chromosomal" disease. The families of these three patients were examined and the extra chromosome was found in 2 of the 3 cases. The third case revealed mosaicism in several members of the family, with some cells having normal karyotype and other cells with abnormal karyotype suggesting direct transmission from generation to generation via unaffected individuals acting as carriers of the extra chromosome. The mosaic nature of this syndrome, especially across generations, also was noted by Gerald[175] as being very unusual since no precedence for this type of genetic-based mosaicism has been reported in any other disease.

However, Neu, et al.[176] was skeptical about the presence of a supernumerary chromosome being responsible for the myriad of clinical characteristics associated with CES. This skepticism was supported by the observation that a patient displaying bilateral colobomata of the irises, imperforate anus and other phenotypic characteristics of CES,

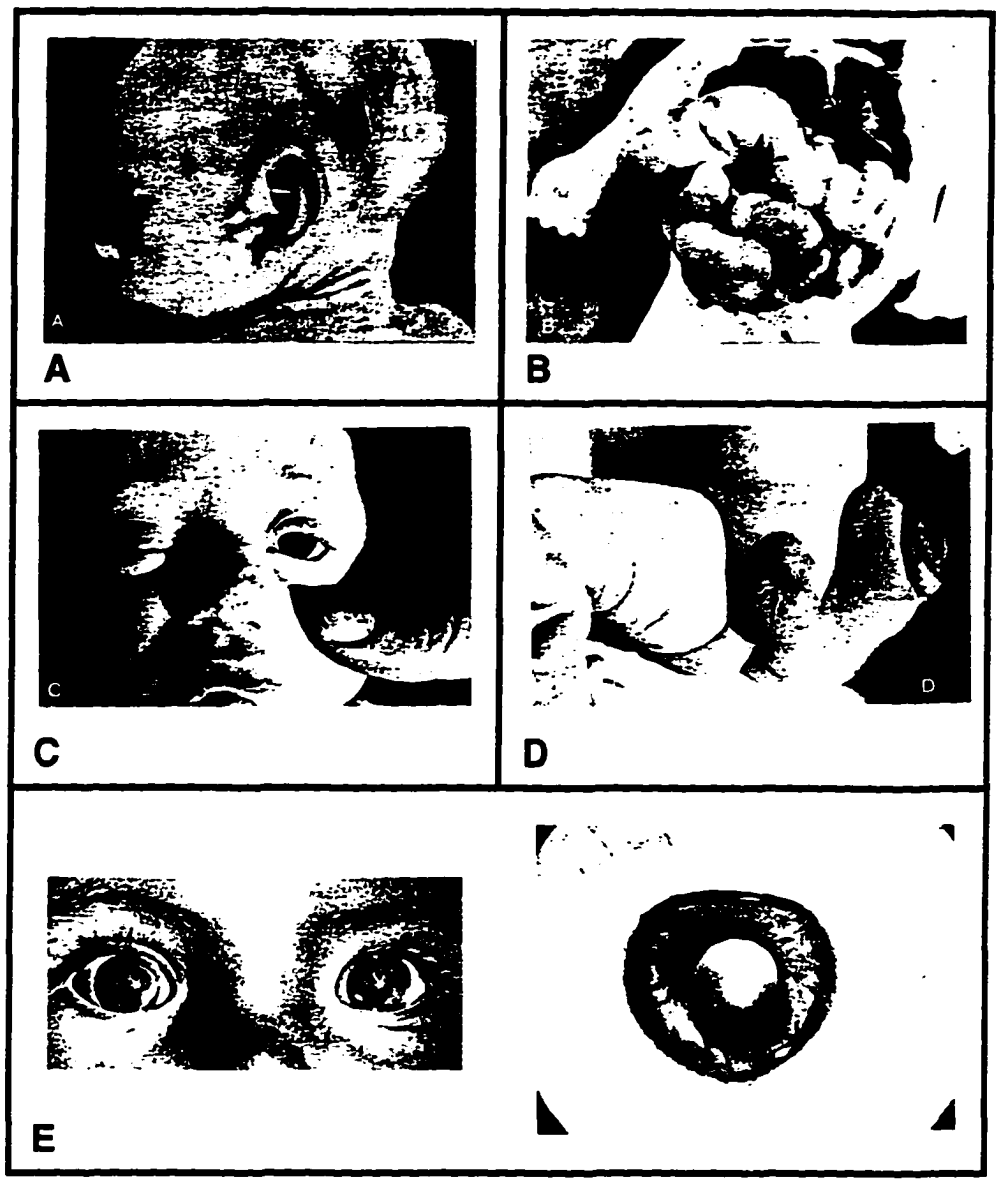


Figure 6: Patients with Cat Eye Syndrome: A) ear malformation, B) intestinal malrotation, C) ocular coloboma, and D) genital malformation, E) ocular coloboma.

lacked any extra chromosome upon karyotype analysis. Both parents of this patient displayed normal karyotypes. Similar results also were presented by Franklin[177].

In support of Schmidt, Fraccaro and Schachermann, others[177,178-187] reported studies that confirmed the presence of a supernumerary chromosome that, by chromosome banding, was slightly smaller (approximately half the size) of a G group chromosome. Gerald[186] and Ginsberg[187] noted that although ocular abnormalities were typical of the patients with trisomy 13-15, 18 and 21, cases of ocular abnormalities associated with additional chromosomes were essentially unknown. They described a patient with a small, extra chromosome and detailed similarities to patients with trisomy of D group chromosomes. A familial inheritance pattern also was demonstrated since the mother of the patient carried the extra chromosome. In several karyotype analyses of the patient in this study, the extra chromosome was characteristically dicentric and occasionally formed a small ring. The authors hypothesized that the additional chromosome could have resulted from any of four possibilities: 1) disruption or deletion of a chromosome; 2) disruption resulting in an isochrome of the short arm of a chromosome; 3) pericentric inversion within a small (probably G group) chromosome; or 4) translocation between two chromosomes. In their discussion, they preferred the first option because of the small size of the chromosome and the lack of visible satellites.

Weber[188] also reported findings of a patient with a small acrocentric chromosome which were consistent with the notion of a "partial trisomy syndrome." Here, observations using marker analysis and chromosome studies of the family showed neither an abnormal inheritance pattern, nor the origin of the extra chromosome. They suggested partial deletion as the formation mechanism of the extra chromosome and speculated that the loss of the long arm of a D or G group chromosome could result in a chromosome of similar size.

The origin of the supernumerary chromosome seen in CES was the source of argument for many years, with different investigators reporting different origins.

Pfeiffer[185] suggested that the Cat Eye Chromosome (CEC) was derived from chromosome 14, based on the observation by Jacobsen[189] of an extra chromosome in a boy with ocular colobomata. The extra chromosome in this case was derived from a translocation event involving chromosomes 14 and 13. Toomey[180], in Q banding studies, observed the CES marker chromosome with satellites and polymorphisms similar to the short arm of a D or G group chromosome. Further comparisons revealed similarity of the marker chromosome with chromosome 22 and based on previous phenotypic comparisons with trisomy 22 syndrome, they implicated chromosome 22 in the origin of the CEC. To this end, they hypothesized that a Robertsonian translocation (i.e. a reciprocal translocation between two acrocentric chromosomes with breaks on opposite sides of the centromeres and subsequent loss of one centromere[190]) between the p arm of 22 and the p arm of another D or G group chromosome (possibly 13) resulted in the marker chromosome found in CES. Other previous studies that had attempted to identify the CEC using banding techniques hinted at the involvement of chromosome 22, but produced limited evidence to support the hypothesis[175,179,191-193]. Toomey[180] also presented a comparison of clinical features of CES compared with trisomy 22 (see table 1).

Schinzel, et al.[194] reported 11 cases of CES, with all cases displaying an extra chromosome that appeared to contain pter->q11 of a normal chromosome. Further cytogenetic evidence suggested, as found by Toomey, that this CEC had derived from chromosome 22. Evidence to support this notion came from a family described by Buhler[183] with two descendants presenting the clinical features of CES. Buhler speculated this was the result of a translocation of chromosome 22 with some other (unidentified) chromosome that resulted in trisomy 22pter-q11. Another supporting study was reported by Niermeijer[195], describing a familial case of trisomy 22pter-q12 with CES characteristics. In further speculation, Schinzel[197-199] suggested that the CEC is probably derived from the centromeric portion of chromosome 22, in combination with

clinical feature	trisomy 22	CES
Mental retardation	+	+
Microcephaly	+	+
Hypertelorism	+	+
Antimongoloid slant	+	+
Beaked nose	+	--
Increased philtrum length	+	--
Cleft palate	+	--
Micrognathia	+	--
Low-set rotated ears	+	--
Preauricular tags or sinuses	+	+
Prominent antihelix	+	+
Deformed upper extremity	+	--
Hip abnormality	+	+
Congenital heart disease	+	+
Urogenital tract anomalies	+	+
Anal atresia	--	+
Coloboma	--	+
Microphthalmia	--	+

Table 1: Comparison of CES clinical features with trisomy 22 clinical features.

Adapted from reference 204.

the centromeric portion of the same 22 or another 22, leading to an inverted dicentric duplication: inv dup (22)(pter-q11::q11-pter). Different mechanisms for how this could be achieved were presented by Wisniewski[196].

In further studies, Schinzel[197-199] examined familial 11/22 translocations, familial trisomy 22q, mosaic trisomy 22 (reported in association with CES), and full trisomy 22. Results indicated that patients with unbalanced 11/22 translocations were trisomic for the centromeric segment of chromosome 22 and for the distal segment of 11 (approximately 11q23-qter). Trisomy 22q is rare (presumably due to lethality), with mosaic-trisomy 22 and full trisomy 22 even more obscure. The existence of full trisomy 22 in humans is still under scrutiny, but Schinzel reported that if it exists, it is only for a short time due to the lethality of the mutation. However, no evidence for lethality has been reported regarding trisomy 22pter-q11 as seen in the CEC. Other studies support the notion of an inv dup (22)(pter-q11::q11-pter) supernumerary chromosome in CES[200-203].

Guanti[204] disputed the claim that the CEC resulted from translocation of 22 by presenting several different arguments in favor of chromosome 13 as the origin of the CEC. First, if chromosome 22 is involved in the CEC, the resulting phenotype should be similar to trisomy 22. Guanti questioned the existence of trisomy 22 and argued that no living patients with inherited trisomy 22 have ever been reported, that carriers of a Robertsonian translocation of 22 frequently have recurrent spontaneous abortions, and that a syndrome has been described involving trisomy 11 which mimics the clinical features of CES and trisomy 22.

With respect to chromosome 13, Guanti pointed out the striking clinical similarities between patients with trisomy 13 (or an additional ring chromosome 13) and CES. However, Guanti speculated that if the CEC resulted from a translocation between 13 and 22, which would account for the presence of fluorescence characteristics of 22, their conclusions that supported a chromosome 13 origin of the CEC still would be valid. A proposed mechanism of translocation between chromosomes 13 and 22, as presented

by Guanti, is shown in figure 7 along with a partial karyotype, illustrating how a rearranged chromosome 13 could be mistaken for chromosome 22 in banding studies.

In response to Guanti's remarks regarding the origin of the CEC, Reiss, et al.[205] reported a patient with a 46, XY, 22q+ karyotype and clinical characteristics of CES. The abnormality of chromosome 22 was interpreted as a tandem duplication of the 22q11.1-q11.2 region. Here, the patient had striking similarities to a patient described by Schinzel. Although Reiss' patient did not have an extra marker chromosome, the duplicated 22q material yielded him trisomic for 22q11. Reiss suggested that the region 22q11 contains genetic material critical to phenotypic characteristics of CES. As a side note, the patient described here had normal levels of β-galactosidase-2, β-galactosidase B and arylsulfatase, the genes for which are located on the 22q11.2-qter region[206-209], leading to the conclusion that the region of duplication in this patient did not contain the above genes. McDermid[210] then suggested that gene dosage studies such as those reported by Reiss could be an important technique to study chromosomal diseases.

In other studies, Magenis[211] recently suggested genomic imprinting in the etiology of CES, because a comparison of Q stained chromosomes from the parents indicated that only the maternal chromosomes were involved in formation of the CEC. These studies are based on the findings of paternal preference in chromosome 15 deletions seen in Prader-Willi syndrome[212] and according to Olsen and Magenis[213], "the majority of de novo structural chromosomal aberrations are paternal in origin, while nondisjunctive events are most often maternal in origin." However, the supposition that CES has parental preference has not been proven conclusively by Magenis and additional studies are required.

Duncan, et al.[214] investigated the supernumerary chromosome in CES for potential breakpoints using marker analysis. *In situ* hybridization was performed using markers that map to the q11 region of chromosome 22 (D22S9, V-lambda and C-lambda). Results indicated that D22S9 was present in CES patients once on both normal

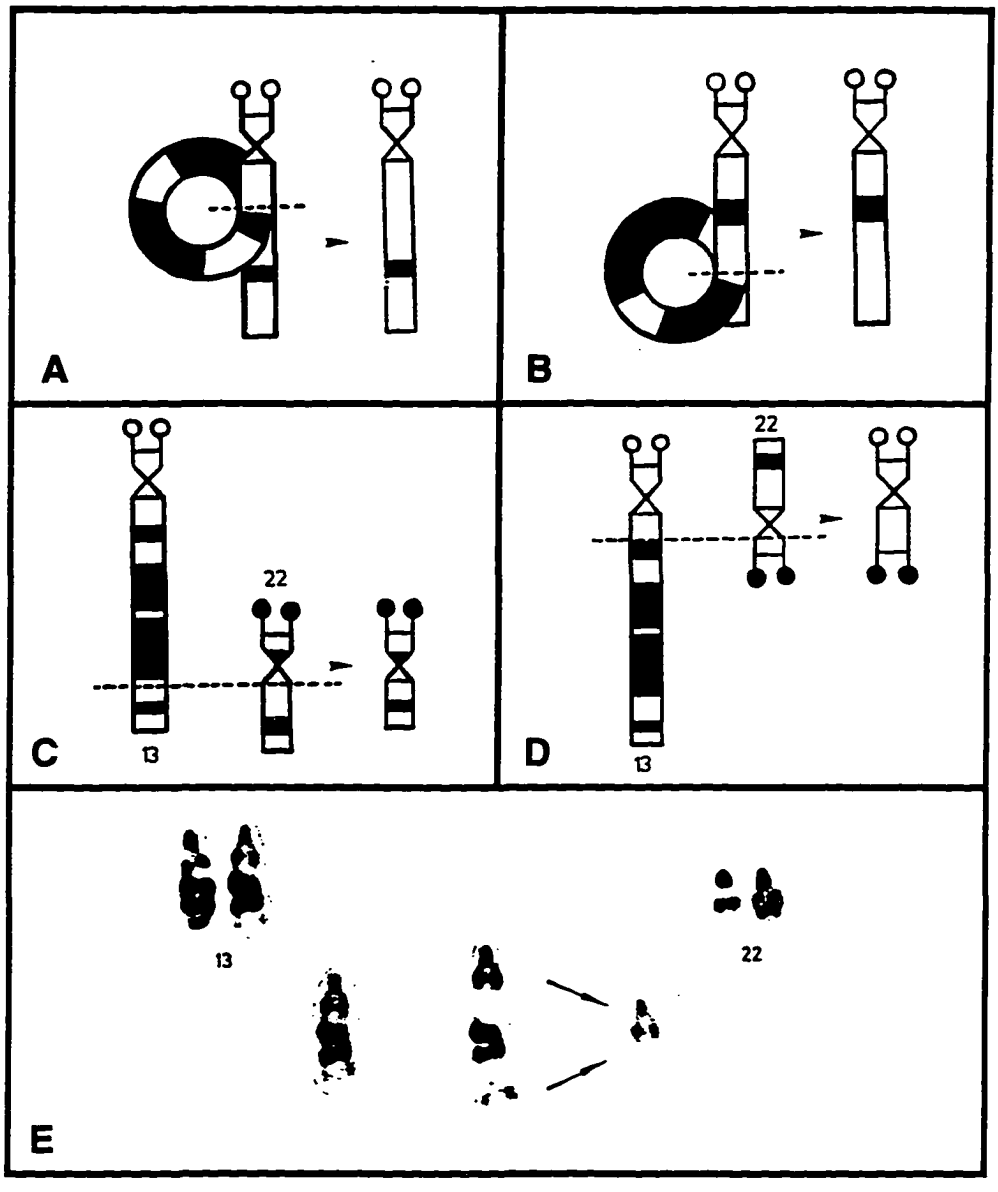


Figure 7: Schematic representation of possible rearrangements of chromosome 13. Interstitial deletion which gives rise to a chromosome 13 showing similar banding pattern as chromosome 22 (A,B); t(13;22) translocation giving rise to a marker chromosome with similar characteristics as chromosome 22 (C, D); partial karyotype illustrating how a rearranged chromosome 13 resembles a chromosome 22 (G banding). Adapted from reference 204.

copies of chromosome 22 but twice in the CES patients, therefore indicating 4 copies of 22q11 in the CES individuals. This observation supports the hypothesis put forth by Schinzel that the CEC is an inverted duplication of 22pter-q11, and also confirmed the increase in copy number as predicted by Southern blot analysis with D22S9 in these same patients. Other recent data has revealed that the breakpoint in generation of the CES chromosome was proximal to the C-lambda and V-lambda regions. Further studies from McDermid's laboratory[215] demonstrate similar results and indicate that some patients have 4 copies of D22S9, while others have 3 copies that are illuminated by this probe. Therefore, McDermid rationalized that CES results from overexpression of the genes present in this duplicated region.

Additional studies performed by Ward[216] confirmed the conclusions drawn by McDermid and Duncan. Here, a patient had 4 copies of the D22S9 marker (i.e., two copies from normal chromosome 22 and 2 copies on the CEC). Luleci[217] and Liehr[203] also reported similar results, and further mapping studies using D22S9, D22S43, D22S57, D22S36 and D22S75, all of which map to 22q11.2, with D22S36 and D22S75 being the most distal (See figure 8). Here, marker analysis was used to distinguish between the breakpoint regions of CES and the gene deletion regions present in DiGeorge Syndrome[218-225]. These results indicate that the DiGeorge critical region (DGCR) is distinct from the CES duplication region with the distal boundary of CES (D22S36) located proximal to the DGCR. Other results in this report also confirmed the presence of 4 copies of 22q11, and indicated that the size of the duplication did not correlate with the severity of CES.

In the latter part of this dissertation research, a series of eight BACs (bacterial artificial chromosomes) which map to the proposed breakpoints in CES were obtained by screening a human BAC library with markers mapping to the CES breakpoint region. The genomic sequence and analysis of one of these BACs is reported in this dissertation to provide direct evidence for the features present in the CES breakpoint region. In

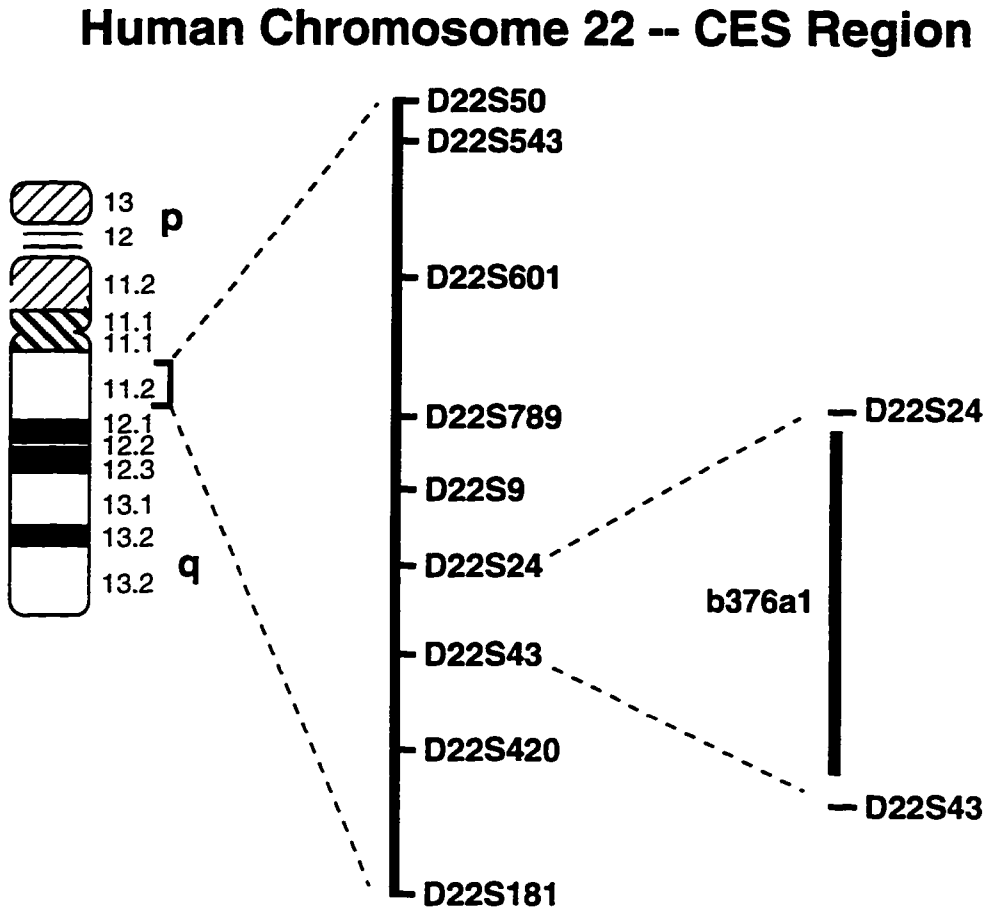


Figure 8: Cat Eye Syndrome critical region and relevant markers on human chromosome 22.

addition, these studies have served as a pilot project for the novel strategy termed "Sequence Scanning," as detailed below.

III. Multiple Endocrine Neoplasia, type I (Wermer's Syndrome)

Multiple endocrine neoplasia. type I (MEN-1 or Wermer's Syndrome)[226.227] is a disease in which multiple benign or malignant tumors are present in the anterior pituitary gland, the pancreatic islets and the parathyroid glands. Other tumors such as adrenal cortical, carcinoid and lipomatous tumors also have been observed in MEN-1. MEN-1 is considered an autosomal dominant inherited disorder, and according to Calendar, et al., more than 80% of individuals with a family history of MEN-1 are affected by the disease by the fifth decade of life[228]. Many different biochemical changes result from the presence of MEN-1 tumors, and include increased levels of parathyroid hormone, growth hormone and vasoactive intestinal peptide production (see table 2)[229].

Mapping studies performed by several groups[230-233] have localized the MEN-1 locus to chromosome 11q13. Recently, detailed tumor deletion mapping studies using restriction fragment length polymorphisms (RFLP) and repetitive (microsatellite) probes, have localized the MEN-1 locus to a region telomeric to the PYGM locus (which encodes human muscle glycogen phosphorylase) and centromeric to the genetic marker D11S146. Since no mitotic recombination was observed between the PYGM locus and the suspected disease locus, Calendar, et al. hypothesize that the predicted MEN-1 locus lies in close proximity to PYGM (see figure 9)[228].

The initial studies of MEN-1 also revealed the mechanism of this disease. Studies focusing on the genetic rearrangements in MEN-1 tumors indicate loss of function suggesting that the MEN-1 gene may be a tumor suppressor gene[234]. Within the 11q13 region localized by the mapping studies, several candidate MEN-1 genes have been identified, including PRAD1(also called BCL1 or CCND1)[229], FAU[229], ZFM1[229],

MEN-1 tumors	Biochemical Features
Parathyroids	**Hypercalcaemia and ↑ PTH**
Pancreatic islets:	
Gastrinoma	**↑Gastrin and ↑ basal gastric output**
Insulinoma	**Hypogylcaemia and ↑ insulin**
Glucagonoma	**Glucose intolerance and ↑ glucagon**
Vipoma	**↑VIP and WDHA**
PPoma	**↑ PP**
Pituitary (anterior):	
Prolactinoma	**Hyperprolactinaemia**
GH-secreting	**↑GH**
ACTH-secreting	**Hypercorisolaemia and ↑ ACTH**
Non-functioning	**Nil**
Associated Tumors:	
Adrenal cortical	**Hypercortisolaemia or primary hyperaldosteronism**
Carcinoid	**↑ 5-HIAA**
Lipoma	**Nil**

Table 2: Biochemical changes resulting from MEN-1 tumor presence. Abbreviations: ↑, increased; PTH, parathyroid hormone; VIP, vasoactive intestinal peptide; WDHA, watery diarrhea, hypocalcemia and achlorhydria; PP, pancreatic polypeptide; GH, growth hormone; ACTH, adrenocorticotrophin; 5-HIAA, 5-hydroxyindoleacetic acid. Adapted from reference 229.

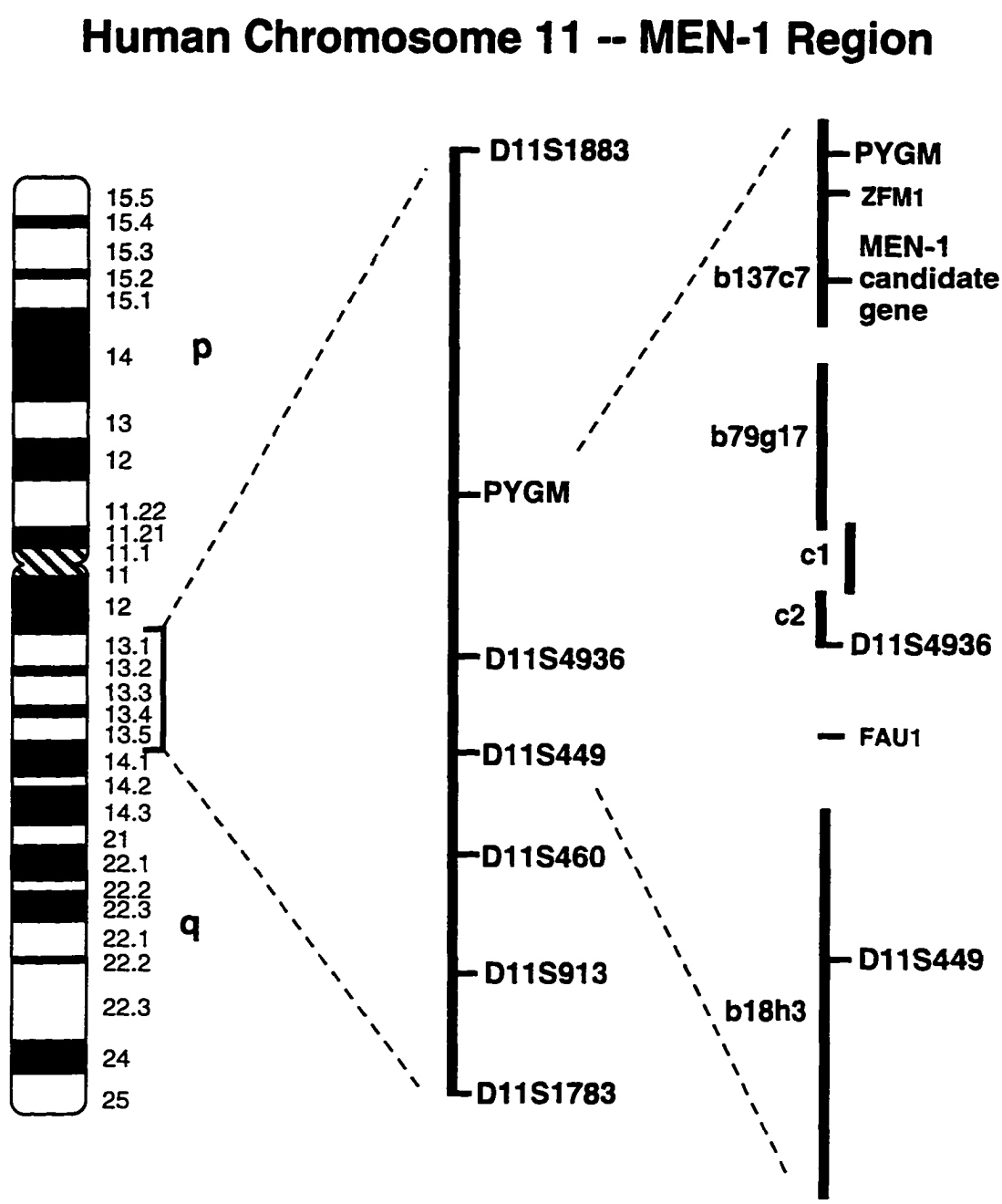


Figure 9: MEN-1 gene region and relevant markers on human chromosome 11. Adapted from Chandrasekharappa, et al. "Positional Cloning of the Gene for Multiple Endocrine Neoplasia I," In press.

4F2HC (or MDU1)[229], INT2[229], HSTF1 (or HST1)[229], PP1α [229], PCLβ3[234. 235], and FKBP2[235].

Recent studies of two pituitary adenomas revealed the presence of a DNA rearrangement involving the first of three exons of the parathyroid hormone (PTH) gene normally on 11p15 juxtaposed with a non-PTH gene region from 11q13. It was found that this non-PTH region, termed the D11S287E region, contains DNA sequence that is conserved in different species and that is expressed not only in the parathyroid adenomas, but in normal parathyroids as well[236.237]. This potential MEN-1 gene was termed PRAD1 for parathyroid adenomatosis, type 1[238] and Arnold, et al.[238] demonstrated that it is overexpressed in tumors with 11q13 rearrangements. The cDNA for PRAD1 was isolated and the predicted protein (of 295 amino acids) demonstrated similarities to the cyclin family of proteins.

Cyclins are a family of proteins involved in cell cycle regulation by activation of cyclin dependent kinases (cdks)[239-241]. In mammals, there are five types of cyclins, classified A, B, C, D, and E, with the D cyclins further subcategorized as D1, D2, and D3. As seen in figure 10, different cyclins are expressed at different stages in the cell cycle. The D class of cyclins regulate progression from G1 phase to the S phase, however, D cyclins behave as end points of mitogenic pathways instead of as an integral part of the cell cycle. D cyclin activity is mediated by growth factors or other nuclear proteins. Further analysis revealed that cyclin D1 and PRAD1 have identical primary sequence, and that chromosomal rearrangements involving the cyclin D1 gene (PRAD1) have been reported in B-cell lymphomas (BCL1)[242.243]. Further linkage studies revealed that the PRAD1 gene was not the gene for MEN-1[238], although this gene plays an important role in cell cycle regulation and potentially, tumorigenesis.

The FAU gene (Finkel-Biskis-Reilly murine sarcoma virus [FBR-MuSv] associated ubiquitously expressed gene) is a second potential MEN-1 gene. The FAU

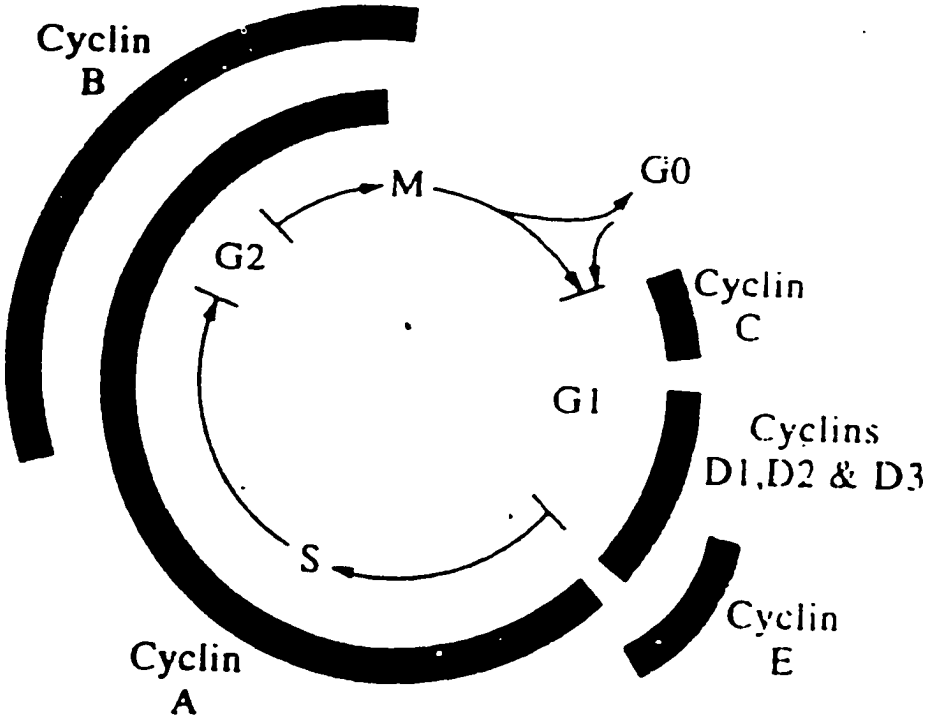


Figure 10: Cell cycle and the different stages of cyclin expression[229]

gene in mouse, when interacting with the viral antisense FAU gene, leads to the disruption of the constitutive expression and results in sarcoma. This correlation between the FAU gene suppression and presence of a sarcoma indicates that the FAU gene could potentially behave as a tumor suppressor gene. However, as with PRAD1, mapping studies place the FAU gene outside the predicted locus for MEN-1[244]. Further evidence reveals that MEN-1 patients do not exhibit mutations or other alterations within the FAU gene region[245].

A third potential MEN-1 gene is XFM1[229], a gene for a zinc finger containing protein in the MEN-1 locus that consists of 14 exons and encodes a 623 amino acid protein which also contains a nuclear transport domain, a metal binding motif and glutamine- and proline-rich regions. This gene is expressed in endocrine organs such as adrenal, thyroid and pancreas, and results in two detectable expressed transcripts. The relationship of ZFM1 with MEN-1 remains to be established.

A fourth potential MEN-1 gene is 4F2HC and it encodes a heavy chain protein of a cell surface glycoprotein whose expression is increased in actively proliferating cells. Because of this increased expression, 4F2 has been suggested to play a role in the G_0 to G_1 transition of the cell cycle. The gene for 4F2 (also called MDU1 for monoclonal Duke University antibody 1) has nine exons and mapping studies have localized it to chromosome 11[246]. Although this gene could be a candidate for tumorigenesis in some capacity based on the role played in the cell cycle regulation and localization at 11q13[247], it maps outside the target region for MEN-1[247]. Therefore, it is unlikely that the 4F2HC gene is the MEN-1 gene.

Two protooncogenes also are located in 11q13. INT2, the human homologue of the murine mammary tumor virus integration site 2, and HSTF1, the heparin secretory transforming factor type 1, formerly termed HST1 for human stomach transforming factor 1, both encode basic fibroblast growth factor (bFGF)- related proteins which could play a role as circulating mitogenic factors[248,249]. However, mapping analysis and

linkage studies ruled out INT2 and HSTF1 as possible candidates for the MEN-1 gene as they are located too far from the mapped MEN-1 locus.

The final gene mapped to 11q13 is PP1α, the protein phosphatase 1 catalytic subunit-encoding cDNA. Protein kinases and phosphatases historically have played a role in cellular regulation and more specifically, signal transduction. For example, in mammals, the conversion of the retinoblastoma protein to its active form requires PP1α-mediated dephosphorylation[250,251] and PP1α therefore plays a critical role in cell proliferation. Recent studies have localized the PP1α gene to the MEN-1 locus between PYGM and D11S97 on chromosome 11q13, and therefore, the PP1α gene now is the leading candidate for the MEN-1 gene.

The results of sequencing a 300 kilobase BAC in the 11q13 region are presented in the results section of this dissertation in an attempt to locate additional genes in the MEN-1 region and to provide the sequences needed for the further mapping and familial studies necessary to pinpoint the MEN-1 gene.

IV. Large scale DNA sequencing

A. Introduction to DNA

1. DNA and inheritance

For centuries, many explanations have been put forth to explain the inheritance of parental characteristics by offspring. Early theories included Aristotle's theory of pangenesis which stated that semen developed in and from all parts of the body by traveling through the blood to the testicles. Because the semen formed in all parts of the body, the characteristic traits of each body part were effectively transferred to the progeny. August Weismann presented a competing theory in the early 1800s, called the germ plasm theory. Weismann's germ plasm theory stated that there was a fundamental difference between the germ plasm (the sex cells and their progenitors) and the somatoplasm (the remainder of the body cells) and that the somatoplasm existed for the

sole purpose of propagation and protection of the germ plasm, which carried the characteristic traits of the organism from parent to offspring. Weismann supported his theory by removing the tails of mice before mating. Results showed that the tail removal did not affect the tail length of the progeny[252].

All such attempts to define heredity focused on the overall similarities between progeny and the parents, albeit still unable to unlock the complex transmission of heritable traits. It was not until the late nineteenth century (1865), when Gregor Mendel[253] focused his study on easily defined traits exhibited by homozygous (true breeding) parental strains of the garden pea (*Pisum sativum*), that the link between the outward appearance of an organism (phenotype) and the genetic makeup of an organism (genotype) was realized. By proposing that "particulate factors" were the carriers of heredity, Mendel formulated his two laws of heredity which include the genetic concepts of dominance, recessivity, segregation and independent assortment[254].

Walter Sutton, building on Mendel's "particulate factors," was the first to relate chromosomal behavior with genetics in his chromosomal inheritance theory. He examined chromosomal movement and noted that the sex cells which formed during meiosis received only one chromosome from each of eleven pairs of chromosomes in grasshoppers[255]. This chromosomal segregation perfectly matched the segregation described by Mendel which led Sutton to the conclusion that Mendel's "particulate factors" must be located directly on the chromosomes. In 1910, Johannsen[256] called these factors genes. Sir Archibald Garrod, in his studies of the inborn errors of metabolism, suggested that these illnesses resulted from a lack of the required enzymes, and proposed that the genes controlled the synthesis of these enzymes[254,257,258]. In 1941, Beadle and Tatum[259] introduced a red bread mold (*Neurospora*) as a genetic model system to study metabolism, since the organism requires only sugar, inorganic salts and biotin to survive. Using X-ray irradiation followed by analysis on several types of media, each containing an added nutritional supplement, Beadle and Tatum ascertained

which gene had been modified by irradiation. Based on this study, they coined the phrase "one gene-one enzyme" and confirmed Garrod's hypothesis. In modern terms, it is more useful to expand Beadle and Tatum's phrase to "one gene-one protein," although it was several decades later that the actual nature of a gene was characterized.

Studies by Fred Griffith on two strains of *Pneumococcus* led to the first evidence for DNA as the genetic material[260]. A pathogenic strain of Pneumococcus (III S), which is able to synthesize a capsular polysaccharide which forms a mucus coat around the bacterium and resists the host defense system, and a mutant strain of Pneumococcus (II R), which had lost the ability to synthesize the capsular polysaccharide, were used. Griffith demonstrated that coinjection of heat killed III S and II R into mice resulted in pneumonic infection and death, whereas injection of only III S or II R individually did not. Furthermore, Griffith was surprised to retrieve virulent III S strain bacteria from the deceased mice. In 1928, Griffith presented his transformation hypothesis that some "principle" was transferred from the heat-killed III S strain to the avirulent II R strain, transforming it such that it could produce the capsular polysaccharide, hence, becoming virulent.

Griffith's experiments were further confirmed by Avery, MacLeod and McCarty[261] in 1944, who purified the transforming "principle" from extracts and found it to be DNA. Further studies with deoxyribonuclease (DNase) confirmed their theory by demonstrating that DNase could destroy the transforming ability of the DNA. This work should have focused all attention on DNA, but because proteins were still considered by many to be the carriers of genetic information, the structure of DNA was uncharacterized, and the genetics of bacterial systems were unclear, DNA remained in the background.

It was not until 1952, that Hershey and Chase confirmed that DNA was the cell component involved in transfer of genetic information by infecting bacteria with radioactively-labeled T2 bacteriophage[262]. The protein coat of the phage was labeled with ^{35}S and the phage DNA was labeled with ^{32}P. Analysis revealed that the empty

phage coats retained the ^{35}S label and the infected bacteria contained the ^{32}P. Upon lysis of the infected bacterial cells, it was found that the progeny phage contained the ^{32}P label indicating that the DNA is directly involved in the transfer of genetic material.

These studies gave credence to the earlier works of Friedrich Miescher[263] who, in the late 1800's, determined that a new cellular organic phosphate compound was located in the nucleus of cells. He had termed this substance nuclein and established that it was characterized as a mixture of DNA and protein, later termed chromatin. Through a series of studies in the first half of this century, the DNA polynucleotide was determined to be a chain of nucleotides, with each nucleotide consisting of three major components: (1) a purine (adenine, guanadine) or pyrimidine (cytosine, thymine, uracil) base (see figure 11), linked by its 1 or 9 ring nitrogen atom, respectively, in an N-glycosidic bond to (2) a 5 carbon sugar in a ring formation (pentose) which is ribose in RNA and 2-deoxyribose in DNA (differing only by the presence or absence of a hydroxyl group at position 2 in the ring) and (3) a phosphate group which is esterified to carbon 5 of the sugar. The DNA polynucleotide is an unbranched polymer bound covalently between the 5' phosphate of one nucleotide to the 3' hydroxyl of the sugar on the adjacent nucleotide. The degrees of rotation freedom are limited to two in DNA, one in the phosphodiester linkage and the other in the glycosyl linkage between the base and the sugar[264]. In 1951, Chargaff, et al.[265], determined that double-stranded DNA consists of equal amounts of adenine (A) and thymine (T) as well as equal amounts of guanosine (G) and cytosine (C). Furthermore, he showed that the molar ratios ([A]+[T]/[G]+[C]) differed among organisms, a finding which suggested that DNA molecules might be more complex than originally considered. Chargaff hypothesized that the variable base compositions reflected differences in the base sequences among organisms, which somewhat revealed the complexity and diversity of genetic material.

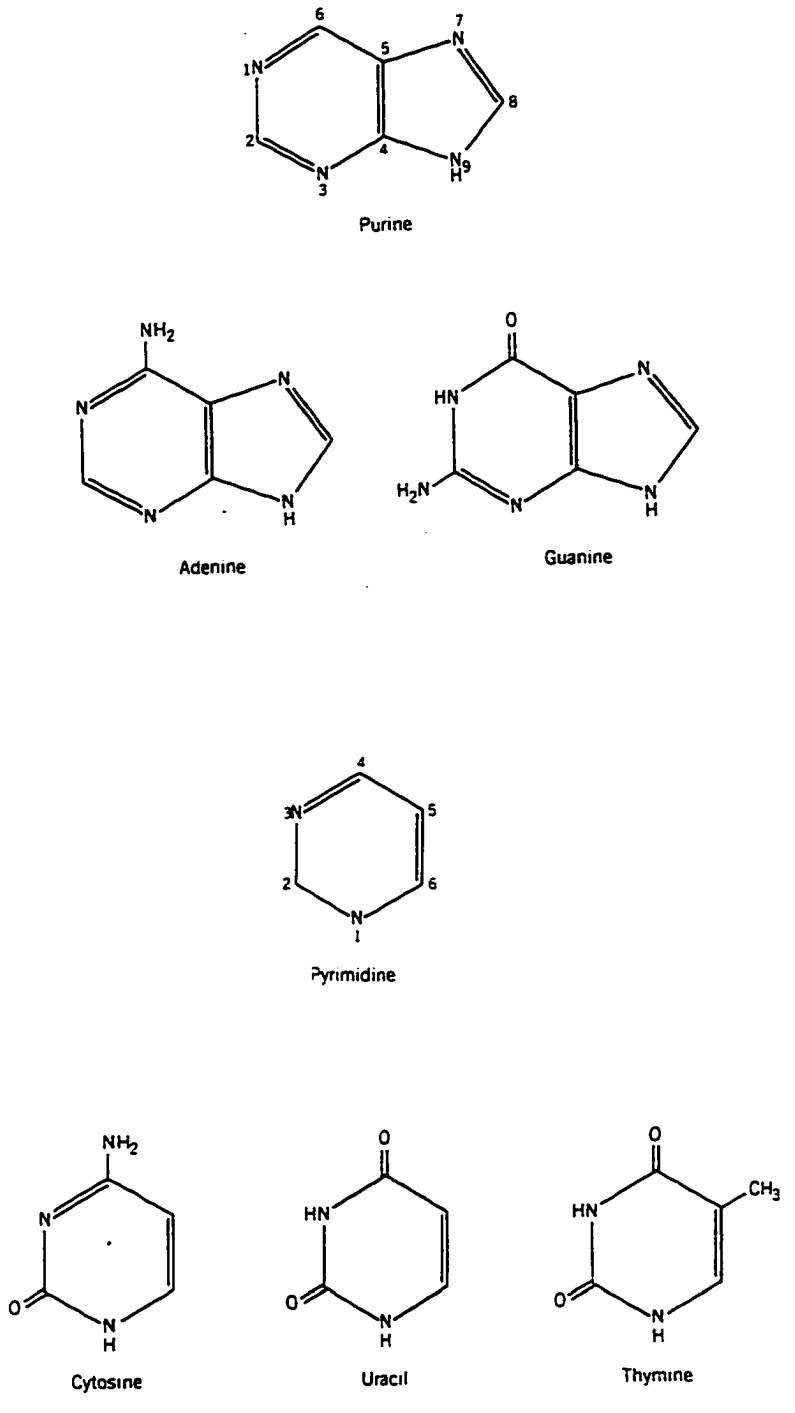


Figure 11: Deoxyribonucleotide structures

2. DNA structure

In 1953, Watson and Crick[266,267] reported the structure of DNA as a hydrogen bonded, double helix (see figure 12), based on three major pieces of evidence: (1) X-ray diffraction data obtained from Maurice Wilkins[268] and Rosalyn Franklin[269] showed that DNA had the form of a regular helix with a complete turn every 34 Angstroms (3.4 nm), with a diameter of 20 angstroms (2 nm). With a distance between nucleotides of 3.4 angstroms, there were 10 bases per turn of the helix. (2) The high density of DNA suggested that 2 chains must be present, and with the X-ray diffraction pattern revealing that the DNA had a constant diameter, Watson and Crick proposed that this could only be achieved if the bases faced inward and a purine always was hydrogen bonded with a pyrimidine. (3) Chargaff's Rules which described that the proportion of G and C is always the same, as is the proportion of A and T, regardless of the actual amount of the base.

To adhere to these three pieces of evidence, the resulting structure must have two DNA polymers coiled together in an antiparallel fashion with the nitrogenous bases facing inward. The two strands are hydrogen bonded between A and T (two hydrogen bonds) and G and C (three hydrogen bonds) generating what is referred to as Watson-Crick base pairing. This structure fulfills the rules originally proposed by Chargaff and formed the basis for a new era of research into the heredity material first described over a century ago.

The crystal structure studied by Watson and Crick is what is now described as B-form, a right handed helix with the distance criteria as listed above. B-form DNA has a major and a minor groove with the surfaces of the hydrogen bonded nitrogenous bases perpendicular to the helix axis[266]. The flat surfaces of the hydrogen bonded bases allows for interaction of the aromatic pi electron clouds, thereby diminishing the electrostatic repulsion expected between the negatively charged phosphate backbones of the helix. This "base stacking" stabilizes the helix. Other forms of DNA, such as the more tightly

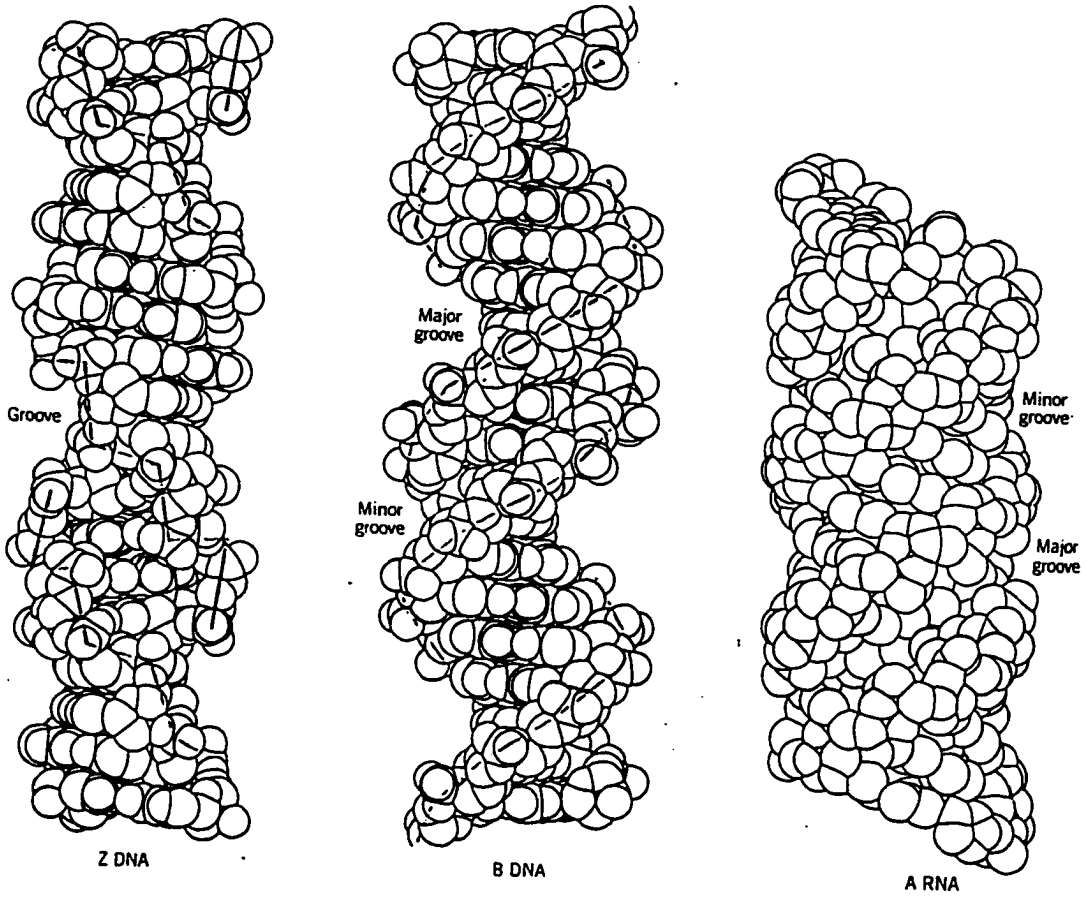


Figure 12: A-form, B-form and Z-form structures[401].

wound right handed helical A-form and the slightly looser C-form also have been described[271].

An alternate form of DNA, Z-DNA, is a left handed double helix with only one deep groove that is similar to the minor groove of B-DNA[272, 273] (see figure 12). Z-DNA contains 12 base pairs per turn with a diameter of approximately 18 angstroms, making this structure much more compact than that seen in B-DNA. Another difference between the two forms lies in the conformation of the nucleotides. In B-DNA, the nucleotides have the *anti* conformation and a C2' endo pucker of the deoxyribose ring, whereas Z-DNA has alternating *anti* and *syn* conformations and a C3' endo pucker of the deoxyribose sugar[273].

Although B-DNA is the most common form, some sequences favor the formation of Z-DNA, such as alternating stretches of purines and pyrimidines as found in promoter regions of housekeeping genes[274]. The formation of the lower energy Z-DNA structure as a result of these alternating purine/pyrimidine elements is induced by negative supercoiling[273,275,276]. Since Z-DNA has been reported in regions of SV40 virus that regulates transcription and replication, Z-DNA is thought to function as an effector of transcriptional activation[277,278]. Miller, et al. reported that Z-DNA can form nucleosomes and that B-DNA nucleosomes can be converted to Z-DNA nucleosomes[279]. This conversion from B-DNA to Z-DNA could change the superhelical density of a DNA segment, altering the gene expression over a large region of DNA[273,280]. Therefore, Z-DNA formation could potentially affect nucleosome phasing, domain activation and local chromatin organization[276,280,281].

3. Gene expression

With the structure of DNA known, the next step in the DNA structure-function question was to find a molecular basis for the biochemical observations of heredity in Garrod, Beadle and Tatum's one gene-one enzyme hypothesis. More specifically, how

is the information contained in DNA used by the cell to carry out metabolic functions and how is this process regulated?

Francis Crick began answering this question in 1957 when he proposed his central dogma of molecular biology. With the assumption that DNA sequence and protein sequence are colinear, Crick hypothesized that the genetic information is stored in DNA, passed on via RNA which acts as an intermediate translator, and translated into the functional protein molecule (see figure 13). It now is widely accepted that the process by which RNA is produced from DNA is termed transcription, and the production of proteins from mRNA is termed translation.

In eukaryotes, transcription of genes which eventually will be translated is facilitated by the presence of specific promoter elements (i.e. specific consensus sequences in the DNA) which are recognized by RNA polymerase II. One such promoter element is the Goldberg-Hogness box, or TATA box, with a consensus sequence of TATAAT, which is located 25-30 base pairs upstream from the transcription start site. The TATA box directs the correct positioning of the RNA polymerase II for transcription initiation and specifies the 5' cap termini in the mature mRNA[282,283]. The TATA box in eukaryotic transcription appears to be similar to the Pribnow box in prokaryote transcription, but differs in that the Pribnow box is essential for prokaryotic transcription. Deletion of the TATA box does not abolish *in vivo* transcription, but is absolutely required for eukaryotic transcription *in vitro* [284].

Further upstream from the transcription start site is a region which can contain one or more upstream regulatory elements such as the CCAAT box[284,285] or the Sp1 consensus binding site GGGCGG[283,284,286]. Levine, et al. reported that methylation at the promoter region was associated with transcriptional inactivity[287]. However, further studies conducted by Hoeller, et al. demonstrated that transcription factor Sp1 could initiate transcription when methylated C residues were present in the Sp1 binding site[288].

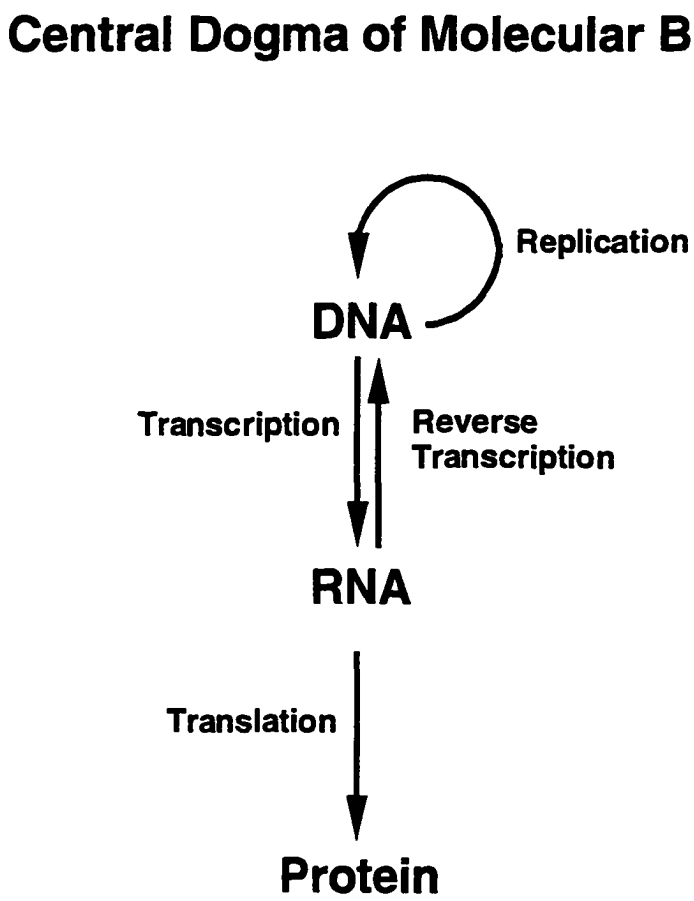


Figure 13: Central dogma of molecular biology

DNA binding proteins play a significant role in both positive and negative control of transcriptional initiation[289]. For example, prokaryotic heat shock genes are positively regulated at the transcriptional level by the DNA binding protein sigma 32, whereas negative regulation is accomplished in SV40 when T antigen binds to the region of SV40 DNA which includes the promoter for the early transcriptional unit[289]. Additional regulatory elements, called enhancers, affect mRNA synthesis from greater than 1000 base pairs from the promoter region. Enhancer regions are positioned upstream, downstream, or within the transcribed region of the gene they regulate[286,274]. Furthermore, enhancer regions have been noted to function even when the 5'-3' orientation is reversed[286].

Whereas the 5' terminus of a mRNA corresponds to the transcription initiation site, transcription does not terminate at the poly A site, but rather at some distance downstream[289,290]. A consensus sequence (AAUAAA) is found 10-30 nucleotides upstream from the poly A addition site for most transcripts[291]. This consensus sequence directs an endonucleolytic cleavage reaction on the nascent RNA transcript, creating the 3' end to which a poly A tail of approximately 200 nucleotides is added[284]. Deletion of this site results in unstable mRNA transcripts, and although polyadenylation still occurs, the cleavage results at both normal and abnormal sites[292].

Because transcription terminates at a point downstream of the polyadenylation consensus sequence, it has been suggested that additional downstream elements are responsible for directing the polyadenylation[284]. However, since eukaryotic transcription termination still is not well characterized, the signals for polyadenylation and termination still are not well defined.

Unlike prokaryotic transcription, eukaryotic transcription produces a pre-mRNA which contains coding sequences (exons) and intervening non-coding sequences (introns). The non-coding sequences are removed from the mRNA through a process called splicing, during which the pre-mRNA is cleaved at a G residue in the 5' donor

consensus splice site (C/A)AG/GT(A/G)AGT, where the underlined sequence is strictly conserved. This G residue at the end of an intron reacts with another base containing region (usually containing an adenosine) that is upstream of the 3' acceptor splice site ((T/C)$_n$N(C/T)AG/G) to form a 5' to 2' phosphodiester bond[284,289,293]. The intermediate is termed a lariat structure and is followed by release of the intron and ligation of the two exons[284].

snRNPs (small nuclear ribonucleoprotein particles) are components of the spliceosome complex where the process of site recognition, lariat formation and cleavage of pre-mRNA occurs. Other components include the pre-mRNA and heterogenous nuclear RNP proteins (or hnRNPs) which associate with the pre-mRNA[294]. It now is known that snRNPs play a role in splicing since it was observed that the sequence of the 5' end of U1 snRNP RNA is complementary to the 5' consensus splice site[295-297] and similarly, that U5 snRNP RNA associates with the 3' consensus splice site[294,298]. Further evidence to link snRNPs with splice interactions suggests an interaction between U2, U5 and U6 snRNP and the branchpoint sequence during lariat formation[299-301].

4. Chromosome organization

The three billion base pairs of the human genome are organized and packaged into a set of 46 chromosomes in a haploid cell; two copies of chromosomes 1-22 and the sex chromosomes XX and XY. The basic structural unit of a chromosome is chromatin[302]. Chromatin consists of densely packed nucleosomes wound in a helical solenoid shape with six nucleosomes per turn of the helix[303]. The nucleosome is a protein octamer of histones (two of each protein: H2A, H2B, H3 and H4) around which a continuous strand of double helical DNA is wound two times, with roughly 200 base pairs of DNA wrapped around each octamer[304]. A fifth histone protein, H1, binds to the DNA between nucleosomes and facilitates chromatin packaging[305,306]. This packaging is believed to be regulated by phosphorylation of histone H1 prior to mitosis, and dephosphorylated following mitosis. It is thought that this covalent modification

regulates the capacity to package DNA into chromatin. As a result of this covalent modification, chromatin is found in different forms depending on the stage of cell division[307]. The compact form is known as heterochromatin and studies indicate that this form is a region of relatively inactive RNA synthesis. Heterochromatin can be further subdivided into constitutive heterochromatin (chromosomes or parts of chromosomes which are not expressed such as the highly repetitive sequences found near the centromeres[307]) or facultative heterochromatin (cell specific and is believed to play a role in transcriptional inactivation of large regions of the chromosome). Euchromatin is a much less compact form of chromatin and is typically equated with active RNA synthesis. These regions have been visualized using an electron microscope and are called "chromosomal puffs" [304].

In females, two copies of the X chromosome are present although only one is transcriptionally active. The inactive chromosome always is present as condensed heterochromatin and is called a Barr body[308].

B. Recombinant DNA

Since the first site-specific restriction endonuclease was isolated from *Haemophilus influenzae* in 1970 [309], molecular biology has moved in the direction of recombinant DNA technology by developing techniques for DNA characterization, cloning and genetic mapping which are central to modern DNA research.

1. Restriction endonucleases

Restriction endonucleases are enzymes which recognize a specific nucleotide sequence and cleave double-stranded DNA by hydrolysis of the phosphodiester bond in the backbone of DNA. Three types of restriction endonucleases have been characterized: types I, II and III. Types I and III are similar in that both recognize sites that are variable distances away from the site of cleavage. Type I endonucleases require ATP for the cleavage, whereas type III does not. Type II enzymes, the most commonly used in

molecular biology, recognize a palindromic sequence present in double-stranded DNA, cleave within the recognition site and do not require ATP for cleavage. Type II restriction endonucleases cleave both strands of the double-stranded DNA substrate and the resulting fragments have either a 5' overhang, a 3' overhang, or blunt ends.

Restriction endonucleases are found to occur naturally in many bacteria and play a major role in the bacterial host defense mechanism. Host DNA is methylated at specific residues (typically A or C), which allows it to be distinguished from invading DNA. Because the host restriction enzymes are unable to cleave methylated DNA, the genomic DNA remains intact while the invading bacteriophage DNA is destroyed. Many restriction enzymes have been isolated from a myriad of bacterial organisms and because they are commercially available, they have become valuable tools that are widely used in molecular biology.

2. DNA cloning

The discovery of restriction enzymes paved the way for the development of recombinant DNA technology. DNA cloning, which requires ligating foreign DNA into a cloning vector, now is one of the basic techniques in a modern molecular biology laboratory. Once foreign DNA is ligated into a vector, the recombinant DNA molecule is transformed into a bacterial host where it is replicated by the host enzymes. A wide variety of viral and plasmid vectors have been engineered to contain a variety of features such as selection mechanisms and cloning sites.

In 1977, Joachim Messing introduced the viral M13-derived cloning system[310-313]. These vectors were derived from the single-stranded genomic DNA of M13 filamentous bacteriophage. It is believed that this bacteriophage infects a host bacteria via an external F pilus on the host cell. The single-stranded M13 DNA and a pilot protein enter the cell where the pilot protein directs the synthesis of the complementary strand of the viral DNA using the host replicating machinery. This circular, double-stranded moiety (called replicative form, or RF) can be isolated from the cells and used as a

cloning vector. Foreign DNA can be ligated into this replicative form that has been cleaved with a restriction endonuclease. This chimeric M13 molecule can be introduced into bacterial cells which then provide the replication machinery to produce progeny phage and infection of other bacterial cells. When plated on an agar plate, cells containing M13 virus grow slower and easily are identified against a background of uninfected cells. The lac Z' coding region of the *E. coli* lac operon was included in the engineering of the M13 vector system because it contains the gene and regulatory elements for the enzyme β-galactosidase which cleaves lactose. Messing also introduced a polycloning site in the M13 vectors into the β-galactosidase gene without altering the function of the gene. Thus, when foreign DNA is inserted into sites in the polycloning site, the insert causes expression of a dysfunctional lac Z' protein. In the presence of isopropyl-β-D-thiogalactopyranoside (IPTG), a lac inducer, and 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (Xgal), a chromogenic substrate for β-galactosidase, cells harboring an intact gene for β-galactosidase will produce a functional protein which cleaves Xgal resulting in blue plaques, whereas those with inserted DNA will produce dysfuntional β-galactosidase and resulting plaques will be colorless. This color selection mechanism allows for differentiation between cells with only vector M13 DNA and cells with M13 vector containing insert DNA.

Plasmids are double-stranded, extrachromosomal DNA naturally found in some bacterial strains. Naturally occurring plasmids typically contain important genes for either antibiotic resistance or bacterial conjugation which give the bacteria a selective advantage under certain growth conditions. As in M13 vectors, pUC-based plasmid vectors also were engineered by Messing to contain the polycloning site, the lac operon for color selection and in addition, the gene for ampicillin resistance[3,14]. These engineered vectors are part of a dual selection system required in plasmid-based cloning. The first selection is the presence of genes on the plasmid which confer antibiotic resistance. When plated on antibiotic-containing media, only those cells which uptake the

plasmid upon transformation can survive. The second selection is the lac operon system as described above for M13 vectors, which identifies colonies containing only vector DNA as blue and those harboring insert-containing pUC-based plasmids as colorless in the presence of IPTG and Xgal.

3. Introduction to DNA sequencing

In 1977, two different methods of DNA sequencing were reported. The first, developed by Maxam and Gilbert in Cambridge, MA, is a chemical cleavage-based approach and the second, developed by Sanger and Coulson in Cambridge, England, is an enzyme-based approach.

The Maxam and Gilbert[315] technique involved chemical cleavage of a 5' radiolabeled DNA strand by exposure to different base modifying chemicals. Since four distinct chemical agents were not available, combinations of chemicals were used resulting in strong A/weak G cleavage, strong G/weak A cleavage, C cleavage only, and C+T cleavage. After separation of the fragments by denaturing polyacrylamide gel electrophoresis, and exposure to film, the DNA sequence could be determined.

Sanger and Coulson[316] developed a sequencing method which utilized the polymerase activity of the Klenow fragment of *E. coli* DNA polymerase I. They described the chain terminating features of 2', 3' dideoxynucleotides which are analogs of nucleotides which lack the 3' hydroxyl group required for polymerization. They also described using four base-specific reactions each containing trace amounts of one of the four dideoxynucleotides, corresponding to adenine, cytosine, guanosine and thymine. DNA polymerization from a defined site beginning with an oligonucleotide primer annealed to single-stranded DNA resulted in a nested fragment set. Incorporation of small amounts of a radiolabeled nucleotide allowed for detection by autoradiography following separation of the nested fragment set by denaturing polyacrylamide gel electrophoresis, with each base-specific reaction loaded in a separate lane on the gel. The sequence of the synthesized strand can be determined by examining the banding pattern

present on the autoradiograph, reading from shorter to longer fragments, and based on the lane assignments of each base-specific reaction. Therefore, in the A-specific lane, a band terminated with a dideoxyATP would be present for every T that occurred in the complementary template sequence.

Many improvements have been made over the years to the original Sanger and Coulson dideoxynucleotide chain termination method, reflecting it's popularity and widespread application in molecular biology. Today, Sanger's enzymatic approach is the DNA sequencing method of choice and was the only sequencing method employed during this dissertation work.

The reported modifications of the Sanger method include the ability to alter the size range of the fragments in the nested fragment set by altering the ratios of deoxynucleotides to dideoxynucleotides, performing extension and termination portions in either one or two step reactions, and the elimination of compression bands or "stops" in reactions resulting from high G/C content or secondary structural elements in the templates by including denaturants such as urea or formamide in the polyacrylamide gels. In addition, the deoxyguanosine triphosphate has been replaced by deoxy-7-deaza guanosine triphosphate (dc^7GTP), deoxyinosine triphosphate or α-thionucleotide triphosphates in reactions to help reduce G/C compressions. Here, the addition of dc^7GTP does not participate in base stacking, thus making some secondary structures energetically unfavorable, while deoxyinosine triphosphate and α-thio nucleotides minimize gel compressions from secondary structure and improve efficiency of the polymerizing enzyme[317]. Also, the introduction of thermostable DNA polymerases. such as the *Thermus aquaticus* DNA polymerase and various mutants of this enzyme[318, 319] have reduced additional template-based secondary structural artifacts typically seen with Sequenase (a modified T7 DNA polymerase) by allowing polymerization at high temperatures.

Additional improvements were reported as the use of double-stranded vectors gained popularity. The challenge of sufficiently denaturing the double-stranded template while at the same time providing optimal conditions for primer annealing with enzymes such as Sequenase, Klenow and *Bst*, resulted in several procedures[320-323] for denaturing double-stranded sequencing templates. Among these protocols, the denaturation technique of Bachmann[324] was the most widely used and entailed denaturing the template by boiling in the presence of primer and detergent, followed by snap cooling on dry ice. However, the advent of cycle sequencing with thermostable enzymes eliminated the need for an initial denaturation step with double-stranded templates prior to enzyme addition and thus, is one of the more widely used modifications of the original Sanger dideoxynucleotide termination method. In cycle sequencing, all the reaction components are combined and incubated for several repeated cycles of denaturation, annealing and extension. These repeated steps allow for more efficient generation of the nested fragment set from either double-stranded or single-stranded templates[325-328].

DNA sequencing was impacted dramatically by the onset of fluorescent detection methods, replacing radioisotopic detection, which entailed 5' labeling of oligonucleotides with fluorescent dyes. The fluorescent-labeled primer methods have several disadvantages in that four separate base-specific reactions must be performed. Following cycle sequencing, the four base-specific reactions are pooled, precipitated and run in one lane of a polyacrylamide gel associated with a fluorescent sequencing instrument. Fluorescent DNA sequencing also has the disadvantage of fold-back compression resulting in abnormal migration of the generated fragments through the polyacrylamide gel during electrophoresis. Dye-labeled terminators were an additional application of fluorescent detection methods, with the fluorescent dye directly attached to the terminating dideoxynucleotide. All four base-specific reactions can be performed in the same tube since the detection moiety is directly attached to the terminating nucleotide. Fluorescent-labeled dideoxynucleotides prevent fold-back compression because the large

fluorescent group attached to the ddNTP does not allow base stacking, and therefore reduces abnormal fragment mobility. Background sequence is greatly reduced since abortive starts and stops go virtually undetected if not terminated with a labeled ddNTP.

All of these improvements of the original Sanger dideoxynucleotide chain termination method have led to the routine generation of DNA sequence data to greater than 400 base pairs (bp) past the priming site in a single fluorescent reaction on DNA sequencing instruments with automated data collection and basecalling capabilities.

C. Introduction to DNA sequencing

Strategies for DNA sequencing basically fall into two major categories: directed sequencing and random (or shotgun) sequencing. Within these two categories, various approaches to sequencing a particular region of DNA are presented.

1. Directed strategies

Directed strategies are based on the notion that in any DNA sequencing project, some information is known, and therefore can be used to generate the unknown information. For example, taken to its extreme, a directed strategy in which the initial reactions are primed with the universal forward or reverse primers with binding sites on M13 or pUC-based vectors and subsequent new sequencing reactions are performed with custom synthetic primers in additional sequencing reactions that result in stepwise generation of 300-400 bp of new extended sequence at a time. This strategy is advantageous to some in that the overlapping region is known and alignment of the sequence data is straightforward, whether or not repeated regions are present. This strategy has been widely used to sequence short, cloned DNA regions (1-5 kb) such as eukaryotic cDNAs. When the sequence of slightly larger cloned fragments was needed, modification of this primer walking strategy was used. For example, the sequence of the human[329], bovine[330] and *Xenopus laevis*[331] mitochondrial genomes and several human

tRNA genes utilized a directed restriction fragment cloning followed by custom synthetic primer walking to successfully complete the sequence of these 10-20 kb cloned regions.

An additional directed strategy eliminated the cost and time requirements of primer synthesis by utilizing a partial template degredation approach with exonuclease III (exoIII). Here, unidirectional deletion of a linear DNA molecule with subsequent ligation to cloning vector, allows systematic sequencing of small DNAs using only the universal sequencing primer[332-333].

2. Random strategies

Random (or shotgun) sequencing strategies employ various techniques to generate a library of sequencing subclones from an initial DNA. Techniques involved are sonication[334,335], nebulization[336,337], transposon insertion[338], mechanical shearing (i.e. French Press)[339, 399] and partial restriction digestion[329,340]. Random sequencing strategies are usually applied to large sequencing projects and as a result, require complex computer programs to align the overlapping subclones into a single contiguous sequence (computer programs discussed in Materials and Methods). The first large scale sequencing projects, bacteriophage lambda (49kb)[341] and the Epstein-Barr virus (170kb)[342], were both sequenced using a random approach.

The number of random subclones required depends on the length of the target DNA to be sequenced. Clarke and Carbon[343] generated a formula for evaluating the completeness of a random sequencing strategy. In this formula, L is the length of the target DNA fragment, s is the average length of gel read, n is the number of gel reads (number of subclones in the case of single-stranded M13-based cloning) required to accumulate a fraction of the gel reads (f_L) in contigs:

$$f_L = 1 - (1-s/L)^n$$

This relationship demonstrates one very important consequence of the random sequencing method. As f_L approaches 1 (or 100% of gel reads present in contigs), the number of gel reads approaches infinity. As mentioned above, for M13-based cloning, n

= number of subclones but in pUC-based cloning, n = number of gel reads, reflecting the 2 for 1 advantage of double-stranded sequencing; the ability to sequence with both the forward and reverse universal primers. In this case, the number of subclones required is n/2.

This inverse proportion of the number of gel reads to f_L emphasizes the importance of long gel reads to reduce the number of subclones required. Redundancy is also a critical component of a shotgun based approach, often expressed as

$$r = (ns)/L$$

where r is the average redundancy, n is the number of gel reads, s is the average read length and L is the length of the target DNA to be sequenced. Many investigators report that six-fold redundancy is required to attain 99.7% of the target genomic DNA sequence. Whereas, at three-fold redundancy, approximately 95% of the target sequence has been obtained.

Of course, both equations assume randomness of subclone generation. Practices have been developed to maximize randomness of the subclone inserts and if all efforts are made to induce randomness, nonrandom subclone distribution rarely inhibits the strategy of shotgun sequencing[329,334-336,344].

When the shotgun sequencing portion of a project reaches three- to six-fold redundancy, what remains is a series of contigs. In the case of an M13-based approach, the order of the contigs and relative position along the genomic target sequence are unknown, although restriction digestion data may be helpful. A plasmid-based random approach gives one additional piece of data which is impossible to glean from the M13-based approach. The presence of forward and reverse primer read pairs from the same subclone provides invaluable information in contig ordering and alignment. Using the forward and reverse read pairs, contigs can be ordered and placed in position on the genomic target sequence. In addition, restriction maps, if available, also are useful to facilitate contig ordering.

Gap closure in random strategies requires a transition to directed sequencing via primer walking. In the case of double-stranded sequencing where forward and reverse reads are positioned in separate contigs that do not overlap, longer reads may facilitate gap closure. In our laboratory, a mixed strategy is employed to close gaps resulting from a shotgun sequencing approach. The first step in closure is to generate longer reads from subclones on the ends of contigs which read into a gap. This is performed using a longer electrophoresis run time and a highly stable polyacrylamide gel matrix. Oftentimes, these long reads are sufficient to close gaps which have been mapped and ordered. Primer walking is employed if long reads do not close the gap. In cases where no subclone spans a gap, the PCR is used (with the original genomic DNA as a template) to generate a PCR fragment spanning the gap. The PCR fragment serves two purposes: first, it reveals the size of the gap and second, it provides a template for primer walking to close the gap. If the gap is large, the PCR product may be fragmented, subcloned and sequenced[345].

The ultimate questions in random strategies remain to be answered. What percentage of shotgun sequence data is required before moving to a directed closure strategy? Is closure necessary for large scale projects such as the human genome initiative, if 98% of genes can be localized with fewer than 3 to 6 fold redundancy, and is the data useful even if gaps of unsequenced regions remain? Finally, is it cost effective to decrease redundancy and close more gaps with primer walking rather than collecting additional shotgun data?

3. Double-stranded sequencing template isolation

The double-stranded subcloning approach implemented in our laboratory during the course of this dissertation work facilitated the need for quality double-stranded DNA preparations suitable for use as DNA sequencing templates. In the early stages, all template preparations were performed according to the alkaline lysis procedure first described by Birnboim and Doly[346], with the addition of purification measures to

eliminate genomic contamination. The original alkaline lysis as described by Birnboim and Doly involves lysing cells containing a plasmid of interest by the addition of high salt or lysozyme, solubilizing cellular lipids, membranes and proteins with detergent, precipitating unwanted cellular components with sodium acetate and clearing the lysate by centrifugation. Typically, most isolation procedures also include an additional step to purify the supercoiled plasmid DNA by binding all nucleic acid to a solid support, washing, and then preferentially eluting the supercoiled plasmid DNA. Subsequently, the purified DNA is concentrated by ethanol or isopropanol precipitation. Solid supports used in this manner include diatomaceous earth[347,348], glass wool or glass beads[349], silica particles[350], magnetic beads[351,352] and other proprietory resins[353].

Modifications of this procedure include alternate methods of cell lysis such as microwave to boiling and others including an organic extraction step. For example, Serghini[355], and later Chowdhury[354], reported a miniprep isolation in which cells were resuspended in buffer, extracted with phenol:chloroform:isoamyl alcohol, centrifuged and the supernatant precipitated with isopropanol. Finally, there is even a report that it was possible to isolate sequence-ready templates after only boiling cells, centrifugation, removing supernatant and precipitating with ethanol[321].

Initially, template preparations used in the studies reported herein were manually performed with up to 24 samples easily processed three hours. More recently, many biotechnology companies have marketed robotic workstations (ABI catalyst, Qiagen, Amersham, Beckman) with automated pipetting capabilities adaptable to double-stranded DNA isolations. As part of this dissertation work, a 96-well automated double-stranded DNA sequencing template isolation was developed for use first on the Beckman Biomek 1000 and then later on the Biomek 2000 laboratory workstations. These automated procedures are based on the original alkaline lysis method described by Birnboim and Doly, with the lysate cleared by filtration followed by ethanol precipitation in the Biomek 1000 preparation and by centrifugation in the Biomek 2000 preparation. With these

automated procedures, it now is possible to isolate 2 x 96 clones in three hours on two Biomek 1000 workstations or 8 x 96 clones in four hours on two Biomek 2000 workstations, with yields from both preparations sufficient for several sequencing reactions and highly reproducible.

4. Enzymes in DNA sequencing

As part of his laboratory's studies on DNA replication, Arthur Kornberg and his associates were the first to isolate and characterized *E. coli* DNA polymerase I in 1959[3]. This enzyme became the model for an entire class of enzymes, the DNA Polymerase I enzymes. DNA polymerase I enzymes perform DNA replication and repair, and require a metal ion cofactor for activity. *E. coli* polymerase I has three domains, each with a separate activity. These three domains include a 5'-3' polymerization activity, 3'-5' exonuclease activity for proofreading and repair, and 5'-3' exonuclease activity that is active in bacterial replication (for removal of RNA primers and replacement with DNA). Because the 5'-3' exonuclease activity would interfere with the DNA sequencing reactions by degrading the common 5' end, Klenow successfully removed the domain responsible for the 5'-3' exonuclease activity by proteolysis with subtilisin. This led to a Klenow fragment of *E. coli* polymerase I that contains only the polymerase and repair domains. It was this enzyme that initially was used by Sanger in developing the dideoxynucleotide chain termination sequencing strategy.

In the following years, additional DNA polymerase enzymes were isolated from a variety of bacteria and modified for use in DNA sequencing. Bacteriophage T7 DNA polymerase was isolated in 1975[358,359] and selectively mutated by *in vitro* mutagenesis in 1989 to inactivate the 3'-5' exonuclease activity[360]. This "modified T7 DNA polymerase" could incorporate nucleotide analogs, a property not found in the native T7 DNA polymerase, which made it an ideal enzyme for DNA sequencing applications including the use of dc^{7}GTP, dITP, alpha thionucleotides or fluorescent dye-labeled dideoxynucleotides. Further studies revealed that the presence of manganese ions rather

than magnesium ions decreased the discrimination of both modified T7 DNA polymerase and *E. coli* DNA polymerase I for dideoxynucleotides by 100 fold and 4 fold, respectively. Thus, the modified T7 DNA polymerase, marketed under the commercial name Sequenase, was the enzyme of choice for DNA sequencing for several years due to high processivity, no 3'-5' exonuclease activity and efficient utilization of nucleotide analogs.

The DNA polymerase from thermophile *Bacillus stearothermophilus* was isolated in 1972 by Stenish and Roe[361] and had the advantage of a high temperature optimum of 65°C. *Bst* polymerase permitted higher temperatures to be used in sequencing reactions to melt out secondary structure[362]. Uniform band intensities, low background on autoradiographs, ability to efficiently incorporate nucleotide analogs, a fast rate of polymerization and the added advantage of high temperature optimum rapidly made this enzyme applicable to DNA sequencing[326]. With a polymerization speed of over 120 nucleotides per second at 65°C, DNA sequencing reactions catalyzed with *Bst* required minimal template DNA (a 10-fold reduction) when compared to other sequencing enzymes available at the time[363].

The DNA polymerase from *Thermus aquaticus* also has been purified, cloned and described for use in DNA sequencing. This enzyme now is the most widely used enzyme due to a high temperature optimum of 75°C. This enzyme, similar to *Bst* polymerase, allows for polymerization at a higher temperature, while simultaneously decreasing the presence of secondary structure in the template DNA and eliminating secondary primer binding sites.

Several mutants of the original *Taq* polymerase have been created and characterized with DNA sequencing applications in mind. The first mutation, marketed by Applied Biosystems as Amplitaq[364], was the original *Taq* polymerase with the domain for 3'-5' exonuclease removed from the protein. Reactions utilizing Amplitaq were optimized for use with dITP to minimize band compression and DMSO to stabilize the

polymerase when sequencing templates with high G/C content in dye terminator and dye primer chemistries. However, Amplitaq exhibited a strong discrimination against incorporation of dideoxynucleotides and required high concentration of dye-labeled terminators in the reactions.

Amplitaq CS+[365] was the second mutation of *Taq* polymerase marketed by Applied Biosystems. This enzyme had all the characteristics of Amplitaq, but also was lacking the 5'-3' exonuclease activity. Removal of the 5'-3' exonuclease activity was reported to result in cleaner sequencing reactions, reduced background noise and fewer band compressions. Amplitaq CS+ included a thermostable inorganic pyrophosphatase to help eliminate "dropout" peaks due to pyrophosphorolysis (reverse of the polymerization process), a result of inorganic pyrophosphate buildup. Reactions utilizing Amplitaq CS+ were optimised by Applied Biosystems containing 7-deaza-dGTP to minimize band compressions in dye primer chemistry.

Klentaq (a *Taq* analog of the Klenow fragment of *E. coli* DNA polymerase I)[366,367] is a third enzyme examined for use in DNA sequencing applications. Klentaq is an N-terminally deleted Taq DNA polymerase I which removes the 5'-3' exonuclease activity in addition to already lacking the 3'-5' exonuclease activity. This enzyme was originally developed to increase the fidelity of *Taq* polymerase in polymerase chain reaction (PCR) applications. Saiki originally described Amplitaq polymerase for use in PCR reactions[368], but noted that Amplitaq introduces base change errors into the sequence at a rate of 1 error in 10,000 bp. This raised a great deal of concern since this error rate per cycle results in an error every 400 bp in a 25 cycle PCR amplification. Experiments performed by Barnes[369] revealed that Klentaq has 2x lower error rate than Amplitaq in PCR. This increased fidelity of Klentaq opened the door for use in DNA sequencing applications.

Taquenase (or Klentaquenase)[318,319], the enzyme used for DNA sequencing in the latter stages of this dissertation work, is the Klentaq enzyme mentioned above which

contains a mutation at a single residue. This residue is critical for distinguishing between deoxynucleotides and dideoxynucleotides and was first recognized as such by Tabor and Richardson. Tabor and Richardson performed comparison studies of several different DNA polymerases in the DNA polymerase I family and noticed four major areas of significant difference. First, T7 DNA polymerase from phage T7 is responsible for replicating the genome whereas *E. coli* polymerase I and *Taq* polymerase I are more utilized for repair and recombination. Second is the difference in exonuclease activity of the enzymes. *Taq* polymerase I has only 5'-3' exonuclease activity, T7 polymerase has only 3'-5' exonuclease activity, and *E. coli* DNA polymerase I has both 5'-3' and 3'-5' exonuclease activities. Third, the thermostability of *Taq* polymerase differs from T7 and *E. coli* polymerases. Fourth is the ability to distinguish between and selectively incorporate deoxynucleotides over dideoxynucleotides. *E. coli* polymerase I and *Taq* polymerase have a strong preference to incorporate deoxynucleotides over dideoxynucleotides, whereas T7 polymerase incorporates dideoxynucleotides with almost equal efficiency to that of deoxynucleotides. Since it was postulated that the incorporation of deoxynucleotides versus dideoxynucleotides is the result of the single residue at position 526 in T7 DNA polymerase, the corresponding position in *Taq* polymerase was mutated from phenylalanine to tyrosine at position 667 and the result is Taquenase, an enzyme with greater ability to incorporate dideoxynucleotides and modified nucleotides such as dye-labeled terminators. A similarly mutated form of *Taq* polymerase was marketed by Applied Biosystems as Amplitaq FS (although Amplitaq FS also contains the thermostable pyrophosphatase as in Amplitaq CS+)[370] and by Amersham/United States Biochemicals as Thermosequenase but each enzyme has a different modification to eliminate the 5' exonuclease activity. Because these enzymes have a reduced discrimination for dideoxynucleotides and modified nucleotides, lower concentrations of dye-labeled terminators could be used, and thereby the cost per reaction could be reduced significantly.

5. Fluorescent DNA sequencing instrumentation

The cloning and characterization of polymerases allowed for developing fluorescent detection methods to the point where hazards associated with radioisotopic detection methods in DNA sequencing could be eliminated. Throughout the course of this dissertation work, several fluorescent DNA sequencing instruments were developed. One DNA sequencer was developed by Ansorge, et al.[371,372], at the European Molecular Biology Laboratory (EMBL) and is marketed by Pharmacia. This instrument requires a single dye label, with which 4 base-specific reactions are performed and electrophoresed in 4 adjacent wells of a denaturing polyacrylamide gel. The fluorescent dyes are excited by a laser beam directed into the gel from the side and the detector is positioned perpendicular to the angle of laser excitation. Hitachi developed a fluorescent sequencer which is similar to the EMBL instrument in that excitation occurs from a laser directed into the side of the gel, detection is from a fixed position perpendicular to the laser angle and use of a single fluorescent dye is required. The detection system differs from the EMBL system in that the emitted fluorescence is screened through bandpass filters, enhanced by an intensifier and detected by a camera before the data is communicated to an associated computer for analysis.

By far, the instruments developed by Applied Biosystems[373,374] are the most widely used. Due to our early involvement with ABI as a university β-test site for the 370 DNA sequencer, the ABI 373 (and an upgraded 370) and later, the ABI 377, were the only instruments used in the course of this dissertation. As mentioned, ABI has developed two different sequencing instruments, the 373 and the 377, with the 377 including features such as a more focused laser, a heat dispersion plate, updated software, and a CCD camera instead of a PMT for detection. Both instruments utilize a principle of four different fluorescent dyes which allows the four base-specific reactions to be pooled and electrophoresed in a single lane of a polyacrylamide gel associated with the instrument, thereby increasing the number of samples which can be processed per

machine run by a factor of four with respect to the other instruments available. Both ABI instruments have a scanning Argon laser for excitation with positioning pins in the electrophoresis chamber which hold the gel at the optimum focal point of the laser beam. As mentioned above, the detection system is the most significant difference between the ABI 373 and 377 (see figure 14). In the 373, the laser traverses the gel and excites the fluorophores through a circular bandpass filter, one for each fluorophore. The emitted photons have characteristic wavelengths, depending on the fluorophore. The photons are detected, amplified by a photomultiplier tube and the raw data stored by the associated computer. In contrast, the 377 detection system uses the same type of argon laser to excite the fluorophores as they enter the read region of the gel. A series of lenses collects and focuses the emitted light onto a spectrograph diffraction grating which separates the light based on wavelength into a predictably spaced pattern across a charged coupled device (CCD) camera. Using the collected light intensities from the CCD camera, the collection software records the amplified emission signals from the CCD camera and the raw data is stored by an associated computer. The CCD detection system is much more sensitive than the 373 PMT detection system.

The raw data collected by the computer must be manipulated to generate the final processed sequence by: (1) suppressing high frequency noise by a digital filter corresponding to fluorescent peak width; (2) multicomponent analysis to determine the relative concentration of the four fluorescent dyes at each time point during the electrophoretic run; (3) mobility shift calculations to accommodate different electrophoretic mobility effects of the four dyes based on pre-determined correction factors; (4) removing low frequency noise and signal enhancement by spectral deconvolution; (5) differential weighting of peaks to provide equal peak intensity and spacing; and (6) DNA sequence nucleotide assignment based on the processed signal.

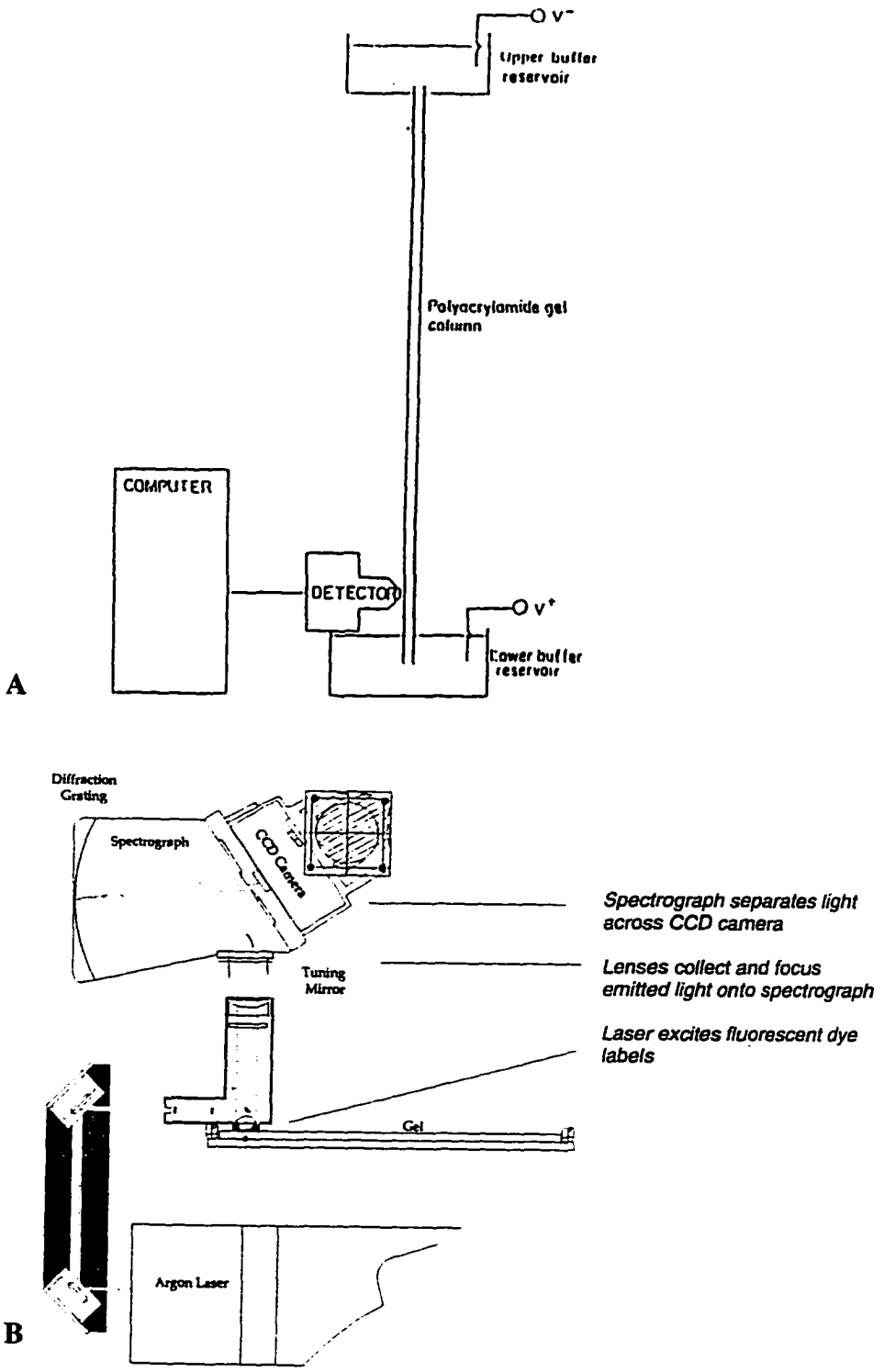


Figure 14: Comparison of ABI 373 and 377 instrumentation. A) ABI 373 instrumentation with Argon laser for excitation and PMT detection. B) ABI 377 instrumentation using Argon laser for excitation and CCD camera for detection.

The processed data is presented as a chromatogram of electrophoretic time versus signal intensity. Electrophoretic time corresponds to the length of the fragments and the four fluorescent dyes appear as four different colors on the chromatogram.

Before processing, the raw data files are transferred via AppleShare from the Macintosh computer associated with the sequencing instruments to Macintosh PowerMac computers dedicated to raw sequence data analysis. The sequence data is stored and analysed on the PowerMacs, then transferred via NSF Share to a Sun Sparcstation for contig assembly (as detailed in Materials and Methods section). This completely computerized system of data handling eliminates the manual sequence data entry and human error associated with the radioisotopic detection systems.

Electrophoresis times and run conditions also differ significantly between the two ABI instruments (See Materials and Methods). The ABI 373 requires a 10 hour electrophoretic run, whereas the ABI 377 requires only three hours to collect the same amount of raw data. ABI 377 instruments use thinner gels (0.2 mm versus 0.4 mm for the 373) and are equipped with a heat dispersion plate and sensors to monitor the gel temperature, allowing for faster run times. The 377 instruments can be run up to 5 times per 24 hours, collecting an average of 400 base pairs of sequence data per sample, with 48 samples run per gel. This results in 96,000 base pairs of raw sequence data generated per 24 hour period on one ABI 377. The ABI 373 is capable of 48 lanes per gel with 2 runs per 24 hour period resulting in 38,400 base pairs of raw data per day on one ABI 373.

6. DNA sequencing chemistries

With the introduction of thermophilic DNA polymerases such as *Bst* and *Taq*, the more recent application of mutant thermophilic DNA polymerases, and the advent of PCR, it became obvious that DNA sequencing could further evolve from the one- and two-step reactions to a more elegant type of sequencing reaction. The invention of PCR by Kary Mullis[375] in 1987 opened the door for the application of temperature cycles,

such as those used in the PCR, to generate multiple copies of the nested fragment sets required for DNA sequencing and thereby greatly improving the final signal intensity.

While on sabbatical in Cambridge, England at the MRC, Dr. Roe learned the linear amplification sequencing method utilizing *Taq* DNA polymerase and a PCR-like cycling of temperatures corresponding to denaturation, primer annealing and extension[325,327]. Cycle sequencing, as it is now called, has two different strategies based on the location of the detection label. Dye-labeled primer cycle sequencing reactions use oligonucleotides labeled on the 5' end with one of 4 fluorescent dye labels, and the dye terminator method uses dideoxynucleotides with one of 4 fluorescent labels at the 3' position in the molecule.

Dye-labeled primer reactions involve incubating four base- and primer-specific reactions; each containing template DNA, a specific fluorescent labeled primer, the four deoxynucleotides, the base-specific chain terminating dideoxynucleotide and thermostable DNA polymerase in a cycling instrument that controls temperature and time. These samples are subjected to cycles of denaturation, primer annealing and extension/termination (temperatures and length of time at each step in the cycle depends on the enzyme of choice) to generate a nested fragment set containing fluorescent labeled fragments. Because of the cyclic nature of these reactions, more primers are extended and the template is used multiple times, thereby decreasing the amount of template required and increasing the number of fragments generated. Following cycling, the four base- and primer-specific reactions then are pooled, precipitated and loaded in one well of an ABI 373 or 377 sequencing gel. This strategy employing four fluorescent dyes attached to the 5' end of oligonucleotides was first developed in the laboratory of Lee Hood. The primers used all have the same DNA sequence and contain the fluorescent dyes linked by a covalent bond through an amino group[376-378]. These dyes, which were marketed by ABI for use in the ABI 373, are derivatives of fluorescein, tetramethylrhodamine, NBD and Texas Red (FAM, JOE, TAMRA and ROX). The dyes

have slightly overlapping emission spectra that are spectrally resolvable. Upon additional investigation by ABI, NBD was replaced with another fluorescein derivative to further improve the spectral resolution.

For several years, dye primer chemistry was the method of choice since the initial phase of random shotgun projects requires only forward and reverse universal primers. Closure, however, requires the shift to a more directed strategy with primer walking to close any remaining gaps in the sequence data. Although a 5' fluorescent label can be attached to any custom oligonucleotide via an aminolink, the cost and time necessary to prepare custom, synthetic, fluorescent-labeled primers prohibits routine labeling for the primer walking directed closure strategy. Thus, the introduction of the dye terminator cycle sequencing chemistry presented an attractive alternative.

Dye terminator reactions have the advantage over dye primer reactions in that dye terminator reactions can be carried out in a single tube with a custom primer since the detection label is directly on the terminating nucleotide. Dye terminator reactions involve incubating unlabeled primer, template DNA, the four deoxynucleotides, the four labeled dideoxynucleotide terminators and the enzyme of choice. These reactions are cycled through steps of denaturation, annealing and extension/termination, just as discussed for dye primer reactions, to generate a fluorescently labeled nested fragment set. After cycling is complete, the residual fluorescent-labeled dideoxynucleotide terminators are removed, either by chromatography on Sephadex-G50 spin columns or by ethanol precipitation prior to electrophoresis on an ABI 373 or 377 sequencing gel.

Dye terminator chemistry with Sequenase[379,380] is an exception to the cycle sequencing rule. Sequenase reactions are isothermal and are carried out at 37°C utilizing α-thionucleotides and terminators labeled with one of four fluorescein derivatives, FAM, HEX, NAN and ZOE. These reactions require an additional bandpass filter for detection on an ABI due to the altered emission spectra of one of the dyes. This requires an ABI equipped with a five filter wheel to accommodate the different spectral emissions.

Taq dye terminator chemistry utilizes the cycle sequencing mentioned above but uses different dyes than Sequenase or the dye primer chemistry. The dyes used with the Taq DNA polymerase are all rhodamine derivatives (R110, R6G, TAMRA and ROX), and all have spectral emission in the range of the standard ABI four-filter wheel.

Cycle sequencing is applicable to both single- and double-stranded DNA templates, although the majority of the reactions performed in this work used double-stranded templates isolated by the robotic procedures as outlined in the Materials and Methods section of this dissertation.

Chapter II

MATERIALS AND METHODS

I. General methods and random subclone generation

The widely accepted methods such as phenol extraction, ethanol precipitation, agarose gel electrophoresis, restriction digestion, elution from agarose, nebulization, calcium chloride competent cell preparation and transformation were performed as described[345]. Protocols for techniques presented in this section are detailed in the Appendix.

A. Cell preparation for electroporation

XL1Blue MRF' cells were grown overnight in YENB media and the overnight culture transferred to 1 L YENB and grown at 37°C with shaking at 250 rpm for 4-5 hours or until A_{600} of 0.5 to 0.9 [400]. Cells then are harvested by centrifugation at 1500 rpm for 10 minutes, washed two times with cold, sterile water followed by centrifugation as before. The cells then were resuspended in 10% sterile glycerol, centrifuged as before, and resuspended in a total of 2 ml 10% glycerol. Forty microliter aliquots of cells are prepared, rapidly frozen on dry ice and stored at -70°C until needed.

B. Electroporation

Cells prepared for electroporation were thawed on ice, followed by incubation on ice for one minute in the presence of the DNA to be transformed. The entire mix of cells and DNA then is transferred to the bottom of an electroporation cuvette and placed between the leads of the electroporator. The cells then are subjected to an electrical pulse which facilitates plasmid DNA uptake by the cells. Following electroporation, additional YENB media is added to the cells to aid in recovery at 37°C for one hour. The cells then

are pelleted by centrifugation, resuspended in a smaller volume of YENB and plated on appropriate agar plates. Agar plates are inverted and incubated at 37°C overnight.

II. Methods for DNA isolation

A. Large scale double-stranded isolation for plasmid and cosmid

The method used for large scale plasmid and cosmid isolations is a diatomaceous earth-based method which was developed[347] to reduce the cost when compared with the traditional cesium chloride ultracentrifugation isolation, to reduce the level of host genomic DNA contamination, and to increase the yield of supercoiled plasmid or cosmid DNA.

In this diatomaceous earth protocol, cells are grown at 37°C in three successive eight-hour incubations of 3 ml, 50 ml, and 4 liters of culture (typically in Lurian Broth including the appropriate antibiotic when necessary) and harvested by centrifugation such that the cells from two liters of culture are pelleted into one 500ml centrifuge bottle. The cells are resuspended in a glucose, Tris-HCl, pH 7.6, EDTA containing buffer, and lysed by the addition of lysozyme. After adding anionic detergent to solubilize membranes and proteins, the unwanted cellular components and detergent are precipitated with sodium acetate and removed by filtration through cheesecloth and subsequent centrifugation. The cleared lysate is treated with RNase A and RNAse T1 to digest the RNA since RNA competes with DNA for binding sites on diatomaceous earth. The supernatant-containing DNA then is bound to diatomaceous earth in the presence of binding agent guanidine hydrochloride and the DNA-associated diatomaceous earth collected by centrifugation, washed with buffer containing ethanol, washed with acetone, and dried under vacuum. The DNA is eluted from the diatomaceous earth by addition of an elution buffer containing Tris-HCl, pH 7.6 and EDTA, incubation at 65°C and separation of the DNA in solution from the diatomaceous earth by centrifugation. The eluted DNA then is

concentrated by ethanol precipitation and assayed for concentration by agarose gel electrophoresis versus known standards.

B. Large scale double-stranded isolation for BAC

The large scale double-stranded isolation for BAC DNA is essentially identical to the large scale double-stranded isolation for plasmid and cosmid with minor exceptions. In short, BAC containing colonies are grown in 3 successive growth steps of eight hours each, beginning with 3 ml which is transferred to 50 ml, of which 25 ml of culture is transferred to one liter (media typically is Lurian Broth containing the appropriate antibiotic when necessary). The cells are harvested such that the cells from one liter of culture are pelleted into one 500 ml centrifuge bottle. Cell lysis, lysate clearing and isopropanol precipitation are carried out using the same reagent volumes as for the plasmid or cosmid preparation. The nucleic acid pellet then is resuspended in 20 ml 10:1 TE and bound by the addition of 50 ml of 100 mg/ml defined diatomaceous earth in 6 M guanadine-hydrochloride. All subsequent washes and elution steps are performed as for the plasmid and cosmid isolations. The final DNA is ethanol precipitated, the BAC pellet resuspended in 1 ml 10:0.1 TE, and an aliquot analyzed by agarose gel electrophoresis. Decreased initial cell concentration results in an effective increase in the binding sites present in the diatomaceous earth which allows for more specific binding of contaminating bacterial genomic DNA and preferential elution of BAC DNA.

C. Midiprep double-stranded DNA isolation for plasmid or cosmid

Midiprep double-stranded DNA isolations were used to obtain double-stranded DNA on a smaller scale. These isolations are similar to the large scale double-stranded DNA isolations described above but scaled back to 50 ml starting culture instead of 4 liters. The pelleted cells then are resuspended in buffer containing Tris-HCl, pH 7.6, EDTA, and RNase A and T1 and subsequently are lysed by the addition of sodium

hydroxide and SDS to solubilize cellular membranes and proteins. The lysate then is cleared by the addition of sodium acetate, filtered through cheescloth, and centrifuged. The cleared lysate is bound to diatomaceous earth as in the large scale prep, washed, eluted, ethanol precipitated and assayed for concentration using agarose gel electrophoresis on a 0.7% agarose gel containing ethidium bromide for visualization.

D. Miniprep double-stranded DNA isolation using diatomaceous earth

Miniprep double-stranded DNA isolations using diatomaceous earth were performed as mentioned above for the midiprep and large scale double-stranded isolations except that the initial cell culture was 6 ml. After collecting the cells by centrifugation, resuspending in buffer, lysing with SDS/sodium hydroxide, precipitating the unwanted cellular components with sodium acetate, the lysate is cleared by centrifugation. The cleared lysate then is incubated with diatomaceous earth in the presence of guanidine hydrochloride to bind the DNA. The DNA-bound diatomaceous earth suspension is collected either by centrifugation, or is separated by filtration through a BioRad Gene Prep vacuum manifold. The diatomaceous earth with bound DNA then is washed twice with wash buffer containing ethanol, centrifuging or filtering between each successive wash, and eluted by addition of elution buffer. During the course of this dissertation work, it was determined that heating the buffer to 65°C slightly improved the yield. The eluted DNA is collected by pelleting the diatomaceous earth and removing the supernatant, or by collecting in vials via filtration. Finally, the eluted DNA is ethanol precipitated to concentrate and assayed for purity using agarose gel electrophoresis.

E. Manual 96-well double-stranded DNA miniprep

The manual 96-well miniprep isolation is similar to the large scale isolation discussed above, as it uses alkaline lysis followed by centrifugation to clear the cell

lysate. However, the manual miniprep does not include the addition of diatomaceous earth to remove genomic DNA. Briefly, the 96-well manual miniprep entails growing plasmid containing cells in a 37°C shaker at 350 rpm for 24 hours in 1.5 ml antibiotic containing media in a 96 well block (2 ml well capacity). Subsequently, the cells are harvested by centrifugation and the cell pellets resuspended in Tris-HCl, pH 7.6, EDTA buffer containing RNase A and RNase T1. The cells then are lysed by the addition of alkaline detergent and the cellular components precipitated by the addition of sodium acetate. The lysate is cleared by centrifugation, the supernatant removed to a clean 96 well block, and the DNA ethanol precipitated. Agarose gel electrophoresis is performed to crudely assay for concentration and purity against known standards. Typical yield for this isolation method is 10-15 ug plasmid DNA per 1.5 ml starting culture.

F. Automated 96 well double-stranded DNA miniprep on the Biomek 1000 laboratory workstation

During the course of this dissertation, an automated double-stranded miniprep isolation was developed using the Beckman Biomek 1000 Automated Laboratory workstation[381]. This procedure is similar to the manual miniprep mentioned above in that the cells are grown 24 hours at 37°C in a 96 well block with 350 shaking and harvested by centrifugation. The Biomek adds buffer and resuspends the cells by repeated cycles of mixing. The cell suspension then is transferred using the Biomek to a Millipore 96 well filter plate. Lysis is achieved by the Biomek addition of anionic detergent and mixing several cycles followed by precipitation by Biomek addition of sodium acetate and mixing at the sides of the well. The lysate is cleared by filtration using the Qiagen filtration apparatus (instead of centrifugation as in the manual prep)[382] and the cleared lysate is collected, ethanol precipitated and assayed for concentration and purity using agarose gel electrophoresis. The Biomek 1000 is capable of isolating 96 subclones in three hours. See figure 15A for tablet configuration.

G. Automated 96 well double-stranded DNA miniprep on the Biomek 2000 laboratory workstation

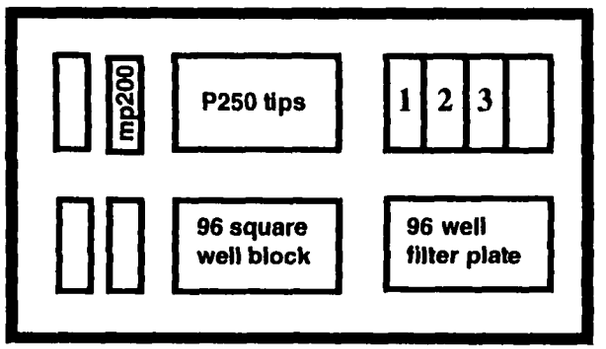
The Biomek 2000 laboratory workstation is the most recent second generation pipetting robot available on the market from Beckman. Improvements in the 2000 include a larger tablet for greater reagent, tube plate and tip storage and improved software accessible via a windows-based interface[383]. The Biomek 2000 double-stranded isolation entails the identical cell growth and lysis as the manual 96-well isolation mentioned above, with the Biomek adding all reagents. The lysis is cleared by centrifugation in a table top centrifuge equipped with deep well microtiter plate carriers. Following centrifugation, the supernatant is removed and transferred to a clean 96 well block and ethanol precipitated. The DNA was resuspended in 10:0.1 TE and assayed by agarose gel electrophoresis. This Biomek 2000 based procedure is capable of simultaneously isolating 4 x 96 subclones in four hours. See figures 15B and 15C for tablet configuration.

III. Methods for DNA Sequencing

A. Fluorescent sequencing gel preparation, loading, electrophoresis on ABI 373 and 377 and preliminary computer analysis

When dealing with large scale DNA sequencing efforts, every attempt must be made to standardize techniques, reagents, and equipment. To this end, polyacrylamide gel mix (containing buffer, urea, acrylamide and bisacrylamide) is prepared in 1 liter batches by dissolving the appropriate amount of urea in water and buffer, heating at 55°C with stirring until urea dissolved, cooling, and then adding acrylamide/bisacrylamide (19:2) solution. This gel mix is filtered and degassed for 10 minutes under vacuum. Before use, ammonium persulfate and TEMED are added to catalyze polymerization of

A. Setup for Biomek 1000 isolation

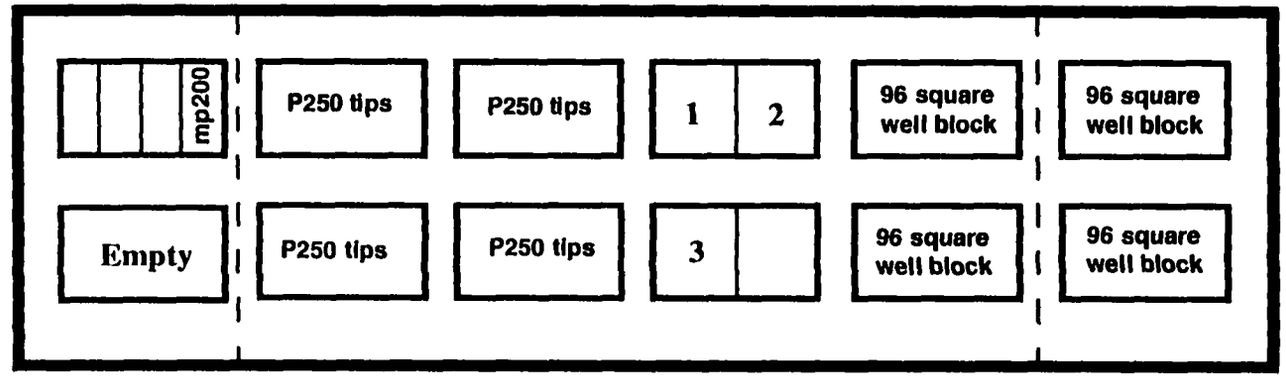


NOTES:

1. **96 square well blocks contain cell pellets**
2. **Reagent reservoirs**
   - **1: TE-RNase A + T1 solution**
   - **2: SDS/NaOH solution**
   - **3: 3M NaOAc, pH 4.5 or 3M KOAc, pH 4.5**

Figure 15: Biomek 1000 tablet configuration for double stranded template isolation

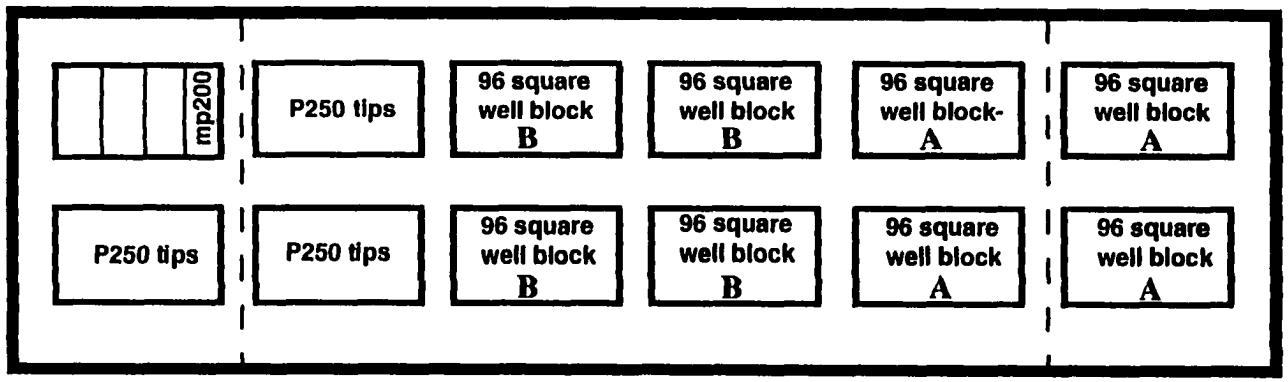
B. Setup for clearing lysate - Biomek 2000 isolation



NOTES:
1. **96 square well blocks contain cell pellets**
2. **Reagent reservoirs**
   - **1: TE-RNase A + T1 solution**
   - **2: SDS/NaOH solution**
   - **3: 3M NaOAc, pH 4.5 or 3M KOAc, pH 4.5**

Figure 15: Biomek 2000 tablet configuration for double stranded template isolation

C. Setup for supernatant transfer - Biomek 2000 isolation



NOTES:
1. **'A' blocks are the initial culture blocks with pelleted precipitate and cleared lysate supernatant.**
2. **'B' blocks are clean blocks into which supernatant is transferred.**

Figure 15: Biomek 2000 tablet configuration for double stranded template isolation

the acrylamide monomers. The gels are cast by pouring between two glass plates separated by 0.4 mm or 0.2 mm thick spacers, for ABI 373 and ABI 377 respectively, and held in place with clamps after inserting a sharkstooth comb. The gels are allowed to polymerize for at least one hour before use and if the gels are to be stored overnight, the ends of the gel first are covered with buffer-soaked paper towels and then with plastic wrap to avoid drying.

Just prior to use, the polymerized gels in glass plates are unclamped, washed carefully with water, rinsed with double distilled water followed by ethanol, and allowed to dry in a fume hood before the sharkstooth comb is inverted, reinserted into the gel to form individual sample wells, and assembled into the ABI machine The gel then is examined for cleanliness using the plate check option in the ABI data collection software and rewashed at this stage if necessary. Clean gel plates are required to obtain a scan chromatogram with all four scan lanes being horizontal and with no extraneous peaks present. If the scan revealed clean gel plates, buffer then is added to the upper and lower buffer chambers and the gel is prerun prior to sample loading.

Dried samples prepared by any of the following techniques are resuspended in a blue dextran/EDTA/deionized formamide solution and heated at 100°C for 5 minutes to denature. Samples then are loaded onto the gel, loading the odd samples first, and after prerunning the gel for 5 minutes, the even samples are loaded. This staggered loading facilitates tracking by the ABI analysis software and avoids abnormal sample mobilities. Once all samples are loaded and briefly electrophoresed into the gel, the run is started and the data is collected automatically on a computer associated with the sequencer. ABI 373 electrophoresis runs are typically for 10 hours at 35 watts (limiting parameter) using a 4.75% polyacrylamide, 8 M urea gel, while the ABI 377 electrophoresis runs are for 3.0 hours at 3 kV (limiting parameter) using a 4.25% polyacrylamide, 8 M urea gel. Alternatively, if longer readlengths are required, the gels are electrophoresed for 7 hours at 1.68 kV (limiting parameter) using a 4.25% polyacrylamide, 8 M urea gel.

At the end of a sequencing run, the gel file is transferred via AppleShare to a Macintosh computer equipped with the ABI sequence analysis software. The gels are tracked, and the data is extracted from each sample lane of the gel. The data produced by both types of DNA sequencers (ABI 373 and 377) exist in two types of files. The two file types represent the signal collected during the run either as raw or analyzed data. Analyzed data is the raw data after correcting for baseline, mobility shift from the fluorescent dyes, signal enhancement and base spacing. Once the data is analyzed, it is transferred to a Sun Microsystems Sparcstation via NSFShare where editing and alignment of the data is performed using TED (trace editor) and XGAP (X windows genome assembly program)[384], FAKII[385], CAP2[386] or PHRED/PHRAP[387] multiple sequence alignment algorithms. During the latter stages of this dissertation research, a script was written by Dr. Bruce Roe and Hong Shing Lai, called ouotto, which automatically performs the TED editing functions and the alignment algorithm became available to remove the tedium associated with the manual, repetitive manipulations.

B. Amplitaq fluorescent-labeled primer cycle sequencing

Cycle sequencing using dye-labeled primers has been reported elsewhere [326,336,345] and was used in the beginning stages of the meningioma project detailed in this work. Briefly this entails combining 100/200ng single-stranded DNA or 200/400 ng of double-stranded DNA for the A/C or G/T reactions, respectively, with buffer, deoxy- and dideoxynucleotides, Amplitaq DNA polymerase and primer labeled with one of four fluorescent dyes. These reactions occur in four separate tubes, with the terminating dideoxynucleotide present in the tube corresponding to the primer fluorescent label for that dideoxynucleotide. These reactions were cycled through 30 cycles of a seven temperature profile of 94°C for 4 seconds, 55°C for 10 seconds, 72°C for 1 minute, 95°C for 4 seconds, 72°C for 1 minute, 95°C for 4 seconds, and 72°C for 1 minute. This cycling profile is linked to a 4°C soak file for storage until retrieval. The four

reactions then are pooled, ethanol precipitated, dried and stored at -20°C prior to loading as detailed above.

C. Amplitaq fluorescent dye terminator cycle sequencing

The majority of the work presented here for the meningioma region and the studies in the Cat Eye Syndrome region were performed using *Taq* fluorescent dye terminator cycle sequencing reactions. These reactions differ from the dye primer chemistry in that the four different fluorescent dyes used for detection were located on the dideoxynucleotides instead of the primer. This single advantage allows all four reactions to take place simultaneously in one reaction tube. Each reaction included 1 ug double-stranded DNA, primer, deoxynucleotides, limited amounts of unlabeled dideoxynucleotides, buffer, Amplitaq DNA polymerase and fluorescent labeled dideoxynucleotides. The addition of 10% DMSO to the reaction mix was used in some instances as it was found to increase the readlength of the reactions, to aid in resolving compressions due to G/C rich templates, and sequencing through regions with polynucleotide repeats. With this chemistry, the reactions are subjected to 25 cycles of a three temperature profile of 96°C for 15 seconds, 50°C for 1 second and 60°C for 4 minutes. This profile also is linked to a 4°C soak file for storage of the completed reactions until retrieval. These reactions then either are ethanol precipitated and washed twice with 70% ethanol, or passed through a Sephadex G-50 column to remove unincorporated terminators (discussed below). In either case, the final reaction products are dried under vacuum and stored at -20°C until loading as described above.

D. Sephadex G-50 dye terminator cleanup

A slurry of Sephadex G-50 and water is prepared by adding 10 g Sephadex G-50 to 100 ml of double distilled water. The slurry is allowed to hydrate at 4°C for at least 3 hours before use, adding additional water if necessary. Columns are prepared by adding

400 ul Sephadex G-50 slurry to each well of a Millipore 96-well filter plate. The filter plate is placed on top of a microtiter plate to collect the eluent, taped together, and centrifuged for 2 minutes at 1500 rpm in a Beckman GPR tabletop centrifuge equipped with microtiter plate carriers. The eluent is discarded and an additional 200 ul Sephadex G-50 slurry is added to the wells of the filter plate. The filter plate is restacked on the microtiter plate, taped, and recentrifuged for 2 minutes at 1500 rpm as before. The resulting eluent is discarded and the collection microtiter plate replaced with a clean microtiter plate for sample collection. The entire 20 ul sample volume is added carefully to the tops of the columns (making sure the reaction is placed in the center of the column so as to not allow the eluent to leak down the sides) and centrifuged as before to collect the DNA in the clean microtiter plate. Samples are covered with kimwipe, dried under vacuum and stored at -20°C until loading.

E. Taquenase dye terminator cycle sequencing

Recently, several groups have used site-directed mutagenesis to investigate various properties of *Taq* polymerase[318,319]. Several different mutants of the DNA polymerase isolated from *Thermus aquaticus (Taq)* now are available commercially and include Amplitaq, Amplitaq CS+, Amplitaq-FS, Klentaq and Taquenase (See Introduction for more detail). Described here is the technique used most recently for sequencing portions of the Cat Eye syndrome region and the multiple endocrine cancer region utilizing the unique properties of Taquenase DNA polymerase.

The Taquenase DNA polymerase cycle sequencing reactions were optimized in the laboratory of Dr. Bruce Roe by Guozhang Zhang. The reactions are performed as follows: Approximately 2 ug of template DNA is added to 96 well sequencing plate and the final volume is brought to 14 ul per sample with double distilled water. The following reagents are added to each sample: 1ul 20 x dNTPs, 1 ul 20 x ddNTPs, 1ul 20 mM $MnSO_4$, 2 ul 10 x $MgCl_2$/Tris-HCl buffer, 1ul 6.5 uM primer and 0.1 ul Taquenase

(1000U/ml). When performing more than one reaction, a reaction premix containing these reagents is made, and 6ul of the reaction premix is added to each sample. The reactions then are capped and cycled in a preheated cycler for 25 cycles using a profile consisting of 97°C for 50 seconds, and 65°C for 4 minutes. The cycling profile is linked to a soak file for storage at 4°C until retrieval. Any unincorporated terminators are removed by filtering the reaction through Sephadex G-50 spin columns as detailed above for Amplitaq dye terminator chemistry.

The reagents used in these reactions are prepared as follows: 20 x ddNTPs: Option 1: if "neat" dye terminator ddNTPs are available, mix 1ul PE/ABI Dyedeoxy A (final 1:100), 2 ul PE/ABI Dyedeoxy G (final 1:50), 0.4 ul PE/ABI Dyedeoxy T (final 1:250), 0.4 ul PE/ABI Dyedeoxy C (final 1:250) and 96.2 ul water to yield the appropriate ratios for 20 x ddNTPs. Option 2: If the PE/ABI PRISM Ready reaction Dyedeoxy terminator mix (part # 401384 for use with Amplitaq) is available, a 1:30 dilution in ddH_2O was used as the 20 x ddNTPs. 10 x $MgCl_2$/Tris-HCl buffer consisted of 500 mM Tris-HCl, pH 9.2, 160 mM ammonium sulfate, and 35 mM $MgCl_2$. 20 x dNTPs were a solution of 3 mM each dATP, dCTP and dTTP and 9 mM dITP diluted in 1x $MgCl_2$/Tris-HCl buffer.

In all instances, an initial DNA template concentration study was performed with representative samples from each 96 well isolation to determine the optimal double-stranded template concentration needed to obtain optimal signal. Studies performed adding 10% DMSO to the reactions with the standard cycling conditions mentioned above concluded that addition of DMSO inhibited the reaction. However, when an annealing temperature is added to the cycling profile (i.e., 25 cycles of 97°C or 93°C for 50 seconds, 50°C for 15 seconds, 65°C for 4 minutes), the reactions are comparable to reactions without DMSO and are useful for resolving compressions.

IV. Methods for shotgun sequence data proofreading, assembly and closure

A. Assembly algorithms

The largest challenge facing investigators using shotgun sequencing as the strategy for large scale DNA sequencing projects is the sequence assembly, proofreading and gap closure. In the era of radiolabeled sequencing, the first sequence assembly software (written by Roger Staden at the MRC in Cambridge, England)[384], designed for use on a VAX computer system, relied on manual entry of sequence data. The algorithms then formed alignments based on homologies between the individual pieces of manually entered sequenced data. This software package relied on the investigator's ability to accurately read the sequencing data and transfer it into the computer. Therefore, data entry and editing were tedious processes, which required continual referral back to the original autoradiographs to compare the sequence data from overlapping gel reads.

The Staden package of software since has been adapted to the current fluorescent sequence data acquisition methods, and has been ported to UNIX-based computers, e.g. Sun Microsystems Sparcstations and Dec Alphas. This package allows users to view the analyzed sequence data on screen at all stages of trace editing and contig alignment. The sequence data chromatograms are stored on hard disks and accessed by the Staden programs TED and XGAP. TED is the trace editing program and allows for masking of vector sequence at the 5' end and ambiguous data at the 3' end of gel reads. XGAP, or X windows based genome assembly program, uses a basic Smith-Waterman algorithm to locate the sequence similarities between the data already in the database when compared with the incoming data. Within these windows-based programs, the user can edit assembled contigs by comparing trace chromatograms, calculate a consensus for the data, and join contigs if the user-defined stringency values are met. The Staden package is useful when performing sequence data alignment on small databases with no repeated elements such as ALU or LINE repeats as found in human genomic DNA. The XGAP

assembly program has no mechanism to deal with repeated sequences or the numerous alignment possibilities that can potentially result from a database containing repeat DNA. XGAP also has no criteria regarding quality of the data entered, and therefore, requires that only high quality data be used as input.

Gene Myers and Susan Larson at the University of Arizona[385] recently have released a sequence alignment program with an altered Smith-Waterman algorithm that is contained in their fragment assembly kernel, or FAKII. This alignment algorithm essentially consists of three phases: the overlap phase, the layout phase and the multialignment phase. In the overlap phase, every gel read is compared against every other gel read in the database, both in forward and reverse orientation. This generates an overlap graph which is entered into the second phase of the assembly, the layout phase, where a series of alternate assemblies are generated. The third phase is the final, optimal assembly from the initial data that, using a modified version of FAKII developed by Hong Shing Lai in Dr. Roe's laboratory, can be viewed and edited using the XGAP interface as described above. FAKII has the advantages of speed and capability to handle larger data sets than the Staden package, but does not include a quality check for the entered data, and has some difficulty resolving differing alignment possibilities that are caused by the presence of repeated sequence regions.

A further modification of the Smith-Waterman algorithm occurred with the CAP2 program written by Huang[386]. This program includes a quality check of the initial data, and a weighting of each nucleotide based on it's position in the gel reading. By including a quality check, CAP2 gives preference to the sequence data in the first 350 nucleotides, where the accuracy of basecalling typically is >98%. Nucleotides on the 3' end of the gel reads are typically less accurate and are not relied upon as heavily in the assembly process. The criterion on which the program chooses alignments is based on this data weighting scheme and thus, CAP2 is better equipped than FAKII or XGAP to align human sequence databases including repeated elements. The output from H. Lai's

modified version of CAP2 assemblies can be viewed and edited with the XGAP program as discussed above.

More recently, Phil Green released an alpha test version of his new PHRED/PHRAP assembly program package[387]. This package also is based on the original Smith-Waterman algorithm and like CAP2, includes a quality check of the data. However, PHRED, the editing portion of this program package, contains a quality check based on the ratio of signal to noise and peak to peak spacing in the trace data. PHRED also accepts longer gel reads (and hence, poorer quality data at the 3' end of the sequence) but in addition, incorporates additional heuristic arguments in an attempt to make the 3' data more usable, although weighted less. PHRAP, the assembly portion of this program package, includes information gathered in assessing the quality of the data. As in CAP2, determining the data quality is an important criterion used by PHRAP to evaluate different possible assemblies. The assembled data from the PHRED/PHRAP alignments is viewed either using the Consed (consensus sequence editing) interface or more recently, Gap4 or Xgap after employing the phrap2gap program developed at Washington University and the Sanger Center, and implemented at the University of Oklahoma by H. Lai.

B. Shotgun sequence data validation

Validation of shotgun sequence data has become a recent concern among those involved in large scale DNA sequencing. To confirm the assembly of shotgun sequences submitted to national databases (e.g. Genbank and Genome Sequence Database), some investigators have suggested that sequencing be repeated by independent laboratories to confirm that the data submitted is in fact, assembled correctly. This notion of resequencing the human genome piece by piece after it initially has been shotgun sequenced and assembled seems to be a costly, not to mention redundant, concept for those involved in the Human Genome Project. Therefore, it was proposed by Dr. Bruce

Roe that the idea of validation be approached by a less expensive, and presently available approach that entails comparing the results of two or three of the aforementioned assembly algorithms with the same input data set. If the results are identical, then the assembly of the data could be assumed correct and therefore, valid. Any discrepancies in this separately assembled data should be manually resolved before the final sequence is released into the public domain. This multiple assembly approach has been implemented in the sequences presented in this dissertation and all conflicts have been resolved by proofreading and/or additional sequencing and PCR experiments.

C. Gap closure strategies

Sequence editing and assembly is only the first step necessary to produce contiguous final data obtained by shotgun sequencing. Closure of the inevitable gaps in shotgun sequencing is an additional challenge. During the course of this dissertation work, several strategies for closure have been investigated. One strategy entailed generating a larger fragment set by restriction digestion, subcloning, isolating and subsequently sequencing the end portion of these fragments. Since a large amount of sequence data already is present in the project database and additional restriction mapping information is available, the sequence of the ends of these restriction fragments could be used to order the contigs and to determine the relative positions of the restriction fragments. Fragments covering a gap region then could serve as templates for primer walking to obtain the final sequence necessary to close the gaps. However, since this strategy requires an accurate restriction map, which not always is available, and is time consuming, it is rarely used. A second strategy entails generating and sequencing additional shotgun subclones, while a third strategy employs custom, synthetic primers to walk for gap closure. However, both of these two strategies are costly in terms of reagents and primer synthesis.

Recently, a strategy that entails a three step closure approach has been developed in Dr. Bruce Roe's laboratory. This strategy utilizes the recent advances in polyacrylamide gel technology in the initial phase of closure. Here, templates at the ends of contigs (i.e., those whose sequences read into the gap) are selected and resequenced with both the forward and reverse universal primers to insure that the correct template clone is chosen. Then, using optimized cycle sequencing conditions, the nested fragment sets are resolved on a Long Ranger 5% gel, instead of the standard 4.25% polyacrylamide gel, because under these conditions, fragments in the range of 500 and 1000 bp from the priming site can be resolved. These longer reads often are sufficient to close many gaps in sequence data. These Long Ranger gels are not routinely employed for shotgun data acquisition because these gels require a 10 hour run on an ABI 377 and manual editing of the long trace data. Following Long Ranger gel runs, PCR is performed on any remaining gaps using universal forward and reverse primers on the subclones known to span the gap as template. The size of the resulting PCR fragment therefore reveals the size of the gap and if small (i.e., less than 2-3 kb) the gap then is closed by primer walking directly off these subclones. However, should the gap prove to be larger, the PCR product can be directly nebulized using a technique developed in our lab by Dr. Zhili Wang[345] and then the resulting subclones can be isolated and sequenced to close the gap. Should the region needed to close any remaining gaps be difficult or impossible to obtain spanning subclones or PCR products, one or more rounds of gap closure by custom synthetic primer walking with the cosmid or BAC as the sequencing template can be used as a last resort.

V. Methods for final sequence data analysis

Once shotgun sequence data is assembled, all the gaps are closed, and the final sequence validated, several computer based tools are employed for final sequence analysis.

In the first phase of analyzing the final sequence data, the sequence is compared to all know sequences present in the international databases using the BLAST program. BLAST, a basic local alignment search tool, allows Alu and other sequences to be identified based on homology with repeat databases at the National Center for Biotechnology Information (NCBI) via the internet utilizing a heuristic search algorithm to identify matches between query sequence and a database sequence[388]. BLASTX is a complementary program, developed by Jean-Michele Claverie at the NCBI[389], that replaces regions of repeat homology with Xs. The intervening, non-repeat sequences then can be extracted and searched using BLAST to find sequence homologies. A script called Autoblast, written at the University of Oklahoma by Dr. Bruce Roe and Yichen Ma, combines BLASTX and BLAST and allows for automated repeat identification, masking, extraction of non-repeat sequences and subsequent database searching.

Recently, Dr. Ed Uberbacher at Oak Ridge National Laboratories has developed an algorithm which can predict various sequence features with a high degree of accuracy[390]. The program which employs these algorithms to facilitate finding features in genomic sequences is called XGrail, an X windows based gene recognition and analysis internet link. During the course of this dissertation research, this XGrail program was used to identify potential exons (based on adherence to consensus intron/exon splice sites and codon usage), open reading frames greater than 100 bp in length, CpG islands, promoter sites, and full length repeated elements using a multiple sensor-neural network.

Finally, in many instances, the regions found by BLAST and XGrail that displayed homology to sequences in the databases, could be further studied using a variety of programs. On the VAX computer system, Dotplot, a Genetic Computer Group (GCG) program[391] and the Staden Sequence Identification Program (SIP)[392] are two programs which can graphically represent nucleotide or protein homology. In addition, the Wilbur and Lipman[393] Nucaln (nucleotide alignment) program, the Higgins Clustal V

program[394] and the GAP, Bestfit and Pileup (GCG) programs also can perform a text comparison of homology[395]. On the Sparcstation, Dotter[396] is available for graphical representation of homology similar to the GCG dotplot program and a UNIX version of Clustal V and FastA can produce a text comparison of the observed homology. Findpatterns, a GCG VAX-based program that searches a query sequence for the presence of simple repeats such as di-, tri- and tetranucleotide repeats employing a user specified string of nucleotides also is available. Staden's Nucleotide Interpretation Program (NIP) and Readseq are two programs that can translate nucleotide sequences into the corresponding protein sequence for analysis by Steve Henikoff's BLOCKS program on either Unix or VMS-based computers via the internet[397]. More recently, a WWW site at Baylor College of Medicine (http://gc.bcm.tmc.edu/search-launcher/launcher.html/) has been useful for performing a variety of internet-based queries of genomic DNA and protein structural features.

Chapter III

RESULTS AND DISCUSSION

I. Shotgun Sequencing Strategy Development

A. Overview

A total of 90,213 base pairs from the meningioma deletion region, approximately 220,000 base pairs form the Cat Eye Syndrome region and approximately 420,000 base pairs from the Multiple Endocrine Neoplasia, type 1 region were sequenced during the course of this dissertation research using a shotgun sequencing strategy[326,336] with subcloning exclusively into plasmid based sequencing template vectors. This shotgun strategy includes large scale isolation of genomic clones, physical shearing of genomic clones by nebulization, fragment end repair, size selection and 2' phosphorylation prior to blunt ended ligation into dephosphorylated pUC vectors[336,337,345,398]. After bacterial electrotransformation, the pUC-based subclones were isolated using an automated alkaline lysis method developed during the course of this dissertation study with the Beckman Biomek pipetting robots. The random subclones were sequenced using one of several mutants of *Taq* polymerase in cycle sequencing reactions[325-328,336]. Following pooling, and removal of unincorporated dye-labeled terminators by chromatography, the sequencing reactions were loaded onto either an ABI 373A or 377 fluorescent DNA sequencer for automated electrophoresis, data collection and basecalling. The sequence data then was transferred via NSF Share to a Sparcstation where it was analyzed using one of several available computer programs for editing, alignment, proofreading and viewing (XGAP, deadted, FAKII, CAP2, PHRED, PHRAP, consed, crossmatch)[384-387]. Gaps in the aligned sequence were closed using either long ranger gel electrophoresis conditions and Amplitaq FS cycle sequencing to obtain longer reads past the priming site, or a by primer walking strategy with custom synthesized oligonucleotide

primers, also using Amplitaq FS catalyzed dye terminator cycle sequencing reactions. In some instances an alternate PCR-based[375] method, developed by Zhili Wang in our lab, was used to close larger sequence gaps. Here, the gap region was amplified from the original genomic clone, sheared in the nebulizer, end-repaired, ligated into pUC vectors and the resulting randomly picked clones were sequenced via *Taq*-based dye terminator chemistry.

Once an entire contiguous sequence of a target genomic clone was obtained, the sequence was validated by PCR with custom oligonucleotide primers, demonstrating that the expected size of PCR fragments correlated with the experimentally obtained fragments. Also, the databases were examined using each of the three assembly algorithms available to confirm the correct alignment of the random subclones. Validated, contiguous sequence was analyzed for database homology[388], potential exons[390], and repeated elements[388,389,395].

Throughout this dissertation work, several stages in the shotgun strategy have been modified and optimized. The large scale diatomaceous earth-based genomic clone isolation was developed and optimized for use in our lab with Kala Iyer[347] to eliminate the CsCl/ethidium bromide equilibrium ultracentrifugation previously required and Angelika Bodenteich optimized the procedures for nebulization in generating shotgun subclone libraries[336]. The plasmid template isolation endured many modifications by colleagues (Steve Toth, Dennis Burian, Guozhang Zhang, Zhili Wang) and the final automation of this optimized method is reported here for use on both the Biomek 1000 and the Biomek 2000 pipetting robots. The lab shifted from dye-labeled primer chemistry for shotgun data collection to dye-labeled terminator chemistry based on experiments performed by Steve Toth and these reactions were optimized using Klentaq TR by Guozhang Zhang.

B. Large Scale Diatomaceous Earth-based genomic clone isolation for cosmid, BAC, PAC and fosmid

In 1994, our lab reported a diatomaceous earth-based procedure for isolating genomic clones, specifically plasmids and cosmids[347]. This method is based on the original alkaline lysis procedure by Birnboim and Doly, followed by binding the DNA to defined diatomaceous earth, washing, and subsequently eluting the DNA. This procedure yielded a high proportion of supercoiled DNA which was ideal for restriction digestion analysis, and for shotgun subclone library production without CsCl/ethidium bromide equilibrium ultracentrifugation. At the time of this report, cosmid-based genomic clones were quite popular for large scale sequencing projects based on insert size availability. Cosmid genomic clones contain 40-50 kilobase pair inserts, much larger than the plasmid-based clones, and were thought to be easily handled, isolated and maintained. However, it was soon realized that cosmids did not always maintain the genomic insert, and showed periodic recombination and insert deletion.

Bacterial artificial chromosomes (BACs) soon became the genomic clone of choice based on several criteria, including low incidence of recombination, high clone stability and increased insert size capacity when compared with cosmids or plasmids. BACs maintain insert sizes from 100 to 300 kb with the average size in most insert libraries around 150 kb. Because of this large insert size, BAC clones presented a problem when isolated by the diatomaceous earth procedure developed earlier and reported by H.Pan, et al. The resulting BAC DNA often was contaminated with host bacterial genomic DNA at levels as high as 50% as evaluated by DNA sequencing. Therefore, modifications to the original diatomaceous earth based large scale isolation were required.

Initially, it was believed that harsh treatment of the cells during resuspension/cell lysis resulted in increased shearing of the host genomic DNA and contaminated BAC preparations. Further studies indicated that this was true to some extent, but the host

contamination was still around 30% as determined by DNA sequencing experiments. To make the resuspension/cell lysis treatment more gentle, the amount of initial cell culture was decreased by half (yielding an effective doubling of the volume of reagents) to allow for easier resuspension of cell pellets, and more gentle treatment. Again, the percentage of host contamination was evaluated by sequencing, and still hovered around 15%. At this stage, it was perceived that perhaps all available DNA binding sites present on the diatomaceous earth were occupied and the resulting contamination was a direct consequence. To test this hypothesis, the absolute amount of diatomaceous earth in the preparation was increased from 4 g to 5 g using the decreased initial cell culture. Host contamination dropped to 6% (+/- 1%) by DNA sequencing analysis (procedure used for isolation number 4 and 5 was identical; see figure 16). Therefore, decreasing the initial cell concentration and increasing the amount of diatomaceous earth resulted in an effective increase in the binding sites present on the diatomaceous earth which allows for more specific binding of contaminating bacterial genomic DNA and preferential elution of BAC DNA. These modifications reproducibly yield supercoiled BAC DNA with low amounts of host genomic contamination, that is sufficient for large scale sequencing applications.

C. Automated Double Stranded Subclone Isolation

Early large scale shotgun sequencing efforts used a single stranded subcloning approach utilizing M13-based cloning vectors. To this end, single stranded DNA template isolation procedures were developed, optimized and finally automated in the Roe laboratory by Elaine Mardis[356], with subsequent modifications by Stephanie Chissoe[326]. The automation of the single stranded template preparation increased productivity and reproducibility lending support to the shotgun approach as a viable strategy for large scale DNA sequencing efforts.

Around 1992, the focus began changing in the direction of double stranded subcloning approach to achieve a two-for-one advantage over the single stranded

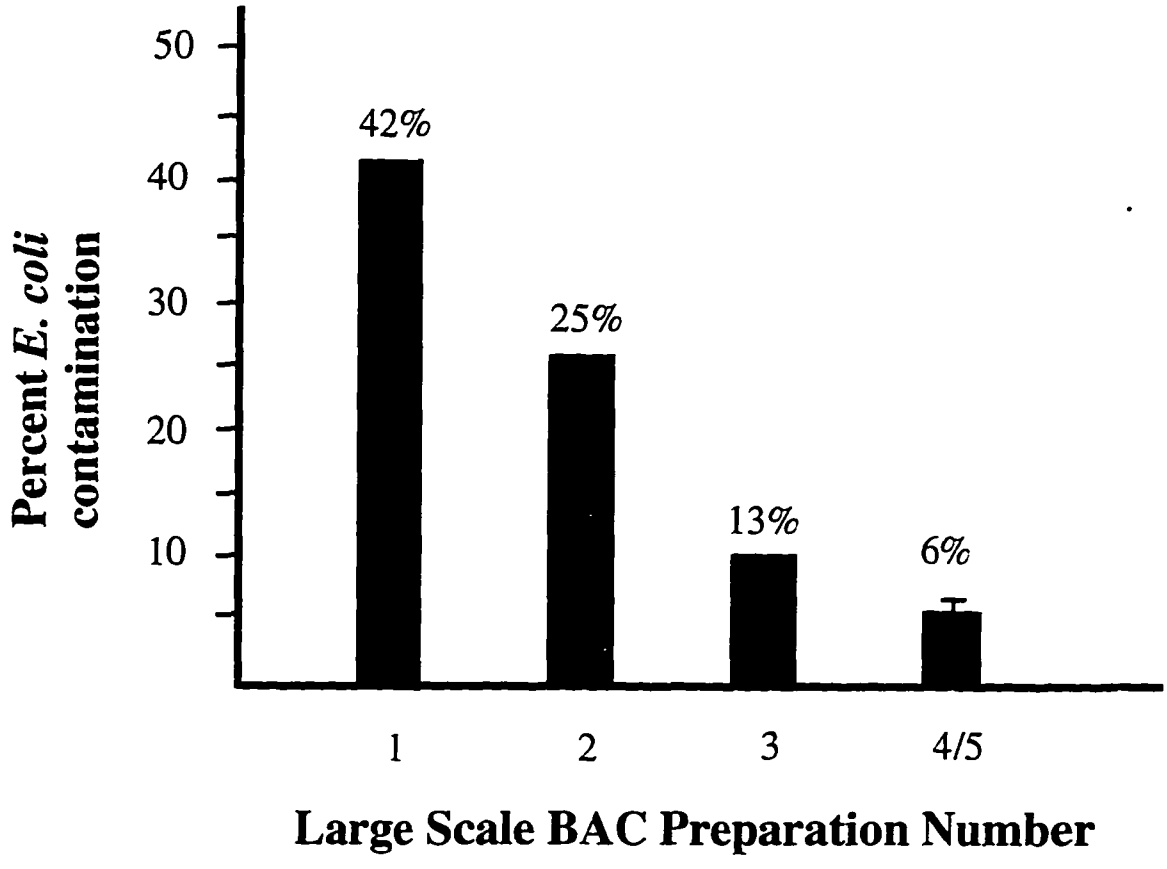


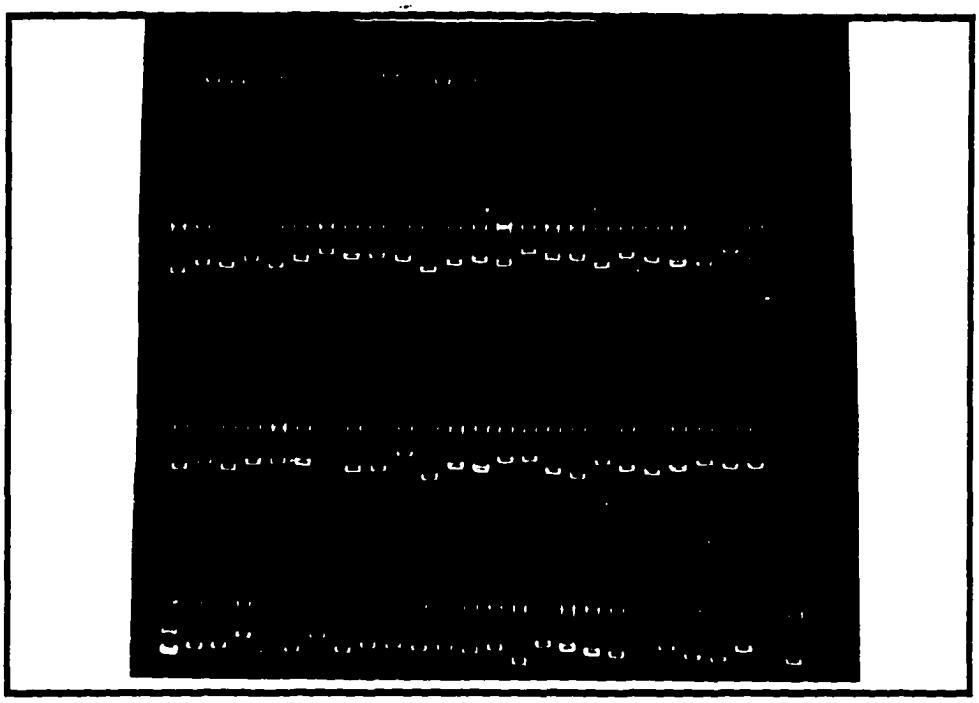
Figure 16: *Escherichia coli* host contamination in preparations of BAC 18H3 as evaluated by DNA sequencing.

methods. With double stranded templates, both ends of the subclone template could be sequenced using the universal forward and universal reverse sequencing primers, as opposed to a single stranded subclone which could only be sequenced with the universal forward primer. This two-for-one advantage decreased the number of subclones required by a factor of two and also provided valuable positional information for the database assembly process. For many years, these advantages were outweighed by one disadvantage, that the double stranded template isolation techniques were not as mature as the single stranded template isolations. Double stranded preparations were time and labor intensive, often involving hazardous organic extraction steps.

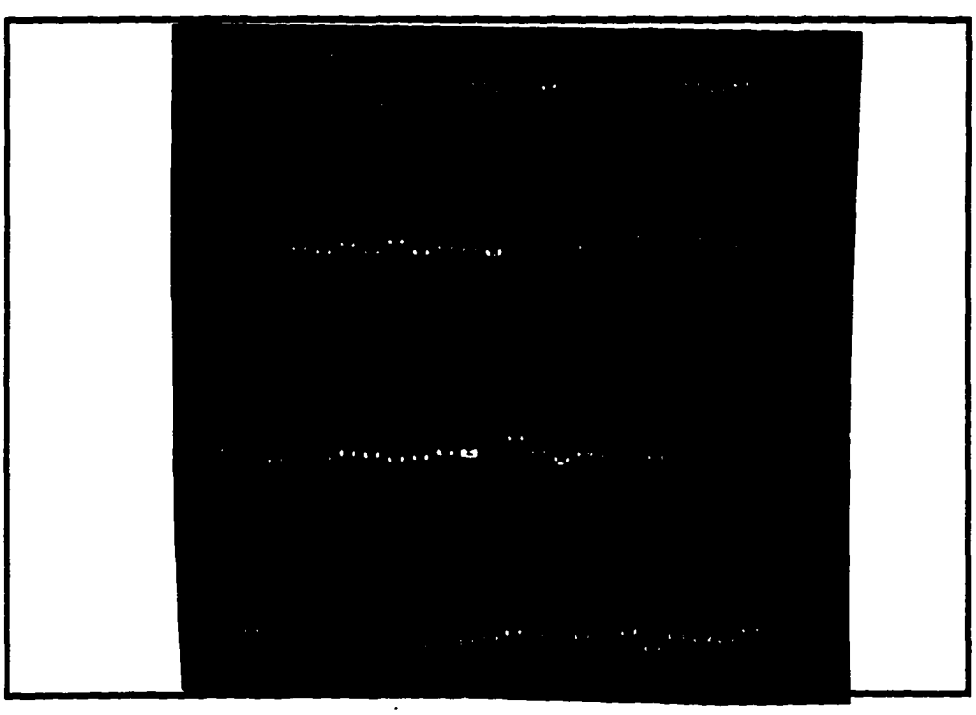
As the strategy matured, the double stranded template isolations became more viable, with many commercial kits available. After investigating a myriad of commercial kits, it was determined that a standard alkaline lysis alone followed by ethanol precipitation worked equally well to generate sequence ready plasmid templates.

Automation of the alkaline lysis procedure initially was performed on the Beckman Biomek 1000 laboratory workstation. Cells containing the plasmid of interest were grown and harvested in a 96-well format, followed by alkaline lysis as described, with the lysate cleared by filtration. The resulting DNA was concentrated by ethanol precipitation, washed and resuspended in buffer for use. This automated procedure allowed for the simultaneous isolation of 96 subclones in three hours. Although this procedure worked relatively well, the cell lysate periodically clogged the filter and resulted in a loss of DNA. A typical agarose gel photograph of Biomek 1000 plasmid DNA is shown in figure 17.

After further experimentation, it became clear that filtration might not be the most effective way to clear the lysate. Replacement of the Biomek 1000 robots with the next generation Beckman Biomek 2000 instruments allowed for reevaluation of the strategy used in the automation procedure. Steve Toth performed several experiments investigating centrifugation of the 96 well plates as an alternate method to clear the lysate



A. Biomek 1000 double stranded template isolation



B. Biomek 2000 double stranded template isolation

Figure 17: Agarose gel photographs of Biomek-isolated double stranded template DNA.

in double stranded preparations. Simultaneously, Dennis Burian performed an experiment which indicated that doubling the reagent volumes increased the final yield of plasmid as obtained by a manual 96 well preparation with centrifugation to clear the lysate. By incorporating all of these minor adjustments into the method, the double stranded template isolation was automated on the Beckman Biomek 2000 such that 4 x 96 subclones were isolated simultaneously in 4 hours. The lysate was cleared by centrifugation, with the Biomek pipetting off the top 400 ul supernatant to a clean 96 well block for ethanol precipitation. This automated procedure is highly reproducible, even by novice users, and generates sufficient template for many sequencing reactions. A typical agarose gel photograph is shown in figure 17.

D. Cycle Sequencing Reactions

Throughout the past several years, there was a shift in the technical strategy of cycle sequencing. With an automated, efficient plasmid-based template preparation, the laboratory switched from the dye-labeled primer cycle sequencing chemistry to the dye-labeled dideoxyribonucleotide terminator cycle sequencing chemistry for shotgun sequence generation. This eliminated the four tube per sample requirement of dye primer chemistry and standardized all manipulations in a 96-well format, allowing more efficient use of thermocyclers, thus eliminating a major bottleneck in the large throughput sequencing strategy.

This increased cycling output made the ABI fluorescent sequencers rate limiting, a problem that was slightly relieved by an increase in capacity from 36 to 48 samples per run. Other modifications to the sequencing instrument include an automatic sample sheet production program initially developed by LaDeana Hillier at Washington University and then adapted for our laboratory by Dr. Roe and Steve Kenton, and decreased run time from 3.5 hours to 3.0 hours, which allowed one additional run to be performed per day, bringing the total to four runs per day per instrument. Current output with 7 instruments

run four times per day, 48 samples per run and an average read length of 400 bp per sample, is over 0.5 Mb of raw sequence data per day.

Alongside the physical improvements gained from the dye terminator chemistry were several chemical improvements such as the introduction of improved thermostable enzymes that increased incorporation of modified nucleotides, and the subsequent base calling fidelity.

II. Strategy Implementation

A. Sequence Assembly, gap closure and sequence validation

The shotgun sequencing data collected from the ABI DNA sequencers for each project was transferred to a Sparcstation and subsequently assembled using a series of programs (FAKII, CAP2, phred/phrap). These programs compare the data from gel reads and calculate the best alignment for assembly of the data into contiguous pieces or contigs. These programs are faster, more accurate and require less manual intervention than the previous generation of assembly programs, XGAP, although FAKII and CAP2 utilize the XGAP windows interface for viewing the assembled data. Consed is the viewing tool for phred/phrap databases, although a phred/phrap database may be converted to an XGAP database following assembly if the XGAP interface is preferred. Gap closure was performed as previously described, and the final contiguous sequence assembly validated by one of three methods. First, the restriction pattern of the cosmid or BAC is compared with the computer generated restriction map using the consensus sequence. Second, the assembly is performed by all three of the above mentioned assembly programs, validating the assembly if all three alignments agree. Any discrepancies must be resolved before the sequence is considered valid. Third, the alignment of known cDNA sequences which correspond to the genomic DNA must be in the correct order and orientation in the assembled sequence. If all of the above criteria are

satisfied, the assembled sequence is considered valid and ready for release into the public databases as completed, level three data.

Each base of the contiguous sequence also must conform to the "rule of three" to assure the highest quality data. Therefore, each base must have at least three fold redundancy and that one of the three gel reads must be in the opposite orientation. This ensures that regions that are difficult to sequence are sequenced accurately. In some cases (for example, G/C rich regions) a region of DNA will not give accurate base resolution, no matter how much redandancy is present. Sequencing in the opposite orientation oftentimes resolves the difficult area, and fulfills the rule of three. In cases when no subclone is available for sequencing in the opposite direction, a custom primer is synthesized to sequence in the opposite orientation.

B. Sequence Scanning

In some cases, complete sequencing to rule of three accuracy is not required to find the majority of genes present in genomic sequence. Upon completion of the shotgun phase in sequencing a cosmid or BAC, the resulting contigs can be ordered using the forward and reverse read pairs, and subsequent sequence analysis can be performed. Typically all of the known genes and most of the unknown genes can be detected using the available computer analysis methods as described below. Database searches and homology to the individual contigs can also aid in ordering the contigs, based on known cDNA sequences present in the databases. Examples of sequence scanning found in this dissertation include b137c7 from the MEN-1 region on chromosome 11q13, and b376a1 from the Cat Eye Syndrome region of chromsome 22q11. As discussed below, b137c7 contains three known human genes and four putative genes, one of which is suspected to be the MEN-1 gene. In contrast, b376a1 contains no known or putative genes based on sequence scanning analysis.

C. Analysis of Completed Sequences

Analysis of the completed sequences requires the use of several different computer programs available both on the Sparcstations and the VAX. These programs allow for placement of genes, ESTs and STSs to obtain genomic organization of known genes, for identification of new genes by comparative study of homologies between human and other organisms, and for prediction of potential genes based on promotor regions, intron/exon splice junctions, polyadenylation sites and codon frequency. One such program, Xgrail (X windows gene recognition and analysis internet link) predicts potential exons by examining the consensus sequence for promotor regions, splice sites, potential open reading frames and start/stop codons. Exons predicted by Xgrail are given a quality rating of excellent, good and marginal based on their adherence to the defined criteria. Xgrail also identifies simple sequence repeats such as di-, tri- tetra-, and pentanucleotide repeat regions, and highlights polyadenylation signals, CpG islands and potential promotor regions on the graphical display.

Powblast, a recent database search tool, combines a filter for repeat sequences and a BLAST (basic local alignment search tool) search with gapped alignment to yield a listing of entries in the Genbank database and the segment of that entry with homology to the genomic input sequence. Powblast, utilizes the BLAST searching algorithm after screening out known elements, for example those in the Alu and Mir repeat families. This pre-screening is useful in that it eliminates all but the unique regions of the database homology. Output from Powblast can be viewed as either a text file which aligns the input sequence with the region of database sequence homology, showing both the direction and position of the homology, or as a pictoral view of the database homology using the Musk program that allows a graphical view of the homology showing both directionality and position. In addition, Musk also has the ability to retrieve the relevant database entries for inspection.

III. Meningioma Sequencing and Analysis

A total of 90,213 bases from the Meningioma Deletion region from human chromosome 22q11 was sequenced using a shotgun based strategy followed by a directed sequencing for gap closure. The sequence of the three cosmids 76A4, 50B11 and 58B12 is in one contiguous piece and has been assigned the Genbank accession number AC000105. These clones were obtained from collaborator Dr. Ellen Zwartoff at Erasmus University in the Netherlands.

A. MN-1 gene

Analysis of the total 90,213 bases from the Meningioma Deletion region from human chromosome 22q11 reveals that it contains both exons for the MN-1 gene. The cDNA for this gene has been characterized[152], and but the genomic sequence only was characterized partially as the two exons present in the MN-1 gene were mapped to the q11 region of chromosome 22 and from these mapping studies, it was estimated that the entire gene spans approximately 70 kb[152]. Through the sequence studies done during the course of this dissertation research, the 22q11 MN-1 gene containing region now could be characterized in much greater detail.

A comparison of the Xgrail and Powblast output shown in figure 18 reveals that both MN-1 exons were accurately predicted by Xgrail in regions corresponding to the publicly available MN-1 cDNA sequence (Genbank accession number X82209). These regions also had significant homology to several human fetal brain ESTs (Genbank accession numbers D80935, Z28469, R59212, M85938, H89207, R59272, H89103, Z28468, D80247, D51775, D51800, D52119, C14502, Z39850, D59763, and D60570), that represent the expression of the MN-1 gene in this tissue. As expected, this region also contains the human STS SHGC-31555 sequence (Genbank accession number G28534), and the first exon of MN-1 has homology with many previously studied human genomic clones isolated based on their CpG island content (Genbank accession numbers Z64626, Z62107, Z79869, R52312, T61559 and Z58982).

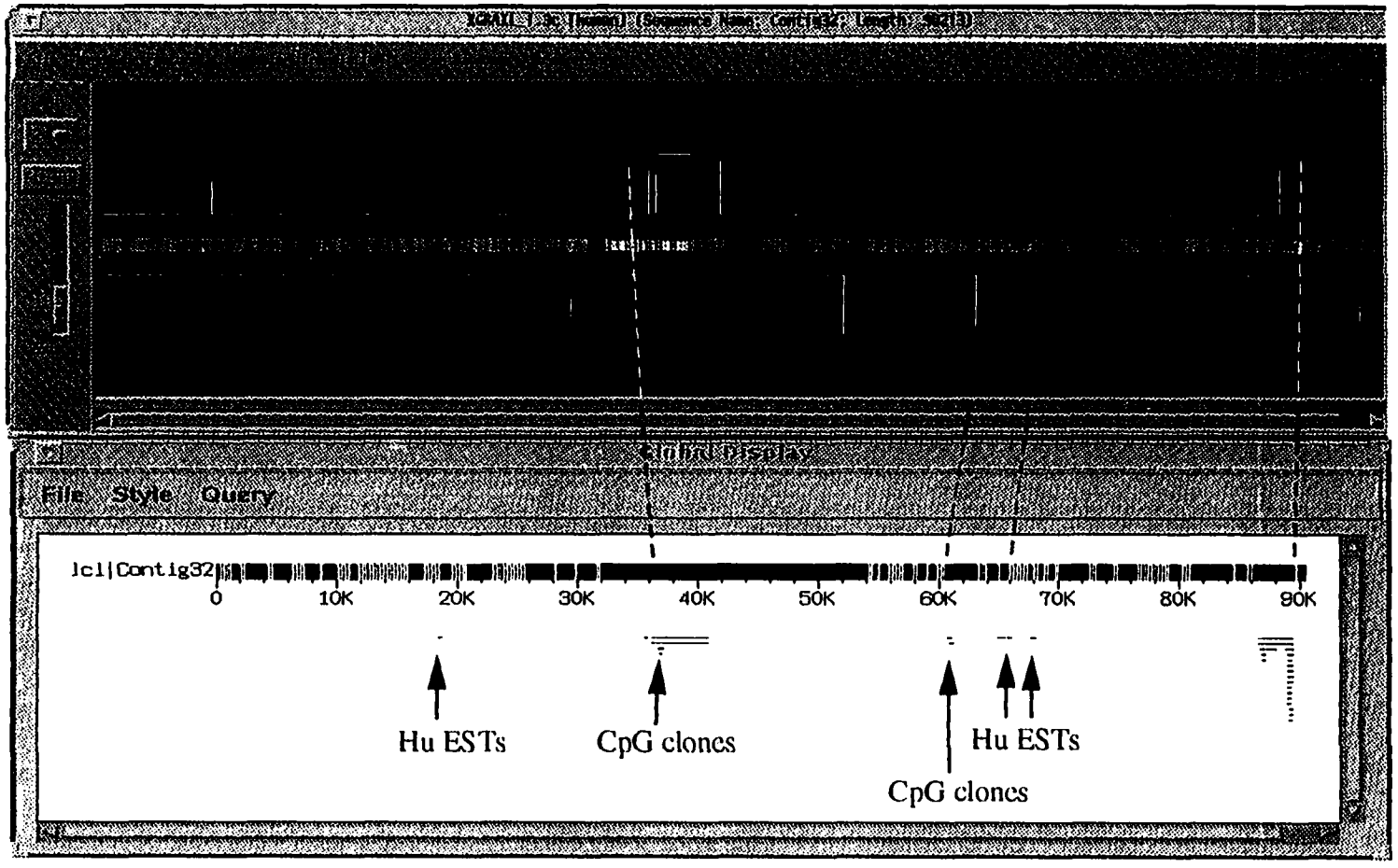

File Style Query
lcl|Contig32
0
10K
20K
30K
40K
50K
60K
70K
80K
90K
Hu ESTs
CpG clones
CpG clones
Hu ESTs
Strong database hits for MN-1
MN-1 gene organization

Figure 18: Top panel: Graphical result of Xgrail gene identification software for the Meningioma Deletion Region from human chromosome 22q11. The peaks present are Xgrail predicted exons with a quality rating of excellent (green), good (blue), or marginal (red). The light blue lollipops are predicted polyadenylation signals and the purple squares are CpG islands. Yellow and red squares indicate promotors and the vertical orange bars denote simple sequence repeats. Bottom panel: Graphical display of the powblast output using Musk. The black bar at the top of the panel represents the consensus sequence with the grey areas indicating repeat regions. The lines under the black bar indicate homology between the consensus sequence and an entry in the public database.

Additional regions of significant database homology also were found (Genbank accession numbers Z63230 and Z63228) indicating the possible presence of an additional gene in this 90 kb rgion.

The genomic organization and consensus splice junctions for the MN-1 gene are shown in tables 3a and 3b, respectively, and the cDNA sequence with the amino acid sequence and exon boundaries is shown in Figure 19. The first MN-1 exon is 4.7 kb which encodes an approximately 887 bp 5' untranslated region and 3846 bp corresponding to the N-terminal 1578 amino acids of the MN-1 protein. The second MN-1 exon is 2.8 kb long and encodes for 159 amino acids of the C-terminal domain of the MN-1 protein (total of 1757 amino acids) and 2644 nucleotides of the 3' untranslated region, following a large 45.6 kb intron. As can be seen from table 3b, the two intron/exon boundaries conform quite nicely to the consensus splice junctions. Interestingly, the MN-1 gene does not contain the usual TATAA promotor region, however, a strong CpG island is present in the region 5' to the first exon. This is not unusual in that approximately 60% of human genes are associated with CpG islands, and many of these genes have been found to be 'housekeeping genes'[403-406]. The MN-1 gene does, however, display a polyadenylation signal at the end of the 3' untranslated region from 89018-89024, and a TGA stop codon at 86738-86740.

The DNA sequence of this 90 kb region also contains several repeated elements, with LINE, ALU, SVA and MIR repeats being the most common. In one very interesting case at 66,000 bp (see figure 20), the EST with Genbank accession number T61559 has significant homology with an ALU repeat. This EST does contain some unique sequence, as seen in the Dotter sequence comparison in figure 21. The entire meningioma deletion region contains 29 SVA, 44 ALU, 58 L1, 9 MLT, 22 MIR, 2 MST, 1 THE and 3 MER repeats as shown by powblast output, especially in the region upstream from the first MN-1 exon (see figure 22).

A. Genomic organization - Meningioma Deletion Region:

exon number	position	size of exon	size of intron
01	36158-40891	4734	45664
02	86555-89374	2820	

B. Splice junctions - Meningioma Deletion Region:

Left Exon	Left splice junction	Right splice junction	Right exon
01	AGCAAAGAAGgtatatgcac		01
02		ctctccacagCCCACGACCT	02

consensus: CAGgtaagt / A t — tncagG / c

polyadenylation signal: 89018-89024
TGAstop: 86738-86740

Table 3: Genomic sequence organization and splice junctions for the two large exons of the MN-1 gene in the Meningioma Deletion Region on human chromosome 22q11.

```
gcgcggaggcagggaaccggagccttggagcgacccaacccctcgtctcgctgccctccc      60
gcgcctgcaacggtgcgcggagactccggcgaactcagacacccaacggcggagaacaga     120
agcggcaggcggcggacgtggcccggaagctgcgcgccgaacgcagcgcacccgctgccg     180
agcagaggagccgcgcctttccccgaccctcggctccagcccccggcgcctgccgcctcg     240
cagcccctctgcgtcctcggctcgggggccggcaccggcgatgccgagcggacgctccag     300
tcctccgacccgctgaagaagcagcagccgctcgcccggagcctacggggattgtgcgag     360
cggatcgtgctcggtggaggctcgggctgcggggcgcggggactccgggggggcgggggga     420
gggaccgctctgtcggtgcccggcgccagccgcggctttgaagggtctccctcccctgcc     480
cttagcagctctgccacggactccgggaggctgcgggcggcgtcctgagggctccccagc     540
agacccaatcggacttgagaaggtgatcgctctgctctcccaacccccttcccctccccat     600
tccccccacttaacttttgtctccgttcatccgcggcttcgtcccctcccccggcagacc     660
cacccgcggctgtgacaaccgcccggggcatgggcccccccaacacggctcctagaggccc     720
cgcggcctcgcaagatgtgagaggccctccccgggcagaatcggagcttcaggagaggag     780
ctaataccccgcccccccgtccctcacatcaggcggggtggaggtgcgcgctgagcccccg     840
                                                  M  A  G  G
cggtgctgagcgtcccggagcgcccaatcctgggctggaacgagtagatggccggaggcg     900
A  P  R  R  A  G  C  H  A  L  L  I  P  L  C  P  P  P  S  M
cgccgcggagagccggctgtcatgccctattgatccccctctgcccccccgccaagtatgt     960
F  G  L  D  Q  F  E  P  Q  V  N  S  R  N  A  G  Q  G  E  R
ttgggctggaccaattcgagccccaggtcaacagcaggaacgctggccagggcgagagga    1020
N  F  N  E  T  G  L  S  M  N  T  H  F  K  A  P  A  F  H  T
actttaacgagaccggactgagcatgaacacccactttaaggccccggctttccacactg    1080
G  G  P  P  G  P  V  D  P  A  M  S  A  L  G  Q  P  P  I  L
gggggccccctggccctgtggatcctgctatgagcgcgctggggcaacccccgatcttgg    1140
G  M  N  M  E  P  Y  G  F  H  A  R  G  H  S  E  L  H  A  G
gcatgaacatggagccctacggcttccacgcgcgcggccactcggagttgcacgcagggg    1200
G  L  Q  A  Q  P  V  H  G  F  F  G  G  Q  Q  P  H  H  G  H
ggctgcaagcgcagcctgtgcacggcttctttggcggccagcagcctcaccacggccacc    1260
P  G  S  H  H  P  H  Q  H  H  P  H  F  G  G  N  F  G  G  P
cgggaagtcatcatccccaccagcatcacccccactttgggggcaacttcggtggcccgg    1320
D  P  G  A  S  C  L  H  G  G  R  L  L  G  Y  G  G  A  A  G
accccggggcctcgtgcctgcacgggggtcgcctgctcggctacggcggcgcagccggag    1380
G  L  G  S  Q  P  P  F  A  E  G  Y  E  H  M  A  E  S  Q  G
gcctgggcagccagccgcccttcgccgagggctatgagcacatggcggagagccaggggc    1440
P  E  S  F  G  P  Q  R  P  G  N  L  P  D  F  H  S  S  G  A
ctgagagcttcggcccgcagcgaccggggaacctcccggacttccacagttcaggtgcct    1500
S  S  H  R  V  P  A  P  C  L  P  L  D  Q  S  P  N  R  A  A
ccagccaccgcgtgccggccccatgcctgccgctggaccagagccctaaccgagccgcct    1560
S  F  H  G  L  P  S  S  S  G  S  D  S  H  S  L  E  P  R  R
ccttccacggcctgccgtcctccagcggctccgattcccacagtctggagccacggaggg    1620
V  T  N  Q  G  A  V  D  S  L  E  Y  N  Y  R  A  R  R  P  R
tgacgaaccaaggagccgtcgactcgctggaatacaattaccgggcgaggcgccctcggg    1680
D  I  L  T  C  F  R  P  L  T  P  K  G  S  C  L  I  M  Q  R
acattttgacatgttttcgccctctgactccgaagggcagctgcctcattatgcagcggg    1740
V  A  R  F  L  G  G  G  F  P  G  A  P  S  A  M  P  R  A  A
tcgccaggttcctgggggggcggctttcccggggcgccctcggccatgcccagagctgcgg    1800
G  M  V  G  L  S  K  M  H  A  Q  P  P  Q  Q  Q  P  Q  Q  Q
gcatggtgggcttgtccaaaatgcacgcccagccaccgcagcagcagccccagcagcagc    1860
Q  Q  P  Q  Q  Q  Q  Q  H  G  V  F  F  E  R  F  S  G  A  R
agcagccccagcagcagcagcagcatggtgtgttctttgagaggttcagtggggccagaa    1920
K  M  P  V  G  L  E  P  S  V  G  S  R  H  P  L  M  Q  P  P
agatgcctgtgggtctggagccctcagtgggctccaggcacccgttaatgcagcctcccc    1980
Q  Q  A  P  P  P  P  Q  Q  Q  P  P  Q  Q  P  P  Q  Q  Q  P
agcaggccccgccacccccctcagcagcagcccccgcagcagccgccacagcagcagccgc    2040
P  P  P  P  G  L  L  V  R  Q  N  S  L  P  A  C  A  P  S  A
cgccgccacccgggcttctagtccgacaaaattcgttgcccgcctgcgctccctcggccc    2100
P  A  G  R  G  G  H  A  Q  R  R  P  A  G  R  R  P  H  A  A
cagcagggcgaggcgggcacgcccagcggcggcctgcaggacggaggccccatgctgccc    2160
```

```
Q  P  A  R  A  I  R  V  S  H  P  P  A  G  E  P  E  H  A  P
agccagcacgcgcaattcgagtatcccatccaccggctggagaaccggagcatgcaccct    2220
L  F  R  A  C  F  Q  H  A  A  S  S  S  A  A  G  A  Q  P  A
tattccgagcctgttttcagcatgcagcatcctcctccgcagcaggcgcccaaccagcgg    2280
A  A  A  F  R  R  A  P  L  H  E  R  G  Q  E  A  R  F  D  F
ctgcagcatttcgacgcgccccctacatgaacgtggccaagaggcgcgcttcgactttc    2340
P  G  S  A  G  V  D  R  C  A  S  W  N  G  S  M  H  N  G  A
cgggcagcgcgggagtggaccgctgcgcttcgtggaacggcagcatgcacaacggcgctc    2400
L  D  N  H  L  S  P  S  A  Y  P  G  L  P  G  E  F  T  P  P
tggataatcacctctccccttccgcctacccaggcctacccggcgagttcacaccgcctg    2460
V  P  D  S  F  P  S  G  P  P  L  Q  H  P  A  P  D  H  Q  S
tgcccgacagcttcccttcggggccgcccctgcagcatccggccccggaccaccagtccc    2520
L  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q  Q
tgcaacagcagcagcagcagcagcagcagcagcaacagcagcagcagcagcagcaacagc    2580
Q  Q  Q  Q  Q  Q  Q  Q  Q  R  Q  N  A  A  L  M  I  K  Q  M
aacagcaacagcagcagcagcagcgccaaaacgcggccctcatgattaagcagatgg      2640
A  S  R  N  Q  Q  Q  R  L  R  Q  P  N  L  A  Q  L  G  H  P
cgtcgcggaatcagcagcagcggctgcgccagcccaacctggctcagctaggccaccccg    2700
G  D  V  G  Q  G  G  L  V  H  G  G  P  V  G  G  L  A  Q  P
gggacgtgggccagggcggcctggtgcatggcggcccggtgggcggcttggcccagccga    2760
N  F  E  R  E  G  G  S  T  G  A  G  R  L  G  T  F  E  Q  Q
actttgagcgcgaaggcggcagcacgggcgccgggcgtctgggcaccttcgagcagcagg    2820
A  P  H  L  A  Q  E  S  A  W  F  S  G  P  H  P  P  P  G  D
cgccgcacttggcgcaagagagcgcgtggttctcaggtccgcatccgccgcccggagacc    2880
L  L  P  R  R  M  G  G  S  G  L  P  A  D  C  G  P  H  D  P
tgctgccccgtaggatgggcggctcgggtctgcccgctgactgtggcccgcacgacccca    2940
S  L  A  P  P  P  P  P  G  G  S  G  V  L  F  R  G  P  L  Q
gcctggcgcccccctcctccgcctggtggctcgggggtgctgttccggggccctctgcagg    3000
E  P  M  R  M  P  G  E  A  T  C  R  A  A  F  T  G  L  Q  F
agccgatgaggatgcccggagaggccacgtgccgcgctgccttcaccggcctgcagttcg    3060
G  G  S  L  G  G  L  G  Q  L  Q  S  P  G  A  G  V  G  L  P
ggggcagtctgggaggcctgggtcagctgcagtcgcccggggcgggcgtggggctcccca    3120
S  A  A  S  E  R  R  P  P  P  P  D  F  A  T  S  A  L  G  G
gcgctgcttcggagcgccggcccccgccgccggactttgctacgtctgcgctcgggggcc    3180
Q  P  G  F  P  F  G  A  A  G  R  Q  S  T  P  H  S  G  P  G
agccgggctttccgtttggtgcagccggccggcagtccacgccgcacagcggtccaggcg    3240
V  N  S  P  P  S  A  G  G  G  G  G  S  S  G  G  G  G  G  G
tgaactcgccccccagcgcgggaggggggcggtggcagctctggtggcggcggtggcgggg    3300
G  A  Y  P  P  Q  P  D  F  Q  P  S  Q  R  T  S  A  S  K  L
gtgcctacccgccgcagcctgatttccagcccagccagcgcacctcggccagtaaattgg    3360
G  A  L  S  L  G  S  F  N  K  P  S  S  K  D  N  L  F  G  Q
gcgcgctctcgctgggctccttcaacaagcccagctccaaggacaacctgttcggccaga    3420
S  C  L  A  A  L  S  T  A  C  Q  N  M  I  A  S  L  G  A  P
gctgcctggctgcgctctccaccgcttgccagaacatgatcgccagcctcggggccccca    3480
N  L  N  V  T  F  N  K  K  N  P  P  E  G  K  R  K  L  S  Q
acctcaacgtgaccttcaacaagaagaacccgccagagggcaagaggaaactgagccaga    3540
N  E  T  D  G  A  A  V  A  G  N  P  G  S  D  Y  F  P  G  G
acgagaccgacggcgcggcagtggccggcaacccgggctcggattacttcccaggaggga    3600
T  A  P  G  G  P  R  T  R  R  P  S  G  T  S  S  S  G  S  K
ctgctcctgggggccccaggacccggaggccgtccgggaccagtagcagcggctccaaag    3660
A  S  G  P  P  N  P  P  A  Q  G  D  G  T  S  L  S  P  N  Y
cctcggggccgcccaaccctccagcccaggggacggcaccagcctctcccccaactaca    3720
T  L  E  S  T  S  G  N  D  G  K  P  V  S  G  G  G  G  R  G
ccctggaatccacgtcggggaatgacggcaagccggtctccgggggcggcggccggggac    3780
R  G  R  R  K  R  D  S  G  H  V  S  P  G  T  F  F  D  K  Y
ggggtcgcagaaaaagggacagtggtcacgtgagccctggcaccttctttgacaagtact    3840
S  A  A  P  D  S  G  G  A  P  G  V  S  P  G  Q  Q  Q  A  S
cggcggctccggacagcgggggcgcacctggggtgagcccagggcagcagcaagcgtcag    3900
```

```
G  A  A  V  G  G  S  S  A  G  E  T  R  G  A  P  T  P  H  E
gcgcagccgtcgggggaagctccgcaggcgagacgcgcggggcaccgacgccccacgaaa    3960
K  A  L  T  S  P  S  W  G  K  G  A  E  L  L  L  G  D  Q  P
aggcgctcacgtcgccatcctgggggaaggggggctgagttgctcctgggggatcagccgg    4020
D  L  I  G  S  L  D  G  G  A  K  S  D  S  S  S  P  N  V  G
acctcattgggtccctggacggcggggccaagtcggacagtagttcgccaaacgtgggtg    4080
E  F  A  S  D  E  V  S  T  S  Y  A  N  E  D  E  V  S  S  S
agttcgcctcggacgaggtgagcacgagctacgccaatgaggacgaggtgtcgtccagct    4140
S  D  N  P  Q  A  L  V  K  A  S  R  S  P  L  V  T  G  S  P
ctgacaacccccaggcactagttaaagcgagcaggagtcccctggtgaccggctcgccca    4200
K  L  P  P  R  G  V  G  A  G  E  H  G  P  K  A  P  P  P  A
aactccctccccgtggggtaggcgccggggaacacggaccgaaggcgcccccgcccgccc    4260
L  G  L  G  I  M  S  N  S  T  S  T  P  D  S  Y  G  G  G  G
tcggcctgggcatcatgtctaactctacctcgacccctgacagctacggcggcggtgggg    4320
G  P  G  H  P  G  T  P  G  L  E  Q  V  R  T  P  T  S  S  S
gcccgggccatccgggcactccgggcctggagcaggtccgcaccccgacgagcagcagcg    4380
G  A  P  P  P  D  E  I  H  P  L  E  I  L  Q  A  Q  I  Q  L
gcgccccgccacccgacgagatccacccccctggagatccttcaggcgcagatccagctac    4440
Q  R  Q  Q  F  S  I  S  E  D  Q  P  L  G  L  K  G  G  K  K
agaggcagcagttcagcatctccgaggaccagcctctggggctgaagggtggcaagaagg    4500
G  E  C  A  V  G  A  S  G  A  Q  N  G  D  S  E  L  G  S  C
gtgagtgcgccgtcggggcctcaggggcgcagaatggcgacagcgagctgggcagctgct    4560
C  S  E  A  V  K  S  A  M  S  T  I  D  L  D  S  L  M  A  E
gctccgaggcggtcaagagcgccatgagcaccattgacctggactcgctgatggcagagc    4620
H  S  A  A  W  Y  M  P  A  D  K  A  L  V  D  S  A  D  D  D
acagcgctgcctggtacatgcccgctgacaaggccctggtggacagcgcggacgacgaca    4680
K  T  L  A  P  W  E  K  A  K  P  Q  N  P  N  S  K  E  A  H
agacgttggcgccctgggagaaggccaaaccccagaaccccaacagcaaagaagcccacg    4740
D  L  P  A  N  K  A  S  A  S  Q  P  G  S  H  L  Q  C  L  S
acctccctgcaaacaaggcctcagcatcccagcctggcagccacttgcagtgcctgtctg    4800
                                                  exon1|exon2
V  H  C  T  D  D  V  G  D  A  K  A  R  A  S  V  P  T  W  R
tccactgcacagacgacgtgggtgacgccaaggctcgagcctccgtgcccacctggcggt    4860
S  L  H  S  D  I  S  N  R  F  G  T  F  V  A  A  L  T  *
ccctgcattctgacatctccaacagatttgggacattcgtggctgccctaacttgaatga    4920
caagaaagatcccctcctctaccaggcccttcctctccccctgtctgtttccttccccct    4980
caaccttaccccaccccctctgttaatttgaaagggccactattgctgagtggatgagttt    5040
ttttttttcctctaggttggtacctgcttagtggcatatggaccggaaagggttaatttt    5100
aaaggggggaacctcaaaagtttttttaaaaaagaaacttgtctgccacagtatgttac     5160
cagtgttaacccttctgcagttagcaaacttttgcttaagccttttcctctagatactc     5220
cccatgtttcggtaatcttggcatacatttttagatgacctctttccttgttttgtttt     5280
catgctgctgtatgtccaagtattgttatttcataataagacaagagttgctttctttttt    5340
tattctttttccttttcttaccccctcccctttttatttttcttttgctttgttcactgct    5400
tattaaaatggaaatcctggagaatagtagttctggaatattgccgggtgaaagtccaat     5460
tgtcatcacaatgttatatattgacaccccagtgtcatcagtcaggcaggagccaaacaa     5520
tgaatgcccctcttaggtattccgcctgggattttgttttgtctgttccctaagaaaata     5580
tattttcattcctgcaaacacagtgctcagccttcagttcccttccacttgagttctctc     5640
ttctcctgctggaagccgcccctctctgcgatggacgtgaggacgtgtccagctctgctc     5700
tgtgggaaggagttggaatgttcgacagcagtgttttctctccttttctgggcctcctcg     5760
caaatgcccaggccctgcattttcacgctgtgctaagcagcctttggtctgcatgggggа     5820
tggtgtgctcccagcctgcagtctttggagcaaggctgctgcccgtgccttgggtgctgg     5880
agttggaggaggctgttctcagccctttccctttctgaaagctgttcctggccgggcat      5940
cccagggaagaaggaggggactgcgtgtatctcctccacctctcccattccatccccagt     6000
ccagcctgggcaaccccaccctgggagggatgaggcaccctcttgctcagcctgctcag     6060
ccttctctgagcctttgcagggatctgcagactcctgagggctagaggacagagaaagag    6120
aatagaatgaaatgactttgattcctgcgccttttagttttgaactctggaattcctctg    6180
ccccctccccaacatttttttggaatctcaccctgttgcaaaactagagccatgtcccaa    6240
gcatctcacaaaggaataactgctctgagcagagatgagtggtggttggcaggggcaggc    6300
```

```
aactttgggtgctgctgatgcctgcaaaagccatttatggcttgtggtgggggcacata    6360
gattccccggtgggttagacaggaagtaactgatatcacttcacccaaatatataaccgt    6420
gatggttatctatttaatttcagttttgttaacgagcgtgtcttactaaaacgctccac    6480
tttgagctcccccaccccctccaggtcctcagagtttgcagatctgggctttctaaagca    6540
agtgacctgaaggctctgggctcaccatacaacacccacgttgtttatttcaaagaactt    6600
ttcagcgaagggagaggagctttcagaaaaacctcactctttcccctcccttctccctc    6660
tttccttctgccggtccttttggctggggtctgagtctgcggttctcgcctgggcagtct    6720
tgacgaggagcaaaccccgccttcagagggcagacaaagcaggtggcatgaattgatcag    6780
cgagaaaggtgtgagccgaggcagttcctgcgttctgctacaaaaggaatggaaagggaa    6840
gggaatttcccccccaccatgggctgtgggagagttgaccgtattctgggcaagactccat    6900
gacccctctgattctgcagtgtacagctgtttgagagcctcatcattttacttttgaaac    7060
aggaatgatttctccttaattgcttaaggccggggagcaaagtgtcttaacttctgtctt    7120
tgactttcccagcgttgagtcatcaacactttgccaattagctcatggtcctggcaacct    7180
cagaaacccctgaagttttaaaaactttctcgctccccacgaccccagaatgaaacagct    7240
ttaaaaatagccttaagcaaaaggatgttatttcattaaatttggtttaatggaaagaat    7300
aaaagtaaatgaaaaacacaccctacacactagactccgaacactggtaatcagtactgc    7360
atagcaaactctttgggaaagaaaacgaaaatgttattgcacatgtaaaatatgaaaact    7420
taactctgctgtgtgttaggcaatcctgtaatctttttgactcttaaaagaaattcatt    7480
tctgaaatgcttggttggaagactgtgacaatagctcatgaaattgagtgttatttttt    7540
ctttctttttaaaaaaatatgtaaagtgcagtcttctgtattcctgcatattgtatata    7600
cctgtatatgttttcctgagcagttaaataacaataaatatgacgttaatggtgaaaaaa    7660
```

Figure 19: MN-1 cDNA sequence and amino acid translation. Genbank accession number X82209.

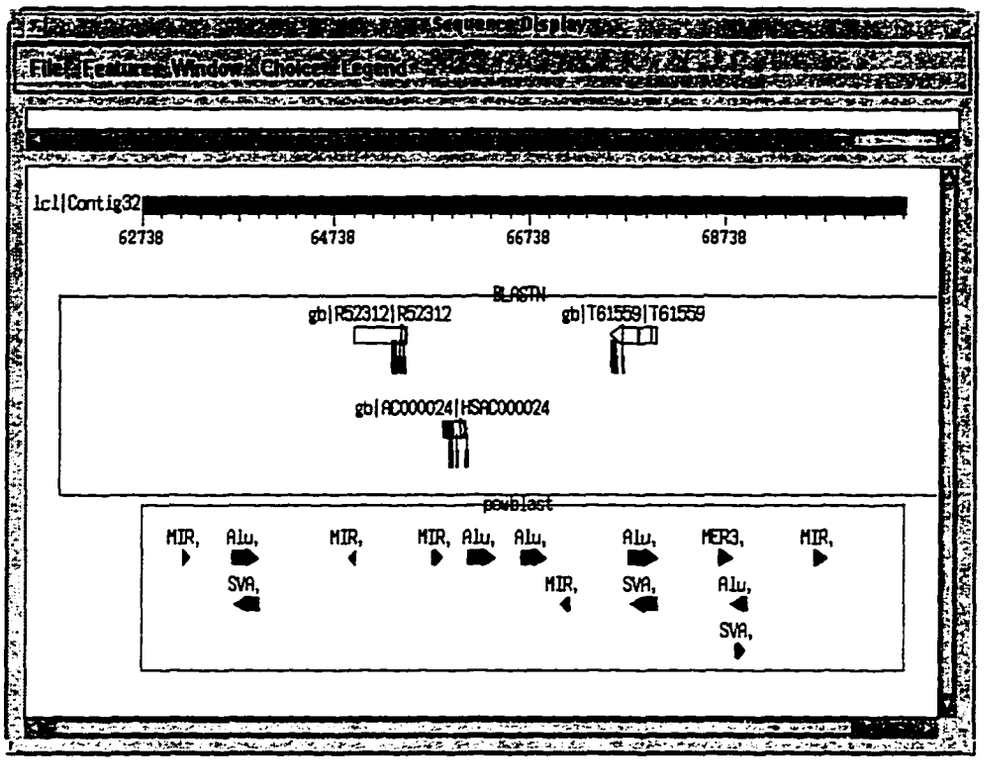


Figure 20: Graphical view of powblast output of the MN-1 gene region from human chromosome 22q11 showing an EST (T61559) containing an ALU family repeat element.

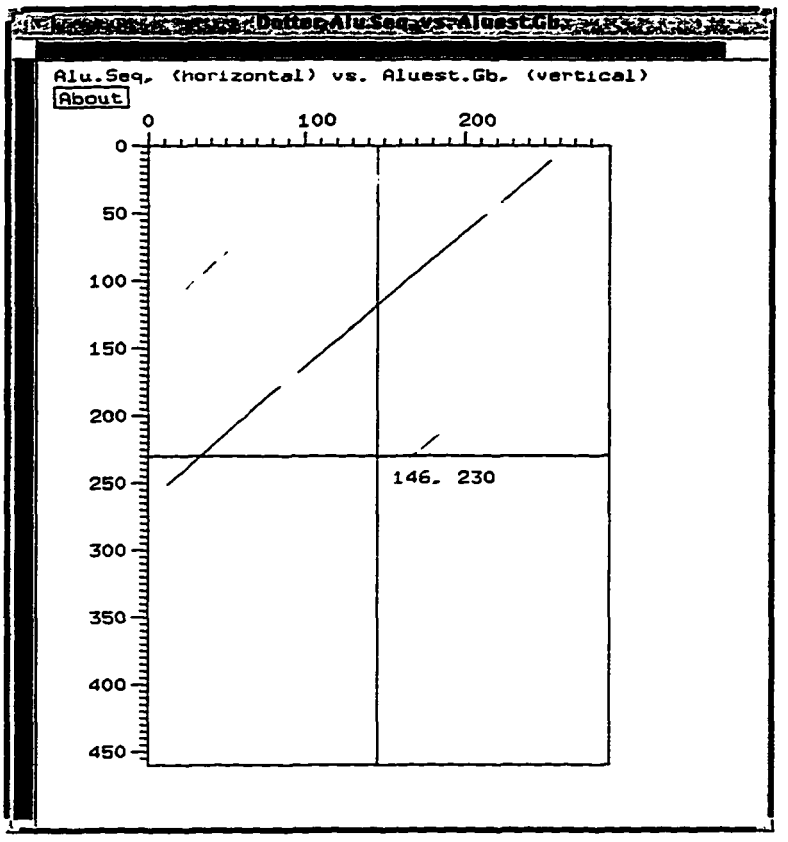


Figure 21: Dotter sequence comparison of the ALU repeat family consensus and EST T61559, demonstrating sequence homology and thus, the presence of an ALU repeat within an EST.

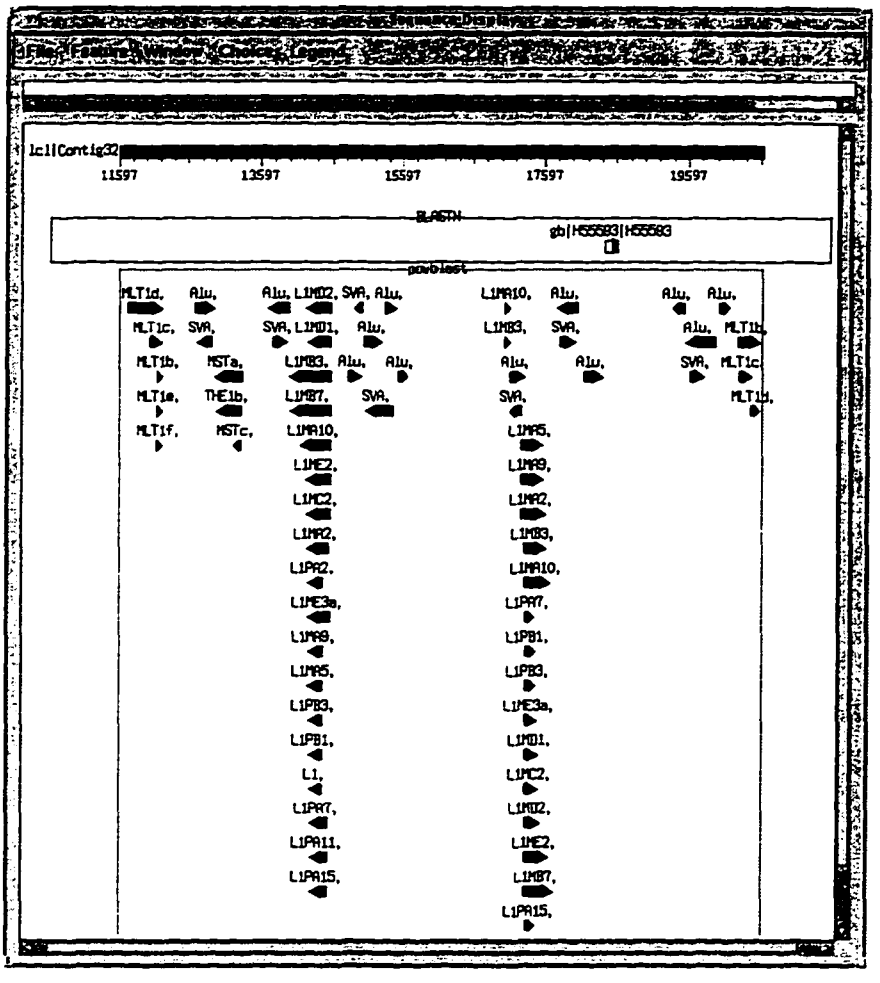


Figure 22: Powblast output of the repeat-rich region upstream of the MN-1 gene from human chromosome 22q11.

This region of chromosome 22 is also rich in simple sequence repeats according to Xgrail, and these are tabulated in Table 4. The simple sequence repeats are evenly dispersed throughout the region. There does, however, seem to be an overabundance of poly A and poly T stretches (13 of the 28 simple sequence repeat regions contain either poly A or poly T of greater than 10 bases).

B. Putative genes

Several additional exons (three excellent and four good in the forward direction; four excellent, five good, and three marginal in the reverse direction) were predicted by Xgrail. Powblast, however, did not report any database homology to these predicted exons and there are no other indications that these regions encode for expressed genes.

IV. Multiple Endocrine Neoplasia, type I Sequencing and Analysis

Approximately 450,000 bases from the MEN-1 gene region from human chromosome 11q13, BACs b137c7 and b18h3, was sequenced as part of this dissertation research. This work was done in collaboration with Dr. Francis Collins at the National Human Genome Research Institute at the National Institutes of Health in Bethesda, MD.

A. b137c7 analysis

Analysis of approximately 140,000 unique bases from b137c7 in the MEN-1 gene region from human chromosome 11q13 (Genbank accession number AC000134) revealed the presence of three genes whose cDNA sequences are known and four putative genes (one of which is suspected to be the gene whose disruption/deletion causes the MEN-1 phenotype). The previously known genes include human β-lymphocyte serine/threonine kinase (or GC kinase) gene, the human zinc finger motif-1 gene, and the muscle glycogen phosphorylase gene. One of the four putative genes in this region displays significant sequence homology with the rat neurexin II alpha gene, as will be discussed below.

Meningioma Deletion Region - Simple Sequence Repeats

position	Repeat
137-417	$C(T)_4$
5554-5634	$(CT)_{21}(CA)_{14}(GA)_{12}$
5609-5659	$(CA)_3(CT)_{24}(CA)_{14}(GA)_{12}$
8447-8464	T_{18}
9744-9786	$(GAA)_9$
10475-10501	T_{18}
10642-10679	T_{25}
14083-14124	$CA_6CA_5CA_{13}$
15288-15338	$(CAAA)_3(CA_6)_4$
15486-15529	$A_4T(A_3T)_2A_4T(A_2T)_2A_3T$
17083-17159	A_{19}
20141-20171	$GA_{10}GA_3GA_4GA_3$
20679-20993	poly GA
22787-22840	$(TG)_{17}(T_4A)_3$
24135-24158	A_{21}
31324-31365	$(TG)_{19}$
31777-31811	$G_4AG_4(TG)_{14}$
38660-38772	$(CAG)_{15}$
42812-42840	A_{14}
52555-52605	$(GT)_{19}$
54495-54525	A_{15}
59735-59764	$T_5CT_4C(T_3C)_2T_{11}$
60073-60109	$T_{15}(CT)_5T_{12}$
66398-66422	A_{25}
68072-68099	A_{28}
69831-69862	$(TC)_9T_8CT_5$
76229-76345	$(GT)_4(GA)_{15}$
89676-89704	A_{21}

Table 4: Simple sequence repeats present in the meningioma deletion region from human chromosome 22q11 as reported by Xgrail.

1. Human β-lymphocyte serine/threonine (GCkinase) gene

The human β-lymphocyte serine/threonine (GC kinase) gene encompasses a total of 14 kb in the genomic sequence as determined by sequence scanning and is present in the 20,448 bp contig comprising the human β-lymphocyte serine/threonine (GC kinase) gene region. Comparison of the genomic sequence with the known cDNA sequence (Genbank accession number U07349; see figure 23) revealed that the gene is composed of 32 exons, although only 31 of the 32 were identified in the genomic sequence (exon 27 is presumably in a sequence gap between two shotgun contigs), with the β-lymphocyte gene using exon 5b. (Exon 5a is included because of homology with the mouse Rab8-interacting protein (discussed below) although it is not present in the reported human β-lymphocyte gene sequence.) The first exon encodes 40 bases of 5' untranslated region, including an SP1 binding site (Tejan GC box), followed by a methionine start codon. The coding region concludes with a TAA stop codon at position 17530-17532 and a polyadenylation signal at 18865-18870, leaving 407 bases in the 3' untranslated region following the stop codon. The 32 exons present in this gene encode a protein of 820 amino acids.

Output from Xgrail and Powblast for this region is displayed in figure 24. Xgrail accurately predicted 24 of the 33 exons (counting exon 05a), but failed to predict exons # 5b, 7, 13-14, 16, 19, 22, and 25 due to their small size (exon 27 also was not predicted because it occurs in a sequencing gap). It is a well documented shortcoming of Xgrail that exons smaller than 100 bp often are not accurately predicted. However, all the predicted exons correlated well with the powblast-revealed database homologies, as seen in figure 24.

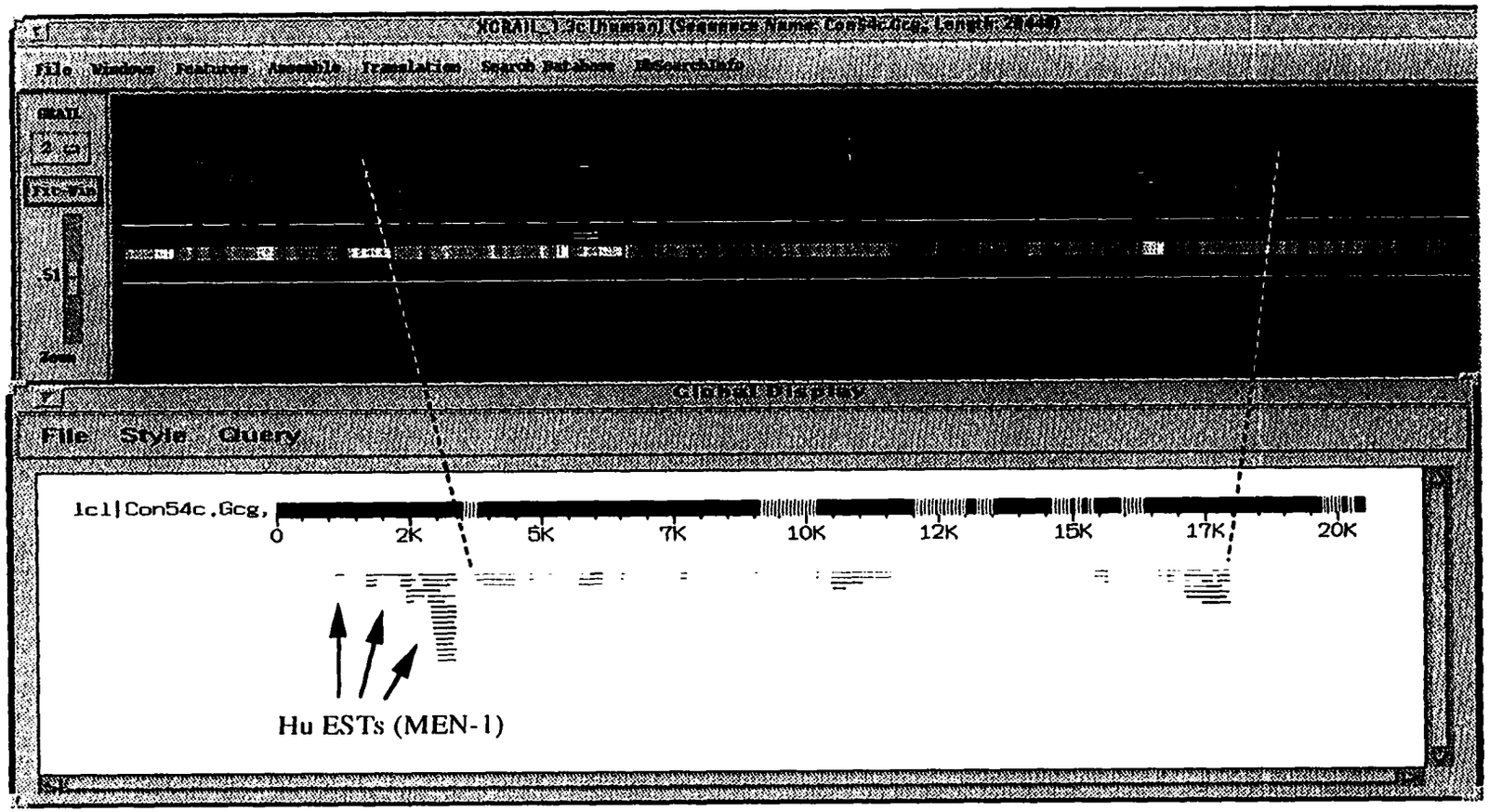
The genomic organization of the β-lymphocyte serine/threonine kinase gene is presented in Table 5. As mentioned above, all exons were identified in the genomic sequence by powblast with the exception of exon 27, which falls in a sequencing gap. An additional sequencing gap is present between exons 11 and 12 but this gap falls in the

```
                                                  M  E  L  R  D  V  S
gctccggcccgccccgctgcccggcccgcgcgccgggccatggagctgcgggatgtgtcg        60
 L  Q  D  P  R  D  R  F  E  L  L  Q  R  V  G  A  G  T  Y  G
ctgcaggacccgcgggaccgcttcgagctgctgcagcgcgtgggggccgggacctatggc       120
    exon1|exon2
 D  V  Y  K  A  R  D  T  V  T  S  E  L  A  A  V  K  I  V  K
gacgtctacaaggcccgcgacacggtcacgtccgaactggccgccgtgaagatagtcaag       180
     exon2|exon3
 L  D  P  G  D  D  I  S  S  L  Q  Q  E  I  T  I  L  R  E  C
ctagacccaggggacgacatcagctccctccagcaggaaatcaccatcctgcgtgagtgc       240
                                    exon3|exon4
 R  H  P  N  V  V  A  Y  I  G  S  Y  L  R  N  D  R  L  W  I
cgccaccccaatgtggtggcctacattggcagctacctcaggaatgaccgcttgtggatc       300
                                          exon4|5a
 C  M  E  F  C  G  G  G  S  L  Q  E  I  Y  H  A  T  G  P  L
tgcatggagttctgcggaggggctccctgcaggagatttaccatgccactgggcccctg       360
                                       exon5a|exon5b
 E  E  R  Q  I  A  Y  V  C  R  E  R  L  K  G  L  H  H  L  H
gaggagcggcagattgcctacgtctgccgagagcgactgaaggggctccaccacctgcat       420
                          exon5b|exon6
 S  Q  G  K  I  H  R  D  I  K  G  A  N  L  L  L  T  L  Q  G
tctcaggggaagatccacagagacatcaagggagccaaccttctcctcactctccaggga       480
     exon6|exon7
 D  V  K  L  A  D  F  G  V  S  G  E  L  T  A  S  V  A  K  R
gatgtcaaactggctgactttggggtgtcaggcgagctgacagcgtctgtggccaagagg       540
                       exon7|exon8
 R  S  F  I  G  T  P  Y  W  M  A  P  E  V  A  A  V  E  R  K
aggtctttcattgggactccctactggatggctcccgaggtggctgctgtggagcgcaaa       600
 G  G  Y  N  E  L  C  D  V  W  A  L  G  I  T  A  I  E  L  G
ggtggctacaatgagctatgtgacgtctgggccctgggcatcactgccattgagctgggc       660
                                    exon8|exon9
 E  L  Q  P  P  L  F  H  L  H  P  M  R  A  L  M  L  M  S  K
gagctgcagccccctctgttccacctgcaccccatgagggccctgatgctcatgtcgaag       720
                                          exon9|exon10
 S  S  F  Q  P  P  K  L  R  D  K  T  R  W  T  Q  N  F  H  H
agcagcttccagccgcccaaactgagagataagactcgctggacccagaatttccaccac       780
 F  L  K  L  A  L  T  K  N  P  K  K  R  P  T  A  E  K  L  L
tttctcaaactggccctgaccaagaatcctaagaagaggccgacagcagagaagctcctg       840
exon10|exon11
 Q  H  P  F  T  T  Q  Q  L  P  R  A  L  L  T  Q  L  L  D  K
cagcacccgttcacgactcagcagctccctcgggccctcctcacacagctgctggacaaa       900
                                             exon11|exon12
 A  S  D  P  H  L  G  T  P  S  P  E  D  C  E  L  E  T  Y  D
gccagtgaccctcatctggggacccccctcccctgaggactgtgagctggagacctatgac       960
 M  F  P  D  T  I  H  S  R  G  Q  H  G  P  A  E  R  T  P  S
atgtttccagacaccattcactcccgggggcagcacggcccagccgagaggacccccctcg      1020
    exon12|exon13
 E  I  Q  F  H  Q  V  K  F  G  A  P  R  R  K  E  T  D  P  L
gagatccagtttcaccaggtgaaatttggcgccccacgcaggaaggaaactgacccactg      1080
    exon13|exon14                                exon14|exon15
 N  E  P  W  E  E  E  W  T  L  L  G  K  E  E  L  S  G  S  L
aatgagccgtgggaggaagagtggacactactgggaaaggaagagttgagtgggagcctg      1140
                          exon15|exon16
 L  Q  S  V  Q  E  A  L  E  E  R  S  L  T  I  R  S  A  S  E
ctgcagtcggtccaggaggccctggaggaaaggagtctgactattcggtcagcctcagaa      1200
exon16|exon17                                          exon17|exon18
 F  Q  E  L  D  S  P  D  D  T  M  G  T  I  K  R  A  P  F  L
ttccaggagctggactccccagacgataccatgggaaccatcaagcgggccccgttccta      1260
```

```
  G  P  L  P  T  D  P  P  A  E  E  P  L  S  S  P  P  G  T  L
gggccactccccactgaccctccagcagaggagcctctgtccagtcccccaggaaccctg    1320
  P  P  P  P  S  G  P  N  S  S  P  L  L  P  T  A  W  A  T  M
cccccacctccttcaggccccaacagctccccactgctgcccacggcctgggccaccatg    1380
                exon18|exon19
  K  Q  R  E  D  P  E  R  S  S  C  H  G  L  P  P  T  P  K  V
aagcagcgggaggatcctgagaggtcatcctgccacgggctccccccaactcccaaggtg    1440
exon19|exon20
  H  M  G  A  C  F  S  K  V  F  N  G  C  P  L  R  I  H  A  A
catatgggcgcctgcttctccaaggtcttcaatggctgccccctgcggatccacgctgct    1500
                     exon20|exon21
  V  T  W  I  H  P  V  T  R  D  Q  F  L  V  V  G  A  E  E  G
gtcacctggattcaccctgttactcgggaccagttcctggtggtaggggccgaggaaggc    1560
                                          exon21|exon22
  I  Y  T  L  N  L  H  E  L  H  E  D  T  L  E  K  L  I  S  H
atctacacactcaacctgcatgaactgcatgaggatacgctggagaagctgatttcacat    1620
                                          exon22|exon23
  R  C  S  W  L  Y  C  V  N  N  V  L  L  S  L  S  G  K  S  T
cgctgctcctggctctactgcgtgaacaacgtgctgctgtcactctcagggaaatccacg    1680
                                       exon23|exon24
  H  I  W  A  H  D  L  P  G  L  F  E  Q  R  R  L  Q  Q  Q  V
cacatctgggcccatgacctcccaggcctgtttgagcagcggaggctacagcaacaggtt    1740
  P  L  S  I  P  T  N  R  L  T  Q  R  I  I  P  R  R  F  A  L
cccctctccatccccaccaaccgcctcacccagcgcatcatccccaggcgctttgctctg    1800
                                  exon24|exon25
  S  T  K  I  P  D  T  K  G  C  L  Q  C  R  V  V  R  N  P  Y
tccaccaagattcctgacaccaaaggctgcttgcagtgtcgtgtggtgcggaacccctac    1860
    exon25|exon26
  T  G  A  T  F  L  L  A  A  L  P  T  S  L  L  L  L  Q  W  Y
acgggtgccaccttcctgctggccgccctgcccaccagcctgctcctgctgcagtggtat    1920
exon26|exon27
  E  P  L  Q  K  F  L  L  L  K  N  F  S  S  P  L  P  S  P  A
gagccgctgcagaagtttctgctgctgaagaacttctccagccctctgcccagcccagct    1980
  G  M  L  E  P  L  V  L  D  G  K  E  L  P  Q  V  C  V  G  A
gggatgctggagccgctggtgctggatgggaaggagctgccgcaggtgtgtgttggggcc    2040
  E  G  P  E  G  P  G  C  R  V  L  F  H  V  L  P  L  E  A  G
gaggggcctgaggggcccggctgccgcgtcctgttccatgtcctgcccctggaggctggc    2100
                     exon27|exon28
  L  T  P  D  I  L  I  P  P  E  G  I  P  G  S  A  Q  Q  V  I
ctgacgcccgacatcctcatcccacctgaggggatcccaggctcggcccagcaggtgatc    2160
                                    exon28|exon29
  Q  V  D  R  D  T  I  L  V  S  F  E  R  C  V  R  I  V  N  M
caggtggacagggacacaatcctagtcagctttgaacgctgtgtgaggattgtcaacatg    2220
  Q  G  E  P  T  A  T  L  A  P  E  L  T  F  D  F  P  I  E  T
cagggcgagcccacggccacactggcacctgagctgacctttgatttccccatcgagact    2280
exon29|exon30
  V  V  C  L  Q  D  S  V  L  A  F  W  S  H  G  M  Q  G  R  S
gtggtgtgcctgcaggacagtgtgctggccttctggagccatgggatgcaaggccgaagc    2340
       exon30|exon31
  L  D  T  N  E  V  T  Q  E  I  T  D  E  T  R  I  F  R  V  L
ctggataccaatgaggtgacccaggagatcacagatgaaacaaggatcttccgagtgctt    2400
     exon31|exon32
  G  A  H  R  D  I  I  L  E  S  I  P  T  D  N  P  E  A  H  S
ggggcccacagagacatcatcctggagagcattcccactgacaacccagaggcgcacagc    2460
  N  L  Y  I  L  T  G  H  Q  S  T  Y  *
aacctctacatcctcacgggccaccagagcacctactaagagcagcgggcctgtccaggc    2520
tccccgccccaccccacgccttagctgcaggccctttgggcaaaggggcccatcctaga    2580
ccagaggagcccaggccctggccctgctggggctgaaggtcagaagtaatcctgagaaat    2640
```

```
gtttcaggcctggggagggaggggagcccccgacgcctctgcaataactggaccaggggg    2700
agctgctgtcactcccccatccccgaggcagcccagtccctagtgcccaaggcagggacc    2760
ctgggcctgggccatccattccattttgttccacatttccttttctactctttctgccaag    2820
agcctgcccctgcatttgtcctgggaaacacggtatttaagagagaactatattggtatt    2880
aaagctggtttgttttaaaaaaaaaa                                      2906
```

Figure 23: Human β lymphocyte serine/threonine (GC kinase) cDNA and amino acid translation. Genbank accession number U07349.



Strong database hits for Rab8 and β lymphocyte (GC kinase)

Gene organization for Rab8 and β lymphocyte (GC kinase)

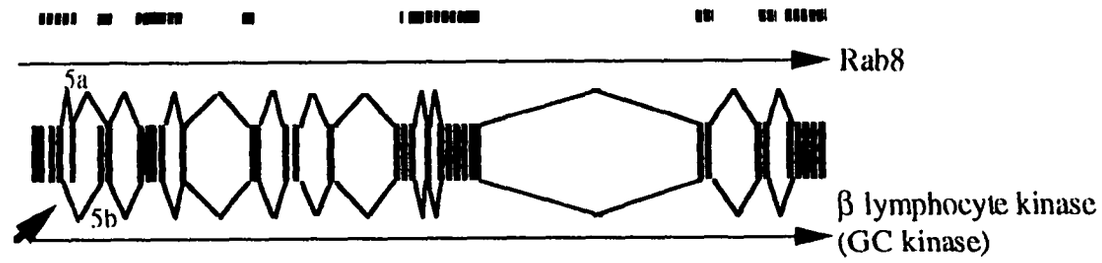


Figure 24: Top panel: Graphical result of Xgrail gene identification software for the mouse Rab8 and β lymphocyte serine/threonine kinase (GC kinase) gene from human chromosome 11q13. The peaks present are Xgrail predicted exons with a quality rating of excellent (green), good (blue), or marginal (red). The light blue lollipops are predicted polyadenylation sites and the purple squares are CpG islands. Yellow and red squares indicate promotors and vertical orange bars denote simple sequence repeats. Bottom panel: Graphical display of the powblast output using Musk. The black bar at the top of the panel represents the consensus sequence with the grey areas indicating repeat sequences. The lines under the black bar indicate homology between the consensus sequence and an entry in the public database.

Genome Organization-Rab8 and β lymphocyte serine/threonine kinase (GC kinase)

Exon number	position	size of exon	size of intron
01	3730-3825	95	112
02	3937-3994	57	232
03	4226-4316	90	80
04	4396-4460	64	298
05a	4758-4813	55	309
05b	5122-5169	47	89
06	5258-5300	42	396
07	5696-5768	72	80
08	5848-5979	131	74
09	6053-6115	62	372
10	6487-6568	81	1057(gap)
11	7625-7732	107	426
12	8158-8236	78	128
13	8364-8420	56	585
14	9005-9048	43	1131
15	10179-10216	37	93
16	10309-10342	33	121
17	10463-10509	46	144
18	10653-10740	87	83
19	10823-10861	38	89
20	10950-11034	84	96
21	11130-11209	79	81
22	11290-11350	60	103
23	11453-11570	117	3853
24	15423-15470	47	59
25	15529-15568	40	28
26	15596-15648	52	958(gap)
27	(absent - in sequence gap)		
28	16606-16674	68	120
29	16794-16882	88	228
30	17110-17180	70	104
31	17284-17339	55	106
32	17445-17532	87	
Tejan GC box	3666-3674		
ATGstart	3730-3732		
TAAstop	17530-17532		
polyA signal	18865-18870		

Table 5: Genome organization of the mouse Rab8 and β lymphocyte serine/threonine kinase (GC kinase) genes from human chromosome 11q13. These genes all contain 32 exons (Rab8 uses exon 5a, whereas β lymphocyte serine/threonine kinase (GC kinase) uses exon 5b) for a total of 2.5 kb, which span roughly 14 kb of genomic sequence. The promotor region has a strong Tejan GC box approximately 50 nucleotides upstream from the ATG start codon at the beginning of exon 1. The polyadenylation signal is 1333 nucleotides downstream of the TAA stop codon at the end of exon 32.

intronic region. Each of the intron/exon boundaries demonstrate good adherence to the consensus splice junctions, as seen in Table 6.

This region of b137c7 also has multiple database hits to human ESTs originating from many different tissue types including pancreas, infant brain, pregnant uterus, adult brain, melanocytes, fetal liver, fetal spleen, fetal lung and placenta. These ESTs also have homology with the human β lymphocyte serine/threonine (GC kinase) gene (Genbank accession numbers H38333, H14310, H14313, R32679, H91638, N42475, N24156, AA147612, N36190, W38527, H97384, AA147620, H41351, H91639, N51774, H38296, H14281, R98414, H14277, R28488, N90077, D56388, Z15229, R35229, R35283, AA213898, D31511, T28384, M62026, R50953, N75361, T88776, W67355, W01805, D20515, AA213819) and occur all along the gene region in conjuction with exon positions as predicted by Xgrail, and confirmed by the cDNA sequence comparison with genomic.

This region also had database homology with the mouse Rab8-interacting protein (Genbank accession number U50595; see figure 24) and the exons are identical with the exception of exon 5. The Rab8 gene uses exon 5a whereas the β-lymphocyte serine/threonine kinase gene uses exon 5b. The gene region for the Rab8 and β-lymphocyte serine/threonine kinase also hits several mouse and rat ESTs at the begining and at the end of this gene (Genbank accession numbers AA168218, AA049922, H34387, W76863, W14310, W67045, and AA204362), as well as multiple repeated elements. This gene region contains one TAR1, eight ALU, six SVA, two L1 and three MIR repeats, with many of these occurring as full length repeats. An interesting feature of repeat sequences is that SVA sequences always occur in the same position as ALUs, but in the opposite orientation.

Simple sequence repeats (see table 7) are not very prevalent in this area of the genome, although the few that are, tend to be large. There are two cases of extensive poly C regions, each over 100 bp in length, positioned at 6630-6772 and

Splice Junctions-Rab8 and β lymphocyte serine/threonine kinase (GC kinase)

Left Exon	Left splice junction	Right splice junction	Right Exon
01	CGTCTACAAGgtgcgacggg	cgccctgcagGCCCGCGACA	02
02	CTAGACCCAGgtgagggccc	tgcttcccagGGGACGACAT	03
03	GCTACCTCAGgtgaggctgc	ttgtccctagGAATGACCGC	04
04	ATTTACCATGgtgaggccag	ttcctttcagCCACTGGGCC	05a
05a	GGCACTGAAGgtagctgggc	ttgtcgccagGGGCTCCACC	05b
05b	AGACATCAAGgtaggcacct	attcccacagGGAGCCAACC	06
06	GTCAAACTGGgtcagtaccc	ctccccacagCTGACTTTGG	07
07	CTCCCTACTGgtgaggctgc	ccacctctagGATGGCTCCC	08
08	ACCCCATGAGgtcagcccaa	ccctccacagGGCCCTGATG	09
09	AGACTCGCTGgtaagggcca	tgccttctagGACCCAGAAT	10
10	GCTCCTGCAGgtgggaggca	cactttccagCACCCGTTCA	11
11	TGAGCTGGAGgtaggtgagg	ttttccttagACCTATGACA	12
12	GAGATCCAGTgtgagtctgc	ggccccccagTTCACCAGGT	13
13	GAATGAGCCGgtgagtgggc	tttcccccagTGGGAGGAAG	14
14	AGTTGAGTGGgtaagtgagg	cactgcccagGAGCCTGCTG	15
15	TGGAGGAAAGgtgagggatg	cctcccacagGAGTCTGACT	16
16	AAAATTCCAGgtgccacaca	ctcctaccagGAGCTGGACT	17
17	AGCGGGCCCCgttcctaggg	gtgcctgcagGAACCCTGCC	18
18	GGATCCTGAGgtaagagggt	ctcccagaggTCATCCTGCC	19
19	CAAGGTGCATgtaagtgttt	tcctctgcagATGGGCGCCT	20
20	GTTACTCGGGgtaggtgggg	aaactggcagACCAGTTCCT	21
21	TCACTCTCAGgtacaaggtg	gtttctgcagGGAAATCCAC	22
22	TCACTCTCAGgtacaaggtg	gtttctgcagGGAAATCCAC	23
23	TCATCCCCAGgtcagggatc	acccctacagGCGCTTTGCT	24
24	GCTGCTTGCAgtgtcgtgtg	tgggggcagCGCAGGCCCA	25
25	TCCCCTGCCAgtgcggaacc	gggtgccaccTTCCTGCTGG	26
26	GGTATGAGCCgggtaccgag	--	27
27	--	ctgcctgcagAGGGGATCCC	28
28	AGCTTTGAACgtgagtgccc	ccctgtgcagGCTGTGTGAG	29
29	GAGACTGTGGgtgagtgagc	ccccccatagTGTGCCTGCA	30
30	TACCAATGAGgtagtggggc	gtcttcccagGTGACCCAGG	31
31	GGGCCCACAGgtaggcggtg	tctcctccagAGACATCATC	32
32	CACCTACTAAgagcagcggg		
consensus:	CAGgtaagt A t	tncagG c	

Table 6: Splice junctions for the exons of the mouse Rab8 and the human β lymphocyte serine/threonine kinase (GC kinase) genes from human chromosome 22q13.

Rab8 and β lymphocyte serine/threonine kinase (GC kinase) - Simple Sequence Repeats

position	Repeat
705-811	poly GC region
3574-3765	poly GC region
3621-3734	poly GC region
6630-6772	C_{125}
8736-8841	poly CT region
9825-9840	A_{15}
10119-10144	A_{25}
12742-12777	A_{19}
15660-15790	C_{104}
15791-15807	A_{14}
15908-15934	$(CT_2)_2T_7G_2T_3GT_3G_2T_2$
20024-20056	$A_{15}(GA)_2A_6(TA)_5$
20159-20170	T_{12}

Table 7: Simple sequence repeats present in the mouse Rab8 and human β lymphocyte serine/threonine kinase (GC kinase) gene from human chromosome 22q11 as reported by Xgrail.

15660-15790 in this 20,448 bp β-lymphocyte serine/threonine (GC kinase) gene region. There also is a poly GC region at 3574-3765, which is highlighted by Xgrail as a CpG island and may represent a possible promoter element.

2. ZFM1 and isoforms

The gene region for zinc finger motif-1, ZFM1 isoform B3, and ZFM isoform B4 encompasses 14 kb of genomic sequence, determined by sequence scanning. Comparison of the genomic sequence with the known cDNA sequence (Genbank accession numbers D26120, L49345, and L49380, respectively; the ZFM1 cDNA and translation is shown in figure 25) revealed that the genes all share exons 2-9, but the presence of exons 1 and 10-14 is variable in different isoforms depending on the splicing pattern. The ZFM1 gene contains all 14 exons and encodes for a protein of 624 amino acids. The 5' untranslated region is 382 nucleotides and the 3' untranslated region is 848 nucleotides long, flanked by the polyadenylation signal at the end of the 3' untranslated region. The genomic organization of the ZFM1 gene region is detailed in Table 8. Each of the intron/exon boundaries (for both the conserved and the alternate splicing exons) conform to the consensus splice juctions as seen in Table 9.

The output from Xgrail and powblast for this 26,040 bp region is displayed in figure 26 along with the splicing patterns of the different isoforms of the zinc finger DNA binding motif. Xgrail perfectly predicted the highly conserved exons 3-9, but failed to predict the exons involved in alternate splicing: 1, 2, 10 and 14. Xgrail did predict exons 12 and 13, however, the quality rating is good or marginal instead of excellent as seen with exons 3-9. Although exons 1 and 2 were not predicted by Xgrail, powblast reveals convincing homology to the cDNAs and, in the case of exon 2, also to ESTs. Exon 2 has homology with two mouse ESTs (Genbank accession numbers AA170957 and W62122), although no human EST homology is present for this exon. Other exons in the ZFM gene region have numerous human

```
cgttgctgtcgaaatgaagtgcgcgctgcgacacctcccagcccaccgaactccgccgcc      60
atttcctcgcttggcctaacggttcggccaatcccagcgcgcatcaagaaggactgaggc     120
tccgccaatcggaggccgccgatttcgacccttcgcctcggcccggcccaatccaggccc     180
cggccccgccgcccccggcccgcccccgcggtgccctctctcctccctctttgtgcgtct     240
cgcgccgccgccgcccgccgcgtgagaggacgggctccgcgcgctccggcagcgcattcg     300
ggtcccctcccccgggaggcttgcgaaggagaagccgccgcagaggaaaagcaggtgcc     360
                                                  exon1|exon2
                          M  A  T  G  A  N  A  T  P  L  D  F  P
ggtgcctgtccccgggggcgccatggcgaccggagcgaacgccacgccgttggacttccc     420
   S  K  K  R  K  R  S  R  W  N  Q  D  T  M  E  Q  K  T  V  I
aagtaagaagcggaagaggagccgctggaaccaagacacaatggaacagaagacagtgat     480
   P  G  M  P  T  V  I  P  P  G  L  T  R  E  Q  E  R  A  Y  I
tccaggaatgcctacagttattccccctggacttactcgagaacaagaaagagcttatat     540
exon2|exon3
   V  Q  L  Q  I  E  D  L  T  R  K  L  R  T  G  D  L  G  I  P
agtgcaactgcagatagaagacctgactcgtaaactgcgcacaggggacctgggcatccc     600
                exon3|exon4
   P  N  P  E  D  R  S  P  S  P  E  P  I  Y  N  S  E  G  K  R
ccctaaccctgaggacaggtccccttcccctgagcccatctacaatagcgaggggaagcg     660
   L  N  T  R  E  F  R  T  R  K  K  L  E  E  E  R  H  N  L  I
gcttaacacccgagagttccgcacccgcaaaaagctggaagaggagcggcacaacctcat     720
                                                  exon4|exon5
   T  E  M  V  A  L  N  P  D  F  K  P  P  A  D  Y  K  P  P  A
cacagagatggttgcactcaatccggatttcaagccacctgcagattacaaacctccagc     780
   T  R  V  S  D  K  V  M  I  P  Q  D  E  Y  P  E  I  N  F  V
aacacgtgtgagtgataaagtcatgattccacaagatgagtacccagaaatcaactttgt     840
                   exon5|exon6
   G  L  L  I  G  P  R  G  N  T  L  K  N  I  E  K  E  C  N  A
ggggctgctcatcgggcccagagggaacaccctgaagaacatagagaaggagtgcaatgc     900
   K  I  M  I  R  G  K  G  S  V  K  E  G  K  V  G  R  K  D  G
caagattatgatccgggggaaagggtctgtgaaagaagggaaggttgggcgcaaagatgg     960
   Q  M  L  P  G  E  D  E  P  L  H  A  L  V  T  A  N  T  M  E
ccagatgttgccaggagaagatgagccacttcatgccctggttactgccaatacaatgga    1020
                         exon6|exon7
   N  V  K  K  A  V  E  Q  I  R  N  I  L  K  Q  G  I  E  T  P
gaacgtcaaaaaggcagtggaacagataagaaacatcctgaagcagggtatcgagactcc    1080
   E  D  Q  N  D  L  R  K  M  Q  L  R  E  L  A  R  L  N  G  T
agaggaccagaatgatctacggaagatgcagcttcgggagttggctcgcttaaatgggac    1140
                      exon7|exon8
   L  R  E  D  D  N  R  I  L  R  P  W  Q  S  S  G  T  R  S  I
ccttcgggaagacgataacaggatcttaagaccctggcagagctcagggacccgcagcat    1200
   T  N  T  T  V  C  T  K  C  G  G  A  G  H  I  A  S  D  C  K
taccaacaccacagtgtgtaccaagtgtggaggggctggccacattgcttcagactgtaa    1260
      exon8|exon9
   F  Q  R  P  G  D  P  Q  S  A  Q  D  K  A  R  M  D  K  E  Y
attccaaaggcctggtgatcctcagtcagctcaggataaagcacggatggataaagaata    1320
   L  S  L  M  A  E  L  G  E  A  P  V  P  A  S  V  G  S  T  S
tttgtccctcatggctgaactgggtgaagcacctgtcccagcatctgtgggctccacctc    1380
   G  P  A  T  T  P  L  A  S  A  P  R  P  A  A  P  A  N  N  P
tgggcctgccaccacacccctggccagcgcacctcgtcctgctgctcccgccaacaaccc    1440
         exon9|exon10
   P  P  P  S  L  M  S  T  T  Q  S  R  P  P  W  M  N  S  G  P
acctccaccgtctctcatgtctaccacccagagccgcccaccctggatgaattctggtcc    1500
   S  E  S  W  P  Y  H  G  M  H  G  G  G  P  G  G  P  G  G  G
ttcagagagttggccctaccacggcatgcatggaggtggtcctggtgggcccggaggtgg    1560
   P  H  S  F  P  H  P  L  P  S  L  T  G  G  H  G  G  H  P  M
cccccacagcttcccacacccattacccagcctgacaggtgggcatggtggacatcccat    1620
   Q  H  N  P  N  G  P  P  P  P  W  M  Q  P  P  P  P  P  M  N
```

```
gcagcacaaccccaatggacccccacccccttggatgcagccaccaccaccaccgatgaa    1680
                                          exon10|exon11
  Q  G  P  H  P  P  G  H  H  G  P  P  P  M  D  Q  Y  L  G  S
ccagggcccccaccctcctgggcaccatggccctcctccaatggatcagtacctgggaag    1740
                                          exon11|exon12
  T  P  V  G  S  G  V  Y  R  L  H  Q  G  K  G  M  M  P  P  P
tacgcctgtgggctctggggtctatcgcctgcatcaaggaaaaggtatgatgccgccacc    1800
  P  M  G  M  M  P  P  P  P  P  P  P  S  G  Q  P  P  P  P  P
acctatgggcatgatgccgccgccgccgccgcctcccagtgggcagcccccacccccctcc    1860
  S  G  P  L  P  P  W  Q  Q  Q  Q  Q  Q  P  P  P  P  P  P  P
ctctggtcctcttcccccatggcaacaacagcagcagcagcctccgccacccccctccgcc    1920
                                          exon12|exon13
  S  S  S  M  A  S  S  T  P  L  P  W  Q  Q  N  T  T  T  T  T
cagcagcagtatggcttccagtacccccttgccatggcagcaaaatacgacgactaccac    1980
  T  S  A  G  T  G  S  I  P  P  W  Q  Q  Q  Q  A  A  A  A  A
cacgagcgctggcacagggtccatcccgccatggcaacagcagcaggcggctgccgcagc    2040
  S  P  G  A  P  Q  M  Q  G  N  P  T  M  V  P  L  P  P  G  V
ttctccaggagcccctcagatgcaaggcaaccccactatggtgcccctgccccccggggt    2100
                                       exon13|exon14
  Q  P  P  L  P  P  G  A  P  P  P  P  P  R  S  I  E  C  L  L
ccagccgcctctgccgcctggggcccctccccctccgccccgtagcatcgagtgtcttct    2160
  C  L  L  S  L  L  T  Q  L  P  L  P  L  P  K  P  G  R  Q  D
ttgtcttctttctctcctcacccaactccctttgcctctccccaaaccgggccgccagga    2220
  P  S  P  R  R  R  W  P  E  P  *
tccctccccgcggcggcgatggcccgagccatgagagtgaggactttccgcgcccattgg    2280
tgacccttccaggcagacagcctcagcaacgcccctggtggacaggatggttcggcaaag    2340
cagcctgagttatttttgtggacggaatcggaacacgctggctccatatcgtgaaatttt    2400
tattaatttttttctttttccttttgttacttctttatcttttccttttcttcagactccgt    2460
ccaaggagatgctctccccggtcttctgctgcaatttagattcctttcccttctctccag    2520
ttctccttcccttaccaaggagaggggagcaaatggtttttgggcaagggctttggccatt    2580
catgtcaagctggttgtgggtttttcaaggtgccatagccacccccaaatatgtttgttt    2640
aaagcgtggggtttttttaatctctgccacccttgtcaagggagtcttgtaaagttgccga    2700
gggtaggttcatctccaggtttcgggattcccatccgtcctggcgatcctgccagcagtg    2760
ggtgggcagcctgagctccctcgggctcgcctgccagcctggagttcttcctgtgctcct    2820
tgatcacctgagctgcctcagattccatttggtcctctccttcctggaaggcttccttttt    2880
atgttttgttttaatcccaaatgtctgaatgttttgcagtgtgtaggggtttgagcccct    2940
tgttcattctccttccttttttcctcccgcttccctctccatgaagtgattctgttgacaa    3000
taatgtatactgcgcgttctcttcactggtttatctgcagaaatttctctgggctttttt    3060
cggtgttagattcaacactgcgctaaagcggggatgttccattgaataaaagagcagtgt    3120
```

Figure 25: ZFM1 cDNA and amino acid translation. Genbank accession number D26120.

Genome Organization-ZFM1, ZFM1 B3 isoform, ZFM1 B4 isoform and ZFM1 alternate splice products

Exon number	gene product(s)	position	size of exon	size of intron
01a	(b3,b4)	8533-8951	418	1736
01b	(zfm,alsp,as)	8550-8951	401	1736
02a	(zfm,as,b3,b4)	10687-10815	128	2994
02b	(alsp)	10687-11266	579	---
03	(zfm,as,b3,b4)	13809-13884	75	3022
04	(zfm,as,b3,b4)	16906-17058	152	203
05	(zfm,as,b3,b4)	17261-17350	89	355
06	(zfm,as,b3,b4)	17705-17888	183	88
07	(zfm,as,b3,b4)	17976-18091	115	94
08	(zfm,as,b3,b4)	18185-18292	107	766
09	(zfm,as,b3,b4)	19028-19208	150	262
10a	(zfm,as)	19470-19743	273	320
10b	(b3,b4)	19470-19747	277	316
11	(zfm,as,b3,b4)	20063-20122	59	113
12a	(b3)	20235-20418	183	1286
12b	(zfm,as,b4)	20235-20414	179	1290
13a	(b4)	21704-22041	337	---
13b	(zfm)	21704-21878	174	390
14a	(zfm)	22268-23253	985	---
14b	(as,b3)	22341-23253	912	---

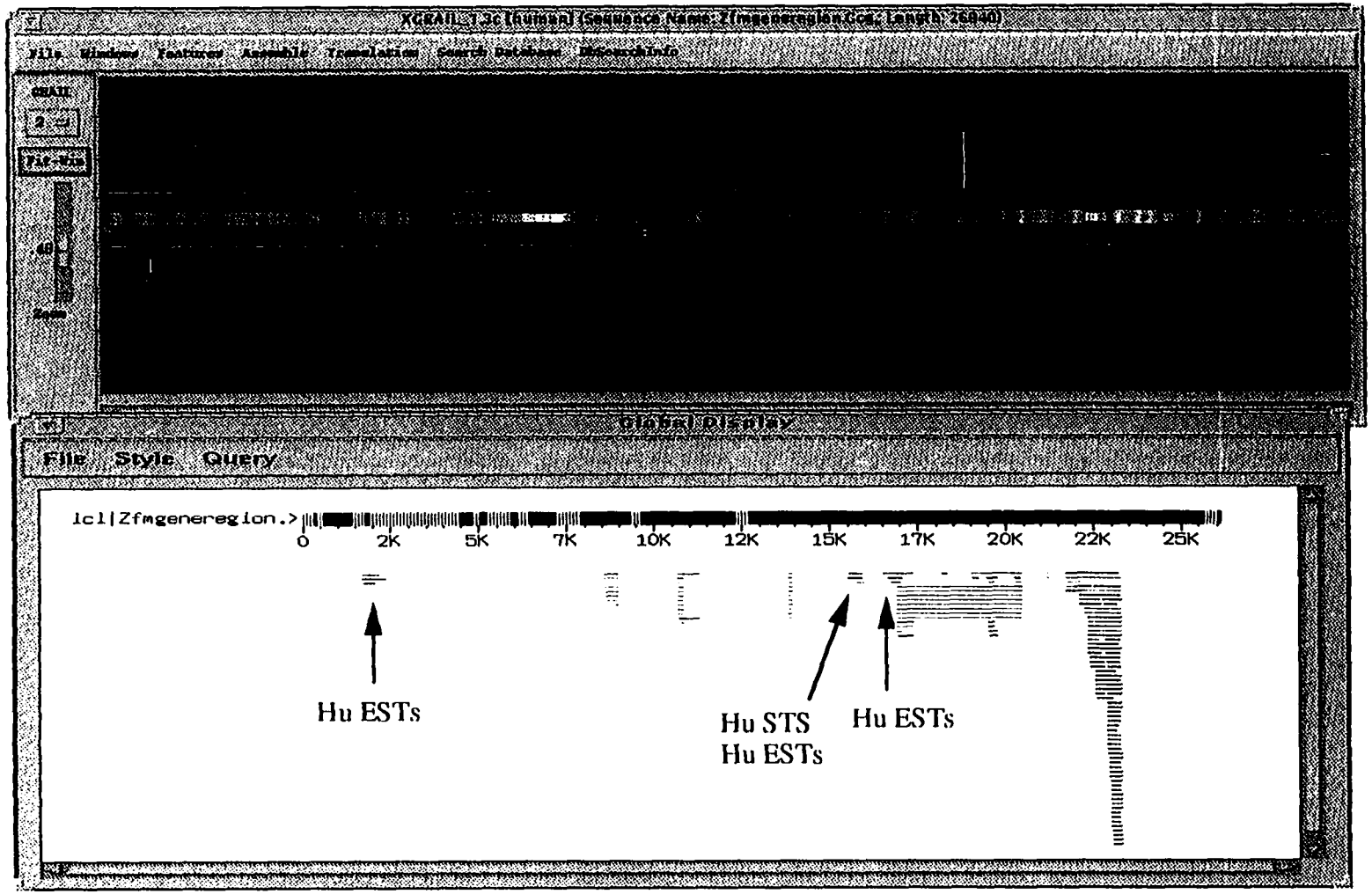
Tejan GC box	8787-8794	
Kosak sequence	8916-8920	
ATGstart	8921-8923	
TAGstop	22039-22041	(b4)
TGAstops	22374-22376	(zfm)
	22403-22405	(b3, as)
polyA signal	23227-23233	

Table 8: Genome organization of ZFM1, ZFM1 B3 isoform, ZFM1 B4 isoform and ZFM1 alternate splice products from from human chromosome 11q13. Abbreviations: zfm, ZFM1 cDNA; as, ZFM1 cDNA alternate splice product; altsp, alternate ZFM1 cDNA alternate splice product; b3, ZFM B3 isoform; b4, ZFM B4 isoform.

Splice Junctions-ZFM1, ZFM B3 isoform, ZFM B4 isoform and ZFM alternate splicing product

Left Exon	Left splice junction	Right splice junction	Right Exon
01ab	ACGCCGTTGGgtaagctggg	tgccttttagACTTCCCAAG	02ab
02a	GCTTATATAGgtaaatatac		
02b	(end of transcript)	tttcttccagTGCAACTGCA	03
03	CTGAGGACAGgttgggaaac	tttcccacagGTCCCCTTCC	04
04	CAGATTACAAgtaagcggag	tctttcttagACCTCCAGCA	05
05	TCGGGCCCAGgtgagtaact	gcctctctagAGGGAACACC	06
06	AGTGGAACAGgtgagcggtg	ttttcactagATAAGAAACA	07
07	ACGATAACAGgtatgtgatc	cctttaatagGATCTTAAGA	08
08	AATTCCAAAGgtgaggggct	cccacttcagGCCTGGTGAT	09
09	ACCTCCACCGgtgagcctgg	tctcacgtagTCTCTCATGT	10ab
10a	TCCTCCATGGgtaagtaagt		
10b	CAATGGGTAAgtaagtgtca	ttccatcaagATCAGTACCT	11
11	CAAGGAAAAGgtaatggctg	ctgtccacagGTATGATGCC	12ab
12a	TGGCAGCAAAgtgagtagaa		
12b	AGCAAAGTGAgtagaatatt	accaccccagATACGACGAC	13ab
13a	(end of transcript)		
13b	CCCCTCCGCCgcctccacca	cgccccgtagCATCGAGTGT	14a
		gggccgccagGATCCCTCCC	14b
consensus:	CAGgtaagt A t	tncagG c	

Table 9: Splice junctions for the 13 exons in ZFM1, ZFM B3 isoform, ZFM B4 isoform and alternately spiced products of ZFM1.



Strong database hits for ZFM1, ZFM1-B3, ZFM1-B4, SF1-Bo, SF1-HL1, CW17R, CW17E and CW.

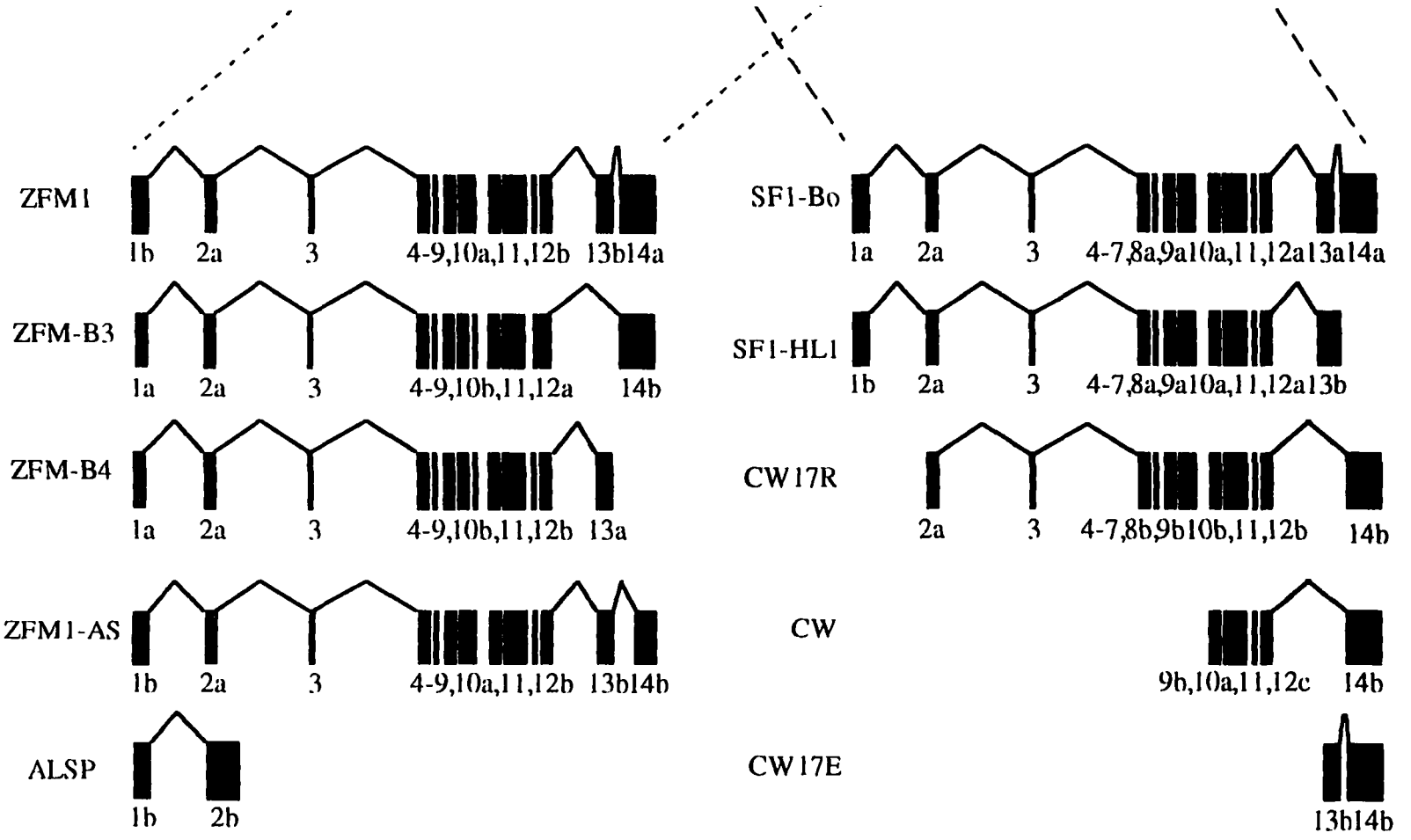


Figure 26: Top panel: Graphical result of Xgrail gene identification software for the ZFM1, ZFM1 B3 isoform, ZFM1 B4 isoform and ZFM1 alternate splice products. The peaks present are Xgrail predicted exons with a quality rating of excellent (green), good (blue), or marginal (red). The light blue lollipops are polyadenylation sites and the purple squares are CpG islands. Yellow and red squares indicate promotors and the vertical orange bars denote simple sequence repeats. Bottom panel: Graphical display of the powblast output using Musk. The black bar at the top of the panel represents the consensus sequence with the grey areas indicating repeat regions. The lines under the black bar indicate database homology between the consensus sequence and the different genes present in this region. Abbreviations: ZFM1, ZFM1 cDNA; ZFM-B3, ZFM1 B3 isoform; ZFM-B4, ZFM1 B4 isoform; ZFM1-AS, ZFM1 cDNA alternate splice product; ALSP, partial ZFM1 cDNA splicing; SF1-Bo, SF1-Bo isoform; SF1-HL1, SF1-HL1 isoform; CW17R, mouse CW17 full length cDNA; CW, mouse CW17 alternate splice product; CW17E, mouse CW17E alternate splice product.

EST hits, specifically on the 3'end of the gene (exons 13 and 14) and include Genbank accession numbers H91057, H91353, H16927, N48866, N46360, T07756, D56431, R64490, AA219560, R88018, H22296, N57427, N91865, R65994, W80863, H72626, and W01543. As expected, an STS (SHGC-35223, Genbank accession number G29780), and mouse and rat EST homologies (Genbank accession numbers AA170957, W62122, AA152851, U50960, W76938, AA171120, W98751, W47865, and W45996) are present in the exons for the ZFM gene, which also is a mouse CW17 gene homologue (Genbank accession numbers X85802 and Y10815, as discussed below) for testosterone-induced immunosuppression of malaria.

Two other genes are encoded by the same coding region as ZFM1, SF1-Bo and SF1- HL1 which are mammalian splicing factor isoforms, as well as a mouse homologue, CW17, a testosterone-induced immunosuppressor for malaria. The genomic organization of the SF1 group of genes is presented in Table 10. This gene family also utilizes alternate splicing and the intron/exon splice junctions from this gene family conform nicely with the consensus splice sequences, as presented in Table 11. Xgrail and powblast information is displayed in figure 26 in conjuction with the ZFM1 family of genes since many of the exons are identical.

This region also contains some repeated elements, most of which are full length, and include 15 ALU repeats (9 of 15 are full length), 23 L1 repeats (10 of 23 are full length), five SVA repeats (5 of 5 are full length) and one full length TAR1 repeat. One area rich in EST homologies also is repeat rich, as shown in the powblast output in figure 27. Three of these ESTs (R94779, H66893 and AA203482) contain some unique sequence along with an L1 repeat sequence (figure 28).

Other repeats in this 26,040 bp region, as listed in table 12, include the simple sequence repeats, several areas of small poly T repeats (6 of 16 are greater than 10 bp in length) and one area of significant poly C located at 21129-21269 which is

Genome Organization-SF1-Bo isoform, SF1-HL1 isoform and mouse CW17 gene alternate splice products

Exon number	gene product	position	size of exon	size of intron
01a	(Bo)	8640-8951	311	1736
01b	(HL1)	8880-8951	71	1736
02a	(Bo,HL1,17r)	10687-10815	128	2994
03	(Bo,HL1,17r)	13809-13884	75	3022
04	(Bo,HL1,17r)	16906-17058	152	203
05	(Bo,HL1,17r)	17261-17350	89	355
06	(Bo,HL1,17r)	17705-17888	183	88
07	(Bo,HL1,17r)	17976-18091	115	94
08a	(Bo,HL1)	18185-18292	107	766
08b	(17r)	18185-18292	107	762
09a	(Bo,HL1)	19058-19208	150	262
09b	(17r,cw)	19028-19208	180	262
10a	(Bo,HL1,cw)	19470-19743	273	320
10b	(17r)	19470-19747	277	316
11	(Bo,HL1,17r,cw)	20063-20122	59	113
12a	(Bo,HL1)	20235-20418	183	1286
12b	(17r)	20235-20418	183	1923
12c	(cw)	20235-20414	179	1927
13a	(Bo)	21704-21939	235	402
13b	(HL1)	21704-22041	337	---
13c	(17e)	21704-21878	174	390
14a	(Bo)	22341-22470	129	---
14b	(17r,17e,cw)	22341-23253	912	---

Tejan GC box	8787-8794	
Kosak sequence	8916-8920	
ATGstart	8921-8923	
TGAstop	22468-22470	(Bo)
	22403-22405	(17r,cw)
	22374022376	(17e)
TAGstops	22039-22041	(HL1)
	22374-22376	
	22403-22405	
polyA signal	23227-23233	

Table 10: Genome organization of SF1-Bo isoform, SF1-HL1 isoform and mouse CW17 gene alternate splice products from human chromosome 11q13. Abbreviations: Bo, SF1-Bo isoform; HL1, SF1-HL1 isoform; 17r, mouse CW17R full length cDNA; 17e, mouse CW17E alternate splice product; cw, mouse CW17 alternate splice product.

Splice Junctions-SF1-Bo isoform, SF1-HL1 isoform and mouse CW17 gene alternate splice products

Left Exon	Left splice junction	Right splice junction	Right Exon
01ab	ACGCCGTTGGgtaagctggg	tgccttttagACTTCCCAAG	02
02	GCTTATATAGgtaaatatac	tttcttccagTGCAACTGCA	03
03	CTGAGGACAGgttgggaaac	tttcccacagGTCCCCTTCC	04
04	CAGATTACAAgtaagcggag	tctttcttagACCTCCAGCA	05
05	TCGGGCCCAGgtgagtaact	gcctctctagAGGGAACACC	06
06	AGTGGAACAGgtgagcggtg	ttttcactagATAAGAAACA	07
07	ACGATAACAGgtatgtgatc	cctttaatagGATCTTAAGA	08ab
08ab	AATTCCAAAGgtgaggggct	ctcaggataaAGCACGGATA	09a
		cccacttcagGCCTGGTGAT	09b
09ab	ACCTCCACCGgtgagcctgg	tctcacgtagTCTCTCATGT	10ab
10a	CCTCCAATGGgtaagtaagt		
10b	CAATGGGTAAgtaagtgtca	ttccatcaagATCAGTACCT	11
11	CAAGGAAAAGgtaatggctg	ctgtccacagGTATGATGCC	12abc
12ab	AGCAAAGTGAgtaatatttt		
12c	TGGCAGCAAAgtgagtagaa	accaccccagATACGACGAC	13abc
13a	TCCGCCTCCCatggaccctt		
13b	(end of transcript)		
13c	ACAGAACTAGacttgttttt	gggccgccagGATCCCTCCC	14ab
consensus:	CAGgtaagt A t	tncagG c	

Table 11: Splice junctions for the SF1-Bo isoform, SF1-HL1 isoform and mouse CW17 gene alternate splice products.

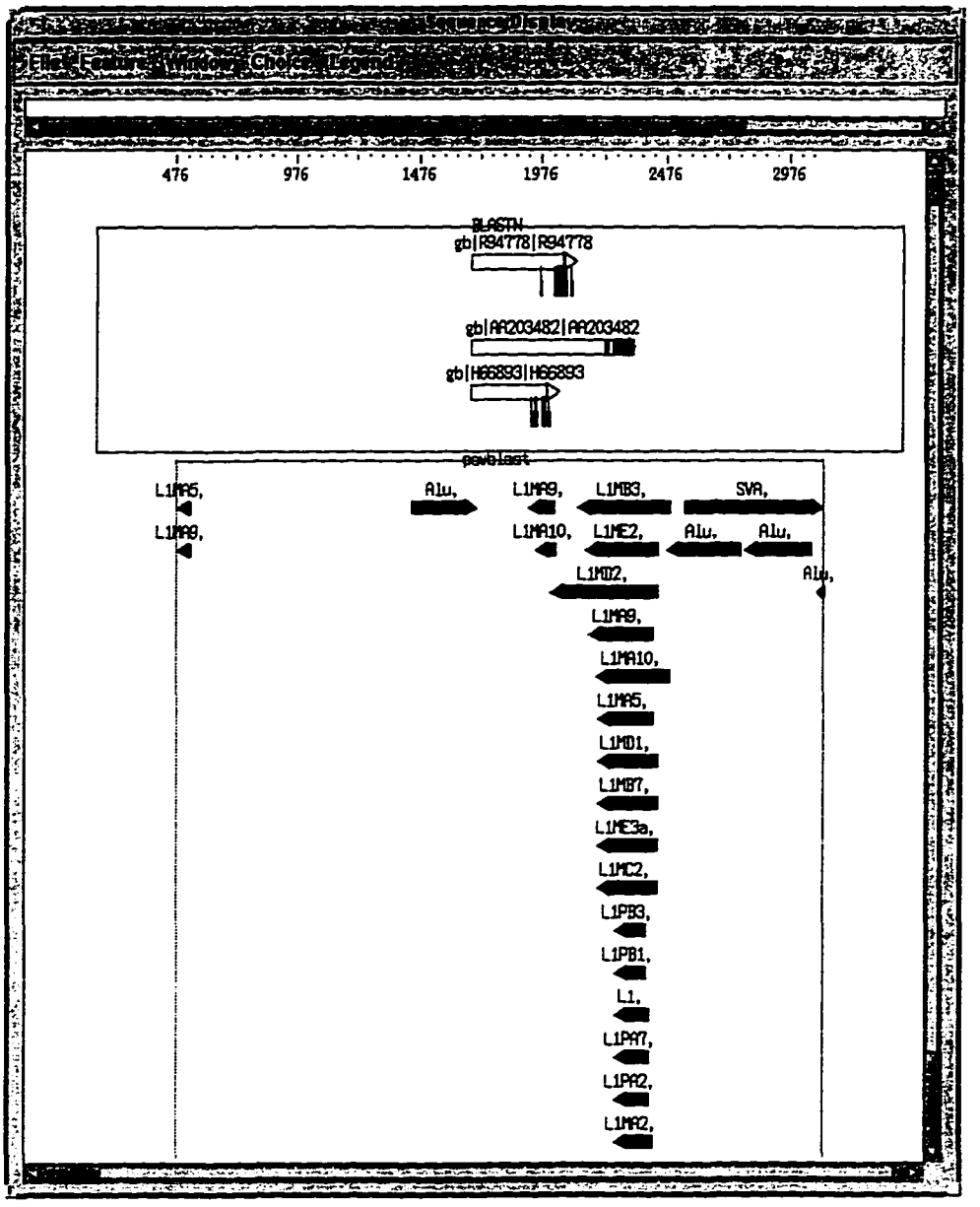


Figure 27: Powblast output of the ZFM1 gene region from human chromosome 11q13. All three ESTs in this area have significant homology to members of the L1 family of repeats.

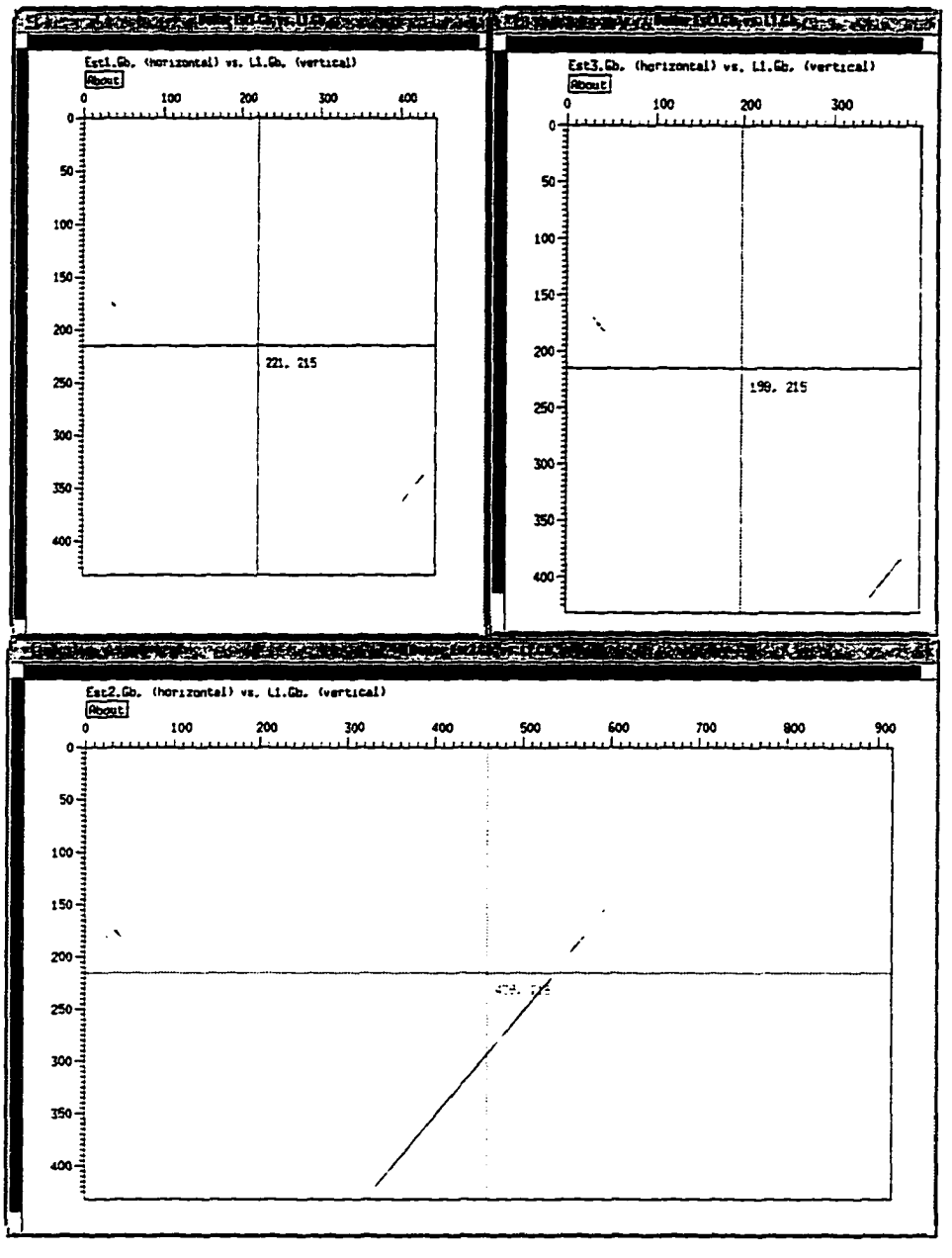


Figure 28: Dotter alignment of human ESTs with an L1 repeat element. Top left: alignment with EST R94779. Top right: alignment with EST H66893. Bottom: alignment with EST AA203482.

ZFM1, ZFM B3, ZFM B4, SF1-Bo, SF1-HL1, CW17 - Simple Sequence Repeats

position	Repeat
263-295	$A_{15}(GA)_2A_6(TA)_5$
398-409	T_{12}
2479-2500	T_{27}
3370-3397	$(AT_6)_3$
3683-3709	T_{27}
4383-4419	A_{16}
4958-4973	$(TAT_6)_2$
5373-5404	$T_{10}AT_6AT_7$
8045-8069	T_{35}
11268-11288	A_3GA_9
12328-12360	T_{17}
14492-14510	T_{11}
20460-20501	$C_{10}GAC_3GC_2G_2C_{10}$
20569-20593	$(C_6G)_2AC_4G_2C_3$
21129-21269	C_{140}
25656-25675	$T_2GT_7GT_3GT_5$

Table 12: Simple sequence repeats present in the ZFM1, ZFM B3, ZFM B4, SF1-Bo, SF1-HL1 and mouse CW17 genes from human chromosome 22q11 as reported by Xgrail.

140 bp long. Interestingly, none of the simple sequence repeats are found within the conserved exon region (exon 3-9 including intronic regions), but instead occur in non-coding areas of the ZFM gene region.

3. Muscle glycogen phosphorylase

The human muscle glycogen phosphorylase (PYGM) gene encompasses a total 16 kb of genomic sequence as determined by sequence scanning and is contained within a 20,185 bp contig of genomic sequence. A comparison of the genomic sequence with the known cDNA sequence (Genbank accession numbers M32579-M32598, numbered sequentially, with each exon submitted separately) revealed that the gene is composed of 20 exons, and all 20 were identified in the genomic sequence. The PYGM gene encodes for a protein of 847 amino acids with 242 nucleotides composing the 5' untranslated region and a 3' untranslated region of 260 nucleotides (see figure 30). Table 13 outlines the position and length of the 20 exons and Table 14 demonstrates the correlation of the intron/exon boundaries with consensus splice junctions. The PYGM gene has an ATG start codon preceded by a strong Kosak sequence, and an in-frame UGA stop codon that delineates a coding region for an 841 amino acid protein. A polyadenylation signal downstream marks the end of the 3' untranslated region.

Xgrail and Powblast output from the PYGM gene region is displayed in figure 29. All exons were identified in the genomic sequence by powblast and 19 of 20 were accurately predicted by Xgrail. The one exon that was not accurately predicted, exon 13, is only 88 bp in length, and does not conform well to the 5' consensus splice juctions sequence.

Powblast detected database homology with many human ESTs on the 3' end of the gene, including the following: F16457, F17281, F16528, F17763, F15633, F16859, AA197179, F18617, AA176345, F18656, F16783, F19679, AA193012, AA180319, C03553, W69291, AA195955, AA086388, AA086351, W30341, F01176 and F01226. These ESTs originated from several different tissue types including skeletal muscle,

```
aagtgagctcacctctccctcccctctcctctcctcccctgccctcaaaatagctccct      60
gagctgccaggccaagaccctcccattgcaggggaggcggcgtgctgcccagagccaggc     120
tttaaatccccttgaggctgggagctccactccttggctgaggcagtgcttgaggccgcc     180
                                                M  S  R  P  L
gtcccctctaccatcagtccagtccggcccgccctcctgcagccatgtcccggcccctgt     240
S  D  Q  E  K  R  K  Q  I  S  V  R  G  L  A  G  V  E  N  V
cagaccaagagaaaagaaagcaaatcagtgtgcgtggcctggccggcgtggagaacgtga     300
T  E  L  K  K  N  F  N  R  H  L  H  F  T  L  V  K  D  R  N
ctgagctgaaaaagaacttcaaccggcacctgcatttcacactcgtaaaggaccgcaatg     360
V  A  T  P  R  D  Y  Y  F  A  L  A  H  T  V  R  D  H  L  V
tggccaccccacgagactactactttgctctggcccataccgtgcgcgaccacctcgtgg     420
                                                exon1|exon2
G  R  W  I  R  T  Q  Q  H  Y  Y  E  K  D  P  K  V  L  L  G
ggcggtggatccgcacgcagcagcactactatgagaaggaccccaaggtgctgctgggga     480
 R  I  Y  Y  L  S  L  E  F  Y  M  G  R  T  L  Q  N  T  M  V
aggatctactacctgtctttagagttctatatgggacggacgctacagaacaccatggtg     540
                                          exon2|exon3
 N  L  A  L  E  N  A  C  D  E  A  T  Y  Q  L  G  L  D  M  E
aacctggccttagagaatgcctgtgacgaggccacctaccagctgggcctggacatggag     600
 E  L  E  E  I  E  E  D  A  G  L  G  N  G  G  L  G  R  L  A
gagctggaggaaattgaggaggatgcggggctgggcaacgggggcctgggccggctggca     660
exon3|exon4
 A  C  F  L  D  S  M  A  T  L  G  L  A  A  Y  G  Y  G  I  R
gcctgctttcttgactccatggcaacactgggcctggctgcctatggctacgggattcgc     720
                                          exon4|exon5
 Y  E  F  G  I  F  N  Q  K  I  S  G  G  W  Q  M  E  E  A  D
tatgagtttgggatttttaaccagaagatctccgggggctggcagatggaggaggccgat     780
 D  W  L  R  Y  G  N  P  W  E  K  A  R  P  E  F  T  L  P  V
gactggcttcgctacggcaacccctgggagaaggcccggcccgagttcacgctacctgtg     840
                                                exon5|exon6
 H  F  Y  G  H  V  E  H  T  S  Q  G  A  K  W  V  D  T  Q  V
cacttctacggccatgtggagcacaccagccagggtgccaagtgggtggacacacaggtg     900
 V  L  A  M  P  Y  D  T  P  V  P  G  Y  R  N  N  V  V  N  T
gtactggccatgccctacgataccccggtgcctggctatcgcaacaatgttgtcaacacc     960
                                          exon6|exon7
 M  R  L  W  S  A  K  A  P  N  D  F  N  L  K  D  F  N  V  G
atgcgcctctggtctgccaaggctcccaatgacttcaacctcaaggacttcaatgtcggt    1020
 G  Y  I  Q  A  V  L  D  R  N  L  A  E  N  I  S  R  V  L  Y
ggctacatccaggctgtgttggaccgaaacctggcggagaacatctctcgtgtcctgtac    1080
      exon7|exon8
 P  N  D  N  F  F  E  G  K  E  L  R  L  K  Q  E  Y  F  V  V
cccaatgataatttcttcgaagggaaggagctgcggctgaagcaggagtatttcgtggtg    1140
 A  A  T  L  Q  D  I  I  R  R  F  K  S  S  K  F  G  C  R  D
gctgccaccctccaggacatcatccgtcgcttcaagtcttccaagttcggctgccgtgat    1200
                                          exon8|exon9
 P  V  R  T  N  F  D  A  F  P  D  K  V  A  I  Q  L  N  D  T
cccgtgcgcacgaacttcgatgccttcccagataaggtggccatccagctcaatgacacc    1260
 H  P  S  L  A  I  P  E  L  M  R  I  L  V  D  L  E  R  M  D
caccctccctggccatccccgagctgatgaggatcctggtggacctggaacggatggac    1320
   exon9|exon10
 W  D  K  A  W  D  V  T  V  R  T  C  A  Y  T  N  H  T  V  L
tgggacaagcgtgggatgtgacagtgaggacctgtgcctacaccaaccacacggtgctgc    1380
P  E  A  L  E  R  W  P  V  H  L  L  E  T  L  L  P  R  H  L
ccgaggccctggagcgctggccggtgcacctcttggagacgctgctgccgcggcacctcc    1440
                                          exon10|exon11
Q  I  I  Y  E  I  N  Q  R  F  L  N  R  V  A  A  A  F  P  G
agatcatctacgagatcaaccagcgcttcctcaaccgggtggcggccgcattcccagggg    1500
D  V  D  R  L  R  R  M  S  L  V  E  E  G  A  V  K  R  I  N
```

```
acgtagaccggctgcggcgcatgtcgctggtggaggagggcgcagtgaagcgcatcaaca     1560
M  A  H  L  C  I  A  G  S  H  A  V  N  G  V  A  R  I  H  S
tggcacacctgtgcatcgcggggtcgcacgccgtcaacggtgtggcccgcatccactcgg     1620
                exon11|exon12
E  I  L  K  K  T  I  F  K  D  F  Y  E  L  E  P  H  K  F  Q
agatcctcaagaagaccatcttcaaagacttctatgagctggagcctcataagttccaga     1680
N  K  T  N  G  I  T  P  R  R  W  L  V  L  C  N  P  G  L  A
ataagaccaacggcatcacccctcggcgctggctggttctgtgtaaccccgggctggcag     1740
          exon12|exon13
E  V  I  A  E  R  I  G  E  D  F  I  S  D  L  D  Q  L  R  K
aggtcattgctgagcgcatcggggaggacttcatctctgacctggaccagctgcgcaaac     1800
                                                exon13|exon14
L  L  S  F  V  D  D  E  A  F  I  R  D  V  A  K  V  K  Q  E
tgctctcctttgtggatgatgaagctttcattcgggatgtggccaaagtgaagcaggaaa     1860
N  K  L  K  F  A  A  Y  L  E  R  E  Y  K  V  H  I  N  P  N
acaagttgaagtttgctgcctacctagagagggaatacaaagtccacatcaaccccaact     1920
S  L  F  D  I  Q  V  K  R  I  H  E  Y  K  R  Q  L  L  N  C
cactcttcgacatccaggtgaagcggattcacgaatataaacgacagctcctcaactgcc     1980
                  exon14|exon15
L  H  V  I  T  L  Y  N  R  I  K  R  E  P  N  K  F  F  V  P
tccatgtcatcaccctgtacaaccgcatcaagagggagcccaataagtttttgtgcctc     2040
                  exon15|exon16
R  T  V  M  I  G  G  K  A  A  P  G  Y  H  M  A  K  M  I  I
ggactgtgatgattggagggaaggctgcacctgggtaccacatggccaagatgatcatca     2100
R  L  V  T  A  I  G  D  V  V  N  H  D  P  A  V  G  D  R  L
gactcgtcacagccatcggggatgtggtcaaccatgacccggcagtgggtgaccgcctcc     2160
R  V  I  F  L  E  N  Y  R  V  S  L  A  E  K  V  I  P  A  A
gtgtcatcttcctggagaactaccgagtctcactggccgagaaagtgatcccagctgcag     2220
D  L  S  E  Q  I  S  T  A  G  T  E  A  S  G  T  G  N  M  K
acctctctgagcagatctccactgcgggcactgaagcctcaggcaccggcaacatgaagt     2280
F  M  L  N  G  A  L  T  I  G  T  M  D  G  A  N  V  E  M  A
tcatgctcaacggggctctgaccattggcaccatggacggggccaatgtggagatggcag     2340
E  E  A  G  E  E  N  F  F  I  F  G  M  R  V  E  D  V  D  K
aagaggcgggagaggaaaacttcttcatctttggcatgcgggtggaggatgtggataagc     2400
L  D  Q  R  G  Y  N  A  Q  E  Y  Y  D  R  I  P  E  L  R  Q
ttgaccaaagagggtacaatgcccaggagtactacgatcgcattcctgagcttcggcagg     2460
V  I  E  Q  L  S  S  G  F  F  S  P  K  Q  P  D  L  F  K  D
tcattgagcagctgagcagtggcttcttctccccaaacaacccgacctgttcaaggaca     2520
                        exon18|exon19
I  V  N  M  L  M  H  H  D  R  F  K  V  F  A  D  Y  E  D  Y
                                 exon19|exon20
ttgtcaatatgctcatgcaccatgaccggtttaaagtcttcgcagattatgaagactaca     2580
I  K  C  Q  E  K  V  S  A  W  Y  K  N  P  R  E  W  T  R  M
ttaaatgccaggagaaagtcagcgcctggtacaagaacccaagagagtggacgcggatgg     2640
V  I  R  N  I  A  T  S  G  K  F  S  S  D  R  T  I  A  Q  Y
tgatccggaacatagccacttctggcaagttctccagtgaccgcaccattgcccagtatg     2700
A  R  E  I  W  G  V  E  P  S  R  Q  R  L  P  A  P  D  E  A
cccgggagatctggggtgtggagccttcccgccagcgcctgccagccccggatgaggcca     2760
I  *
tctgagtctcagaccagaccccaaaccatcccttgagtctgtcacactctcttgggccag     2820
ccccacacctcatgcagagggtggggtactggagttagatctctaccaccctcctggaac     2880
cctcattaaccccactctcaatgtccagtgtccagcgtgactaaggacacgggccccctt     2940
ccgtgcctggctcccggtacccctcctatttatggggtctgaccaactgcaccactccct     3000
aataaattcatctccattgggaaa                                         3024
```

Figure 29: Human muscle glycogen phosphorylase cDNA and amino acid translation.

Genbank accession numbers M32579-M32598.

Genome Organization - Human muscle glycogen phosphorylase (PYGM) gene and homologues

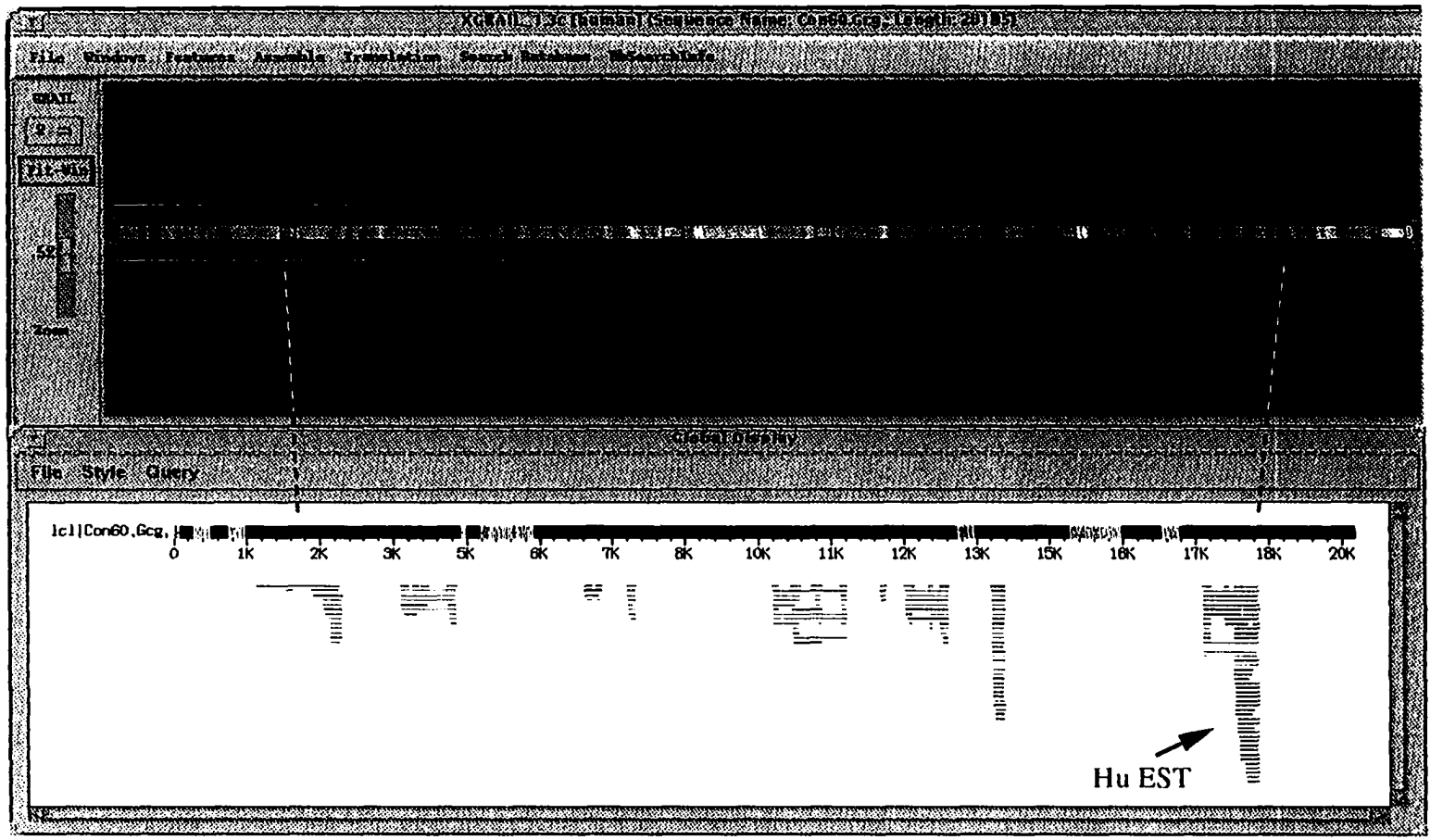
Exon number	position	size of exon	size of intron
01	2319-2860	541	1026
02	3886-3989	103	87
03	4076-4155	79	88
04	4243-4346	103	334
05	4680-4816	136	2217
06	7033-7144	111	91
07	7235-7318	83	445
08	7763-7895	132	2357
09	10252-10348	96	228
10	10576-10720	144	199
11	10919-11082	163	331
12	11413-11528	115	582
13	12110-12198	88	327
14	12525-12676	151	268
15	12944-13002	58	128
16	13130-13266	130	747
17	14013-14226	213	3391
18	17617-17755	138	235
19	17990-18058	68	111
20	18169-18323	154	---
ATGstart	2618-2620		
Kosak	2613-2617		
UGAstop	18316-18318		
PolyA	18557-18562		

Table 13: Genome organization of the human muscle glycogen phosphorylase (PYGM) gene region from human chromosome 11q13. The gene contains 20 exons for a total of 2.8 kb, which span roughly 16 kb of genomic sequence. The start codon is located within exon 1, preceded by a strong Kosak sequence, and the stop codon at the end of exon 20.

Splice Junctions-Human muscle glycogen phosphorylase (PYGM) gene and homologues

Left Exon	Left splice junction	Right splice junction	Right Exon
01	GGACCCCAAGgtgctgctgg	ctctgggcagAGGATCTACT	02
02	CACCTACCAGgtgtgtgtgg	actcctccagCTGGGCCTGG	03
03	CGGCTGGCAGgtaagaagac	acacccccagCCTGCTTTCT	04
04	GGGCTGGCAGgtgagcagcc	ttgggtgcagATGGAGGAGG	05
05	GGACACACAGgtgaggatga	ccccccgcagGTGGTACTGG	06
06	CTCAAGGACTgtgagttcat	tgtccctcagTCAATGTCAG	07
07	CAATGATAATgtgcgtcacc	ctgtccccagTTCTTCGAAG	08
08	CCCAGATAAGgtaccatgcg	cctaccccagGTGGCCATCC	09
09	TGGGCTTCAGggcccctctg	cgtgtgccagGCGTGGGATG	10
10	GATCAACCAGcgcttcctca	ctgggcacagCGGGTGGCGG	11
11	AGAAGACCATgtgagccccg	catcctgcagCTTCAAAGAC	12
12	CATTGCTGAGgtgagaggcc	aaagccatctTCCCATGGCG	13
13	GAGATGCAATgtgagacagc	ctgcctccagGAAAACAAGT	14
14	CTGTACAAGGgtgagtggca	ttctccacagGCATCAAGAG	15
15	TGGAGGGAAGgtgagaagcc	gtcccctcagGCTGCACCTG	16
16	GCCGAGAAAGgtgggtgctg	caccccacagTGATCCCAGC	17
17	GGTGGAGGATgtggataagc	gtcctggcagGTACAATGCC	18
18	ACCATGACCGgtgagctggt	tcccctccagGTTTAAAGTC	19
19	CTTGTACAAGgtgaggggtc	agtgaaccagAGCTTCCCTT	20
consensus:	CAGgtaagt A t	tncagG c	

Table 14: Splice junctions for human glycogen phosphorylase (PYGM) gene from human chromosome 11q13.



Muscle glycogen phosphorylase (PYGM) gene

Figure 30: Top panel: Graphical result of Xgrail gene identification software for the Human muscle glycogen phosphorylase (PYGM) gene region from human chromosome 11q13. The peaks present are Xgrail predicted exons with a quality rating of excellent (green), good (blue) or marginal (red). The light blue lollipops are predicted polyadenylation signals and the purple squares are CpG islands. Yellow and red squares indicate promotors and the vertical orange bars denote simple sequence repeats. Bottom panel: Graphical display of the powblast output using Musk. The black bar at the top represents the consensus sequence with the grey areas indicating repeat regions. The lines under the black bar indicate homology between the consensus sequence and an entry in the public database.

pregnant uterus, and adult heart (reflecting the specialized function of the gene) and all have homology with the PYGM gene at exons 19 and 20. This region also contains sequences with database homology to the rat (Genbank accession numbers L18752 and L10669), cow (S82859) and rabbit (D00040 and X04265) homologues of the human PYGM gene, indicating a strong conservation of sequences of this gene among various species.

The PYGM gene region contains only a few repeat elements, however, those present are typically full length, with eight of ten ALU repeats full length, four of six SVA repeats full length being observed. Additionally, there is one region of partial MIR repeat homology. However, simple sequence repeats (see table 15) are quite prevalent in this region, and a particularly interesting simple sequence repeat occurs in the intron between exons 8 and 9. The first part of this region is composed entirely of a poly CT repeat, stretching from 8014-8996. Between 9001-10039, the repeat changes to a poly GA repeat, with one area of poly G stretching for 38 bases. One small area within this intronic region was marked by Xgrail as a CpG island, and this region encompasses the transition area between the poly CT repeat and the poly GA repeat.

4. Putative genes

a. MEN-1 candidate gene

The putative MEN-1 gene region is contained within a 14,837 nucleotide long contig that also contains a portion of the β lymphocyte serine/threonine (GC kinase)/Rab8 gene region of human chromosome 11q13. This region contains the putative gene candidate whose deletion/disruption is paramount in the tumorigenesis of MEN-1. This putative MEN-1 gene spans 9 kb of genomic sequence and encodes a 2.8 kb transcript, resulting in a predicted protein of 610 amino acids which has no homology to previously known proteins. The cDNA for this putative gene has been isolated and sequenced in the lab of Dr. Francis Collins and is reported in figure 31 with intron/exon boundaries indicated.

Human muscle glycogen phosphorylase (PYGM) gene region - Simple Sequence Repeats

position	Repeat
303-332	$T_4AT_2C(T_3A)_2T_7AT_6$
1200-1216	A_{17}
1466-1478	$T_{11}CT$
5607-5625	$A_{10}T_3A_6$
8014-8214	poly CT
8374-8700	poly CT
8738-8755	C_9TC_3T
8874-8996	poly CT
9001-9157	$(CT)_5C_5(CT)_9C_{133}$
9160-9515	poly GA
9509-9699	poly GA
9707-9996	poly GA
10001-10039	G_{38}
10816-10919	poly GT
14750-14885	poly GA
15001-15124	C_{123}
15165-15181	A_{16}
16917-16940	T_{15}
20073-20103	C_{13}
20129-20152	$C_8G_4C_2G_4C_2G_2C_6$

Table 15: Simple sequence repeats present in the human muscle glycogen phosphorylase (PYGM) gene region from human chromosome 11q13 as reported by Xgrail.

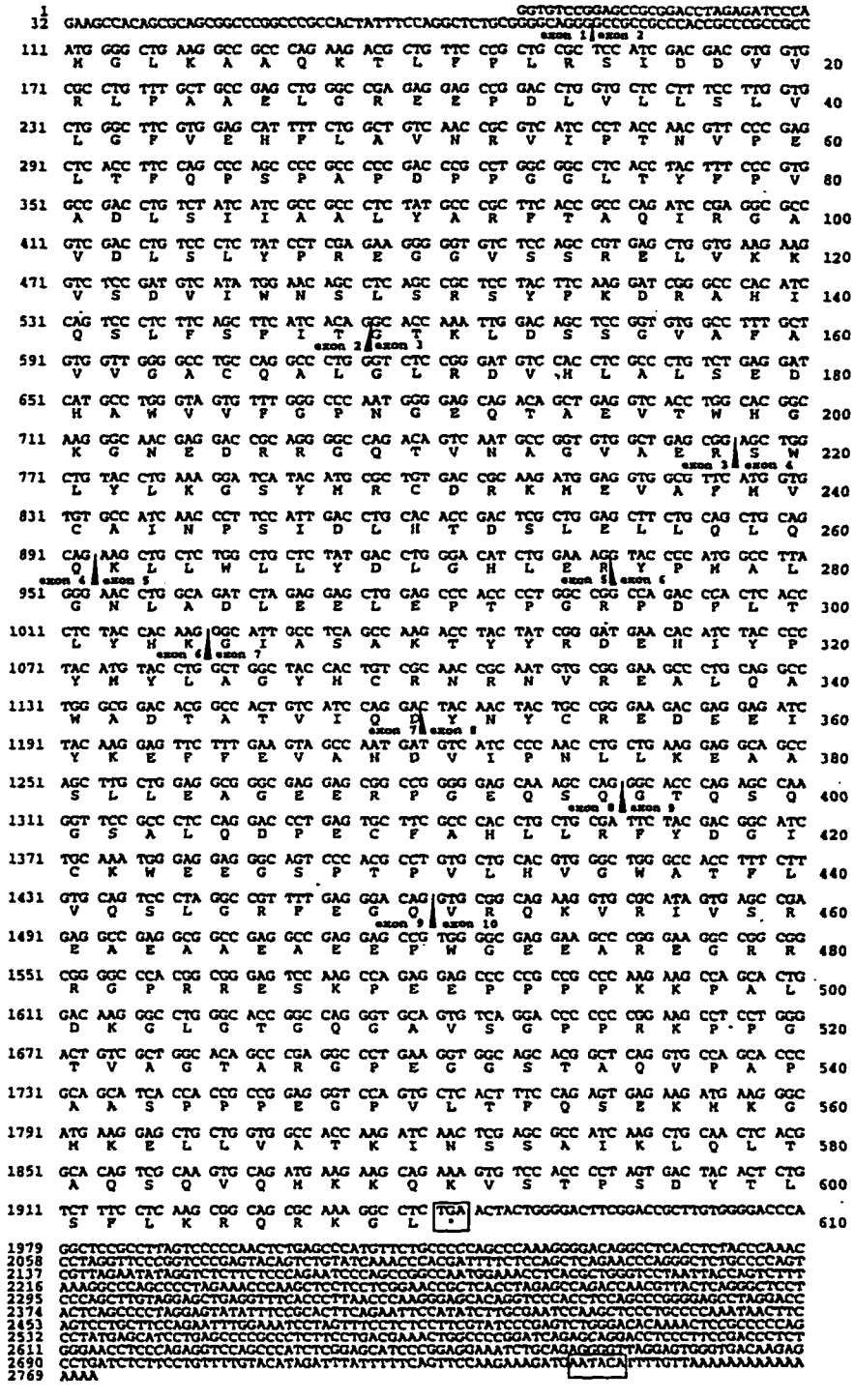
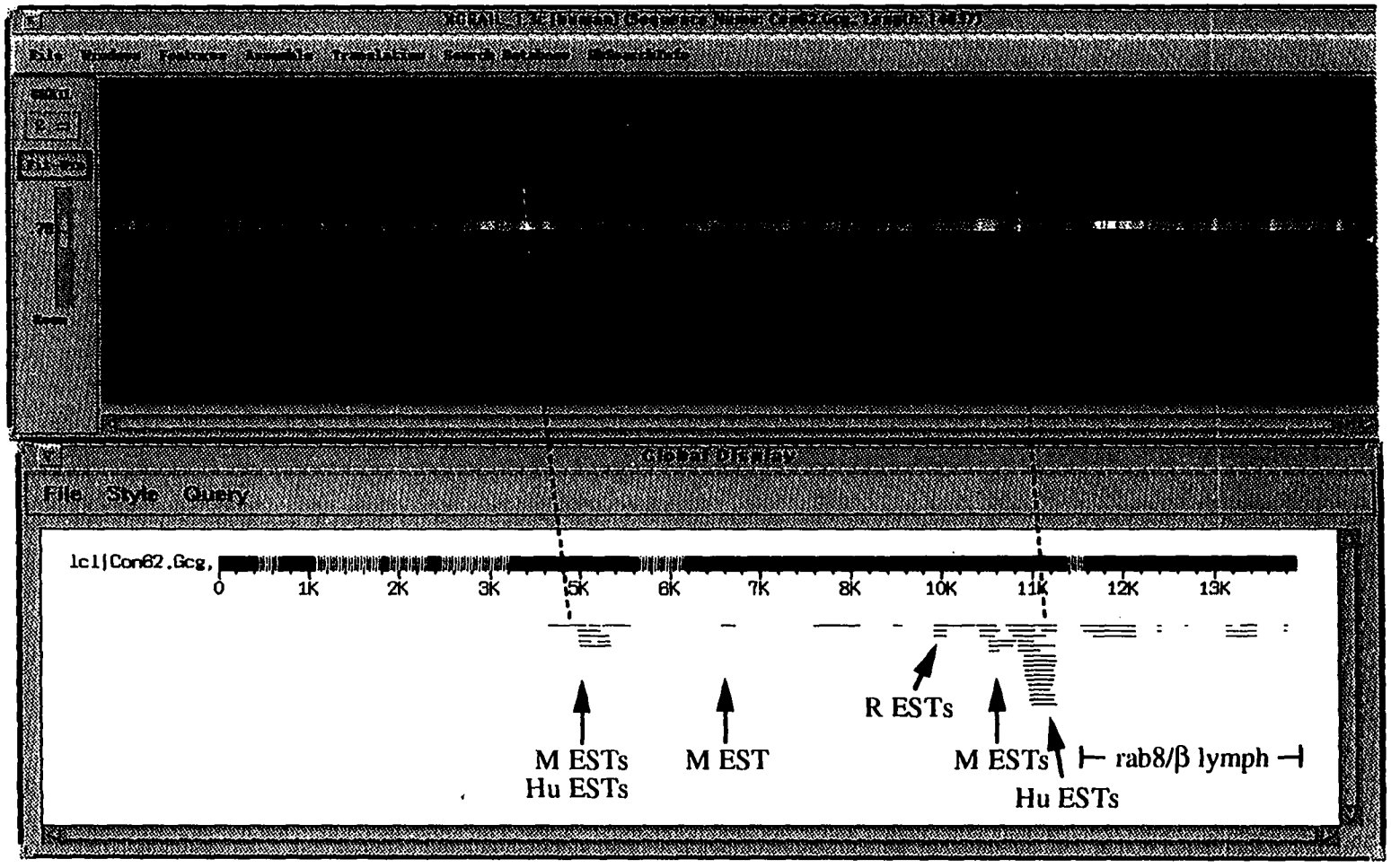


Figure 31: cDNA seqeunce of the MEN-1 gene. Exons are marked with arrow

The putative MEN-1 gene was predicted by a comparison of Xgrail and powblast results, which shows a correlation between the ten Xgrail predicted exons (Xgrail accurately predicted all but exon 1) and the human EST database homologies as seen in figure 32. The position and genomic organization of the ten exons of the MEN-1 gene are listed in table 16, and the adherence to the consensus intron/exon splice sequence is detailed in table 17. The putative MEN-1 gene does not contain the typical TATAA promoter, but instead has a strong CpG island in the 5' untranslated region, indicating that this gene most likely is a ubiquitously expressed "housekeeping" gene. The ubiquitous expression of the MEN-1 gene also is cofirmed by the observance of significant homologies with human ESTs (Genbank accession numbers AA211377, AA209475, AA211877, AA157873, H16350, H14300, H38333, H14313, R32679, R98414, H91638, N90077, AA147612, W38527, N36190, 24156, H97384, AA147620, AA157374, H41351, H91639, N51774, H38296, H14281, H14277, and R28488) originating from a variety of tissue sources including pancreas, adrenal medulla, thyroid, adrenal cortex, testis, thymus, small intestine, stomach, spleen, prostate, ovary, colon and leukocyte. The translation start methionine codon is located within exon 2, with the first exon being untranslated. A stop codon occurs in frame at position 10748-10750 followed by a polyadenylation signal at 11549-11554 denoting the end of the 3' untranslated region. Xgrail did predict one additional exon in this region with a good quality rating located at position 136-226, however, powblast did not report any significant database homology in this area.

Several mouse ESTs (Genbank accession numbers AA105533, AA031132, AA000099, AA048913, W89897, W36791, AA168218, W76863, AA049922, and W14310) and one rat EST (H34387) also have homology in this region and several repeat elements also are observed in the last 10 kb of the sequence. The region upstream of the predicted MEN-1 gene also is quite rich is repeat elements, with seven ALU


File Style Query
lcl|Con82.Gcg,
0
1K
2K
3K
5K
6K
7K
8K
10K
11K
12K
13K
M ESTs
Hu ESTs
M EST
R ESTs
M ESTs
Hu ESTs
rab8/β lymph

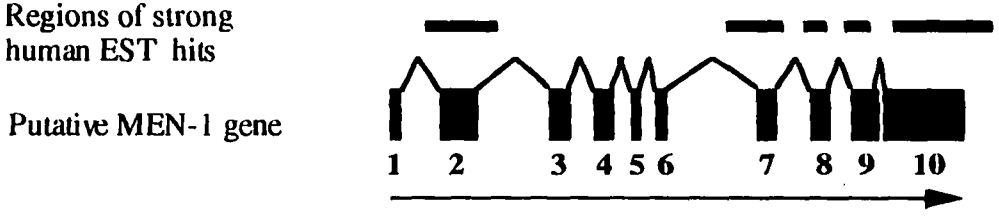

Regions of strong human EST hits
Putative MEN-1 gene
1 2 3 4 5 6 7 8 9 10

Figure 32: Top panel: Graphical result of Xgrail gene identification software for the Multiple Endocrine Neoplasia, type 1 gene region from human chromosome 11q13. The peaks present are Xgrail predicted exons with a quality rating of excellent (green), good (blue) or marginal (red). The light blue lollipops are polyadenylation signals and the purple squares are CpG islands. Yellow and red squares indicate promotors and the vertical orange bars denote simple sequence repeats. Bottom panel: Graphical display of the powblast output using Musk. The black bar at the top of the panel represents the consensus sequence with the grey areas indicating repeat regions. The lines under the black bar indicate database homology between the consensus sequence and entries in the public database.

Genome Organization - Multiple Endocrine Neoplasia, type 1 candidate gene

Exon number	position	size of exon	size of intron
01 (UTR)	4370-4457	87	497
02	4954-5421	467	1565
03	6986-7194	208	211
04	7405-7533	128	333
05	7866-7906	40	81
06	7987-8074	87	640
07	8714-8850	136	462
08	9312-9447	135	438
09	9885-10049	164	218
10	10267-10750	483	
ATGstart	4977-4979		
UGAstop	10748-10750		
polyA	11549-11554		

Table 16: Genome organization of the MEN-1 candidate gene region from human chromosome 11q13. The candidate gene contains 10 exons (the first being untranslated) for a total of 2.8 kb, which span roughly 9 kb of genomic sequence. The start codon is located within exon 2, the stop codon at the end of exon 10 and the polyadenylation signal 799 nucleotides downstream of the stop codon.

Splice Junctions-Multiple Endocrine Neoplasia, type 1 candidate gene

Left Exon	Left splice junction	Right splice junction	Right Exon
01	CGGGGCAGGGtgggcccaga	tgccttgcagGCCGCCGCCC	02
02	TTCATCACAGgttggagccc	ccttccacagGCACCAAATT	03
03	GGCTGAGCGGgtattgttcc	ccccccctaagAGCTGGCTGT	04
04	GCTGCAGCAGgtgagggctg	cggctcctagAAGCTGCTCT	05
05	ATCTGGAAAGgtcagtagag	cctttcctagGTACCCCATG	06
06	CTACCACAAGgtgggggcat	ctccatccagGGCATTGCCT	07
07	TCATCCAGGAgtgaggatcc	gctctcacagCTACAACTAC	08
08	GCAAAGCCAGgtgaaaggct	cctcgtccagGGCACCCAGA	09
09	TGAGGGACAGgtgagggaca	actggcccagGTGCGGCAGA	10
consensus:	CAGgtaagt A t	tncagG c	

Table 17: Splice junctions for the Multiple Endocrine Neoplasia, type I candidate gene from human chromosome 11q13.

(four with SVA in same position in reverse direction), two MER repeats and one full length TAR1. The repeat regions at 2 kb and 5.6-6 kb are full length ALU repeats and the full length TAR1 repeat is located after the MEN-1 gene at 11.5 kb (prior to the exons for the β lymphocyte serine/threonine (GC kinase)/Rab8 genes.

Another unique feature of this genomic region is the relative lack of simple sequence repeats (see table 18). There only are three regions of poly T tracts, but all are less than 20 nucleotides (20, 14 and 7 nucleotides), and there is one region which contains a poly A tract of 27 nucleotides.

b. Neurexin II alpha

Detailed analysis of a 60,615 bp region of b137c7 revealed the presence of a putative human gene with significant homology to the first 15 exons of the rat neurexin II alpha gene (Genbank accession number M96376) and slight homology (homology to two exons or fewer) to the rat neurexin I alpha cDNA (Genbank accession number M96374), the rat neurexin III alpha cDNA (Genbank accession number L14851), Bos taurus neurexin I alpha (Genbank accession number L14855) and the mouse alpha adaptin cDNA (Genbank accession number X14971). Thirteen of the first 15 exons of the rat neurexin II alpha gene are present in this gene region and encompass a 58 kb region as determined by sequence scanning. Two of the 15 exons, exons 2 and 11, are presumed to be located in sequencing gaps present in this region. An interesting feature of this gene region is the varying size of the introns, ranging from 218 bp to 9848 bp in length. A detailed analysis of the rat neurexin II alpha genome organization including the length of introns and exons is presented in table 19. The remaining exons of the rat neurexin II alpha gene are not found in b137c7 as determined by Crossmatch, since this 60,651 bp region is located at one end of the b137c7 genomic clone insert. However, the intron/exon boundaries are quite conserved for the 13 exons present in the genomic sequence (see table 20) and indicate that a human homologue for this gene may be located

Multiple Endocrine Neoplasia, type 1 candidate gene - Simple Sequence Repeats

position	Repeat
772-805	$A_5CA_4CA_5GA_{10}$
1311-1340	$T_4AT_6A_2T_{13}AT_3$
1784-1809	$T_3A_4TA_2T_{16}$
2986-2996	T_7
3709-3754	$A_{27}(TA)_9(TG)_{13}$
5803-5823	T_{20}
6118-6129	T_{14}
11782-11973	poly GC region
11829-11942	poly GC region

Table 18: Simple sequence repeats present in the Multiple Endocrine Neoplasia, type 1 gene region from human chromosome 11q13 as reported by Xgrail.

Genome Organization - rat neurexin II alpha gene region

Exon number	position	size of exon	size of intron
01	2522-3398	876	8857(gap)
02	(absent--in sequencing gap)		
03	12255-12287	32	4700
04	16987-17058	71	4458
05	21516-21817	301	8667
06	30484-30528	44	9848
07	40376-40538	162	755
08	41293-41731	438	7622
09	49353-49739	386	218
10	49957-50160	203	6743(gap)
11	(absent--in sequencing gap)		
12	56903-57031	128	399
13	57430-57812	382	617
14	58429-58619	190	1541
15	60160-60333	173	---

Table 19: Genome organization of the rat neurexin II alpha gene region from human chromosome 11q13. This region has homology to the first 15 exons of the rat neurexin II alpha gene for a total of 3.4 kb, which span roughly 58 kb of genomic sequence and indicates a putative human homologue for this gene family based on sequence similarity.

Splice Junctions-Rat neurexin II alpha gene region

Left Exon	Left splice junction	Right splice junction	Right Exon
01	TGCAGCGAAGgtgagccccc	(sequence gap)	02
02	(sequence gap)	tccccttccAAGGTCCGGC	03
03	AACAGCGAAGgtatggtttg	tgcttctccaGTAGGGTCCT	04
04	CCAACAAAAGgtcagtgacc	cggcatccagGCAAGGAGGA	05
05	CCTGCGCCAGgtaggaggag	tccggtgaagCACGCAGGGA	06
06	GCATTATCTGgtagatatca	ggctggacagGTGACCATCT	07
07	CCTCAAGGACgtgagtaggg	ctccctgcagGTGGTCTATA	08
08	GGGCGAAAAGgtcaggaggc	tgattcccagGCTCCATCTC	09
09	TGTGAGAGAGgtgaggctgg	ctgtgggcagAGGCCACGGT	10
10	GTCAACCTCGgtaaccaccc	(sequence gap)	11
11	(sequence gap)	tccccccttgGTGCCGCAGG	12
12	ACTGTGGAGGgtaggtggtc	tcctgagcagGACAGATGGC	13
13	TGGTCAAGGGgtgaggggct	tgtggggcagGTACATCCAC	14
14	GATCTCAAAGgtggggctgg	cccttgcagGGGAGTTGTA	15
15	GGCTGTGATGgtgagtggag	----------	
consensus:	CAGgtaagt A t	tncagG c	

Table 20: Splice junctions for rat neurexin II alpha gene from human chromosome 11q13, which indicates a putative human homologue for this gene family by sequence similarity.

distal to this region of chromosome 11. The rat neurexin II alpha gene has an ATG start codon at position 2669-2671 and an SP1 binding site (Tejan GC box) upstream in the 147 bp 5' untranslated region at position 2592-2603 that may represent a promoter enhancer for this gene.

Output from Xgrail and powblast for this region is displayed in figure 33. Xgrail accurately predicted 9 of the 13 exons present in the genomic sequence data, with exons 3, 4, and 15 not predicted (exons 2 and 11 presumably are located in sequencing gaps). Each of the Xgrail predicted exons correlates well with the powblast database homology for the rat neurexin II alpha gene as seen in figure 33. This region of the genome also has database homology with two human ESTs, N220878 and R85242, and the 5' untranslated region of the mouse alpha adaptin cDNA, X14971, however none of the exons for this gene are present. The database homology of the rat neurexin II alpha gene discussed above, with the rat neurexin I alpha (M96374), the Bos taurus neurexin III alpha (L14855), and the rat neurexin III alpha (L148510) cDNA sequences is located around 40 kb. This indicates cross-species conservation of gene sequence within the neurexin gene family, and alludes to the possiblity of a human neurexin II alpha gene in this region.

To investigate this possibility, the Xgrail output was examined for other exons located within the genomic rat neurexin II alpha gene region, but which were not present in the rat neurexin II alpha gene. This inquiry revealed five exons (three excellent, one good and one marginal) which could potentially be part of a human neurexin II alpha gene. The genomic organization of a putative human neurexin II alpha gene is listed in table 21 and includes both the rat exons and the additional five exons predicted by Xgrail. These exons all have the conserved intron/exon boundaries (see table 22) and are included in the pictoral representation of the putative human neurexin gene found in figure 33. These exons all encode in-frame amino acids and in combination with the exons from the rat neurexin II alpha gene, a human putative gene model was constructed

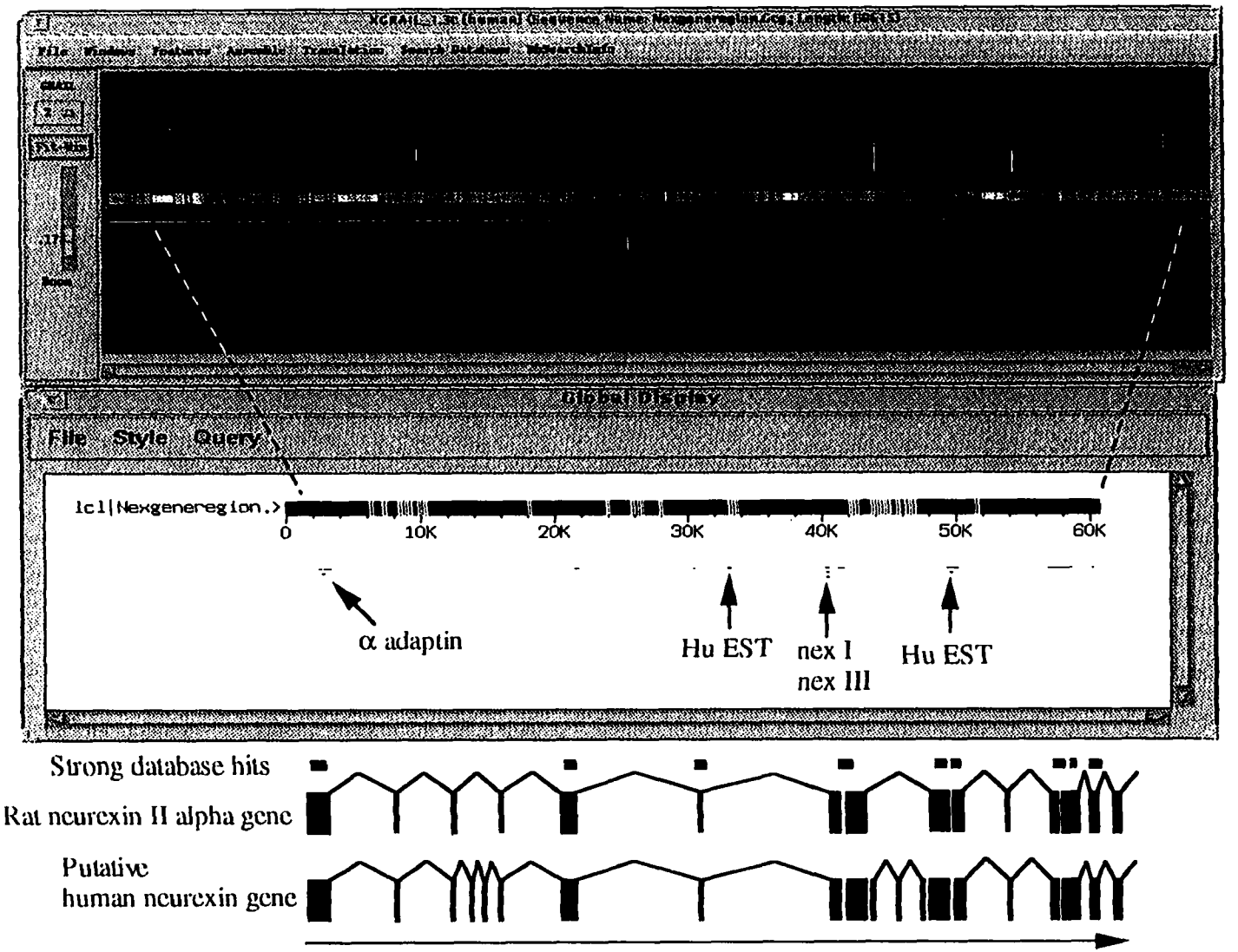

File Style Query
lcl|Nexgeneregion.>
0
10K
20K
30K
40K
50K
60K
α adaptin
Hu EST
nex I
nex III
Hu EST
Strong database hits
Rat neurexin II alpha gene
Putative
human neurexin gene

Figure 33: Top panel: Graphical result of Xgrail gene identification software for the neurexin II alpha gene region from human chromosome 11q13. The peaks present are Xgrail predicted exons with a quality rating of excellent (green), good (blue) or marginal (red). The light blue lollipops are predicted polyadenylation signals and the purple squares are CpG islands. Yellow and red squares indicate promotors and the vertical orange bars denote simple sequence repeats. Bottom panel: Graphical display of the powblast output using Musk. The black bar at the top represents the consensus sequence with the grey areas indicating repeat regions. The lines under the black bar indicate homology between the consensus sequence and an entry in the public database.

Genome Organization - putative human neurexin II alpha gene region

Exon number	position	size of exon	size of intron
01	2522-3398	876	8857 (gap)
02	(absent--in sequencing gap)		
03	12255-12287	32	648
03a	12935-13126	191	3621
03b	16747-16806	59	181
04	16987-17058	71	4458
05	21516-21817	301	8667
06	30484-30528	44	9848
07	40376-40538	162	755
08	41293-41731	438	81
08a	41812-41975	163	5027
08b	47002-47136	134	2077
08c	49213-49252	386	101
09	49353-49739	386	218
10	49957-50160	203	6743 (gap)
11	(absent--in sequencing gap)		
12	56903-57031	128	399
13	57430-57812	382	617
14	58429-58619	190	1541
15	60160-60333	173	---

Table 21: Genome organization of the putative human neurexin II alpha gene region from chromosome 11q13. This region has homology to the first 15 exons of the rat neurexin II alpha gene for a total of 3.4 kb, which span roughly 58 kb of genomic sequence.

Splice Junctions-putative human homologue of the rat neurexin II alpha gene region

Left Exon	Left splice junction	Right splice junction	Right Exon
01	TGCAGCGAAGgtgagccccc	(sequence gap)	02
02	(sequence gap)	tccccttccAAGGTCCGGC	03
03	AACAGCGAAGgtatggtttg	ctcgcaccagGTCGCTGTGC	03a
03a	GACCACGCGGgtaatgcctg	tcatggtgagGCCTCTGTTC	03b
03b	TTCCTTAGAGgtcttgaagg	tgcttctccaGTAGGGTCCT	04
04	CCAACAAAAGgtcagtgacc	cggcatccagGCAAGGAGGA	05
05	CCTGCGCCAGgtaggaggag	tccggtgaagCACGCAGGGA	06
06	GCATTATCTGgtagatatca	ggctggacagGTGACCATCT	07
07	CCTCAAGGACgtgagtaggg	ctccctgcagGTGGTCTATA	08
08	GGGCGAAAAGgtcaggaggc	ctctcttcagCGGGTCTCTT	08a
08a	CTTGCTGCATgtgagactct	gcccggccaaGCACCCTCTA	08b
08b	TGGCTGAGGGgtggagagtg	tccaccccagGACTGCTGCT	08c
08c	CGTGCTGCAGggcacaaggt	tgattcccagGCTCCATCTC	09
09	TGTGAGAGAGgtgaggctgg	ctgtgggcagAGGCCACGGT	10
10	GTCAACCTCGgtaaccaccc	(sequence gap)	11
11	(sequence gap)	tccccccttgGTGCCGCAGG	12
12	ACTGTGGAGGgtaggtggtc	tcctgagcagGACAGATGGC	13
13	TGGTCAAGGGgtgaggggct	tgtggggcagGTACATCCAC	14
14	GATCTCAAAGgtggggctgg	ccctttgcagGGGAGTTGTA	15
15	GGCTGTGATGgtgagtggag	----------	
consensus:	CAGgtaagt A t	tncagG c	

Table 22: Splice junctions for putative human homologue of the rat neurexin II alpha gene from human chromosome 11q13.

```
cccagcggctgggtccatggcattgtgaatcgacgagggagccccgctccgcgcgagggg      60
cgcccggcccctccccgccccatgcgcccgcggctctgaagcctgagcggggccgggggg     120
                                  M  A  L  G  S  R  W  R  P  P  P
ccgggcggcgtcggggccgccggaggcatggctctcgggagtcggtggcgaccgccaccc     180
 Q  L  P  P  L  L  L  L  L  A  L  V  A  G  V  R  G  L  E  F
cagttgccgccgctgctgttgttgctggcgctggtggcaggcgtccgtggcttggagttt     240
 G  G  G  P  G  Q  W  A  R  Y  A  R  W  A  G  A  A  S  T  G
ggcggcggccccgggcagtgggctcgctacgcgcgctgggcaggcgcggcgagcaccggc     300
 E  L  S  F  S  L  R  T  N  A  T  R  A  L  L  L  Y  L  D  D
gagctcagcttcagcctgcgcaccaacgccacgcgcgcgctgctgctctacctggacgac     360
 G  G  D  C  D  F  L  E  L  L  L  V  D  G  R  L  R  L  R  F
ggcggcgactgcgacttcctggagctgctgctggtggacgggcgcctgcggctgcgcttc     420
 T  L  S  C  A  E  P  A  T  L  Q  L  D  T  P  V  A  D  D  R
acgctgtcttgcgccgagcccgccacgctgcagttggacacgccggtggccgacgaccgc     480
 W  H  M  V  L  L  T  R  D  A  R  R  T  A  L  A  V  D  G  E
tggcacatggtgctgctgacccgcgacgcgcggcgcacggcgctggcggtggacggcgaa     540
 A  R  A  A  E  V  R  S  K  R  R  E  M  Q  V  A  S  D  L  F
gcccgtgcggccgaggtgcgctcaaagcggcgcgagatgcaggtggccagcgacctgttc     600
 V  G  G  I  P  P  D  V  R  L  S  A  L  T  L  S  T  V  K  Y
gtgggcggcatcccacccgacgtgcgcctgtctgcgctcacgctcagcaccgtcaagtac     660
 E  P  P  F  R  G  L  L  A  N  L  K  L  G  E  R  P  P  A  L
gagccgcctttccgcggcctcctggccaacctgaagctgggcgagcggccgccggcgctg     720
 L  G  S  Q  G  L  R  G  A  A  A  D  P  L  C  A  P  A  R  N
ctgggtagccagggtctgcgcggtgcggccgccgacccccctgtgcgcgcccgcacgcaat     780
 P  C  A  N  G  G  L  C  T  V  L  A  P  G  E  V  G  C  D  C
ccctgcgccaacggcggcctctgcaccgtgctagcccccggcgaggtgggctgcgactgc     840
                              exon1|exon2       exon2|exon3
 S  H  T  G  F  G  G  K  F  C  S  E  E  E  H  P  M  E  G  P
agccacactggcttcggcggcaagttctgcagtgaagaggaacaccccatggaaggtccg     900
                         exon3|exon3a
 A  H  L  T  L  N  S  E  V  A  V  P  R  V  R  S  H  H  R  S
gctcacctgacgttaaacagcgaagtcgctgtgccccgagtgaggtcgcaccatcgctcc     960
 E  R  R  A  Y  R  S  L  R  L  S  A  T  Q  S  V  I  E  A  A
gagcggcgtgcgtaccggagcctacgcctgtccgcgacacagtcggtcatcgaggcagcg    1020
 G  A  Q  S  Q  S  D  R  D  A  A  V  S  L  A  A  Q  H  A  R
ggggcgcagtcccaatctgatcgtgacgccgcagttagcctggcggctcagcacgctcga    1080
                   exon3a|exon3b
 V  S  C  E  T  T  R  A  S  V  P  L  P  S  L  P  T  P  G  K
gtcagctgcgagaccacgcgggcctctgttcctctgcccagtctccccacccccaggcaaa    1140
 E  W  K  S  V  I  M  N  L  M  G  S  V  G  L  I  A  L  G  L
gagtggaagtcggtgatcatgaatctgatggggtctgtggggctgatcgccctcggtttg    1200
                      exon3b|exon4
 S  S  G  S  F  L  R  V  G  S  L  L  F  S  E  G  G  A  G  R
tcaagcggctccttccttagagtagggtccttactgttctccgaggggggggccgggaga    1260
                                  exon4|exon5
 G  G  A  G  N  V  H  Q  P  T  K  G  K  E  E  F  V  A  T  F
ggaggagccggcaatgtgcaccagccaacaaaaggcaaggaggaatttgtggcaaccttc    1320
 K  G  N  E  S  F  C  Y  D  L  S  H  N  P  I  Q  S  S  T  D
aagggcaatgagtccttctgctacgacctgtcacacaacccgatccagagcagcactgat    1380
 E  I  T  L  A  F  R  T  L  Q  R  N  G  L  M  L  H  T  G  K
gagatcacactggccttccgcaccctgcagcgcaacgggctgatgctgcacacggggaag    1440
 S  A  N  Y  V  N  L  S  L  K  S  G  A  V  W  L  I  I  N  L
tcggctaactacgtcaacctgtccctcaagtccggggctgtctggctgatcatcaaccta    1500
 G  S  G  A  F  E  A  L  V  E  P  V  N  G  K  F  N  D  N  A
ggctcaggtgccttcgaggccctcgtggaacccgtcaatggcaagttcaacgacaacgcc    1560
                                     exon5|exon6
 W  H  D  V  R  V  T  R  N  L  R  Q  H  A  G  I  G  H  A  M
tggcacgacgtccgggttacccgaaacctgcgccagcacgcagggattggacacgctatg    1620
```

exon6|exon7
V N K L H Y L V T I S V D G I L T T T G
gtaaacaaactgcattatctggtgaccatctcggtggacgggatcctgaccaccacaggc 1680
Y T Q E D Y T M L G S D D F F Y I G G S
tacacgcaggaggattacaccatgctgggctctgatgacttcttctacattgggggcagc 1740
P N T A D L P G S P V S N N F M G C L K
cccaacacagccgacctgcctggctcccctgtcagcaacaacttcatgggctgtctcaag 1800
exon7|exon8
D V V Y K N N D F K L E L S R L A K E G
gacgtggtctacaagaataatgacttcaagctggagctatcccgactggctaaggaaggg 1860
D P K M K L Q G D L S F R C E D V A A L
gacccgaagatgaagctgcaggggatttgtctttccgttgtgaggacgtggctgccttg 1920
D P V T F E S P E A F V A L P R W S A K
gaccctgtgaccttcgagagtcctgaggcctttgtcgcactgccccgctggagcgccaag 1980
R T G S I S L D F R T T E P N G L L L F
cgcactggttctatctccctggacttcagaaccacggagcccaatgggttgctgctcttc 2040
S Q G R R A G A G V G S H S S S Q R A D
agccagggccggcgggctggggccggggtaggcagccacagttccagccagagggccgac 2100
Y F A M E L L D G Y L Y L L L D M G S G
tactttgccatggagctattggatggctacctctatcttctgctggacatgggctctggg 2160
G I K L R A S S R K V N D G E W C H V D
ggcatcaagctgcgggcgtccagccgcaaggtcaacgacggcgaatggtgccacgtggac 2220
exon8|exon8a
F Q R D G R K A G L F Y T I S Y I L P C
ttccagagggatgggcgcaaagcgggtctcttctacaccatttcctatattcttccctgc 2280
R F S S T P L T S S Q L S F L S S D I G
cgcttctcctccacccccctcacgtcttcccaattgtcctttctgtcatcggacattggc 2340
R K S T A F G I S Q T W I T P Q L P Y L
agaaagagcacagcatttggcatcagtcagacctggattacacctcagctcccctacttg 2400
exon8a|exon8b
L H A P S T S E E L F Y L K I L C L P P
ctgcatgcaccctctacttctgaagaactcttctacctaaaaattctgtgcctacctcct 2460
F S S L L S C P A G P G A N L L V H L A
ttctcctcactgctgtcctgccccgcaggccctggagccaatctgcttgtgcaccttgct 2520
exon8b|exon8c exon8c|exon9
G R R V G S W L R G L L L V L L S V L Q
ggccgcagagtgggctcttggctgaggggactgctgctcgtcctcctgagcgtgctgcag 2580
G S I S V N S R S T P F L A T G E S E V
ggctccatctctgtgaacagccgcagtacaccattcttggccacaggagagagtgaggtc 2640
L D L E S E L Y L G G L P E G G R V D L
ctggacctggagagtgagctgtacctgggcggtctccctgaggggggccgagtggacctg 2700
P L P P E V W T A A L R A G Y V G C V R
ccattgcctcctgaggtgtggacagctgctctccgggctggctacgtgggctgcgtgcga 2760
D L F I D G R S R D L R G L A E A Q G A
gacctcttcatagatggacggagtcgagatctccggggcctggctgaggcacagggagct 2820
V G V A P F C S R E T L K Q C A S A P C
gtgggcgttgcacctttctgctctcgggagaccctgaagcagtgtgcatcggcccctgt 2880
R N G G I C R E G W N R F V C D C I G T
cgtaatggtggcatctgtcgagagggctggaaccggttcgtctgtgattgcatcgggacc 2940
exon9|exon10
G F L G R V C E R E A T V L S Y D G S M
ggctttctgggtcgggtctgcgagagagaggccacagtcttaagctatgatggctccatg 3000
Y M K I M L P N A M H T E A E D V S L R
tacatgaagatcatgctgcccaatgccatgcacacggaagcagaggatgtgtccctgcga 3060
F M S Q R A Y G L M M A T T S R E S A D
ttcatgtcccagagggcttatggactcatgatggccaccacctctagggagtcggccgac 3120
exon10|exon11
T L R L E L D G G Q M K L T V N L D C L

```
actctgcgtctcgagctggacgggggcagatgaagctcacagtcaaccctcgactgcctg    3180
          exon11|exon12
 R  V  G  C  A  P  S  A  A  A  K  G  P  E  T  L  F  A  G  H
cgcgtcggctgcgcacccagtgctgcagctaaaggccccgagaccctctttgcggggcac    3240
 K  L  N  D  N  E  W  H  T  L  R  V  V  R  R  G  K  S  L  Q
aagctcaacgacaatgagtggcacacgctgagggtggtacggcgtggcaagagcctgcag    3300
             exon12|exon13
 L  S  V  D  N  V  T  V  E  G  Q  M  A  G  A  H  T  R  L  E
ctgtctgtggacaacgtgactgtggagggacagatggcaggagcccacacgcggctggag    3360
 F  H  N  I  E  T  G  I  M  T  E  R  R  F  I  S  V  V  P  S
ttccacaacatcgagacaggcatcatgacagagcgacggtttatctctgtggtgccctcc    3420
 N  F  I  G  H  L  S  G  L  V  F  N  G  Q  P  Y  M  D  Q  C
aacttcatcggccacctgagcgggctggtgttcaatggtcaaccctacatggaccagtgc    3480
 K  D  G  D  I  T  Y  C  E  L  N  A  R  F  G  L  R  A  I  V
aaagatggagacatcacctactgtgagcttaatgcccgctttgggctgcgtgccatcgtg    3540
 A  D  P  V  T  F  K  S  R  S  S  Y  L  A  L  A  T  L  Q  A
gctgatcctgttaccttcaagagtcgcagtagctacctggcgttggccacgctccaagcc    3600
 Y  A  S  M  H  L  F  F  Q  F  K  T  T  A  P  D  G  L  L  L
tatgcctccatgcacctcttcttccagttcaagaccacagcccctgatggacttctgctg    3660
                          exon13|exon14
 F  N  S  G  N  G  N  D  F  I  V  I  E  L  V  K  G  Y  I  H
ttcaactcaggcaacggcaatgacttcattgtcatcgagctggtcaaggggtacatccac    3720
 Y  V  F  D  L  G  N  S  P  S  L  M  K  G  N  S  D  K  P  V
tacgtgtttgacctgggaaatagcccgtccttgatgaagggaaactcagacaaaccagtt    3780
 N  D  N  Q  W  H  N  V  V  V  S  R  D  P  G  N  V  H  T  L
aatgacaaccagtggcacaacgtggtggtgtccagggacccaggcaacgtgcacacgctg    3840
 K  I  D  S  R  T  V  T  Q  H  S  N  G  A  R  N  L  D  L  K
aagattgactcccgcacggtcacgcagcattccaacggtgcccgaaatctggatctcaaa    3900
exon14|exon15
 G  E  L  Y  I  G  G  L  S  K  N  M  F  S  N  L  P  K  L  V
ggggagttgtacatcggtggcctgagcaagaacatgttcagcaacctgccaaagctggtg    3960
 A  S  R  D  G  F  Q  G  C  L  A  S  V  D  L  N  G  R  L  P
gcctctcgggatggctttcagggctgcctggcttctgtggacctcaacgggcgcctccca    4020
 D  L  I  A  D  A  L  H  R  I  G  Q  V  E  R  G  C  D
gacctcatcgcagacgccctgcaccgaatcgggcaggtggagaggggctgtgatg         4075
```

Figure 34: cDNA and amino acid translation for the putative human neurexin gene.

(taking into account that the 3' end of the gene is absent from the sequencing data) and translated (see figure 34). The translation of this partial putative human gene results in a partial gene encoding for a protein of 1309 amino acids, assuming the start codon and promotor region used in humans are the similar to those in the rat gene. The Xgrail analysis also reveals the presence of three exons (one excellent, one good and one marginal) in the reverse direction, although these three exons do not have any significant powblast database homology.

This region of the genome contains several repeat elements, including the ALU, SVA, L1 and MIR families. There are 19 regions of homology with ALU sequences (and the same 19 regions are homologous with SVA repeats in the opposite direction), 2 regions of MIR homology and two areas of significant L1 repeat family homology. Again, this region has homology to a human EST (N20878) which contains an ALU repeat element. A graphical view of this EST region as displayed by Musk is seen in figure 35, and a dotter alignment of the EST with the ALU sequence is shown in figure 36 to demonstrate the presence of both unique and repeat sequences in this EST.

Although this region is not very rich in homology to the large families of repeat elements, it is very rich in simple sequence repeats (table 23). Several areas which are marked by Xgrail as CpG islands are also poly GA regions. Of significant interest is the presence of nine areas with long poly C tracts which vary in length from 16 nucleotides to 175 nucleotides, and the six areas of poly T tracts which are 13 to 21 nucleotides in length.

c. Kappa

Xgrail analysis of a 38,355 bp region of b137c7 revealed the presence of a putative human gene consisting of 23 exons. This region of chromosome 11q13 contains many Xgrail predicted excellent exons, as well as regions with database homology that include several human ESTs (Genbank accession numbers H01451, AA219223, T48828 and R52045), one mouse EST (Genbank accession number W16079) and one STS

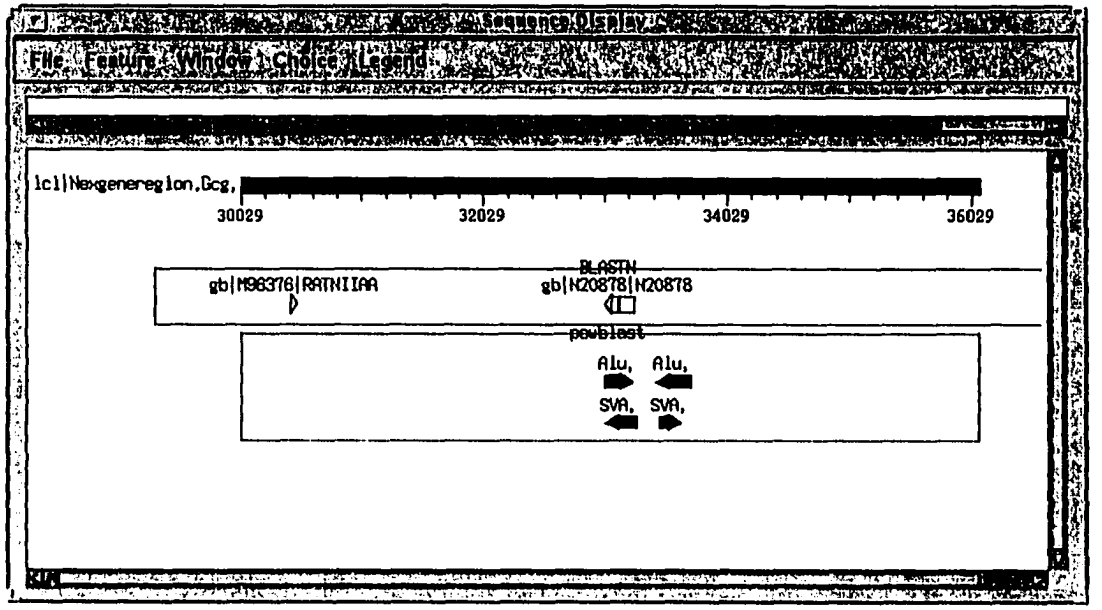


Figure 35: Musk display of powblast database homology between human EST N20878 and the ALU repeat element.

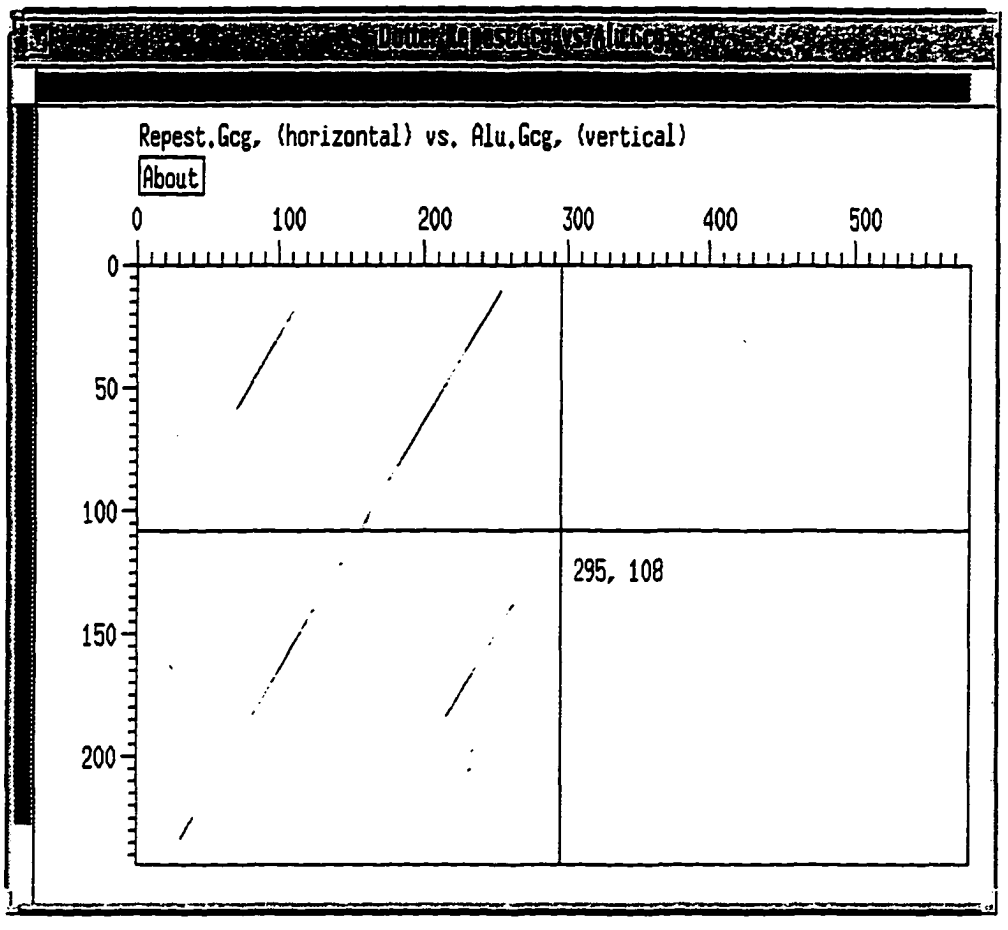
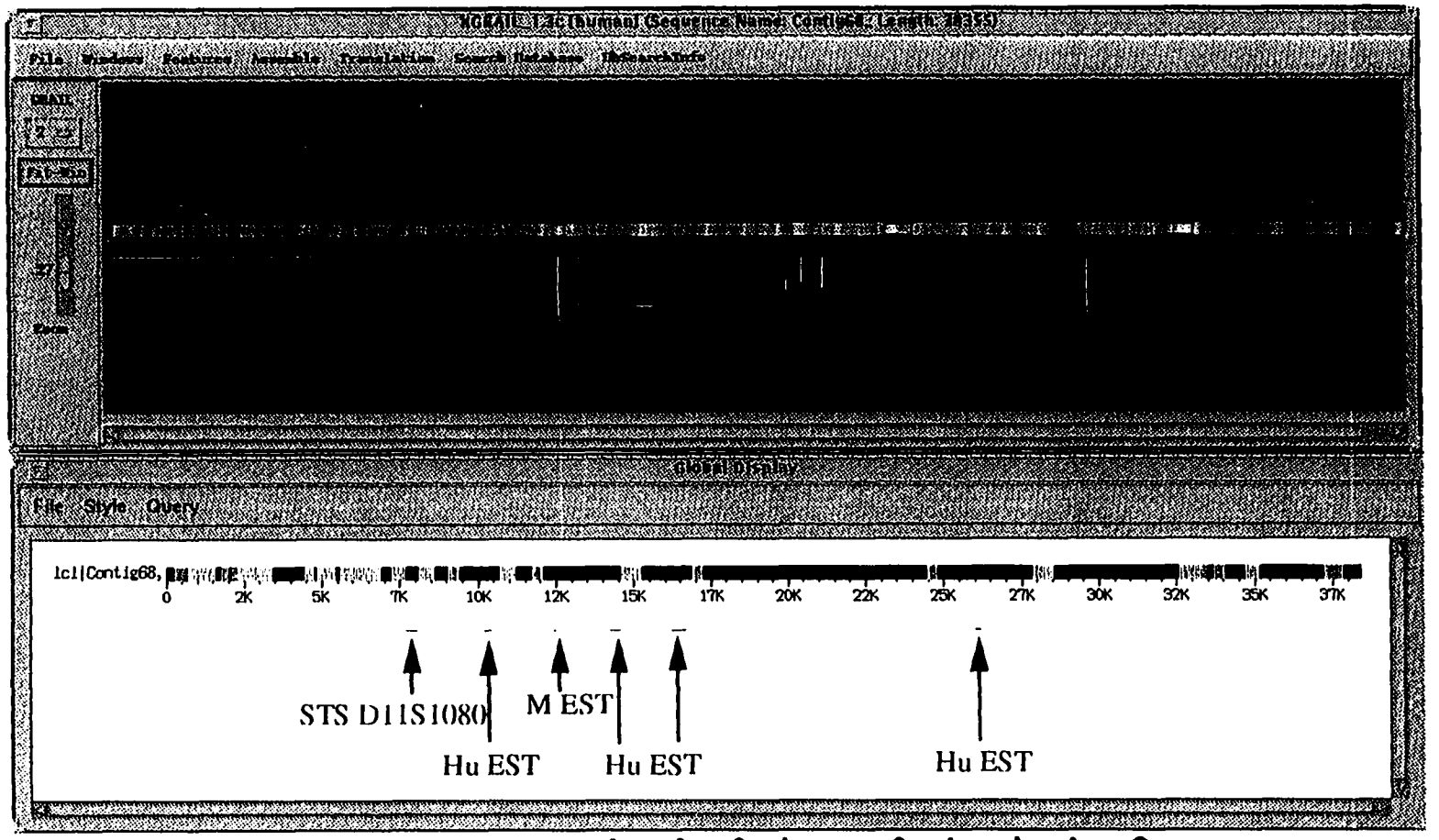


Figure 36: Dotter alignment of Human EST N20878 homology with the ALU repeat element.

Putative human and rat neurexin II alpha gene region - Simple Sequence Repeats

position	Repeat
2617-2732	poly GC
6445-6487	A_{17}
7157-7184	T_{21}
8962-8985	T_{14}
10167-10326	$(TA)_{78}$
10462-10497	poly GA
10965-10996	C_{27}
11001-11115	C_{114}
11374-11478	poly GA
11716-11853	poly GA
11828-11942	poly GA
12137-12259	poly CT
13594-13733	C_{139}
13726-13981	poly GA
14210-14834	poly GAC
23813-23843	$T_3(CT_4)_3T_{12}$
26339-26369	$T_{13}ACT_{15}$
30333-30475	C_{142}
33463-33481	T_{16}
33811-33860	$A_4(CA)_{14}A_7$
36188-36395	poly GA
36344-36650	poly GA
36440-36591	poly GA
36655-36831	C_{175}
37312-37453	poly GA
43677-43706	$(CT_3)_4T_9$
44486-44500	T_{13}
46172-46199	$(T_3A)_4T_{13}$
46870-46900	T_{20}
47606-47753	poly GA
48817-48996	C_{159}
49001-49043	C_{16}
49112-49135	C_{21}
51863-51996	C_{133}

Table 23: Simple sequence repeats present in the putative human and rat neurexin II alpha gene region from human chromosome 11q13 as reported by Xgrail.



Putative gene "kappa"

Figure 37: Top panel: Graphical result of Xgrail gene identification software for the putative gene "kappa" from human chromosome 11q13. The peaks present are Xgrail predicted exons with a quality rating of excellent (green), good (blue) or marginal (red). The light blue lollipops are polyadenylation signals and the purple squares are CpG islands. Yellow and red squares indicate promotors and the vertical orange bars denote simple sequence repeats. Bottom panel: Graphical display of the powblast output using Musk. The black bar at the top of the panel represents the consensus sequence with the grey areas indicating repeat regions. The lines under the black bar indicate database homology between the consensus sequence and entries in the public database.

(D11S1080; Genbank accession number G28839) revealed by powblast analysis (see figure 37). However, except for R52045, none of the other ESTs correlate with Xgrail predicted exons. This EST homology matches in the same position as exon 7 in the predicted gene. This gene region also was inspected by collaborators in Dr. Collins' laboratory and was given the name "kappa."

The genomic organization, i.e. putative exons, their size and the size of the intronic regions, of the kappa gene is detailed in table 24. In further support for the notion that kappa represents a likely gene, Xgrail predicts a promotor region upstream from exon 1 which encodes a methionine start codon that immediately follows a consensus Kosak sequence. In addition, the first occurrence of a stop codon is at the end of exon 23 and it is followed by four polyadenylation signals in the downstream region. The area including exon 1 also contains a CpG island. The intron/exon junctions of the predicted protein adhere quite well to the consensus splice junctions (see table 25) and therefore, taken together, the above evidence strongly suggests that a gene occurs in this region of chromosome 11.

The kappa gene has limited repeated elements, although the genomic sequences on either side of this putative gene are very dense in repeat elements. The entire 38,355 bp region contains 31 ALU, three MER, one MIR, and 15 SVA repeat elements (SVA usually occurring in the same position as an ALU, but in the reverse orientation). Simple sequence repeats also are found throughout this region (see table 26), with 7 of 27 poly A tracts greater than 12 nucleotides long, 10 of 27 poly T tracts greater than 11 nucleotides long, and one region of poly C which stretches for 125 nucleotides.

d. Nu

A putative gene named "Nu" and containing 12 Xgrail-predicted exons, has been identified in a 19,699 base pair region of b137c7. This region also contains several sequences with significant database homology, including human ESTs (Genbank accession numbers T15445, T78563, T96795, AA035643 and T09103), a rat EST

Genome Organization - putative gene "kappa"

Exon number	position	size of exon	size of intron
01	31392-31551	159	2506
02	28773-28886	113	58
03	28632-28715	83	1001
04	27536-27631	95	286
05	27042-27250	208	489
06	26348-26553	205	133
07	26015-26216	200	131
08	25636-25884	248	491
09	25066-25145	79	2822
10	22052-22244	192	685
11	21295-21367	72	194
12	20982-21101	119	71
13	20797-20911	114	272
14	20377-20525	148	75
15	20214-20302	361	430
16	19686-19784	98	2017
17	17453-17669	216	1436
18	15460-16017	557	1726
19	13637-13734	97	75
20	13478-13562	84	129
21	13159-13349	190	41
22	13001-13118	117	256
23	12622-12745	123	--
ATGstart	31553-31555		
UAAstop	12622-12624		
polyA	9335-9340		
	9013-9018		
	8629-8634		
	7327-7332		

Table 24: Genome organization of the putative gene "kappa" from human chromosome 11q13. The candidate gene contains 23 exons for a total of 3.9 kb, which span roughly 9 kb of genomic sequence. The start codon is located within exon 1, the stop codon at the end of exon 23 and four polyadenylation signals follow downstream of the stop codon. Xgrail also predicted a large promoter region in the upstream region from this putative gene at 33231-33709.

Splice Junctions-putative gene "kappa"

Left Exon	Left splice junction	Right splice junction	Right Exon
01	GCGGCGAGGTAccgacgccg	gaggggggtgGATGGGGTCC	02
02	TTCCCCGACCgactcccctg	ggtgtgactgGAGTCGGGAG	03
03	GGTGCCACTGgattccggtg	cataggagtgGTCCATGAGG	04
04	TGTCCCGACAgacgtcgtcc	aggggtcgtgTCGACCGTCC	05
05	AGTGTTCCATgacctccacc	acccagagtgGGTGGAAGGG	06
06	GAACTGTAGGgacccccta	cacaggtctgGAGCACCAAA	07
07	TACTTAGGTGgacacccacc	acggtgagtgACCTACCAAA	08
08	CGTCCACCAGgacacccagc	gatcggagtgGACGGACTCC	09
09	CGTCCCAGGGgacgtctccc	gggatgggtgGGACCAGTGA	10
10	GCCGGTCGAGgactcctccg	gggacgagtgGGCGAGGACG	11
11	TGGTGAGCGGgacccctcct	gggctgggtgGGGACCGTGG	12
12	GACGAAGATCgaccccggcc	accccgagtgGTTACCGGCC	13
13	TAGTAAGTGGgacttaccct	gatcggagtgGCGAGTCGGA	14
14	TGACCACCAGgaccaccggt	tcggtgagtgGGTGACCGGA	15
15	GACGTCTGCCgacccctctt	gacaggagtgGGGGTACCGC	16
16	GTCTTAGGAAgacccccacc	gggtcgagtgGCTCCTACCG	17
17	TCCAACAGTGgacacccgtt	agcaccggtgCTCTGAACGC	18
18	ACCTTCTACAgacgtcccaa	cctgggagtgGAAGAACTCG	19
19	CGGCGTATGGgacgcccgta	tgagggagtgGACGGTCGAC	20
20	TCCCGGCGTGgactccgtca	cttcgaggtgGGAGGGAAGC	21
21	AGCAGGAAGAgacaccctct	ctcctgattgGCCCCTGAAC	22
22	GTCGTAGAGGgacacccgcc	cgacgggaatGACCCCAGAC	23
23	GGAAGCCTCGgacaacccac	---------	
consensus:	CAGgtaagt A t	tncagG c	

Table 25: Splice junctions for the putative gene "kappa" from human chromosome 11q13.

Putative gene "Kappa" - Simple Sequence Repeats

position	Repeat
253-285	T_{10}
576-624	$T_6(T_2G)_6(T_3G)_5$
1206-1265	$T_5GT_3GT_5AT_5CTAT_7AT_5AT_9$
1947-1977	A_{19}
2097-2113	T_{13}
2977-2991	$A_{12}T_2A$
4411-4458	$T_9GT_4GT_6GT_{10}G$
4934-4959	A_{26}
5567-5611	T_{40}
5888-5910	T_{17}
6053-6076	$TA_2T_{10}ATGT_5AGT$
6832-6860	A_{13}
7130-7169	$GT_3(T_4G)_3T_5GT_2G(T_3G)_3$
8112-8131	T_{11}
8419-8437	$T_7C_2T_9$
9010-9052	$T_3C_2(T_4C)_2T_{16}$
10691-10713	$C_2T_{10}C_7T$
10862-10888	$T_2A_2(T_5G)_2T_{11}G$
12019-12046	A_{19}
17166-17190	$(A_7G)_2A_{13}G$
22131-22240	poly CT
22656-22673	G_{18}
23567-23700	C_{125}
28435-28475	$A_{15}T_3A_6T_2A_3TA_3$
31327-31428	poly GC (CpG)
32573-32590	T_{18}
37836-37860	G_{15}

Table 26: Simple sequence repeats present in the putative gene "Kappa" region from human chromosome 11q13 as reported by Xgrail.

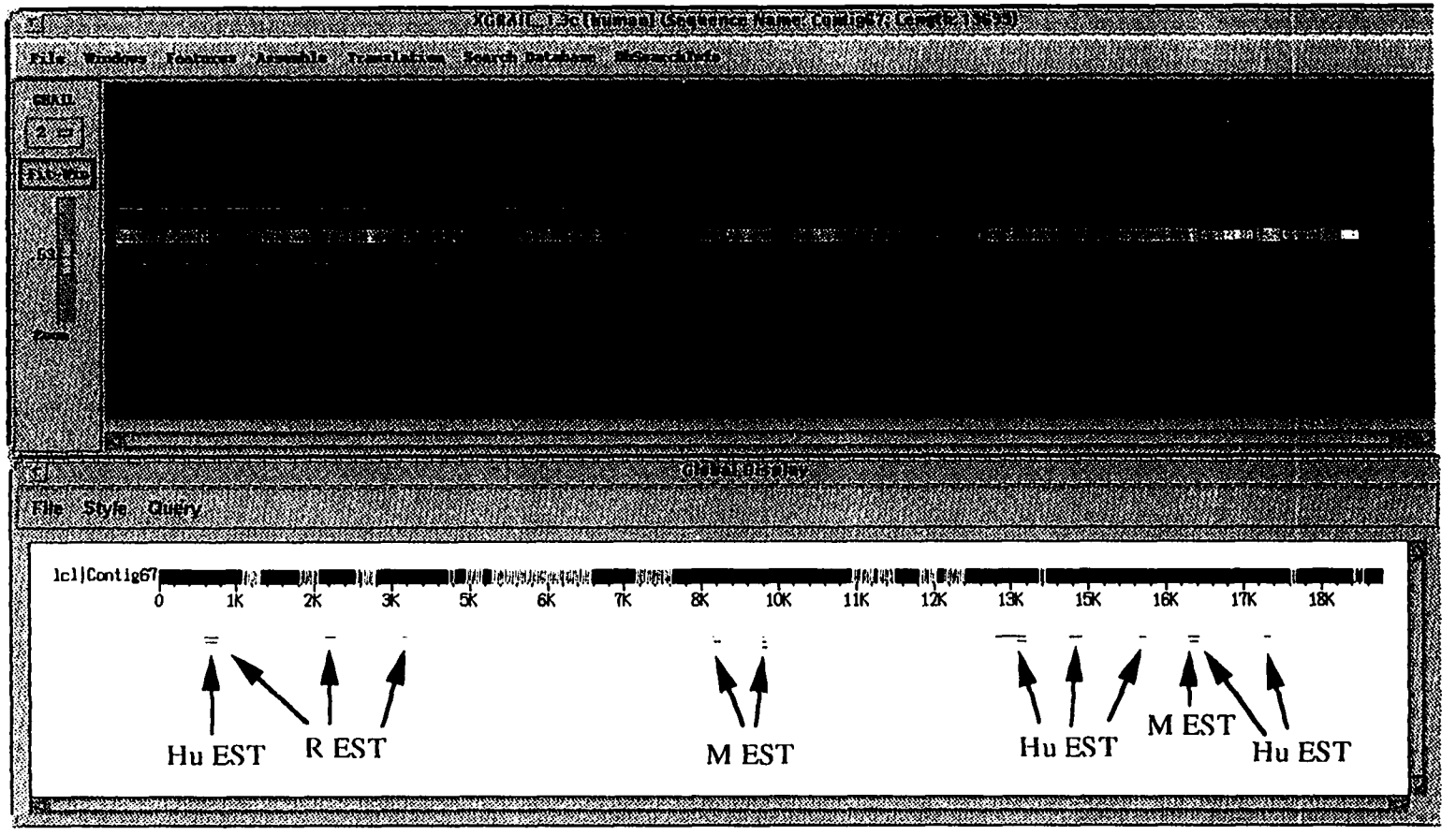
(Genbank accession numbers C06861) and three mouse ESTs (Genbank accession numbers AA04893, W70955, and W71787). A comparison of both Xgrail and powblast results indicates that the human and mouse EST database homologies correlate well with the Xgrail predicted exons in the first portion of the putative gene (see figure 38).

Closer analysis of "Nu" indicates an ATG start codon located at 16678-16760, and an UGA stop codon following exon 12 at position 3559-3561. The positions, the sizes of the exons and the sizes of introns are detailed in table 27. The 12 exons present in this region all were predicted by Xgrail with 11 of the 12 assigned an excellent quality rating (the final exon has a good quality rating). All the intron/exon boundaries conformed to the consensus splice junction sequences, as seen in table 28.

This region also contains several repeat elements, including ALU, SVA, MIR, and MER repeats. There are 17 ALU repeats (11 of which have an accompanying SVA repeat in the same position, but opposite direction), three MIR repeats, one full length L1 repeat, one MER repeat, and several areas with simple sequence repeats (see table 29). Finally, a region located at the end of this 19,699 base pair contig is highly AT rich, and was scored by Xgrail as four large poly AT simple sequence repeats, there are several regions containing poly A tracts 12 to 24 nucleotides long, and other regions that are rich in GA repeats.

B. b18h3 analysis

Analysis of approximately 240,000 bases from b18h3 in the MEN-1 gene region from human chromosome 11q13 (Genbank accession number AA000353) reveals the presence of six genes. The first is the DNA polymerase a gene which previously was mapped to 11q13 and whose cDNA has been sequenced. The second is a putative requiem gene located by it's homology to the mouse requiem gene, a zinc finger gene essential for apoptosis in myeloid cells. The remaining four genes have not previously been described and these putative genes have been named Zeta, Eta, Theta, and Epsilon/Beta.



Putative gene "Nu"

Figure 38: Top panel: Graphical result of Xgrail gene identification software for the putative gene "Nu" region from human chromosome 11q13. The peaks present are Xgrail predicted exons with a quality rating of excellent (green), good (blue) or marginal (red). The light blue lollipops are polyadenylation signals and the purple squares are CpG islands. Yellow and red squares indicate promotors and the vertical orange bars denote simple sequence repeats. Bottom panel: Graphical display of the powblast output using Musk. The black bar at the top of the panel represents the consensus sequence with the grey areas indicating repeat regions. The lines under the black bar indicate database homology between the consensus sequence and entries in the public database.

Genome Organization - Putative gene "Nu"

Exon number	position	size of exon	size of intron
01	16617-16760	143	682
02	15764-15935	171	862
03	14771-14902	131	785
04	13835-13986	151	204
05	13458-13631	173	160
06	13182-13298	116	2327
07	10574-10855	281	700
08	9725-9874	149	237
09	9366-9488	122	315
10	8936-9051	115	4919
11	3876-4017	141	255
12	3585-3621	36	---
ATGstart	16678-16680		
UGAstop	3559-3561		

Table 27: Genome organization of the putative gene "Nu" from human chromosome 11q13. The candidate gene contains 12 exons (the first being untranslated) for a total of 1.7 kb, which span roughly 13 kb of genomic sequence. The start codon is located within exon 1, and the stop codon following exon 12.

Splice Junctions-Putative gene "Nu"

Left Exon	Left splice junction	Right splice junction	Right Exon
01	ACGGGTCCTAgacgttatct	cctgagagtgGTCCCGAAAC	02
02	CGTCGGTAGGgacacccctc	gggaggagtgGACCAAGACC	03
03	CCTTGTGTCTgaccccgtgc	tcctcgagtgGAGGTGGTAG	04
04	CCGACCAACCgataccttta	tgggggagtgGCTCCTGAAC	05
05	TTTCCAGCTGgatctcggtg	acgaggagtgGGCACTCGTC	06
06	GGAGGGTCTCgacccctgtt	agggtccgtgGAACTACCAG	07
07	CATCGTCGAAgaccttccgt	cccgtccgtgGAGCGGTGCA	08
08	TCAGGACTTTgacccccgta	gctcggcgtgGTCCTAGAAC	09
09	CATCCATCCCgacccccgt	ggggtgcgtgTGCGACAGAT	10
10	CTAGGTCATGgacctcctcg	gtaagacgtgGACCCGACCC	11
11	GGCCTCAGTAgactgaggct	tcggtgagtgGCTTCCGAAG	12
12	CTGGCTTGGGgacagggccc	----------	
consensus:	CAGgtaagt A t	tncagG c	

Table 26: Splice junctions for the putative gene "Nu" from human chromosome 11q13.

Putative Gene "Nu" - Simple Sequence Repeats

position	Repeat
24-165	poly GA
325-334	G_{11}
1332-1353	T_{21}
2231-2265	$T_3AT_5A_2T_{11}$
2415-2440	$T_{11}GTAT_5$
3132-3179	$G_4T_4G_2T_4G_5T_7GT_2T_{12}$
3526-3549	$(CA_3)_5$
5183-5223	$A_{24}GA_4$
6606-6625	A_{12}
6777-6799	$A_8GA_5GA_9$
6967-6993	$(A_6G)_3$
7196-7266	$(CA)_{23}$
7654-7701	$CT_3AT_3CT_2C_2T_9CT_{16}$
8137-8153	$A_5(TA_2)_4$
11501-11584	$A_{16}GA_3GA_5(GA_3)_2(GA_4)_2$
12780-12815	A_{15}
12964-12982	A_{15}
13370-13471	poly GC
17039-17157	poly CT
18694-18889	poly AT
19098-19262	poly AT
19306-19596	poly AT
19575-19694	poly AT

Table 29: Simple sequence repeats present in the putative gene "Nu" region from human chromosome 11q13 as reported by Xgrail.

1. DNA polymerase α

Analysis of a 101,449 bp region of b18h3 revealed the presence of a known gene, DNA polymerase alpha (Genbank accession number L24559). This gene is composed of 18 exons which encompass appproximately 35 kb of genomic sequence, and code for a protein of 599 amino acids in length. The first exon encodes 76 bases of a 5' untranslated region, followed by a methionine start codon. The coding region concludes with a TGA stop codon at position 51838-51840, and a polyadenylation signal at 52169-52176, 329 nucleotides downstream from the stop codon, delineates the end of the 3'untranslated region. This gene does not have an SP1 binding site or Kosak sequence, but Xgrail predicts a strong CpG island upstream of exon 1. As mentioned above, this is typical in the promotor regions of housekeeping genes. The genomic organization including the size and position of introns and exons is displayed in table 30.

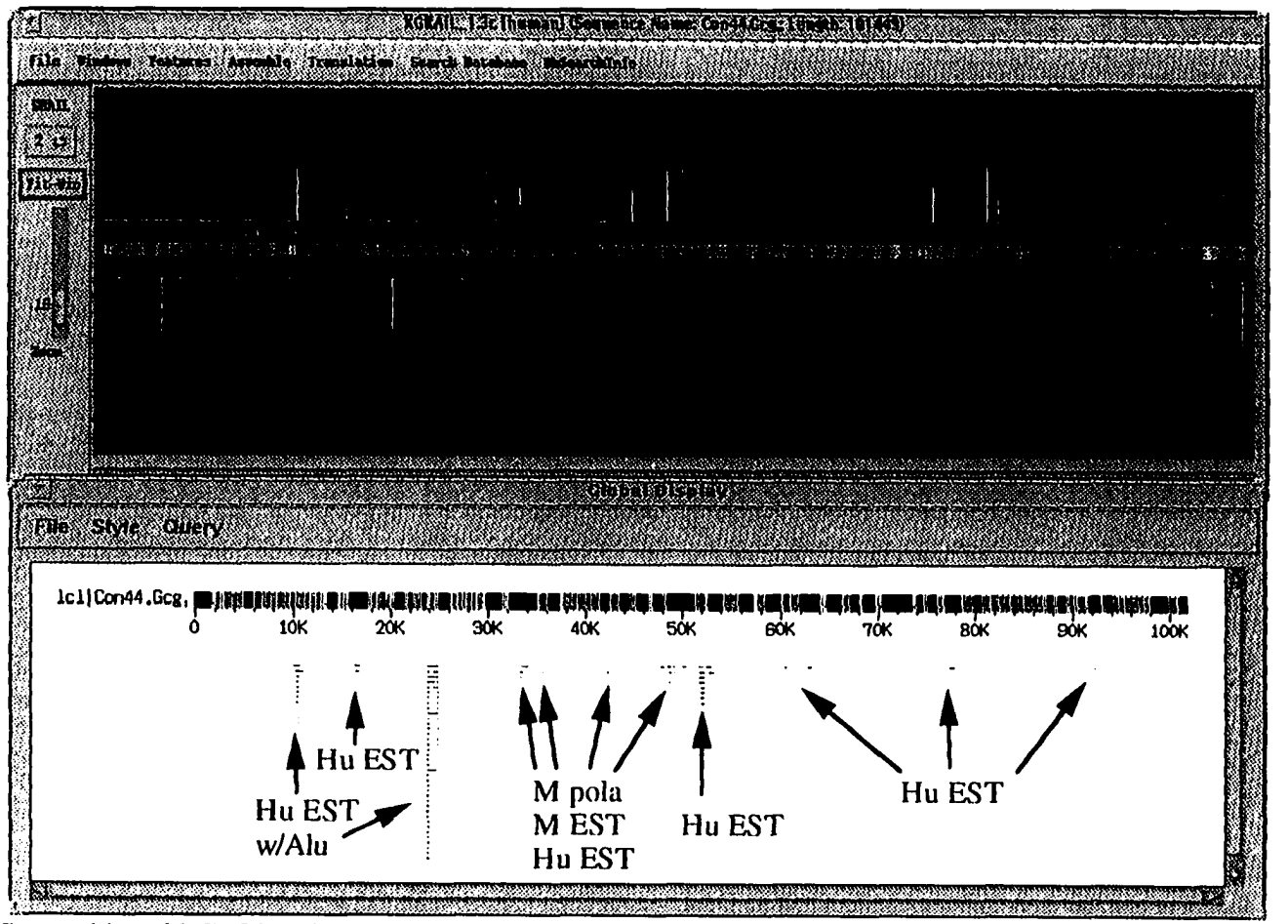
Xgrail and powblast analyses of this region are shown in figure 39. Xgrail accurately predicted 13 of the 18 exons, with exons 3,5,13,14 and 18 not being predicted, most likely because of their small size. However, all 18 intron/exon boundaries for this gene adhere to the consensus splice junctions (see table 31). The exons predicted by Xgrail correlate perfectly with the cDNA sequence and it's genomic organization will be a valuable asset for those interested in human DNA replication.

Along with the obvious homology to the mouse DNA polymerase alpha (Genbank accession number D13546), this region of 11q13 also has homology to numerous human ESTs from many different tissues. These ESTs include Genbank accession numbers T77743, AA214047, R99160, H90258, AA143403, AA159455, L78765, W26175, T63904, T63254, AA071113, AA159312, T81239, T28918, C02161, T54494, H954494, H95849, T54442, W23797, and N94590, and are derived from pancreas, lung, neuroepithelium, liver, spleen, retina, placenta, fetal liver/spleen, breast, fetal heart, breast and neuron indicating the ubiquitous nature of the DNA polymerase α gene.

Genome Organization - DNA polymerase α gene

Exon number	position	size of exon	size of intron
01	16807-16942	135	4211
02	21153-21277	124	769
03	22046-22139	93	1104
04	23243-23301	58	716
05	30461-30569	108	273
06	33299-33494	195	613
07	34107-34198	91	1363
08	35561-35720	159	518
09	36238-36304	66	744
10	37048-37092	44	5193
11	42285-42411	126	1209
12	43620-43659	39	709
13	44368-44443	75	4279
14	48722-48832	110	283
15	49115-49223	108	886
16	50109-50170	61	263
17	50433-50562	129	1127
18	51689-52158	469	--
ATGstart	16864-16866		
UGAstop	51838-51840		
polyA	52169-52176		

Table 30: Genome organization of the DNA polymerase α gene region from human chromosome 11q13. The gene contains 18 exons for a total of 2.2 kb, which span roughly 35 kb of genomic sequence. The start codon is located within exon 1, the stop codon at the end of exon 18 and a polyadenylation signal 329 nucleotides downstream of the stop codon.



Strong hits with DNA pol α

DNA pol α gene

Figure 39: Top panel: Graphical result of Xgrail gene identification software for the DNA polymerase α gene region from human chromosome 11q13, also including the putative gene, "Zeta". The peaks present are Xgrail predicted exons with a quality rating of excellent (green), good (blue) or marginal (red). The light blue lollipops are polyadenylation signals and the purple squares are CpG islands. Yellow and red squares indicate promotors and the vertical orange bars denote simple sequence repeats. Bottom panel: Graphical display of the powblast output using Musk. The black bar at the top of the panel represents the consensus sequence with the grey areas indicating repeat regions. The lines under the black bar indicate database homology between the consensus sequence and entries in the public database.

Splice Junctions-DNA polymerase α gene

Left Exon	Left splice junction	Right splice junction	Right Exon
01	ATTGAGAAATgtgagtcccg	ctttctacagTGGTAGAGCT	02
02	TGAGCATGAGgtaagaacaa	tctttttcagTTTCTGAGCA	03
03	TTCAAGAGCTgtatccttttt	gtgttcttagAATTGAAGTG	04
04	ACCTTCAAAGgtaagtaaac	ttttccctagGGTTCTCAGA	05
05	TCTCTCCAAGgtatggatca	ttccaattagTGCTACTCCC	06
06	ATTCGAGAAGgtgattgttt	gggattgtagTTCTGACCTG	07
07	CCCAGCACAGgtaagagttg	tcctttggtaGGAGCCTGTC	08
08	TCCTGGACAGgtgagattgg	tgtctttcagGTTGTAATTA	09
09	ACTCTACGAGgtacacagca	ctgtttttagGGTGTGCCAC	10
10	GAGGATGCAGgtgagtttcg	tctgtcttagACTTTGAGCA	11
11	CTGCATCCTGgtaagaggca	ccccttgcagTTTGGCCCTT	12
12	ACAGGTGGAGgtgagtgggc	tttcttgcagAATTGTCTAC	13
13	GCACAAGAAGgtcagatttc	gccatactagCTCCGGCTCC	14
14	GGACAAAAAGgtagcagcac	ctcattttagCAAGTACAGT	15
15	AGATCAGTAGgtaagaagtg	tctttcccagTTCTTCCGGA	16
16	CCCAGAGGAGgtgagcttgc	gtctctgtagCTACTACCCA	17
17	CTTCGTGAAGgtaggtttga	tcctccccagGATGTCCTCG	18
consensus:	CAGgtaagt A t	tncagG c	

Table 31: Splice junctions for the DNA polymerase α gene from human chromosome 11q13.

Repeat elements also are abundant throughout this contig, with ALU, SVA, LTR, L1, MER, MIR, MLT, MST, and THE repeat elements present. This region contains 96 ALU sequences (partial or full) with 48 of the 96 containing SVA repeats in the opposite direction. There are 3 L1 repeats, 2 LTRs, 2 MERs, 1 THE, 4 MIR and 4 MLT repeats present in this region, and in several regions, the EST hits reported by powblast include an ALU or L1 repeat element (indicated in figure 39 as Hu EST w/ALU).

Simple sequence repeats (see table 32) also are quite prevalent in this region as there are multiple polynucleotide repeat regions of greater than 14 nucleotides, a poly C tract of 126 nucleotides, and the TA repeat of 144 nucleotides.

2. Zeta

The 11q13 region containing the DNA polymerase alpha gene from human chromosome 11q13 also contains a putative gene, named Zeta, that consists of 10 exons totalling 1.7 kb, and spans approximately 33 kb of genomic sequence. The genomic organization of Zeta is detailed in table 33. This Zeta gene region has a start methionine codon at 56183-56185, an in-frame UGA stop codon at 88784-88786 and a polyadenylation signal 806 nucleotides downstream at 89592-89597.

The Xgrail and powblast analyses presented in figure 40, predicted the presence of 10 exons in the Zeta region, which correlate to four human EST hits as predicted by powblast (Genbank accession numbers AA055031, R944475, AA214296, and AA205945). This allowed construction of a gene model for the putative Zeta gene, based on the splice junctions tabulated in table 34. Further evidence kindly provided by collaborators at the NHGRI, indicate that the Zeta gene produces a transcript of 2.0 kb (close to the 1.7 predicted by Xgrail exons) indicating that Xgrail may not have predicted all the exons present in this gene. Further sequence analysis of the full length Zeta cDNA would be useful to resolve this size discrepancy.

DNA polymerase α gene - Simple Sequence Repeats

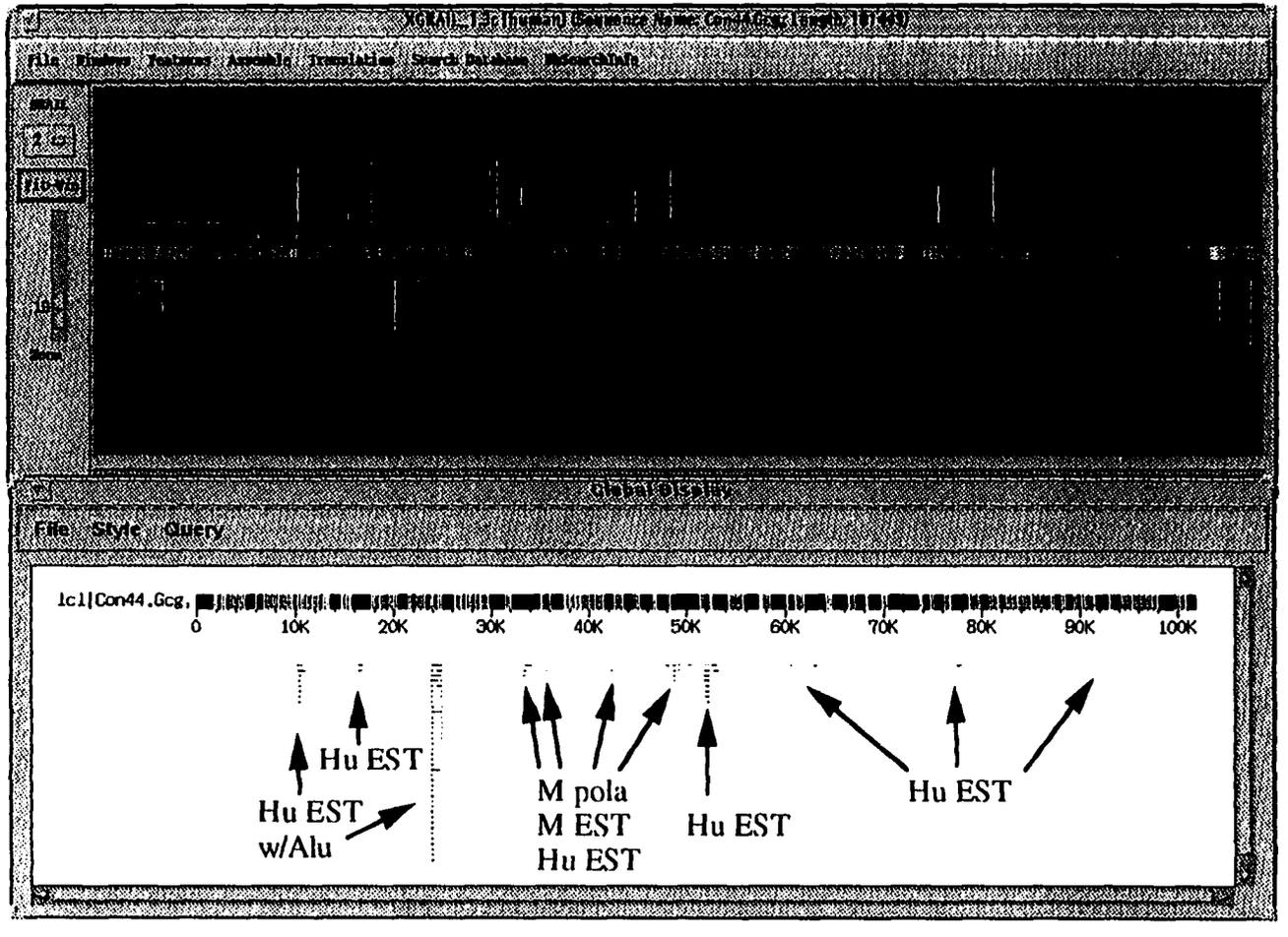
position	Repeat
247-347	poly GC
831-989	$(TA)_{29}$
3712-3772	$(A_4T)_{12}$
4831-4945	$A_7(GA_3)_2(GA)_7(GA_3)_{15}$
5771-5845	$(TA)_{14}(CA)_4(TA)_8(GA)_{11}$
7048-7074	A_{21}
10715-10787	$(T_3C)_4(T_4C)_2(T_3C)_2T_{24}$
11190-11229	$T_9GT_5GT_{11}$
17768-17821	$(CT)_7T_{20}$
13583-13692	poly CT
14761-14793	$T_3CT_{10}CT_{18}$
18620-18664	poly GT
22912-22945	$(TTA)_5T_{13}$
24058-24098	$A_{16}(CA_2)_4$
26410-26437	A_{19}
29938-29960	A_{18}
31349-31385	T_{18}
31716-31765	$T_9CT_4CT_3CT_{14}$
39090-39116	$T_{10}CT_{14}$
40146-10187	$A_{15}GA_5GA_8GA_4$
43060-43100	$T_4AT_2GT_2CGT_3GT_6GT_{14}$
44820-44837	A_{17}
57914-57956	$A_{19}C_2(GA_2)_2G_2TG_4$
58798-58905	$(TCC[A/C])_{28}$
61517-61552	$TA_2(T_2A)_7T_4AT_2AT_4$
66673-66704	A_{18}
67001-67127	C_{126}
70757-71389	$(TG[A/G]G)_{97}$
76318-76337	A_{19}
78805-78827	A_{21}
84452-84599	poly CT
85409-85538	$A_5(TA)_6T_6(AT)_6(AC)_8(TA)_{25}(CA)_9$
88881-88995	$(TA)_{57}$
89001-89302	$(TA)_{144}T_{12}$
90190-90260	$A_{14}(GA_3)_3(GA_5)_2(GA_3)_2$
94405-94445	$T_7A_2TC_2TCT_{19}$
97003-97109	poly GA
97880-97916	$A_{18}(GA_3)_3$

Table 32: Simple sequence repeats present in the DNA polymerase α gene region, including the putative gene "Zeta," from human chromosome 11q13 as reported by Xgrail.

Genome Organization - Putative gene "Zeta"

Exon number	position	size of exon	size of intron
01	56183-56262	79	10125
02	66387-66441	54	2366
03	68807-69363	556	2613
04	71976-72129	153	138
05	72267-72345	78	3092
06	75437-75626	189	1288
07	75914-77037	123	869
08	77906-78013	107	9611
09	87624-87828	214	843
10	88681-88785	104	--
ATGstart	56183-56185		
UGAstop	88784-88786		
polyA	89592-89597		

Table 33: Genome organization of the putative gene "Zeta" from human chromosome 11q13. The gene contains 10 exons for a total of 1.7 kb, which span roughly 33 kb of genomic sequence. The start codon is located within exon 1, the stop codon at the end of exon 10 and a polyadenylation signal 806 nucleotides downstream of the stop codon.



Putative "Zeta" gene

Figure 40: Top panel: Graphical result of Xgrail gene identification software for the DNA polymerase α gene region from human chromosome 11q13, also including the putative gene, "Zeta". The peaks present are Xgrail predicted exons with a quality rating of excellent (green), good (blue) or marginal (red). The light blue lollipops are polyadenylation signals and the purple squares are CpG islands. Yellow and red squares indicate promotors and the vertical orange bars denote simple sequence repeats. Bottom panel: Graphical display of the powblast output using Musk. The black bar at the top of the panel represents the consensus sequence with the grey areas indicating repeat regions. The lines under the black bar indicate database homology between the consensus sequence and entries in the public database.

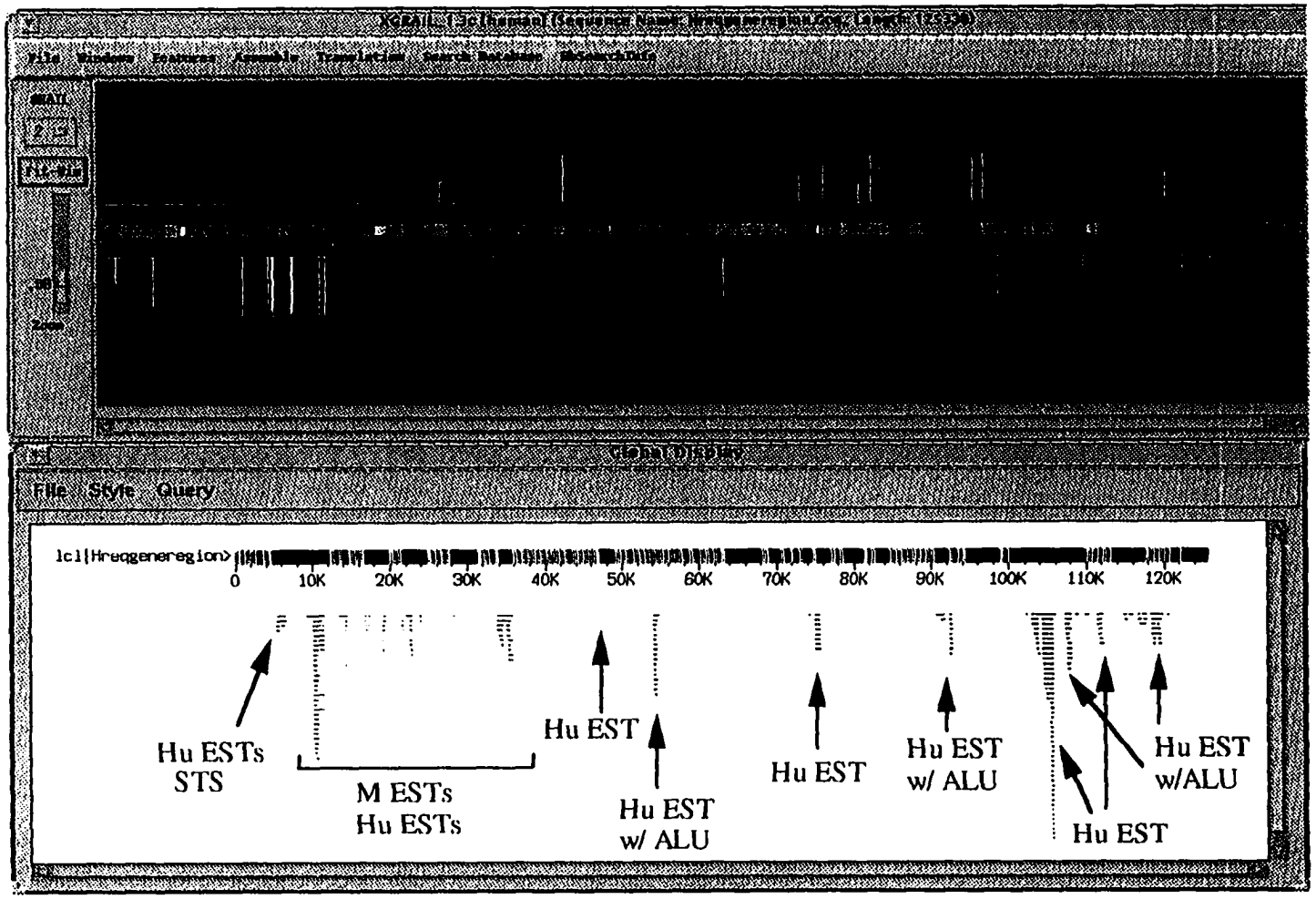
Splice Junctions-Putative gene "Zeta"

Left Exon	Left splice junction	Right splice junction	Right Exon
01	AGGCAAAAGGgtgggactgg	aggtcaacatGGGGCTGAGC	02
02	GGAGGCCAAGgtgggtggat	cccctcccagGGCCCCTCCC	03
03	GGCGCAGTCCgtctacattg	ctggactcagGCTGGGCTGC	04
04	CCTCAACTCCgtctccctgg	tctcatccagCAACCTCAAC	05
05	CCTGCTGGCTgtgggtgggg	ttctccccagTTGTGGAGTG	06
06	ACAGAGTGGGgtgggggagg	tcggtcccagGTGGCTTCCA	07
07	GACCAAGGAGgtaagcgagc	tcttacctagGTGATGAGCT	08
08	CTGTCTCATGgtgatttggt	cctctcccagGTTCTCCAAC	09
09	GTGCCAGAAGgtgagtcacc	gtgagtccagCAAGATGCCA	10
consensus:	CAGgtaagt A t	tncagG c	

Table 34: Splice junctions for the putative gene "Zeta" from human chromosome 11q13.

3. Mouse requiem gene

Xgrail analysis of b18h3 also revealed the presence of a putative human gene homologue of the mouse requiem gene, termed HREQ, that consists of 10 exons. Exon 4 is absent in the human genomic sequence, based on sequence comparison with the mouse requiem mRNA sequence (Genbank accession number U10435) and spans 25 kb of human chromosome 11q13. This region contains many Xgrail predicted excellent exons which correspond to the mouse requiem gene exons, as well as regions of homology with numerous human ESTs (Genbank accession numbers H10344, N64757, F10384, H10525, N76165, F12776, N53620, N51478, H10681, R39084, H15370, N98417, H69223, AA115487, D20048, AA025238, Z39650, F02914, R65599, F01706, R36865, R36877, W01541, R74322, T30759, H69222, W30705, T34079, T34099, AA114998, R64504, R69903, R59870, R57893, W23195, H15369, H10680, T80399, R18759, R94803, H48837, F06642, and F06645), several mouse ESTs (Genbank accession numbers AA033220, W80017, AA239744, W12880, AA058295, AA026025, and W57432) and one STS (WI-30608; Genbank accession number G21758) as revealed by powblast (see figure 41). Three regions also have homology to human ESTs containing repeat sequences (Genbank accession numbers R74416, R70582, and R94721). All of the human ESTs predicted by powblast correspond to the Xgrail predicted exons for the mouse requiem gene and thus, it is highly likely that this region encodes a human homolog of the mouse requiem gene. The genome organization for the human homologue of the mouse requiem gene is detailed in table 35, and has a methionine start codon at position 22725-22727, an in-frame stop codon at 11377-11379, and two polyadenylation signals downstream at 10570-10575 and 8891-8896. The mouse requiem intron/exon boundaries conform well to the conserved splice junctions (see table 36).


File Style Query
lcl|Hreqgeneregion
0
10K
20K
30K
40K
50K
60K
70K
80K
90K
100K
110K
120K
Hu ESTs
STS
M ESTs
Hu ESTs
Hu EST
Hu EST
w/ ALU
Hu EST
Hu EST
w/ ALU
Hu EST
w/ALU
Hu EST
HREQ
Eta
Theta
Epsilon/Beta

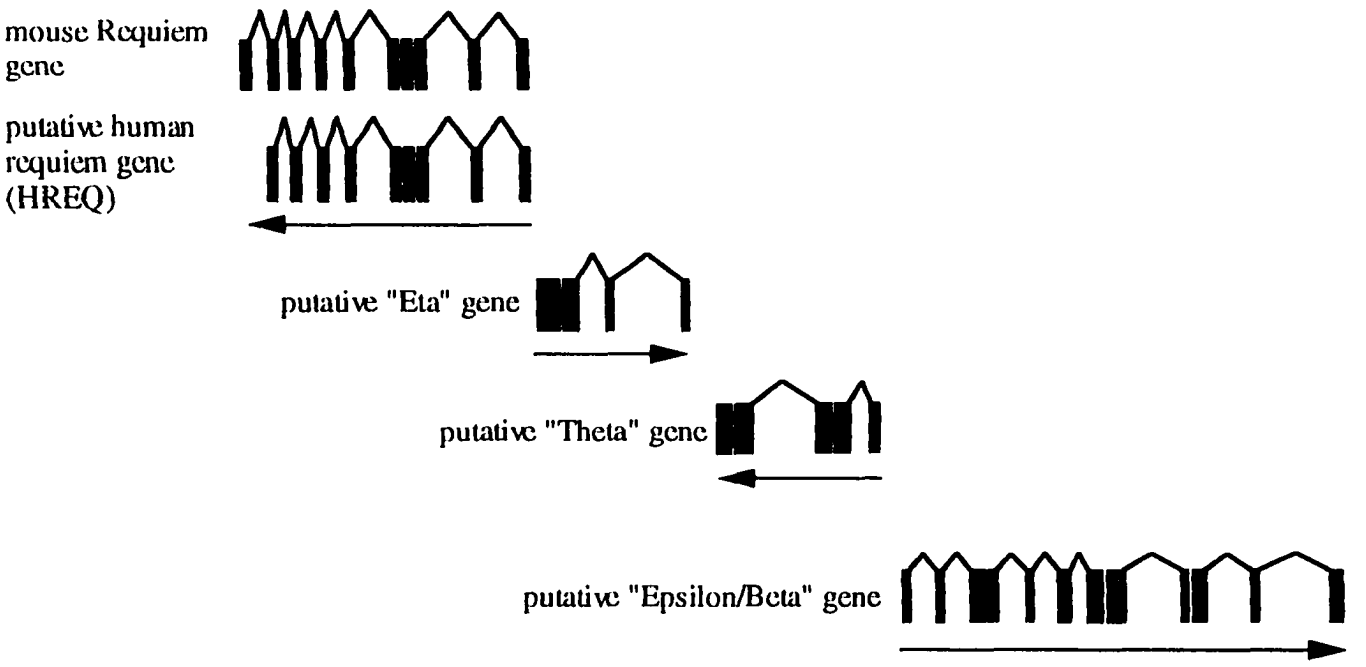


Figure 41: Top panel: Graphical result of Xgrail gene identification software for the mouse Requiem gene region from human chromosome 11q13, and putative genes "Eta," "Theta," and "Epsilon/Beta". The peaks present are Xgrail predicted exons with a quality rating of excellent (green), good (blue) or marginal (red). The light blue lollipops are polyadenylation signals and the purple squares are CpG islands. Yellow and red squares indicate promotors and the vertical orange bars denote simple sequence repeats. Bottom panel: Graphical display of the powblast output using Musk. The black bar at the top of the panel represents the consensus sequence with the grey areas indicating repeat regions. The lines under the black bar indicate database homology between the consensus sequence and entries in the public database.

Genome Organization - Mouse requiem gene

Exon number	position	size of exon	size of intron
01	22594-22755	161	417
02	22067-22177	110	323
03	21578-21744	166	2430
04	(absent in human)		
05	19071-19148	77	1597
06	17337-17474	137	127
07	17081-17210	129	185
08	16785-16896	111	2496
09	14200-14289	89	2735
10	11328-11465	137	---
ATGstart	22725-22727		
UGAstop	11377-11379		
polyA	10570-10575		
	8891-8896		

Table 35: Genome organization of the mouse requiem gene region from human chromosome 11q13. The gene contains 10 exons for a total of 1.1 kb, which span roughly 11.5 kb of genomic sequence. The start codon is located within exon 1, the stop codon within exon 10 and the polyadenylation signals 753 and 2432 nucleotides downstream of the stop codon.

Splice Junctions-Mouse Requiem gene

Left Exon	Left splice junction	Right splice junction	Right Exon
01	CGGGGTCCAGgtgagggtcc	cctgccacagGATTGGCCTC	02
02	ATTAAGCCAGgtaaggcaca	ttgcttgcagACACAGACCA	03
03	GAGCGCGAAAggtacaggat	(absent in human)	04
04	(absent in human)	aactcctcagGCCCAGCTAC	05
05	CCTGTGACAGtgagtgcctc	tcccactcagTTTGTGGAAA	06
06	TGAGGAGCAGaaatgtaagc	ctctctgtagCCAAAAAGGG	07
07	GGCCGCTCAGgtactgcttc	cccactacagGGCATCCATC	08
08	CGAGAATGACgtgtgtatcc	tcaggaccagTTGCTCTTCT	09
09	CTGAAGGTAAgttgcccaga	ctctctgcagGAAGTTGGAG	10
consensus:	CAGgtaagt A t	tncagG c	

Table 36: Splice junctions for the mouse requiem gene from human chromosome 11q13.

Using the mouse requiem gene sequence as a guideline, a putative human homologous gene model was constructed by combining the human EST homologies from powblast, Xgrail predicted exons, and the splice juction consensus adherence. The putative gene organization for the human requiem gene (HREQ) is shown in table 37 and the Xgrail predicted intron/exon boundaries conform well to the consensus splice junctions (see table 38). Additional evidence for the HREQ gene in 11q13 is the presence of an Xgrail predicted promoter upstream from exon 1, and an Xgrail predicted polyadenylation signal downstream of exon 9 (mouse exon 10 was not predictedby Xgrail and therefore, assumed to be absent in the human homologue). The construction of a putative HREQ gene is shown in figure 41.

This entire 125,330 bp region of b18h3 also contains representatives of several different repeat families, including ALU, SVA, MIR, MER, L1, LTR and MLT. There are 93 ALU repeats (of which 52 have SVA repeats in the opposite direction), three MIR repeats, five MER repeats, seven L1 repeats, one LTR repeat and one MLT repeat. In addition, this region contains many simple sequence repeats that are listed in table 39. Several regions of polynucleotide tracts, including eight regions of poly A greater than 15 nucleotides, four regions of poly C of greater than 18 nucleotides (with one region stretching for 139 nucleotides), 11 regions of poly T tracts greater than 15 nucleotides, a 62 bp CT repeat, a 134 bp TA repeat, a 56 bp TG repeat, and a 208 bp TG repeatalso are present in this region.

4. Eta

The 11q13 BAC b18h3 also encodes for a putative gene "Eta" that contains four Xgrail-predicted exons (see figure 41) and spans approximately 13 kb of genomic sequence. This gene also was identified by its database homology to several human ESTs (Genbank accession numbers Z43708, F07645, F12871, R14699, T75138, Z20586, R38559, W68275, Z39767, R42421, F10475, W8264, F03895, H40104, W81196, W68276, R38595, and R84576) as well as several mouse ESTs

Genome Organization - Human homologue of the mouse requiem gene

Exon number	position	size of exon	size of intron
01	22594-22755	161	417
02	22067-22177	110	323
03	21578-21744	166	2185
04	19302-19393	91	154
05	19071-19148	77	1597
06	17337-17474	137	127
07	17081-17210	129	185
08	16785-16896	111	2496
09	14200-14289	89	---
ATGstart	22725-22727		
UGAstop	14210-14212		
polyA	11862-11867		

Table 37: Genome organization of a human homologue of the mouse requiem gene from human chromosome 11q13. The gene contains 9 exons for a total of 1.1 kb, which span roughly 8.5 kb of genomic sequence. The start codon is located within exon 1, the stop codon within exon 9 and the polyadenylation signal 2345 nucleotides downstream of the stop codon.

Splice Junctions-Human homologue of the mouse Requiem gene

Left Exon	Left splice junction	Right splice junction	Right Exon
01	CGGGGTCCAGgtgagggtcc	cctgccacagGATTGGCCTC	02
02	ATTAAGCCAGgtaaggcaca	ttgcttgcagACACAGACCA	03
03	GAGCGCGAAAggtacaggat	Gtccactcag CGGATCCTAG	04
04	GAAATCCAAGgtgaggggcc	aactcctcagGCCCAGCTAC	05
05	CCTGTGACAGtgagtgcctc	tcccactcagTTTGTGGAAA	06
06	TGAGGAGCAGaaatgtaagc	ctctctgtagCCAAAAAGGG	07
07	GGCCGCTCAGgtactgcttc	cccactacagGGCATCCATC	08
08	CGAGAATGACgtgtgtatcc	tcaggaccagTTGCTCTTCT	09
09	CTGAAGGTAAgttgcccaga	---------	
consensus:	CAGgtaagt A t	tncagG c	

Table 38: Splice junctions for a human homologue of the mouse requiem gene from human chromosome 11q13.

Mouse Requiem gene region - Simple Sequence Repeats

position	Repeat
89-134	A_{17}
4285-4314	$(T_4C)_3T_{15}$
12747-12806	$A_{15}(A_3G)_3(A_5G)_5$
14508-14528	T_{20}
19817-19840	T_{23}
25319-25368	$A_8TA_4T_2A_8CA_4(CA_3)_3$
25593-25655	$(CT)_{31}$
26064-26098	$T_4CT_8CT_{14}$
31373-31426	$T_4A_3T_{10}A_2T_8AT_8$
31736-31760	T_{21}
32183-32225	T_{45}
36524-36551	$T_4(T_2G)_6(T_4G)_3T_7$
36619-36666	poly GT
42706-42731	A_{25}
45194-45231	A_{19}
46260-46423	$(TA)_{67}T_{29}$
46579-46623	$CT_5CT_4CT_7C_2(T_3A)_5$
50090-50113	A_{23}
50637-50776	C_{139}
52075-52100	A_{22}
52666-52706	T_{19}
55078-55120	$T_5G_3T_8G_3T_{23}$
57060-57099	$A_4(A_3T)_5A_4TATA_3TA_5$
63109-63152	A_{22}
68592-68623	T_{24}
70796-70831	A_{35}
74938-74996	C_{58}
75001-75072	C_{71}
77061-77095	$T_5(CT_4)_2T_{13}$
80473-80620	poly CT
81007-81037	T_4C_{18}
85194-85239	$(TA)_6T_8(A_2T_2)_5AT_5$
87071-87106	A_{15}
99949-99978	$T_{14}CT_3CAT_6$
111827-111896	$T_{10}(TG)_8T_{12}$
116342-116551	$(TA)_{104}$
119721-119775	poly AC
120103-120150	A_{20}
123525-123581	$(TG)_{28}$
124669-124802	poly CT
125010-125197	poly CT

Table 39: Simple sequence repeats present in the mouse requiem gene region, including putative genes "Eta," "Theta," and "Epsilon/Beta," from human chromosome 11q13 as reported by Xgrail.

(Genbank accession numbers AA002649, AA168660, AA030239, R75085, AA241289, and AA036536). The genome organization of this gene, found in table 40, details the size and position of the introns and exons, as well as the possible regulatory elements present. The putative gene "Eta" has four potential start methionine codons at positions 34212-34215, 34297-34299, 34363-34365, and 34417-34419. There also are two potential stop codons present in this gene region at position 47215-47217 and 47223-47225, and a polyadenylation signal downstream from the stop codon at position 48659-48664. The intron/exon boundaries for Eta correlate well with the consensus splice junctions, and are tabulated in table 41. The potential Eta gene structure is shown in figure 41.

5. Theta

A third putative gene named "Theta" also is located in this 125,330 bp region of chromosome 11q13. Theta is predicted by Xgrail to consist of five exons which span a total of 13 kb of genomic sequence. The "Theta" gene region has powblast homology with several human ESTs which contain repeat elements as well as the unique sequence (Genbank accession numbers T54254, Z69719, U73023, T41147, H05039, R49432, N34500, AA017402, M64936, T41177, W15274, R02550, F15197, T41176, AA089966, T84227, N58335, T40311, and U91321). The genomic organization of this gene, shown in table 42, indicates the size and position of the Xgrail predicted exons. The intron/exon boundaries for these predicted exons are given in table 43. "Theta" contains one potential methionine start codon at position 66928-66930, one potential in-frame stop codon at position 59840-59842 and a polyadenylation signal at position 58420-58425. Taken together, these results allow for the construction of a putative Theta gene model as shown in figure 41.

6. Epsilon/Beta

The fourth putative gene, named "Epsilon/Beta," was located in the 11q13 BAC b18h3, and shown to contain 12 Xgrail predicted exons that span 35 kb of genomic

Genome Organization - putative "Eta" gene

Exon number	position	size of exon	size of intron
01	34074-34486	412	156
02	34642-34845	203	1146
03	35991-36036	45	11154
04	47190-47221	31	---
potential ATGstarts	34212-34215		
	34297-34299		
	34363-34365		
	34417-34419		
potential UGAstop	47215-47217		
	47223-47225		
poly A signal	48659-48664		

Table 40: Genome organization of the putative "Eta" gene from human chromosome 11q13. The gene contains 4 exons for a total of 0.7 kb, which span roughly 13 kb of genomic sequence. The potential start codons all are located within exon 1, the stop codons within exon 4 and the polyadenylation signal is 1444 or 1436 nucleotides downstream of the stop codon, depending which stop codon is used.

Splice Junctions-putative "Eta" gene

Left Exon	Left splice junction	Right splice junction	Right Exon
01	GCCACCGTGTgtgggcggga	ttccccacagGAGGGAGGGA	02
02	CCCCACATAGgacacgaggc	tgatttccagCAGTGGTAAA	03
03	TGCCTAGCAGggaggtcagg	aacagggaagATGGCGGCTG	04
04	TAGTGAAGCTgtgagtgggt	----------	
consensus:	CAGgtaagt A t	tncagG c	

Table 41: Splice junctions for putative "Eta" gene from human chromosome 11q13.

Genome Organization - putative "Theta" gene

Exon number	position	size of exon	size of intron
01	66861-66974	113	264
02	66412-66597	185	2300
03	63854-64112	258	143
04	63453-63711	258	3412
05	59829-60041	212	---
potential ATGstart	66928-66930		
potential UGAstop	59840-59842		
poly A signal	58420-58425		

Table 42: Genome organization of the putative "Theta" gene from human chromosome 11q13. The gene contains 5 exons for a total of 1.0 kb, which spans roughly 13 kb of genomic sequence. The potential start codon is located within exon 1, the stop codon within exon 5 and the polyadenylation signal is 1420 nucleotides downstream of the stop codon.

Splice Junctions-putative "Theta" gene

Left Exon	Left splice junction	Right splice junction	Right Exon
01	TTTCGGGGAGgtatctgggg	ctctctgcagCTCCTGGGCT	02
02	GTTCCTGCAGgtgaggggggc	ccggtaccagGGGCCCGTGC	03
03	CAGAATCCCAgtaagtaaag	tctgccccagGCTGCCACGG	04
04	TCCTCCGCTGgtggggatga	cccgcctcagCCTCCCAAAG	05
consensus:	CAGgtaagt A t	tncagG c	

Table 43: Splice junctions for putative "Theta" gene from human chromosome 11q13.

sequence. This gene region has homology to many human ESTs (Genbank accession numbers C18833, W32038, G21695, U60873, T86516, N22644, H01421, H91303, R68990, W57744, T78464, N41457, N64365, AA169520, AA053165, T96710, T86426, AA160043, T18972, D16859, W76407, W72389, R23859, D16982, W32317, N28235, W56683, W39338, W56108, AA010989, W56012, W38857, W38907, N23466, W37277, W37893, AA054357, AA016190, R71357, AA053611, N67461, AA0101866, AA045252, AA045143, W57753, N92350, W37790, AA039604, N98703, D17151, D63288, H86877, R53723, R78782, R31243, AA24835, C18790, and AA224926), and several human ESTs which contain repeat elements (Genbank accession numbers R19205, R61134, N76852, N55476, R44467, T70741, T98980, AA191318, T98936, R61852, R61852, N20806, H56441, H56442, N31062, AA214047, T81440, T69164, T68489, AA233267, T69227, N21656, AA243867, R95099, R95100, H02561, T98359, T68422, T59151, and U90223).

The genomic organization of the putative "Epsilon/Beta" gene is shown in table 44. The gene contains 12 exons which correspond to a coding region of 1.4 kb. Two possible methionine start codons are present, one at 71623-71625 and the other at 71629-71631, three possible in-frame stop codons are observed at positions at 107074-107076, 107085-107087, and 107099-107101, and a polyadenylation signal is seen downstream at position 118629-118634. All of the intron/exon boundaries for this gene conform well to the consensus splice junctions as shown in table 45.

V. Cat Eye Syndrome Region Sequencing and Analysis

One BAC of approximately 220,000 bp (b376a1) from the Cat Eye Syndrome region was analyzed after sequence scanning. A blast search versus the pALU database reveals that 85% of this BAC consists of repeat elements, with several repeats being full length. With such a high percentage of repeat sequences, this BAC is nearly impossible to assemble using the current shotgun assembly programs. Statistically, even though sufficent shotgun data has been generated to complete the entire BAC with 4-fold

Genome Organization - putative "Epsilon/Beta" gene

Exon number	position	size of exon	size of intron
01	71618-71680	62	2398
02	74078-74162	84	1835
03	75997-76218	221	364
04	76582-76697	115	1037
05	77734-77813	79	1117
06	78930-79053	123	865
07	79918-80087	169	3971
08	84058-84262	204	5193
09	89455-89502	47	928
10	90430-90548	118	5971
11	96519-96558	39	10370
12	106928-107101	173	---
potential ATGstarts	71623-71625		
	71629-71631		
potential UGAstop	107074-107076		
	107085-107087		
	107099-107101		
poly A signal	118629-118634		

Table 44: Genome organization of the putative "Epsilon/Beta" gene from human chromosome 11q13. The gene contains 12 exons for a total of 1.4 kb, which span roughly 35 kb of genomic sequence. The potential start codons all are located within exon 1, the stop codons within exon 12 and the polyadenylation signal is 11569 nucleotides downstream of the stop codon. However, three additional polyadenylation sites are present in the intronic region before exon 12 (100087-100092, 101014-10109, and 105693-105698), although no stop codon is present at the end of exon 11.

Splice Junctions-putative "Epsilon/beta" gene

Left Exon	Left splice junction	Right splice junction	Right Exon
01	CACACAGCAAgtgagcagag	gcctttgcagATGGACGGGA	02
02	GGAGCCCGAGgtgggtccca	tcctccacagGGAGAAGCTG	03
03	CGTTTTATGGgtaagagcca	cctgagccagAGACACCGTG	04
04	AGAGGATGGGgtgtggcttg	cctagcgcatGCCCCGCTTG	05
05	GATGGAGACAgtgagtgtga	ccctccccagGTGTGCCTTC	06
06	CAGAGAGAAGgtactgcccc	tgctggccagCATGTCCTGC	07
07	CTGCACACGGgtggggaggc	tcctccccagGCCGAACTGA	08
08	GTGGAGGACGgtgagcagcc	ggaggccaagGATGGGTGGA	09
09	GGCCAGCATGgtgaaacccc	tccctcgcagGCAAGGGGAT	10
10	GCAGGGCTGAggacgctgca	gccctggcacCATGGCCAAC	11
11	TCAGACCAAGgtgggttggg	tttaccccagCACCTCTATA	12
consensus:	CAGgtaagt A t	tncagG c	

Table 45: Splice junctions for putative "Epsilon/beta" gene from human chromosome 11q13.

redundancy, the BAC still remains in 32 contigs (including only the contigs greater than 1.9 kb, since smaller contigs could potentially be contaminating *E. coli* sequence). The largest contig is 44.9 kb and has 17-fold coverage, while the additional 31 contigs range in size from 1.9 to 8.1 kb and have an average of 3-fold coverage. Since this BAC contains mostly repetitive elements, the assembly programs tend to combine all gel reads with these elements into the same contig, i.e. the 44.9 kb contig (contig 111) in this case, and the remaining contigs cannot be ordered (see figure 42).

Xgrail and powblast analysis was performed on the largest contig, contig111, and the results are shown in figure 43. The lone database homology highlighted by powblast (Genbank accession number Z54334) does not have a corresponding exon in Xgrail, but this is not surprising since the region of database homology is with a clone containing the Huntington's Disease expandable CAG repeat. This repeat is flagged by an erroneously predicted Xgrail CpG island this sequence is not a perfect CAG repeat. Furthermore, the match in this region of contig 111 also is part of an L1 repeat element. Contig 111 contains a large number of polyadenylation sites and the positions and direction of these sites are given in table 46. In addition to repeat elements such as ALU, L1, MER and MIR, this contig also contains many simple sequence repeats (see table 47). Interestingly, the majority of the simple repeats found in this contig are tracts of poly T or poly A. Ten of the 30 simple sequence repeats recognized by Xgrail are poly T tracts of greater than 9 nucleotides, while 13 of the 30 are poly A tracts of greater than 10 nucleotides.

Several other regions of database homology are present in the other contigs of b376a1, and the Xgrail predicted exons, features and powblast homologies for all b376a1 contigs is detailed in table 48. One particular region of interest is the database homology to mouse and human U6 small nuclear RNA in contig 102. These database homologies could indicate a potential gene since the same gene in mouse and human has homology in this region, but the lack of any open reading frame, or potential splice junctions predicted

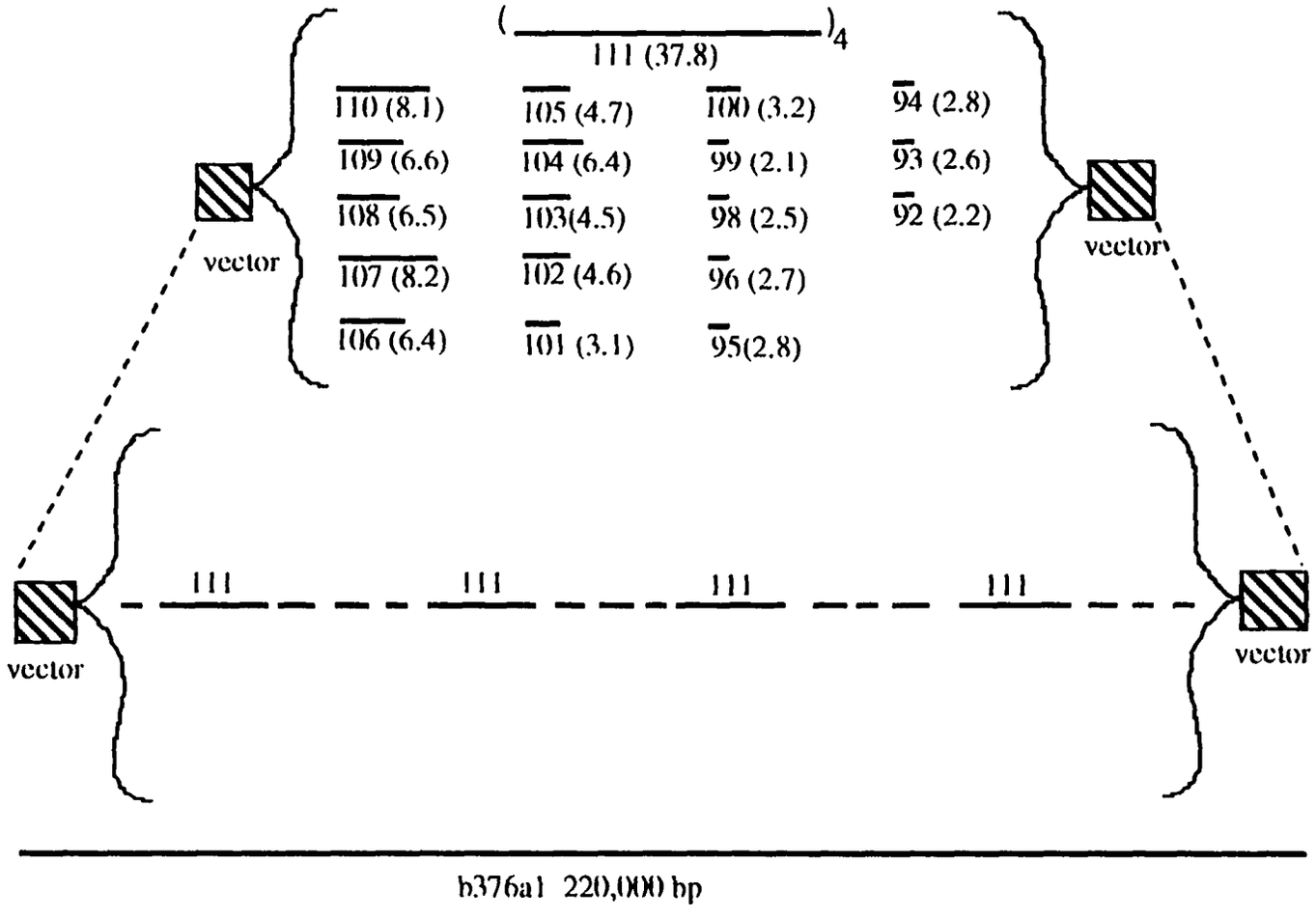


Figure 42: Contig map of b376a1. This BAC has a 37.8 kb repeated region which does not assemble properly with any of the assembly programs available.

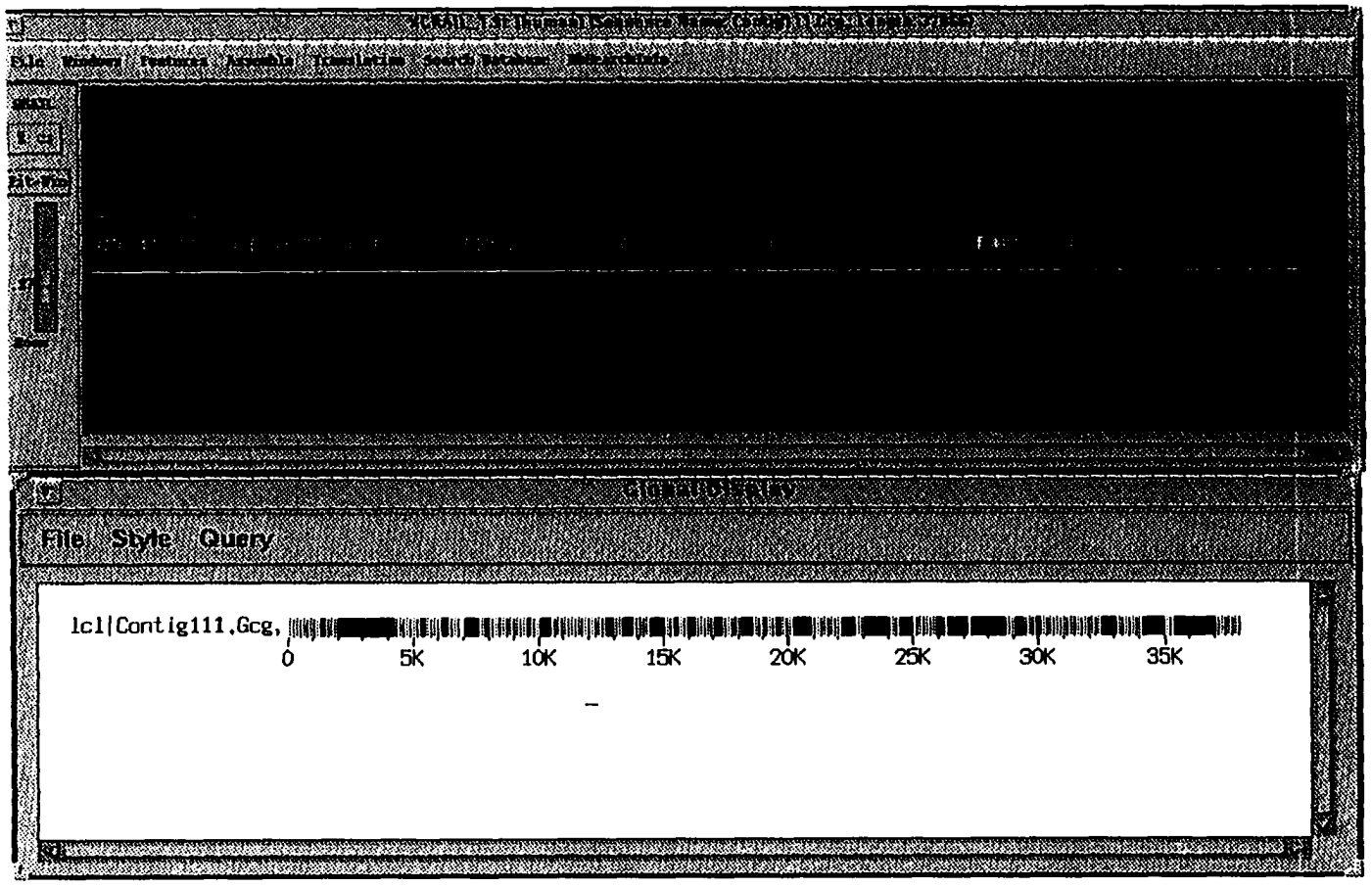

File Style Query
lcl|Contig111.Gcg,
0
5K
10K
15K
20K
25K
30K
35K

Figure 43: Top panel: Graphical result of Xgrail gene identification software for the repeat contig from human chromosome 22q11. The peaks present are Xgrail predicted exons with a quality rating of excellent (green), good (blue) or marginal (red). The light blue lollipops are polyadenylation signals and the purple squares are CpG islands. Yellow and red squares indicate promotors and the vertical orange bars denote simple sequence repeats. Bottom panel: Graphical display of the powblast output using Musk. The black bar at the top of the panel represents the consensus sequence with the grey areas indicating repeat regions. The lines under the black bar indicate database homology between the consensus sequence and entries in the public database.

b376a1 contig 111-
polyadenylation signals

position	strand
8505-8510	f
13769-13774	f
20848-20853	f
25430-25435	f
33542-33547	f
34645-34650	f
6482-6487	r
13787-13792	r
23364-23369	r
26120-26125	r
26127-26132	r
26542-26547	r
26842-26847	r
26846-26851	r
28427-28432	r
30274-30279	r
34315-34320	r
34464-34469	r

Table 46: Polyadenylation signals present in the b376a1 contig 111 from human chromosome 22q11 as reported by Xgrail.

b376a1 contig 111-
simple sequence repeats

position	Repeat
95-117	$(CA_2)_5$
239-261	T_{21}
867-901	$T_4GT6GT_5G(T_3G)_3T_4G$
1847-1879	A_{23}
4450-4474	T_{16}
4935-4977	$TGT_4GT_8GT_{15}G_3TG_2$
5746-5775	A_{25}
6063-6101	A_9CA_{13}
6358-6369	A_{12}
6949-6982	$CA_7(CA_5)_3CA_9$
8022-8058	T_{16}
8787-8824	$(A_3T)_7$
10647-10662	T_{11}
11571-11585	T_{14}
12426-12457	A_{20}
12891-12991	poly CA
13266-13276	A_{11}
15197-15219	T_9
16601-16629	A_{16}
18631-18664	T_{14}
21387-21421	A_{19}
21870-21918	$CA_7CA_6CA_4(CA_6)_2CA_4C_2A(CA_2)_2$
22927-22940	A_{10}
25464-25481	T_{17}
26460-26481	$A_{11}CA_6$
28657-28688	$T_4CT_3CT_{21}$
30376-30424	A_{18}
31247-31283	A_{23}
33834-33857	$(A_2T)_3A_3TA_2T$
35155-35170	T_{15}

Table 47: Simple sequence repeats present in the b376a1 contig 111 from human chromosome 22q11 as reported by Xgrail.

b376a1 Xgrail predicted exons/powblast predicted database homology

contig #	Xgrail exons	Powblast hits	Features
111	3 excellent	Z54334 (CA repeat)	18 PA
(44.9)	3 good		3 CpG
	1 marginal		30 SSR
110	1 excellent	none	2 PA
(8.1)	1 marginal		9 SSR
109	1 good	none	2 PA
(6.6)			1 CpG
			11 SSR
108	1 good	none	2 PA
(6.5)			1 CpG
			7 SSR
107	1 good	none	1 SSR
(8.2)	2 marginal		
106	1excellent	G29765 (STS)	2 PA
(6.4)	1 good	H89730 (Hu EST)	1 SSR
		Z78021 (STS)	
		X78694 (M EST)	
		7 EST w/LTR	
105	1 marginal	none	3 PA
(4.7)			4 SSR
104	1 excellent	none	5 SSR
(5.0)	1 good		
103	1 excellent	none	1 PA
(4.5)			4 SSR
102	none	V00853 (M U6 snRNA)	2 PA
(4.6)		J00648 (M U6 snRNA)	1 CpG
		X07425 (Hu U6 snRNA)	9 SSR
101	none	U16738 (Hu EST w/CAG)	4 PA
(3.1)		8 EST w/Mer	1 prom
			1 CpG
			2 SSR
100	none	Z77249 (Hu cosmid)	1 PA
(3.2)		H59740 (Hu EST)	2 SSR
		Z85990 (BRCA2 GR)	
		R31871 (Hu EST)	
99	none	none	1 SSR
(2.1)			
98	none	none	1 PA
(2.5)			4 SSR
97	none	none	1 PA
(1.8)			2 SSR
96	1 marginal	T27264 (Hu EST)	1 PA
(2.7)			1 prom
			1 SSR
95	2 good	67 Hu ESTs	1 prom
(2.8)			1 CpG
			2 SSR

94 (2.8)	1 good	none	1 PA 1 CpG 6 SSR
93 (2.6)	none	none	1 PA 7 SSR
92 (2.2)	1 excellent	none	1 PA 3 SSR
91 (1.5)	none	none	5 SSR
90 (2.1)	1 good	none	1 PA 2 CpG 2 SSR
89 (1.6)	1 excellent	none	none
88 (1.9)	none	3 Hu ESTs	2 PA 3 SSR
87 (1.8)	1 good	none	1 PA 1 SSR
86 (1.9)	none	none	1 PA 1 CpG 1 SSR
85 (1.9)	1 good	none	1 PA 2 SSR
84 (1.9)	none	none	1 CpG 2 SSR
82 (3.4)	1 marginal	none	2 PA 2 SSR
77 (2.0)	none	none	1 CpG 3 SSR
71 (1.9)	1 good	none	1 PA
70 (2.0)	none	none	1 SSR

Table 48: Simple sequence repeats present in the b376a1 contig 111 from human chromosome 22q11 as reported by Xgrail. Abbreviations: PA, polyadenylation signals; CpG, CpG islands; prom, promoter region,; SSR, simple sequence repeat; Hu, human; M, mouse; GR, gene region; CAG, CAG repeat region. Numbers in parenthesis are contig size in kb.

by Xgrail make this unlikely, and in addition, the contigs containing unique sequence arenot sufficiently large to reveal any pattern of predicted exons or gene organization.

Chapter IV

CONCLUSIONS

A. Technique development and sequence scanning

The results presented here demonstrate that a plasmid-based, shotgun sequencing approach for large scale sequencing is a feasible strategy to obtain over 0.67 Mb of sequence data. Completing this task in a reasonable amount of time was due to the improvements in the overall shotgun sequencing strategy and newly developed methods that were implemented during the past year. These methods include a large scale diatomaceous earth-based BAC isolation and an automated double-stranded template isolation on the Biomek robots developed as part of this disserttation research and these techniques have greatly increased the laboratory efficiency by minimizing human error and resulted in increased quality of the template samples.

The search for the genes involved in inherited disorders has prompted new strategies for DNA sequencing which focus not on gaining complete, final sequence to 100% accuracy, but on gaining information about particular a gene's location, organization and regulation. A strategy, called Sequence Scanning, developed to find genes responsible for inherited diseases, was implemented and tested as part of this dessertation research. The studies presented here clearly show the feasibility of a scanning strategy in which the genomic DNA is sequenced to approximately 98% completion, and then analysed for gene content. This approach has been shown to be extremely efficient as the results of these studies demonstrate that for previously known genes, i.e. genes whose mRNA has been sequenced, but the genomic organization was unknown, only three out of 84 known exons (3.6%) were not found in the genomic sequence data generated by sequence scanning. The three missing exons are located in gaps between the contigs assembled from the shotgun data, but, even so, the genes easily

were identified by the presence of the overwhelming majority of exons, database homologies, and regulatory regions present in the resulting contigs. That this approach yielded the localization of nine putative genes in the approximately 0.5 Mb chromosome 11q13 region also supports the notion that sequence scanning is a viable method for finding putative genes when teamed with powerful analysis programs such as Xgrail and powblast.

B. MN-1 gene region, 22q11

The 90,213 bp meningioma deletion region of human chromosome 22q11 was sequenced to a confidence level of greater than 99.99% (NCBI-NLM Phase 3) using an inital shotgun based sequencing approach followed by directed sequencing for gap closure. The sequence of three overlapping cosmids 76A4, 50B11 and 58B12 form a contiguous region of 22q11 and has been assigned Genbank accession number AC000105. Within this meningioma deletion region, the MN-1 gene was localized, the genomic organization was fully characterized, and a repeat density of 30% was observed in both intergenic and intragenic regions. No other genes in addition to the MN-1 gene were observed in this 90 kb region.

C. b137c7, 11q13

Approximately 140,000 bases of unique sequence (Genbank accession number AC000134) was generated from b137c7 in the MEN-1 gene region from human chromosome 11q13 by sequence scanning. The sequence of this BAC is contained in 21 ordered contigs (NCBI-NLM Phase 2) ranging in size from 2.5 kb to 21.1 kb (see figure 44). Using the Xgrail and powblast analysis tools, three known genes were located (PYGM, ZFM1, and GC kinase) and the genomic organization of these genes was fully characterized. Four previously unknown genes (human neurexin II α homologue, Nu, MEN-1 and Kappa) also were identified and characterized by the Xgrail and powblast analysis, and this gene-rich BAC from 11q13 contains an average repeat density of 30%. It is important to note that the gene named MEN-1 subsequently has been shown to be the

Contig Alignment Map - b137c7 from human chromosome 11q13

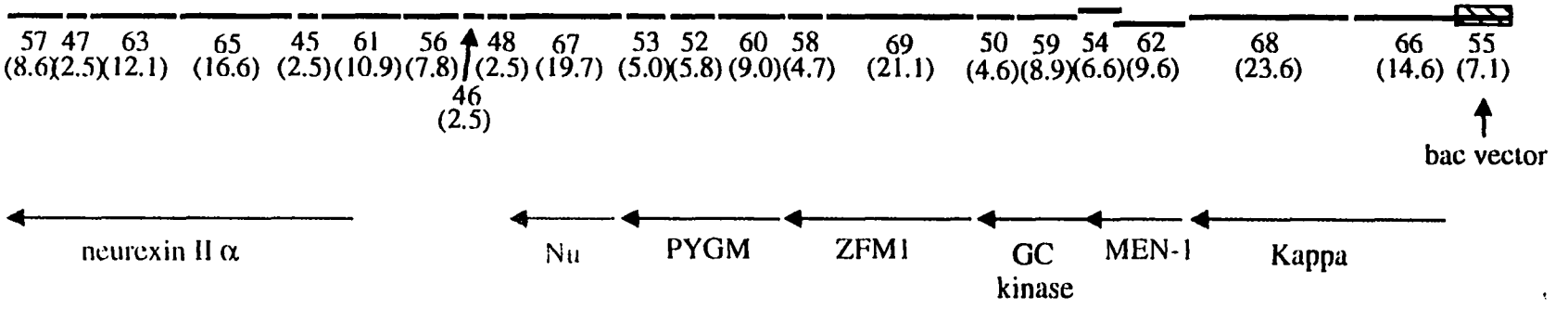


Figure 44: Contig alignment map of bac 137c7 from human chromosome 11q13. The black dashed line represents the size and location of the contigs. Numbers under the black dashed line indicate the contig number, and the number in parenthesis is the contig size in kilobases. Gene/putative gene positions are indicated in color at the bottom of the figure.

gene responsible for the MEN-1 phenotype, as determined by disruption/deletion experiments performed in Dr. Collins' laboratory at the NHGRI.

D. b18h3, 11q13

The BAC b18h3 from the human chromosome 11q13 region, contains approximately 240,000 bp (Genbank accession number AC000353), as determined after 3-fold sequence coverage via sequence scanning, and could be organized into seven ordered contigs ranging in size from 0.2 kb to 67 kb (see figure 45). The DNA polymerase α previously mapped to 11q13, is located within this 11q13 BAC and its genomic structure is now reported for the first time. Additionally, five previously unknown genes (Zeta, Eta, Theta, HREQ and Epsilon/Beta) were identified in this gene rich-region of 11q13, and an average repeat density is approximately 45% was observed. The previously unknown genes were characterized by analysis with Xgrail and powblast, and gene models were developed to reveal their genomic organization.

E. CES b376a1, 22q11

Analysis of the 220,000 bp (Genbank accession number AC000360) BAC b376a1 by sequence scanning was completely devoid of genes, and had a repeat element frequency of 80%. The sequence of this BAC lends further support to the notion that sequence scanning is an effective method for gene hunting, based on the knowledge that coding regions do not typically contain repeat elements such as ALU or L1. The significant percentage of repeat elements found in this BAC indicate that this clearly is not a gene rich region of the chromosome, and subsequently, in the search for genes, the inquiry could continue with other adjacent genomic clones in the region by sequence scanning at a lower cost and effort than is required for complete closure of the shotgun data into one contig.

F. Summary

During the course of this dissertation research, methods were developed and implemented that resulted in the sequence of approximately two thirds of a million bases

Contig Alignment Map - b18h3 from human chromosome 11q13

Figure 45: Contig alignment map of bac 18h3 from human chromosome 11q13. The black dashed line represents the size and location of the contigs. Numbers under the black dashed line indicate the contig number, and the number in parenthesis is the contig size in kilobases. Gene/putative gene positions are indicated in color at the bottom of the figure.

of human genomic DNA. Computer based analysis methods were employed and indicate the presence of 14 genes, both previously characterized and newly revealed. Based on earlier studies[326], this present work confirms the 30-50% repeat sequence density in gene rich regions of human DNA. Finally, the long sought after gene for Multiple Endocrine Neoplasia, type 1 (MEN-1) now has been characterized at the genomic sequence level, and now will allow for early detection of this familial cancer.

Chapter V

LITERATURE CITED

1. Cushing, H. The meningiomas (dural endotheliomas). Their source and favored seats of origin (Cavendish Lecture). Brain 45:282-316 (1922).

2. Al-Mefty, O. "Meningiomas." Raven Press, Ltd., New York, NY (1991).

3. Cushing, H., and Eisenhardt, L. "Meningiomas. Their Classification, Regional Behaviour, Life History and Surgical End Results." Charles C. Thomas Publishing. Springfield, IL (1938).

4. Kepes, J.J. "Meningiomas: Biology, Pathology and Differential Diagnosis." Masson Publishing USA, Inc. New York, NY (1982).

5. Kersting, G., and Lennarzt, H. *In vitro* cultures of human meningioma tissue. J. Neuropathol. Exp. Neurol. 16:507-513 (1957).

6. Grant, F.C. A study of the results of surgical treatment in 2326 consecutive patients with brain tumor. J. Neurosurg. 13:479-488 (1956).

7. Zulich, K.J. "Brain tumors. Their Biology and Pathology." Springer-Verlag, New York. pg. 57 (1957).

8. Hoessly, G.F., and Olivecrona, H. Report on 280 cases of verified parasagittal meningioma. J. Neurosurg. 12:614-626 (1955).

9. Zimmerman, H.M. Brain tumors: their incidence and classifications in man and their experimental production. Ann. NY Acad. Sci. 159:337-359 (1969).

10. Schoenberg, B.S., Christine, B.W., and Whisnant, J.P. The descriptive epidemiology of primary intracranial neoplasms: the Connecticut experience. Am. J. Epidemiol. 104:499-510 (1976).

11. Dastur, D.K., Lalitha, V.S., and Prabhakar, V. Pathological analysis of intracranial space-occupying lesions in 1000 cases including children. Part I. Age, Sex and Pattern; and the tuberculomas. J. Neurol. Sci. 6:575-592 (1968).

12. Balasubramaniam, V., and Ramamurthi, B. Meningiomas. Neurology India 18 (Suppl.1):81-88 (1970).

13. Katsura, S., Suzuki, J., and Wada, T. A statistical study of brain tumors in the neurosurgical clinics in Japan. J. Neurosurg. 16:570-580 (1959).

14. Giordano, C., and Lamouche, M. Meningiomas in Cote D'Ivoire. Afr. J. Med. Sci. 4:249-263 (1973).

15. Levy, L.F. Brain tumors in Malawi, Rhodesia and Zambia. Afr. J. Med. Sci. 4:393-397 (1973).

16. Odeku, E.L., and Adeloye, A. Cranial meningiomas in Nigerian Africa. Afr. J. Med. Sci. 4:275-287 (1973).

17. Froman, C., Phil, D., and Lipschitz, R. Demography of tumors of the central nervous system among the Bantu (African) population of the Transvaal, South Africa. J. Neurosurg. 32:660-664 (1970).

18. Manfredonia, M. Tumours of the nervous system in the African in Eritrea (Ethiopia). Afr. J. Med. Sci. 4:383-387 (1973).

19. Russell, D.S., and Rubenstein, L.J. "Pathology of Tumors of the Nervous System." Edition 5. Edward Arnold Publishing. London (1989).

20. Elsberg, C.A. "Tumors of the Spinal Cord." Hoeber Publishing, New York, NY. (1925).

21. Rohringer, M., Sutherland, G.R., Louw, D.F., and Sima, A.F. Incidence and clinicopathological features of meningioma. J. Neurosurg. 71:665-672 (1989).

22. Taptas, J.N. Intracranial meningioma in a four month old infant simulating subdural hematoma. J. Neurosurg. 18:120-121 (1961).

23. Crouse, S.K., and Berg, B.O. Intracranial meningiomas in childhood and adolescence. Neurology (Minneapolis) 22:135-141 (1972).

24. Deen, H.G., Jr., Scheithauer, B.W., and Ebersold, M.J. Clinical and pathological study of meningiomas of the first two decades of life. J. Neurosurg. 56:317-322 (1982).

25. Nakamura, Y., and Becker, L.E. Meningeal tumors of infancy and childhood. Pediatr. Pathol. 3:341-358 (1985).

26. Cooper, M., and Dohn, D.F. Intracranial meningiomas in childhood. Cleveland Clin. Q. 41:197-204 (1974).

27. Paillas, J.E., Pellet, P., Guillermain, P., and Lavieille, J. Les meningiomes intra-craniens de l'enfant et de l'adolescent. Neurochirugia (Stuttg.) 14:41-53 (1971).

28. Wood, W.M., White, R.J., and Kernohan, J.W. One hundred intracranial meningiomas found incidentally at necropsy. J. Neuropathol. Exp. Neurol. 16:337-340 (1957).

29. Cooney, L.M., and Solitare, G.B. Primary intracranial tumors in the elderly. Geriatrics 27:94-104 (1972).

30. Gaist, G., and Piazza, G. Meningiomas in two members of the same family (with no evidence of neurofibromatosis). J. Neurosurg. 16:110-113 (1959).

31. Joynt, R.J., and Perret, G.E. Meningiomas in a mother and a daughter. Cases without evidence of neurofibromatosis. Neurology (Minneapolis) 11:164-165 (1961).

32. Joynt, R.J., and Perret, G.E. Familial Meningiomas J. Neurol. Neurosurg. Psychiat. 28:163-164 (1965).

33. Delleman, J.W., De Jong, J.G., and Bleeker, G.M. Meningiomas in five members of a family over two generations, in one member simultaneously with acoustic neurinomas. Neurology (Minneapolis) 28:567-570 (1978).

34. Memon, M.Y. Multiple and familial meningiomas without evidence of neurofibromatosis. Neurosurg. 7:262-264 (1980).

35. Sedzimer, C.B., Frazer, A.K., and Roberts, J.R. Cranial and spinal meningiomas in a pair of identical twin boys. J. Neurol. Neurosurg. Psychiat. 36:368-376 (1973).

36. Walsh, J., Gye, R., and Connelly, T.J. Meningioma: A late complication of head injury. Med. J. Aust. 1:906-908 (1969).

37. Gardeur, D. Allal, R., Sichez, U.P., and Metzger, J. Post traumatic intracranial meningiomas: Recognition by computed tomography in three cases. J. Comput. Assist. Tomogr. 3:103-104 (1979).

38. Von Holander, H., Kosmaoglou, V., and Sturm, W. Intraspinales meningeom nack wirbelfraktur. Kasuistische mitteilung. Zentralbl. Neurochir. 32:179-185 (1971).

39. Preston-Martin, S., Paganini-Hill, A., Henderson, B.E., Pike, M.C., and Wood, C. Case control study of intracranial meningiomas in women in Los Angeles County, California. J. Natl. Cancer Inst. 65:67-69 (1980).

40. Preston-Martin, S., Yu, M.C., Henderson, B.E., and Roberts, C. Risk factors for meningiomas in men in Los Angeles County. J. Natl. Cancer Inst. 70:863-866 (1983).

41. Dunsmore, R.H., and Roberts, M. Trauma as a cause of brain tumor. A medicolegal dilemma. Conn. Med. 38:521-523 (1974).

42. Choi, N.W., Schuman, L.M., and Gullen, W.H. Epidemiology of primary central nervous system neoplasms. II. Case control study. Am. J. Epidemiol. 91:467-485 (1970).

43. Dimant, I.N., Loktionov, G.M., and Sataev, M.M. Induction of spinal cord meningeal tumor in rabbit with radioactive cobalt. Vopr. Onkol. 11:46-53 (1965).

44. Bogdanowitz, W.M., and Sachs, E., Jr. The possible role of radiation in oncogenesis of meningioma. Surg. Neurol. 2:379-383 (1974).

45. Norwood, C.W., Kelly, D.L., Jr., Davis, C.H., and Alexander, E., Jr. Irradiation-induced mesodermal tumors of the central nervous system: Report of two meningiomas following X-ray treatment of gliomas. Surg. Neurol. 2:161-164 (1964).

46. Waga, S., and Handa, H. Radiation induced meningioma: With review of literature. Surg. Neurol. 5:215-219 (1976).

47. Pagni, C.A., Canavero, S., Fiocchi, F., and Ponzio, G. Chromosome 22 monosomy in a radiation-induced meningioma. Ital. J. Neurol. Sci. 14:377-379 (1993).

48. Horanyi, B. Rontgen besugarzas hatasara keletkezett meningeoma gyermekben. Magy. Radiol. 17:1-7 (1961).

49. Munk, J., Peyser, E., and Gruszkiewica, J. Radiation induced intracranial meningiomas. Clin. Radiol. 20:90-94 (1969).

50. Waterson, K.W., Jr., and Shapiro, L. Meningioma cutis: Report of a case. Int. J. Dermatol. 9:125-129 (1970).

51. Feiring, E.H., and Foer, W.H. Meningioma following radium therapy. Case report. J. Neurosurg. 29:192-194 (1968).

52. Watts, C. Meningioma following irradiation. Cancer 38:1939-1940 (1976).

53. Soffer, D., Pittaluga, S., Feiner, M., and Beller, A.J. Intracranial meningiomas following low dose irradiation to the head. J. Neurosurg. 59:1048-1053 (1983).

54. Giaquinto, S., Massi, G., Ricolfi, A., and Vitali, S. On six cases of radiation meningiomas from the same community. Ital. J. Neurol. Sci. 5:173-175 (1984).

55. Rubenstein, A.B., Shalit, M.N., Cohen, M.L., Zandbank, U., and Reichenthal, E. Radiation induced cerebral meningioma: a recognizable entity. J. Neurosurg. 61:966-971 (1984).

56. Sussman, S., TerBrugge, K.G., Solt, L.C., and Deck, J.H.N. Thorotrast-induced meningioma. J. Neurosurg. 52:834-837 (1980).

57. Kyle, R.H., Oler, A., Lasser, E.C., and Rosonoff, H.L. Meningioma induced by thorium dioxide. N. Engl. J. Med. 268:80-82 (1963).

58. Meyer, M.W., Powell, H.C., Wagner, M., and Niwayama, G. Thorotrast induced adhesive arachnoiditis assiciated with meningioma and schwanoma. Hum. Pathol. 9:366-370 (1978).

59. Weiss, A.F., Portmann, R., Fischer, H., Simon, J., and Zang, K.D. Simian virus 40-related antigens in three human meningiomas with defined chromosome loss. Proc. Natl. Acad. Sci., USA 72:609-613 (1975).

60. Theile, M., Strauss, M., Luebbe, L., Scherneck, S., Krause, H., and Geisler, E. SV40 induced somatic mutations: possible relevance to viral transformation. Cold Springs Harbor Symp. Quant. Biol. 44:377-382 (1980).

61. Scherneck, S., Rudolph, M., Geissler, E, Vogel, F., Lubbe, L., Wahlte, H., Nisch, G., Weickmann, F., and Zimmerman, W. Isolation of a SV40-like papovavirus from a human glioblastoma. Int. J. Cancer 24:523-531 (1979).

62. Zang, K.D., May, G., and Fischer, H. Expression of SV40-related T antigen in cell cultures of human meningiomas. Naturwissenschaften 66:59 (1979).

63. Wold, W.S.M., Mackey, J.K., Brackmann, K.H., Takemori, N., Rigden, P., and Green, M. Analysis of human tumors and human malignant cell lines for BK virus-specific DNA sequences. Proc. Natl. Acad. Sci., USA 75:454-458 (1978).

64. Fiori, M., and Di Mayorca, G. Occurrence of BK virus DNA in DNA obtained from certain human tumors. Proc. Natl. Acad. Sci., USA 73:4662-4666 (1976).

65. Israel, M.A., Markin, M.A., Takemoto, K.K., Howley, P.M., Aaronson, S.A., Solomon, D., and Khoury, G. Evaluation of normal and neoplastic human tissue for BK virus. Virology 90:187-196 (1978).

66. Rachlin, J.R., Wollmann, R., and Dohrmann, G. SV40 viral DNA in human CNS tumors. J. Neuropathol. Exp. Neurol. 43:301 (1984).

67. Bickerstaff, E.R., Small, J.M., and Guest, I.A. The relapsing course of certain meningiomas in relation to pregnancy and menstruation. J. Neurol. Neurosurg. Psychiat. 21:89-91 (1958).

68. Hagedoorn, A. The chiasmal syndrome and retrobulbar neuritis in pregnancy. Am. J. Opthalmol. 20:690-699 (1937).

69. Walsh, F.B. "Clinical Neuro-Opthalmology." Williams and Wilkins Press, Baltimore (1947).

70. Rucker, C.W., and Kearns, T.P. Mistaken diagnoses in some cases of meningioma. Am. J. Opthalmol. 51:15-19 (1951).

71. Smith, F.P., Slavik, M., and MacDonald, J.S. Association of breast cancer with meningioma: Report of two cases and review of the literature. Cancer 42:1992-1994 (1978).

72. Schoenberg, B.S., Christine, B.W., and Whisnant, J.P. Nervous system neoplasms and primary malignancies of other sites. Neurology (Minneapolis) 25:705-712 (1975).

73. Burns, P.E., Jha, N., and Bain, G.O. Association of breast cancer with meningioma. A report of five cases. Cancer 58:1537-1539 (1986).

74. Jacobs, D.H., McFarlane, M.J., and Holmes, F.F. Female patients with meningioma of the sphenoid ridge and additional primary neoplasms of the breast and genital tract. Cancer 60:3080-3082 (1987).

75. Donnell, M.W., Meyer, G.A., and Donegan, W.L. Estrogen receptor protein in intracranial meningiomas. J. Neurosurg. 50:499-502 (1979).

76. Schnegg, J.F., Gomez, R., LeMarchand-Beraud, T., and Tribolet, N. Presence of sex steroid hormone receptors in meningioma tissue. Surg. Neurol. 15:415-418 (1981).

77. Fontaine, B., Rouleau, G.A., Seizinger, B., Jewell, A.F., Hanson, M.P., Martuza, R.L., and Gusella, J.G. Equal parental origin of chromosome 22 losses in human sporadic meningioma: No evidence of genomic imprinting. Am. J. Hum. Genet. 47:823-827 (1990).

78. Reik, W., Surani, M.A. Genomic imprinting and embryonal tumors. Nature 338:112-113 (1989).

79. Schroeder, W.T., Chao, L.Y., Dao, D.D., Strong, L.C., Pathak, S., Riccardi, V., Lewis, W.H., and Saunders, G.F. Nonrandom loss of maternal chromosome 11 alleles in Wilms tumors. Am. J. Hum. Genet. 40:413-420 (1987).

80. Scrabble, H, Cavenee, W., Ghavimi, F., Lovell, M., Morgan, K., and Sapienza. C. A model for embryonal rhabdomyosarcoma tumorigenesis that involves genomic imprinting. Proc. Natl. Acad., Sci. USA 86:7480-8484 (1989).

81. Toguchida, J., Ishizaki, K., Sasaki, M.S., Nakamura, Y., Ikenaga, M., Kato, M., Sugimot, M, Kotoura, Y., and Yamamuro, T. Preferential mutation of paternally derived RB gene as the initial event in sporadic osteosarcoma. Nature 338:156-158 (1989).

82. Nowell, P.C., Emanuel, B.S., Finan, J.B., Erikson, J., and Croce, C.M. Chromosome arrangements on oncogenesis. Microbiol. Sci. 1:223-228 (1984).

83. Bailly, R.A., Boselut, R., Zucman, J., Cormier, F., DeLattre, O., Roussel, M., Thomas, G., and Ghysdael, J. DNA-binding and transcriptional activation properties of the EWS-FLI-1 fusion protein resulting from the t(11;22) translocation in Ewing Sarcoma. Mol. Cell Biol. 14:3230-3241 (1994).

84. Alcalay, M., Zangrilli, D., Fagoli, M., Pandolfi, P.P., Mencarelli, A., Lo Coco, F.. Biondi, A., Grignani, F., and Pelicci, P.G. Expression patterns of the RARα-PML fusion gene in acute promyelocytic leukemia. Proc. Natl. Acad. Sci., USA 89:4840-4844 (1992).

85. Golub, T.R., Barker, G.F., Lovett, M., and Gilliland, D.G. Fusion of PDGF receptor β to a novel ets-like gene, tel, in chronic myelomonocytic leukemia with t(5;12) chromosomal translocation. Cell 77:307-316 (1994).

86. Chissoe, S.L., Bodenteich, A., Wang, Y.F., Wang, Y.P., Burian, D., Clifton, S.W., Crabtree, J., Freeman, A., Iyer, K., Jian, L., Ma., Y., McLaury, H.J., Pan, H.Q., Sarhan, O.H., Toth, S., Wang, Z., Zhang, G., Heisterkamp, N., Groffen, J., and Roe, B.A. Sequence and analysis of the human ABL gene, the BCR gene anad regions involved in the Philadelphia chromosomal translocation. Genomics 27:67-82 (1995).

87. Zang, K.D., and Singer, H. Chromosomal constitution of meningiomas. Nature (letter) 216:84-85 (1967).

88. Zankl, H., and Zang, K.D. The role for acrocentric chromosomes in nucleolar organization I. Correlation between the loss of acrocentric chromosomes and a decrease in number of nucleoi in meningioma cell cultures. Virchows Arch. (Cell Pathology) 11:251-256 (1972).

89. Mark, J., Levan, G., and Mitelman, F. Identification by fluorescence of the G chromosome lost in human meningiomas. Hereditas 71:163-172 (1972).

90. Paul, B., and Porter, I.H. Giemsa banding in an established line of human malignant meningioma. Humangenetik 18:185-187 (1973).

91. Zankl, H., Weiss, A.F., and Zang, K.D. Cytological and cytogenetical studies on brain tumors. VI. No evidence for a translocation in 22-monosomic meningiomas. Humangenetik 30:343-348 (1975).

92. Zankl, H., Seidel, H., and Zang, K.D. Cytological and cytogenetical studies on brain tumors. V. Preferential loss of sex chromosomes in human meningiomas. Humangenetik 27:119-128 (1972).

93. Katsuyama, J., Papenhausen, P.R., Herz, F., Gazivoda, P., Hirano, A., and Koss, L.G. Chromosome abnormalities in meningiomas. Cancer Genet. Cytogenet. 22:63-68 (1986).

94. Casalone, R., Granata, P., Simi, P., Tartantino, E., Butti, G., Buonaguidi, R., Faggionato, F., Knerich, R., and Solero, L. Recessive cancer genes in meningiomas? An analysis of 31 cases. Cancer Genet. Cytogenet. 27:145-159 (1987).

95. Mark, J. The human meningioma: A benign tumor with specific chromosome characteristics in "Chromosomes and Cancer," German, J., ed. John Wiley and Sons, New York pp. 497-517 (1974).

96. Zang, K.D. Cytological and cytogenetical studies on human meningioma. Cancer Genet. Cytogenet. 6:249-274 (1982).

97. Zankl, H., and Zang, K.D. Cytological and cytogenetical studies on brain tumors. IV. Identification of the missing G chromosome in human meningiomas as no.22 by fluorescence technique. Hum. Genet. 14:167-169 (1972).

98. Mark, J. The chromosomal findings in seven human meningiomas and one neurosarcoma. Acta. Pathol. Microbiol. Scand. 80:61-70 (1972).

99. Yamada, K., Kondo, T., Yoshioda, M., and Oami, H. Cytogenetic studies in twenty human brain tumors: Association of no. 22 chromosome abnormalities with tumors of the brain. Cancer Genet. Cytogenet. 2:293-307 (1980).

100. Rey, J.A., Bello, J.M., de Campos, J.M., Benitez, T., Ayuso, M.C., and Valcarel, E. Chromosome studies in two human tumors. Cancer Genet. Cytogenet. 10:159-165 (1983).

101. Mark, T. Chromosomal abnormalities and their specificity in human neoplasms. An assessment of recent observations by banding techniques. Adv. Cancer Res. 24:165-222 (1977).

102. Zankl, H., and Zang, K.D. Correlations between clinical and cytogenetical data in 180 meningiomas. Cancer Genet. Cytogenet. 1:351-356 (1980).

103. Sandberg, A. "The Chromosomes in Human Cancer and Leukemia." Elsevier, New York pp.535-543 (1980).

104. Maltby, E.L., Ironside, J.W., and Battersby, R.D.E. Cytogenetic studies in 50 meningiomas. Cancer Genet. Cytogenet. 31:199-210 (1988).

105. Zankl, H., and Huwer, H. Are NOR's easily translocated to deleted chromosomes? Hum. Genet. 42:137-142 (1978).

106. Levy, M.Z., Allsopp, R.C., Futcher, A.B., Greider, C.W., and Harley, C.B. Telomere end replication problems and cell aging. J. Mol. Biol. 225:951-960 (1992).

107. Vaziri, H., Schachter, G., Uchida, I., Wei, L., Zhu, X., Effros, R., Cohen, D.. and Harley, C.B. Loss of telomeric DNA during aging of normal and trisomy 21 human lymphocytes. Am. J. Hum. Genet. 52:661-667 (1993).

108. Counter, C.M., Avilion, A.A., Le Feuvre, C.E., Stewart, N.G., Greider, C.W., Harley, C.B., and Bacchetti, S. Telomere shortening associated with chromosome instability is arrested in immortal cells which express telomerase activity. EMBO J. 11:1921-1929 (1992).

109. Stoll, C., and Roth, M.P. Segregation of a 22 ring chromosome in three generations. Hum. Genet. 63:294-296 (1983).

110. Hunter, A.G.W., Ray, M., Wang, H.S., and Thompson, D.R. Phenotypic correlations in patients with ring chromosomes. Clin. Genet. 12:239-249 (1977).

111. Arinami, T., Kondo, I., Hamaguchi, H., and Nakajima, S. Multifocal meningiomas in a patient with a constitutional ring chromosome 22. J. Med. Genet. 23:178-180 (1986).

112. Benn, P.A. Specific chromosome aberrations in senescent fibroblast cell lines derived from human embryos. Am. J. Hum. Genet. 28:465-473 (1976).

113. Seizinger, B.R., de la Mont, S., Atkins, L., Gusella, J.F., and Martuza, R.L. Molecular genetic approach to human meningioma: Loss of genes on chromosome 22. Proc. Natl. Acad. Sci., USA 84:5419-5423 (1987).

114. Seizinger, B., Rouleau, G., Ozelius, L.J., Lane, A.H., St. George-Hyslop, P., Huson, S., Gusella, J.F., and Marguza, R.L. Common pathogenic mechanism for three tumor types in bilateral acoustic neurofibromatosis. Science 236:317-319 (1987).

115. Dumanski, J.P., Carlbom, E., Collins, V.P., Nordenskjold, M. Deletion mapping of a locus on human chromosome 22 involved in the oncogenesis of meningioma. Proc. Natl. Acad. Sci., USA 84:9275-9279 (1987).

116. Okasaki, M., Nishisho, I., Tateishi, H., Motomura, K., Yamamoto, M., Miki, T., Hayakawa, T., Takai, S., Honjo, T., and Mori, T. Loss of genes on the long arm of chromosome 22 in human meningiomas. Mol. Biol. Med. 5:15-22 (1988).

117. Schneider, G., Lutz, S., Henn, W., Zang, K.D., and Blin, N. Search for putative suppressor genes in meningioma: significance of chromosome 22. Hum. Genet. 88:579-582 (1992).

118. Dumanski, J.P., Geurts van Kessel, A.H.M., Ruttledge, M., Wladis, A., Sugawa, N., Collins, V.P., and Nordenskjold, M. Isolation of anonymous, polymorphic DNA fragments from human chromosome 22q12-qter. Hum. Genet. 84:219-222 (1990).

119. Dumanski, J.P., Rouleau, G.A., Nordenskjold, M., and Collins, P. Molecular genetic analysis of chromosome 22 in 81 cases of meningioma. Cancer Res. 50:5863-5867 (1990).

120. Rouleau, G.A., Haines, J.L., Bazanowski, A., Collela-Crowley, A., Trofatter, J.A., Wexler, N.S., Conneally, P.M., and Gusella, J.F. A genetic linkage map of the long arm of human chromosome 22. Genomics 4:1-6 (1989).

121. Rouleau, G.A., Seizinger, B.R., Wertelecki, W., Haines, J.L., Superneau, D.W., Martuza, R.L., and Gusella, J.F. Flanking markers bracket the neurofibromatosis type 2 (NF-2) gene. Am. J. Hum. Genet. 46:323-328 (1990).

122. Pulst, S.M., Rouleau, G.A., Marineau, C., Fain, P., and Sieb, J.P. Familial meningioma is not allelic to neurofibromatosis 2. Neurology 43:2096-2098 (1993).

123. Cogen, P.H., Daneshvar, L., Bowcock, A.M., Metzger, A.K., and Cavalli-Sforza, L.L. Loss of heterozygosity for chromosome 22 DNA sequences in human meningioma. Cancer Genet. Cytogenet. 53:271-277 (1991).

124. Nigro, J.M., Baker, S.J., Preisinger, A.C., Jessup, J.M., Hostetter, R., Cleary, K., Bigner, S.H., Davidson, N., Baylin, S., Devilee, P., Glover, T., Collins, F.S., Weston, A., Modali, R., Harris, C.C., and Vogelstein, B. Mutations in the p53 gene occur in diverse human tumour types. Nature 342:705-708 (1989).

125. Sanson, M., Richard, S., Delattre, O., Poliwka, M., Mikol, J., Philippon, J., and Thomas, G. Allelic loss on chromosome 22 correlates with histopathological predictors of recurrence of meningiomas. Int. J. Cancer 50:391-394 (1992).

126. Bello, M.J., de Campos, J.M., Vaquero, J., Kusak, M.E., Sarasa, J.L., Rey, J.A., and Pestana, A. Chromosome 22 heterozygosity is retained in most hyperdiploid and psuedodiploid meningiomas. Cancer Genet. Cytogenet. 66:117-119 (1993).

127. Lekanne Deprez, R., Groen, N.A., van Biezen, N.A., Hagemeijer, A., van Drunen, E., Koper, J.W., Avezaat, C.J.J., Bootsma, D., and Zwartoff, E. A t(4;22) in a meningioma points to the localization of a putative tumor-suppressor gene. Am. J. Hum. Genet. 48:783-790 (1991).

128. Al Saadi, A., Latimer, R., Madercic, M., and Robbins, T. Cytogenetic studies of human brain tumors and their clinical significance. II. Meningioma. Cancer Genet. Cytogenet. 26:127-141 (1987).

129. Meese, E., Blin, N., and Zang, K.D. Loss of heterozygosity and the origin of meningioma. Hum. Genet. 77:349-351(1987).

130. Rey, J.A., Bello, M.J., de Campos, J.M., and Kusak, M.E. Incidence and origin of dicentric chromosomes in cultured meningiomas. Cancer Genet. Cytogenet. 35:55-60 (1988).

131. Casartelli, C., Rogatto, S.R., and Barbieri Neto, J. Karyotypic evolution of human meningioma. Progression on through malignancy. Cancer Genet. Cytogenet. 40:33-45 (1989).

132. Poulsgard, L., Ronne, M., and Schroder, H.D. Cytogenetic studies of 19 meningiomas and their clinical significance. I. Anticancer Res. 9:109-112 (1989).

133. Rey, J.A., Bello, M.J., de Campos, J.M., Vaquero, J., Kusak, M.E., Sarasa, J.L., and Pestana, A. Abnormalities of chromosome 22 in human brain tumors determined by combined cytogenetic and molecular genetic approaches. Cancer Genet. Cytogenet. 66:1-10 (1993).

134. Ruttledge, M.H., Sarrazin, J., Rangaratnam, S., Phelan, C.M., Twist, E., Merel, P., Delattre, O.,Thomas, G., Nordenskjold, M., Collins, V.P., Dumanski, J.P., and Rouleau, G.A. Evidence for the complete inactivation of the NF2 gene in the majority of sporadic meningiomas. Nature Genetics 6:180-184 (1994a).

135. Wellenreuther, R., Kraus, J.A., Lenartz, D., Menon, A.G., Schramm, J., Louis, D.N., Ramesh, V., Gusella, J.F., Wiestler, O.D., and von Deimling, A. Analysis of the neurofibromatosis 2 gene reveals molecular variants of meningioma. Am. J. Pathol. 146(4):827-832 (1995).

136. Leto, T.L., and Marchesi, V.T. A structural model of human erythrocyte protein 4.1. J. Biol. Chem. 259:4603-4608 (1984).

137. Conboy, J., Kan, Y.W., Shohet, S.B., and Mohandas, N. Molecular cloning of protein 4.1, a major structural element of the human erythrocyte membrane skeleton. Proc. Natl. Acad. Sci., USA 83:9512-9616 (1986).

138. Tchernia, G., Mohandas, N., and Shohet, S.B. Deficiency of skeletal membrane protein band 4.1 in homozygous hereditary ellipotcytosis. Implication for erythrocyte membrane stability. J. Clin. Invest. 68:454-460 (1981).

139. Rees, D.J., Ades, S.E., Singer, S.J., and Hynes, R.O. Sequence and domain structure of talin. Nature 347:685-689 (1990).

140. Luna, E.J., and Hitt, A.L. Cytoskeleton-plasma membrane interactions. Science 258:955-964 (1992).

141. Bianchi, A.B., Hara, T., Ramesh, V., Gao, J., Klein-Szanto, A.J.P., Morin, F., Menon, A.G., Trofatter, J.A., Gusella, J.F., Seizinger, B.R., and Kley, N. Mutations in transcript isoforms of the neurofibromatosis 2 gene in multiple human tumour types. Nature Genetics 6:185-192 (1994).

142. Gould, K.L., Bretscher, A., Esch, F.S., and Hunter, T. cDNA cloning and sequencing of the protein-tyrosine kinase substrate, ezrin, reveals homology to band 4.1. EMBO J. 8:4133-4142 (1989).

143. Turunen,O., Winqvist, R., Pakkanen, R., Grzeschik, K.H., Wahlstrom, T., and Vaheri, A. Cytovillin, a microvillar Mr 75,000 protein cDNA sequence, prokaryotic expression and chromosomal localization. J. Biol. Chem. 264:16727-16732 (1990).

144. Funayama, N., Nagafuchi, A., Sato, N., Tsukita, S., and Tsukita, S. Radixin is a novel member of the band 4.1 family. J. Cell Biol. 115:1039-1048 (1991).

145. Lankes, W.T., and Furthermayr, H. Moesin: a member of the protein 4.1 talin-ezrin family of proteins. Proc. Natl. Acad. Sci., USA 88:8297-8301 (1991).

146. Furthermayr, H., Lankes, W., and Amieva, M. Moesin, a new cytoskeletal protein and constituent of filopodia: its role in cellular functions. Kidney Int. 41:665-670 (1992).

147. Sato, N., Funayama, N., Nagafuchi, A., Yonemura, S., Tsukita, S., and Tsukita, S. A gene family consisting of ezrin, radixin and moesin. Its specific localization at actin filament/plasma membrane association sites. J. Cell Sci. 103:131-143 (1992).

148. Sekido,Y., Pass, H.I., Bader, S., Mew, D.J.Y., Christman, M.F., Gazdar, A.F., and Minna, J.D. Neurofibromatosis type 2 (NF2) gene is somatically mutated in mesothelioma but not in lung cancer. Cancer Res. 55:1227-1231 (1955).

149. Bourn, D., Carter, S.A., Evans, D.G.R., Goodship, J., Coakham, H., and Strachan, T. A mutation in the neurofibromatosis type 2 suppressor gene, giving rise to widely different clinical phenotypes in two unrelated individuals. Am. J. Hum. Genet. 55:69-73 (1994).

150. Rubio, M.P., Correa, K.M., Ramesh, V., MacCollin, M.M., Jacoby, L.B., von Deimling, A., Gusella, J.F., and Louis, D.N. Analysis of the neurofibromatosis 2 gene in human ependymomas and astrocytomas. Cancer Res. 54:45-47 (1994).

151. Bianchi, A.B., Mitsunaga, S.I., Cheng, J.Q., Klein, W.M., Jhanwar, S.C., Seizinger, B., Kley, N., Klein-Szanto, A.J.P., and Testa, J.R. High frequency of inactivating mutations in the neurofibromatosis, type 2 (NF2) gene in primary malignant mesotheliomas. Proc. Natl. Acad. Sci., USA 92:10854-10858 (1995).

152. Lekanne Deprez, R., Reigman, P.H.J., Groen, N.A., Warringa, U.L., van Biezen, N.A., Molijn, A.C., Bootsma, D., DeJong, P.J., Menon, A.G., Kley, N.A., Seizinger, B.R., and Zwartoff, E.C. Cloning and characterization of MN1, a gene from chromosome 22q11, which is disrupted by a balanced translocation in meningioma. Oncogene 10:1521-1528 (1995).

153. Kosak, M. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. Nucl. Acids Res. 12:857-872 (1984).

154. Kosak, M. An analysis of 5' noncoding sequences from 699 vertebrate messenger RNAs. Nucl. Acids Res. 15:8125-8148 (1987).

155. Warren, S.T., and Nelson, D.L. Trinucleotide repeat expansions in neurological disease. Curr. Opin. Neurobiol. 3:752-759 (1993).

156. Wooster, R., Cleton-Jansen, A.M., Collins, N., Mangion, J., Cornelis, R.S., Cooper, C.S., Gusterson, B.A., Ponder, B.A.J., von Deimling, A., Wiestler, O.D., Cornelisse, C.J., Devilee, P., and Stratton, M.R. Instability of short tandem repeats (microsatellites) in human cancers. Nature Genetics 6:152-156 (1994).

157. Gerber, H.P., Seipel, K., Georgiev, O., Hofferer, M., Hug, M., Rusconi, S., and Schaffner, W. Transcriptional activation modulated by homopolymorphic glutamine and proline stretches. Science 263:808-811 (1994).

158. Ponnambalam, S., Robinson, M.S., Jackson, A.P., Peiperl, L., and Parham, P. Conservation and diversity in families of coated vesicle adaptins. J. Biol. Chem. 265:4814-4820 (1990).

159. Kirchhausen, U., Nathanson, K.L., Matsui, W., Vaisberg, A., Chow, E.P., Burne, C., Keen, J.H., and Davis, A.E. Structural and functional division into two domains of the large (100- to 115-kDa) chains of the clathrin-associated protein complex AP-2. Proc. Natl. Acad. Sci., USA 86:2612-2616 (1990).

160. Goldstein, J.L., Brown, M.S., Anderson, R.G.W., Russell, D.W., and Schneider, W.J. Receptor-mediated endocytosis: concepts emerging from the LDL receptor system. Ann. Rev. Cell Biol. 1:1-39 (1985).

161. Burgess, T.L., and Kelly, R.B. Constutive and regulated secretion of proteins. Ann. Rev. Cell Biol. 3:243-293 (1987).

162. Robinson, M.S. Cloning and Expression of γ-adaptin, a component of clathrin coated vesicles associated with the golgi apparatus. J. Cell Biol. 111:2319-2326 (1990).

163. Pearse, B.M.F. Receptors compete for adaptors found in plasma membrane coated pits. EMBO J. 7:3331-3336 (1988).

164. Glickman, J.N., Conibear, E., and Pearse, B.M.F. Specificity of binding of clathrin adaptors to signals on the mannose-6-phosphate/insulin-like growth factor II receptor. EMBO J. 8:1041-1047 (1989).

165. Keen, J.H., Willingham, M.C., and Pastàn, I.H. Clathrin coated vesicles: isolation, dissociation and factor-dependent reassociation of clathrin baskets. Cell 16:303-312 (1979).

166. Pearse, B.M.F., and Robinson, M.S. Purification and properties of 100kD proteins from coated vesicles and their reconstitution with clathrin. EMBO J. 3:1951-1957 (1984).

167. Keen, J.H. Clathrin assembly proteins: affinity purification and a model for coat assembly. J. Cell Biol. 105:1989-1998 (1987).

168. Keen, J.H., and Beck, K.A. Identification of the clathrin binding domain of assembly protein AP-2. Biochem. Biophys. Res. Commun. 158:17-23 (1989).

169. Jackson, A.P., Seow, H.F., Holmes, N., Drickamer, K., and Parham, P. Clathrin light chains contain brain-specific insertion sequences and a region of homology with intermediate filaments. Nature 326:154-159 (1987).

170. Kirchhausen, T., Scarmato, P.. Harrison, S.C., Monroe, J.J., Chow, E.P., Mattaliano, R.J., Ramachandran, K.L., Smart, J.E., Ahn, A.H., and Brosius, J. Clathrin light chains LCA and LCB are similar, polymorphic, and share repeated heptad motifs. Science 236:320-324 (1987).

171. Jackson, A.P., and Parham, P. Structure of human clathrin light chains: conservation of light chain polymorphism in three mammalian species. J. Biol. Chem. 263:16688-16695 (1988).

172. Roe, B.A., and Dumanski, J.P. Personal communication, 1996.

173. Schachenmann, G., Schmid, W., Fraccaro, M., Mannini, A., Tiepolo, L., Perona, G.P., and Sartori, E. Chromosomes in coloboma and anal atresia. The Lancet 2:290 (1965).

174. Haab, O. Beitrage zu den angeborenen fehlern des auges. Albrecht von Graefes Arch. Opthalmol. 24:257-284 (1878).

175. Gerald, P.S., Davis, C., Say, B., and Wilkins, J. Syndromal association of imperforate anus: the Cat Eye syndrome. Birth Defects Orig. Art. Ser VIII(2):79-84 (1972).

176. Neu, R.L., Assemany, S.R., and Gardner, L.I. Cat eye syndrome with normal chromosomes. Lancet 1:949 (1970).

177. Franklin, R.C., and Parslow, M.I. The cat eye syndrome. Review and two further cases occurring in female siblings with normal chromosomes. Acta Paediatr. Scand. 61:581-586 (1972).

178. Peterson, R.A. Schmidt-Fraccaro syndrome (cat's eye syndrome). Arch. Opthalmol. 90:287-291 (1973).

189. Cory, C.C., and Jamison, D.L. The cat eye syndrome. Arch. Opthalmol. 92:259-262 (1974).

180. Toomey, K.E., Mohandas, T., Leisti, J., Szalay, G., and Kaback, M.M. Further delineation of the supernumerary chromosome in the cat eye syndrome. Clin. Genet. 12:275-284 (1977).

181. Zellweger, H., Mikamo, K., and Abbo, G. Two cases of multiple malformations with an autosomal chromosomal aberration--partial trisomy D? Helv. Paediat. Acta. 17:290-300 (1962).

182. Beyer, P., Ruch, J.V., Rumpler, Y., and Girard, J. Observation d'un enfant debile mental et polymalforme dont le caryotype montre le presence d'un petit extra chromosome mediocentrique. Pediatrie 23:439-442 (1968).

183. Buhler, E.M., Mehes, K., Muller, H.J., and Stalder, G.R. Cat eye syndrome, a partial trisomy 22. Humangenetik 15:150-162 (1972).

184. Noel, B., and Quack, B. Petit metacentrique surnumeraire chez du polymalforme. J. Genet. Hum. 18:45-56 (1970).

185. Pfeiffer, R.A., Heimann, K., and Heiming, E. Extra chromosome in cat eye syndrome. Lancet 1:97 (1970).

186. Gerald, P.S., Davis, D., Say, B.M., and Wilkins, J.L. A novel chromosomal basis for imperforate anus (the cat eye syndrome). Ped. Res. 2:297 (1968).

187. Ginsberg, J., Dignan, P., and Soukup, S. Ocular abnormality associated with extra small autosome. Am. J. Ophthalmol. 65:740-746 (1968).

188. Weber, F.M., Dooley, R.R., and Sparkes, R.S. Anal atresia, eye anomalies, and an additional small abnormal acrocentric chromosome (47, XX, mar+): report of a case. J. Pediatr. 76:594-597 (1970).

189. Jacobsen, P., Mikkelsen, M., Froland, A., and DuPont, A. Familial transmission of a translocation between two non-homologous large acrocentric chromosomes. Clinical, cytogenetic and autoradiographic studies. Ann. Hum. Genet. 29:391-402 (1966).

190. Niebuhr, E. Dicentric and monocentric Robertsonian translocations. Humangenetik 16:217-226 (1972).

191. Petit, P. Identifying the extra chromosome in "Cat Eye syndrome" with Q, GC techniques. Bull. Europ. Soc. Hum. Genet. October, 70-73 (1973).

192. De Chieri, R., Malfatti, C., Stanchi, F., and Albores, J.M. Cat eye syndrome: Evaluation of the extra chromosome with banding techniques: case report. J. Genet. Hum. 22:101-107 (1974).

193. Kunze, J., Tolksdorf, M., and Wiedemann, H.R. Cat eye syndrome. Hum. Genet. 26:271-289 (1975).

194. Schinzel, A., Schmid, W., Fraccaro, M., Tiepolo, L., Zuffardi, O., Opitz, J.M., Lindsten, J., Zetterqvist, P., Enell, H., Baccichetti, C., Tenconi, R., and Pagan, R.A. The "Cat Eye syndrome": dicentric small marker chromosome probably derived from no. 22 (tetrasomy 22pter -> q11) associated with a characteristic phenotype. Hum. Genet. 57:148-158 (1981).

195. Niermeijer, M.F., Sachs, E.S., Johodova, M., Tichelaar-Klepper, C., Kleijer, W.J., and Galjaard, H. Prenatal diagnosis of genetic disorders. J. Med. Genet. 13:182-194 (1976).

196. Wisniewski, L. Hasold, T., Heffelfinger, J., and Higgins, J.V. Cytogenetic and clinical studies in five cases of inv dup (15). Hum. Genet. 50:259-270 (1979).

197. Schinzel, A., Schmid, W., Auf der Maur, P., Moser, H., Degenhardt, K.H., Geisler, M., and Grubisic, A. Incomplete Trisomy 22. I. Familial 11/22 translocation with 3:1 meiotic disjunction. Delineation of a common clinical picture and report of nine new cases from six families. Hum. Genet. 56:249-262 (1981).

198. Schinzel, A. Incomplete Trisomy 22. II. Familial trisomy of the distal segment of chromosome 22q in two brothers from a mother with a translocation, t(6;22)(q27;q13). Hum. Genet. 56:263-268 (1981).

199. Schinzel, A. Incomplete Trisomy 22. III. Mosaic-trisomy 22 and the problems of full trisomy 22. Hum. Genet. 56:269-273 (1981).

200. Hoo, J.J., Robertson, A., Fowlow, S.B., Bowen, P., Lin, C.C. Letter to the editor: Inverted duplication of 22pter -> q11.21 in Cat Eye Syndrome. Am. J. Med. Genet. 24:543-545 (1986).

201. Ing, P.S., Lubinsky, M.S., Smith, S.D., Golden, E., Sanger, W.G., and Duncan, A.M.V. Cat-Eye Syndrome with different marker chromosomes in a mother and daughter. Am. J. Med. Genet. 26:621-628 (1987).

202. Magenis, R.E., Sheehy, R.R., Brown, M.G., McDermid, H.E., White, B.N., Zonana, J., and Weleber, R. Parental origin of the extra chromosome in the Cat Eye syndrome: Evidence from heteromorphism and *in situ* hybridization analysis. Am. J. Med. Genet. 29:9-19 (1988).

203. Liehr, T., Pfeiffer, R.A., and Trautman, U. Typical and partial cat eye syndrome: identification of the marker chromosome by FISH. Clin. Genet. 42:91-96 (1992).

204. Guanti, G. The aetiology of the cat eye syndrome reconsidered. J. Med. Genet. 18:108-118 (1981).

205. Reiss, J.A., Weleber, R.G., Brown, M.G., Bangs, C.D., Lovrien, E.W., and Magenis, R.E. Tandem duplication of proximal 22q: A cause of cat eye syndrome. Am. J. Med. Genet. 20:165-171 (1985).

206. Bruns, G.A., Mintz, B.J., Leary, A.C., Regina, V.M., and Gerald, P.S. Expression of human arylsulfatase-A in man-hamster somatic cell hybrids. Cytogenet. Cell Genet. 22:182-185 (1978).

207. Hors-Cayla, M.C., Henertz, S., Van Cong, N., Weil, D., and Frezal, J. Confirmation of the assignment of the gene for arylsulfatase-A to chromosome 22 using somatic cell hybrids. Hum. Genet. 49:33-39 (1979).

208. Fryns, J.P., Jaeken, J., van den Berghe, H. Partial trisomy 22q with elevated arylsulfatase-A activity. Ann. Genet. 22:168-170 (1979).

209. Hors-Cayla, M.C., Junien, C., Henertz, S., Mattei, J.F., and Frezal, J. Regional assignment of arylsulfatase-A , mitochondrial aconitase and NADH-cytochrome b5 reductase by somatic cell hybridization. Hum. Genet. 58:140-143 (1981).

210. McDermid, H.E., Duncan, A.M.V., Brasch, K.R., Holden, J.J.A., Magenis, R.E., Sheehy, R., Burn, J., Kardon, N., Noel, B., Schinzel, A., Teshima, I., and White, B.N. Characterization of the supernumerary chromosome in cat eye syndrome. Science 232:646-648 (1986).

211. Magenis, R.E., Sheehy, R.R., Brown, M.G., McDermid, H.E., White, B.N., Zonana, J., and Weleber, R. Parental origin of the extra chromosome in the cat eye syndrome: evidence from heteromorphism and *in situ* hybridization analysis. Am. J. Med. Genet. 29:9-19 (1988).

212. Butler, M.G., Meany, F.J., and Palmer, C.G. Clinical and cytogenetic survey of 39 individuals with Prader-Labhard-Willi syndrome. Am. J. Med. Genet. 23:793-809 (1986).

213. Olsen, S.B., and Magenis, R.E. Preferential paternal origin of *de novo* chromosome structural rearrangements. In Daniel, A. (ed): "Cytogenetics of Mammalian Autosomal Rearrangements." Alan R. Liss, Inc. Publishing, New York (1987).

214. Duncan, A.M.V., Hough, C.A., White, B.N., and McDermid, H.E. Breakpoint localization of the marker chromosome associated with the cat eye syndrome. Am. J. Hum. Genet. 38:978-980 (1986).

215. Mears, A.J., Duncan, A.M.V., Budarf, M., Emanuel, B.S., Sellinger, B., Siegel-Bartelt, J., Greenberg, C.R., and McDermid, H.E. Molecular characterization of the marker chromosome associated with cat eye syndrome. Am. J. Hum. Genet. 55:134-142 (1994).

216. Ward, J., Sierra, I.A., and D'Croz, E. Cat eye syndrome associated with aganglionosis of the small and large intestine. J. Med. Genet. 26:647-648 (1989).

217. Luleci, G., Bagci, G., Kivran, M., Luleci, E., Bektas, S., and Basaran, S. A hereditary bisatellite-dicentric supernumerary chromosome in a case of cat eye syndrome. Hereditas 111:7-10 (1989).

218. DiGeorge, A.M. Discussions on a new concept of the cellular basis of immunology. J. Pediatr. 67:907-908 (1965).

219. Robinson, H.B. DiGeorge's or the III-IV pharyngeal pouch syndrome: pathology and a theory of pathogenesis. Perspect. Pediatr. Pathol. 2:173-206 (1975).

220. Conley, M.E., Beckwith, J.B., Mancer, J.F.K., and Tenckhoff, L. The spectrum of the DiGeorge syndrome. J. Pediatr. 94:883-890 (1979).

221. Greenberg, F., Elder, F.F.B., Haffner, P., Northrup, H., and Ledbetter, D.H. Cytogenetic findings in a prospective series of patients with DiGeorge anomaly. Am. J. Hum. Genet. 43:605-611 (1988).

222. Heisterkamp, N., Mulder, M.P., Langerveld, A., ten Hoeve, J., Wang, Z., Roe, B.A., and Groffen, J. Localisation of the human mitochondrial citrate transporter protein gene to chromosome 22q11 in the DiGeorge Syndrome critical region. Genomics 29:451-456 (1995).

223. Driscoll, D.A., Budarf, M.L., and Emanuel, B.S. A genetic etiology for DiGeorge syndrome: consistent deletions and microdeletions of 22q11. Am. J. Hum. Genet. 50:924-933 (1992).

224. Driscoll, D.A., Spinner, N.B., Budarf, M.L., McDonald-McGinn, D.M., Zackai, E.H., Goldberg, R.B., Shprintzen, R.J., Saal, H.M., Zonana, J., Jones, M.C., Mascarello, J.T., and Emanuel, B.S. Deletions and microdeletions of 22q11.2 in velo-cardio-facial syndrome. Am. J. Med. Genet. 44:261-268 (1992).

225. Kelly, D., Goldberg, R., Wilson, D., Lindsay, E., Carey, A.H., Goodslip, J., Burn, J., Cross, I., Shprintzen, R.J., and Scrambler, P.J. Confirmation that the velo-cardio-facial syndrome is associated with the haploinsufficiency of genes at chromosome 22q11. Am. J. Med. Genet. 45:308-312 (1993).

226. Wermer, P. Genetic aspects of adenomatosis of endocrine glands. Am. J. Med. 16:363-371 (1954).

227. Wermer, P. Multiple endocrine adenomatosis; multiple hormone-producing tumors, a familial syndrome. Clin. Gastroenterol. 3:671-684 (1974).

228. Calendar, A., Giraud, S., Cougard, P., Chanson, P., Lenoir, G., Murat, A., Hamon, P., and Proye, C. Multiple endocrine neoplasia, type 1 in France: clinical and genetic studies. J. Int. Med. 238:263-268 (1995).

229. Pang, J.T., and Thakker, R.V. Multiple endocrine neoplasia, type 1 (MEN-1). Eur. J. Cancer 30A:1961-1968 (1994).

230. Friedman, E., Sakaguchi, K., Bale, A.E., Falchetti, A., Streeten, E., Zimering, M.B., Weinstein, L.S., McBride, W.O., Nakamura, Y., Brandi, M.L., Norton, J.A., Aurbach, G.D., Spiegel, A.M., and Marx, S.J. Clonality of parathyroid tumors in familial multiple endocrine neoplasia, type 1. New Engl. J. Med. 321:213-218 (1989).

231. Bystrom, C., Larsson, C., Blomberg, C., Sandelin, K., Falkmer, U., Skogseid, B., Oberg, K., Werner, S., and Nordenskjold, M. Localisation of the MEN1 gene to a small region within chromosome 11q13 by deletion mapping in tumors. Proc. Natl. Acad. Sci., USA 87:1968-1972 (1990).

232. Friedman, E., Bale, A.E., Marx, S.J., Norton, J.A., Arnold, A., Tu, T., Aurbach, G.D., and Spiegel, A.M. Genetic abnormalities in sporadic parathyroid adenomas. J. Clin. Endocrinological Metabol. 71:481-515 (1990).

233. Thakker, R.V., Bouloux, P., Wooding, C., Chotai, K., Broad, P.M., Spurr, N.K., Besser, G.M., and O'Riordan, J.L. Association of parathyroid tumors in multiple endocrine neoplasia, type 1 with loss of alleles on chromosome 11. New Engl. J. Med. 321:218-224 (1989).

234. Teh, B.T., Cardinal, J., Shepherd, J., Hayward, N.K., Weber, G., Cameron, D., and Larsson, C. Genetic mapping of the multiple endocrine neoplasia, type 1 locus at 11q13. J. Int. Med. 238:249-253 (1995).

235. Lagercrantz, J., Larsson, C., Grimmond, S., Skogseid, B., Gobl, A., Friedman, E., Carson, E., Phelan, C., Oberg, K., Nordenskjold, M., Hayward, N.K., and Weger, G. Candidate genes for multiple endocrine neoplasia, type 1. J. Int. Med. 238:245-248 (1995).

236. Arnold, A., Kim, H.G., Gaz, R.D., Eddy, R.L., Fukushima, Y., Byers, M.G., Shows, T.B., and Kronenberg, H.M. Molecular cloning and chromosomal mapping of DNA rearranged with parathyroid hormone gene in a parathyroid adenoma. J. Clin. Invest. 83:2034-2040 (1989).

237. Rosenberg, C.L., Kim, H.G., Shows, T.B., Kronenberg, H.M., and Arnold, A. Rearrangements and overexpression of D11S287E, a candidate oncogene on chromosome 11q13 in benign parathyroid tumors. Oncogene 6:449-453 (1991).

238. Bale, A.E., Wong, E., and Arnold, A. The parathyroid breakpoint locus on 11q13 maps close to BCL1 and is not a candidate gene for MEN1. Am. J. Hum. Genet. 47: A3 (abstract) (1990).

239. Roberts, J.M., Koff, A., Polyak, K., Firpo, E., Collins, S., Ohtsubo, M., and Massagne, J. Cyclins, cdks and cyclin kinase inhibitors. Cold Springs Harbor Symp. Quant. Biol. LIX:31-38 (1994).

240. Sherr, C.J., Kato, J., Quelle, D.E., Matsuoka, M., and Roussel, M.F. D-type cyclins and their cyclin-dependent kinases: G1 phase integrators of the mitogenic response. Cold Springs Harbor Symp. Quant. Biol. LIX:11-20 (1994).

241. Sherr C.J. Mammalian G1 cyclins. Cell 73:1059-1065 (1993).

242. Erikson, J., Finan, J., Tsujimoto, Y., Nowell, P.C., and Croce, C.M. The chromosome 14 breakpoint in neoplastic B cells with the t(11;14) translocation involves the immunoglobulin heavy chain locus. Proc. Natl. Acad. Sci., USA 81:4144-4148 (1984).

243. Tsujimoto Y., Yunis, J., Onorato-Showe, L., Erikson, J., Nowell, P.C., and Croce, C.M. Molecular cloning of the chromosomal breakpoint of B-cell lymphomas and leukemias with the t(11;14) chromosomal translocation. Science 224:1403-1406 (1984).

244. Kas, K., Schoenmakers, E., van de Ven, W., Weber, G., Nordenskjold, M., Michiels, L., Merregaert, J., and Larsson, C.. Assignment of human FAU gene to a subregion of chromosome 11q13. Genomics 17:387-392 (1993).

245. Kas, K., Weber, G., Merregaert, J., Michiels, L., Sandelin, K., Skogseid, B., Thompson, N., Nordenskjold, M., Larsson, C., and Friedman, E. Exclusion of the FAU gene as the multiple endocrine neoplasia, type 1 (MEN1) gene. Hum. Mol. Genet. 2:349-353 (1993).

246. Gottesdiener, K.M., Karpinski, B.A., Lindsten, T., Strominger, J.L., Jones, N.H., Thompson, C.B., and Lieden, J.M. Isolation and structural characterization of the human 4F2 heavy chain gene, an inducible gene involved in T-lymphocyte activation. Mol. Cell Biol. 8:3809-3819 (1988).

247. Rochelle, J.M., Watson, M.L., Oakey, R.L., and Seldin, M.F. A linkage map of mouse chromosome 19: definition of comparative mapping relationships with human chromosomes 10 and 11 including the MEN1 locus. Genomics 14:26-31 (1992).

248. Brandi, M.L., Aurbach, G.D., Fitzpatrick, L.A., Quarto, R., Spiegel, A.M., Bliziotes, M.M., Nortan, J.A., Doppman, J.L., and Marx, S.J. Parathyroid mitogenic activity in plasma from patients with familial multiple endocrine neoplasia type 1. New Engl. J. Med. 314:1287-1293 (1986).

249. Zimering, M.B., Brandi, M.L., DeGrange, D.A., Marx, S.J., Streeten, E., Katsumata, N., Murphy, P.R., Sato, Y., Friesen, H.G., and Aurbach, G.D. Circulating fibroblast growth factor-like substance in familial multiple endocrine neoplasia type 1. J. Clin. Endocrinol. Metab. 70:149-154 (1990).

250. Alberts, A.S., Thorburn, A.M., Shenolikar, S., Mumby, M.C., and Feramisco, A. Regulation of cell cycle progression and the nuclear affinity of the retinoblastoma protein by protein phosphatases. Proc. Natl. Acad. Sci., USA 90:388-392 (1993).

251. Ludlow, J.W., Glendening, C.L., Livingston, D.M., and DeCaprio, J.A. Specific enzymatic dephosphorylation of the retinoblastoma protein. Mol. Cell Biol. 13:367-372 (1993).

252. Ayala, F.J., and Kiger, J.A., Jr. "Modern Genetics." The Benjamin/Cummings Publishing Co., Inc. Menlo Park, CA (1980).

253. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J.D. "Molecular Biology of the Cell, third edition." Garland Publishing, New York (1989).

254. Singer, M., and Berg, P. The inheritance of single traits in "Genes and Genomes" University Science Books, Mill Valley, CA. p.17-22 (1991).

255. Sutton, W.S. The chromosomes in heredity. Biol. Bulletin 4:231-251 (1903).

256. Johannsen, W. Elemente der exakten erblichkeitslehre. Fischer. Jena (1909).

257. Garrod, A.E. Inborn errors of metabolism. Lancet 2:1616-1620 (1902).

258. Garrod, A.E. "Inborn Errors of Metabolism." Frowde, Hodder, and Stoughton, London (1909).

259. Beadle, G.W., and Tatum, E.L. Genetic control of biochemical reactions in *Neurospora*. Proc. Natl. Acad. Sci., USA 27:499-506 (1941).

260. Griffith, F. The significance of Pneumococcal types. J. Hyg. 27:113-159 (1928).

261. Avery, O.T., MacLeod, C.M., and McCarty, M. Studies on the chemical nature of the substance inducing transformation of Pneumococcal types. J. Exp. Med. 79:137-158 (1944).

262. Hershey, A.D., and Chase, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. J. Gen. Physiol. 36:39-52 (1952).

263. Meicher, F. Uber die chemische zusammensetzung der eiterzellen. Hoppe-Seyler's Medizinish-Chemoscen Untersuchungen 4:441-460 (1871).

264. Kornberg, A. "DNA Replication." W.H. Freeman and Co. San Franscisco, CA. (1980).

265. Chargaff, E., Lipschitz, R., Green, C., and Hodes, M.E. The composition of the deoxyribonucleic acid of salmon sperm. J. Biol. Chem. 192:223-230 (1951).

266. Watson, J.D., and Crick, F.H.C. Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. Nature 171:737-738 (1953).

267. Watson, J.D., and Crick, F.H.C. Genetical implications of the structure of deoxyriboneucleic acid. Nature 171:964-967 (1953).

268. Wilkins, M.H.F., Stokes, A.R., and Wilson, H.R. Molecular structure of deoxypentose nucleic acids. Nature 171:738-740 (1953).

269. Franklin, R.E., and Gosling, R.G. Molecular configuration in sodium thymonucleate. Nature 171:740-742 (1953).

271. Kang, H., and Johnson, W.C., Jr. Infrared linear dicroism reveals that A-, B- and C-DNAs in films have bases highly inclined from perpendicular to the helix axis. Biochemistry 33:8330-8338 (1994).

272. Wang, A.H.J., Quigley, G.J., Kolpak, F.J., Crawford, J.L., van Boom, J.H., van der Marel, G., and Rich, A. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. Nature 282:680-686 (1979).

273. Rich, A., Nordheim, A., and Wang, A.H.-J. The chemistry and biology of left handed Z-DNA. Ann. Rev. Biochem. 53:791-846 (1984).

274. Micklos, D.A., and Freyer, G.A. "DNA Science: A first course in recombinant DNA technology." Cold Springs Harbor Laboratory Press, New York (1990).

275. Haniford, D.B., and Pulleyblank, D.E. Facile transition of poly [d(TG)•d(CA)] into a left handed helix in physiological conditions. Nature 302:632-634 (1983).

276. Nordheim, A., and Rich, A. The sequence $(dC\text{-}dA)_n \bullet (dG\text{-}dT)_n$ forms left handed Z-DNA in negatively supercoiled plasmids. Proc. Natl. Acad. Sci., USA 80:1821-1825 (1983).

277. Nordheim, A., and Rich, A. Negatively supercoiled simian virus 40 DNA contains Z-DNA segments within transcriptional enhancer sequences. Nature 303:674-679 (1983).

278. Rich, A., Nordheim, A., and Azorin, F. Stabilization and detection of natural left-handed Z-DNA. J. Biomol. Struct. Dyn. 1:1-19 (1983).

279. Miller, F.D., Rattner, J.B., and van de Sande, J.H. Nucleosome-core assembly on B and Z forms of poly[d(G-m^5C)]. Cold Spring Harbor Symp. Quant. Biol. 47:571-576 (1983).

280. Smith, G.R. DNA supercoiling: another level for regulating gene expression. Cell 24:191-193 (1981).

281. Rich, A. Right-handed and left-handed DNA: conformational information in genetic material. Cold Spring Harbor Symp. Quant. Biol. 47:1-12 (1983).

282. Weil, P.A., Luse, D.S., Segall, J., and Roeder, R.G. Selective and accurate initiation of transcription at the Ad2 major late promoter in a soluble system dependent on purified RNA polymerase II and DNA. Cell 18:469-484 (1979).

283. Manley, J.L., Fire, A., Cano, A., Sharp, P.A., and Gefter, M.L. DNA-dependent transcription of adenovirus genes in a soluble whole-cell extract. Proc. Natl. Acad. Sci., USA 77:3855-3859 (1980).

284. Leff, S.E., and Rosenfeld, M.G. Complex transcriptional units: diversity in gene expression by alternative RNA processing. Annu. Rev. Biochem. 55:1091-1117 (1986).

285. Breathnach, R., and Chambon, P. Organization and expression of eucaryotic split genes coding for proteins. Annu. Rev. Biochem. 50:349-383 (1981).

286. Dynan, W.S., and Tijan, R. Control of eukaryotic messenger RNA synthesis by sequence-specific DNA-binding proteins. Nature 316:774-778 (1985).

287. Levine, A., Cantoni, G.L., and Razin, A. Inhibition of promoter activity by methylation: possible involvement of protein mediators. Proc. Natl. Acad. Sci., USA 88:6515-6518 (1991).

288. Holler, M., Westin, G., Jiricny, J., and Schaffner, W. Sp1 transcription factor binds DNA and activates transcription even when the binding site is CpG methylated. Genes Dev. 2:1127-1135 (1988).

289. Nevins, J.R. The pathway of eukaryotic mRNA formation. Annu. Rev. Biochem. 52:441-446 (1983).

290. Nevins, J.R., and Darnell, J.E., Jr. Steps in the processing of Ad2 mRNA: poly(A)+ nuclear sequences are conserved and poly (A) addition precedes splicing. Cell 15:1477-1493 (1978).

291. Proudfoot, N.J., and Brownlee, G.G. 3' non-coding region sequences in eukaryotic messenger RNA. Nature 263:211-214 (1976).

292. Montell, C., Fisher, E.F., Caruthers, M.H., and Berk, A.J. Inhibition of RNA cleavage but not polyadenylation by a point mutation in mRNA 3' consensus sequence AAUAAA. Nature 305:600-605 (1983).

293. Mount, S.M. A catalogue of splice juction sequences. Nucl. Acids Res. 10:459-472 (1982).

294. Maniatis, T., and Reed, R. The role of small nuclear ribonucleoprotein particles in pre-mRNA splicing. Nature 325:673-678 (1987).

295. Lerner, M.R., Boyle, J.A., Mount, S.M., Wolin, S.L., and Steitz, J.A. Are snRNPs involved in splicing? Nature 283:220-224 (1980).

296. Rogers, J., Wall, R. A mechanism for RNA splicing. Proc. Natl. Acad. Sci., USA 77:1877-1879 (1980).

297. Mount, S.M., and Steitz, J.A. Sequence of U1 RNA from *Drosophila melanogaster*: implications for U1 secondary structure and possible involvement in splicing. Nucl. Acids Res. 9:6351-6368 (1981).

298. Chabot, B., Black, D.L., LeMaster, D.M., and Steitz, J.A. The 3' splice site of pre-messenger RNA is recognized by a small nuclear ribonucleoprotein. Science 230:1344-1349 (1985).

299. Black, D.L., Chabot, B., and Steitz, J.A. U2 as well as U1 small nuclear ribonucleoproteins are involved in pre-messenger RNA splicing. Cell 42:737-750 (1985).

300. Lesser, C.F., and Guthrie, C. Mutations in U6 snRNA that alters splice site specificity: Implications for the active site. Science 262:1982-1988 (1993).

301. Sontheimer, E.J., and Steitz, J.A. The U5 and U6 small nuclear RNAs as active site components of the spliceosome. Science 262:1989-1996 (1993).

302. Kornberg, R.D. Chromatin structure, a repeating unit of histones and DNA. Science 184:868-871 (1974).

303. Finch, J.T., and Klug, A. Solenoidal model for superstructure in chromatin. Proc. Natl. Acad. Sci., USA 73:1897-1901 (1976).

304. Pederson, D.S., Thoma, F., and Simpson, R.T. Core particle fiber and transcriptionally active chromatin structure. Annu. Rev. Cell Biol. 2:117-147 (1986).

305. Widom, J., and Klug, A. Structure of the 300 Å chromatin filament: X-ray diffraction from oriented samples. Cell 43:207-213 (1985).

306. Sobell, H.M., Tsai, C.C., Jain, S.C., and Sakore, T.D. Conformational flexibility in DNA structure and its implications in understanding the organization of DNA in chromatin. Philos. Trans. R. Soc. Lond. Biol. 283:295-298 (1978).

307. Voet, D., and Voet, J.G. Chromosome structure in "Biochemistry," John Wiley and Sons, NY pp. 1032-1041 (1990).

308. Moore, K.L., and Barr, M.L. Smears from the oral mucosa in the detection of chromosomal sex. Lancet 2:57 (1955).

309. Smith, H.O., and Wilcox, K.W. A restriction enzyme from *Haemophilus influenzae*. I. Purification and general properties. J. Mol. Biol. 51:379-391 (1970).

310. Messing, J., Gronenborn, B., Muller-Hill, B., and Hofschneider, P.H. Filamentous coliphage M13 as a cloning vehicle: Insertion of a Hind II fragment of the lac regulatory region in M13 replicative form *in vitro*. Proc. Natl. Acad. Sci., USA 74:3642-3646 (1977).

311. Gronenborn, B., and Messing, J. Methylation of single-stranded DNA *in vitro* introduces new restriction endonuclease cleavage sites. Nature 272:375-377 (1978).

312. Messing, J., and Vieira, J. A new pair of M13 vectors for selecting either DNA strand of double-digest restriction fragments. Gene 19:269-276 (1982).

313. Messing, J. New M13 vectors for cloning. Meth. Enzymol. 101:20-78 (1983).

314. Yanish-Perron, C., Vieira, J., and Messing, J. Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. Gene 33:103-119 (1985).

315. Maxam, A.M., and Gilbert, W. A new method for sequencing DNA. Proc. Natl. Acad. Sci., USA 74:560-564 (1977).

316. Sanger, F., Nicklen, S., and Coulson, A.R. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci., USA 74:5463-5467 (1977).

317. Lee, L.G., Connell, C.R., Woo, S.L., Cheng, R.D., McArdle, B.F., Fuller, C.W., Halloran, N.D., and Wilson, R.K. DNA sequencing with dye-labeled

terminators and T7 DNA polymerase: effect of dyes and dNTPs on incorporation of dye-terminators and probability analysis of termination fragments. Nucl. Acids Res. 20:2471-2483 (1992).

318. Tabor, S., and Richardson, C.C. A single residue in DNA polymerases of the *Escherichia coli* DNA polymerase I family is critical for distinguishing between deoxy- and dideoxynucleotides. Proc. Natl. Acad. Sci., USA 92:6339-6343 (1995).

319. Korolev, S., Nayal, M., Barnes, W.M., DiCera, E. and Waksman, G. Crystal structure of the large fragment of *Thermus aquaticus* DNA polymerase I at 2.5-Å resolution: structural basis for thermostability. Proc. Natl. Acad. Sci., USA 92:9264-9268 (1995).

320. Chen, J., Saheta, A., Stambrook, P.J., and Tischfield, J.A. Polymerase chain reaction amplification and sequence analysis of human mutant adenine phosphoribosyl transferase genes: the nature and frequency of errors caused by Taq DNA polymerase. Mutat. Res. 249:169-176 (1991).

321. Holmes, D.S., and Quigley, M. A rapid boiling method for the preparation of bacterial plasmids. Anal. Biochem. 114:193-197 (1981).

322. Chen, E.Y., and Seeburg, P.H. Supercoil sequencing: a fast and simple method for sequencing plasmid DNA. DNA 4:165-176 (1985).

323. Siemenak, D.R., Sieu, L.C., and Slightom, J.L. Strategy and methods for directly sequencing cosmid clones. Anal. Biochem. 192:441-448 (1991).

324. Bachmann, B., Luke, W., and Hunsmann, G. Improvement of PCR amplified DNA sequencing with the aid of detergents. Nucl. Acids Res. 18:1309 (1990).

325. Craxton, M. Linear amplification sequencing, a powerful method for sequencing DNA. Methods: Companion Methods Enzymol. 3:20-26 (1991).

326. Chissoe, S.L., Wang, Y.F., Clifton, S.W., Ma, N., Sun, H.J., Lobsinger, J.S., Kenton, S.M., White, J.D., and Roe, B.A. Strategies for rapid and accurate DNA sequencing. Methods: Companion Methods Enzymol. 3:55-65 (1991).

327. McBride, L.J., Koepf, S.M., Gibbs, R.A., Salser, W., Mayrand, P.E., Hunkapiller, M.W., and Kronick, M.N. Automated DNA sequencing methods involving polymerase chain reaction. Clin. Chem. 35:2196-2201 (1989).

328. Protocol included with the ABI PRISM Taq-FS double stranded DNA sequencing kit.

329. Anderson, S., Bankier, A.T., Barrell, B.G., deBruijn, M.H.L., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., Schreier, P.H., Smith, A.J.H., Staden, R., and Young, I.G. Sequence and organization of the human mitochondrial genome. Nature 290:457-465 (1981).

330. Anderson, S., deBruijn, M.H., Coulson, A.R., Eperon, I.C., Sanger, F., Young I.G. Complete sequence of bovine mitochondrial DNA. Conserved features of the mammalian mitochondrial genome. J. Mol. Biol. 156:683-717 (1982).

331. Roe, B.A., Ma, D.P., Wilson, R.K., and Wong, J.F. The complete nucleotide sequence of the *Xenopus laevis* mitochondrial genome. J. Biol. Chem. 260:9759-9774 (1985).

332. Henikoff, S. Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene 28:351-359 (1984).

333. Bankier, A.T., and Barrell, B.G. in Nucleic Acids Sequencing: A practical approach, C.J. Howe and E.S. Ward, eds. IRL Press, Oxford, pp.37-78 (1989).

334. Deininger, P.L. Random subcloning of sonicated DNA: application to shotgun sequence analysis. Anal. Biochem. 129:216-223 (1983).

335. Bankier, A.T., Weston, K.M., and Barrell, B.G. Random cloning and sequencing by the M13/dideoxynucleotide chain termination method. Meth. Enzymol. 155:51-93 (1987).

336. Bodenteich, A., Chissoe, S., Wang, Y.F., and Roe, B.A. Shotgun cloning as the strategy of choice to generate templates for high-throughput dideoxynucleotide sequencing. In "Automated DNA Sequencing and Analysis Techniques." C. Venter, ed. Academic Press, London, UK (1994).

337. Surzycki, S. Department of Biology, Indiana University, personal communication (1994).

338. Phadnis, S.H., Huang, H.V., and Berg, D.E. Tn5supF, a 264 base pair transposon derived from Tn5 for insertional mutagenesis and sequencing DNAs cloned in phage λ. Proc. Natl. Acad. Sci., USA 86:5908-5912 (1989).

339. Ordahl, C.P., Johnson, T.R., and Caplan, A.I. Sheared DNA fragment sizing: comparison of techniques. Nucl. Acids Res. 3:2985-2999 (1976).

340. Fitzgerald, M.C., Skowron, P., Van Etten, J.L., Smith, L.M., and Mead, D.A. Rapid shotgun cloning utilizing the two base recognition endonuclease *Cvi*JI. Nucl. Acids Res. 20:3753-3762 (1992).

341. Sanger, F., Coulson, A.R., Hong, G.F., Hill, D.F., and Petersen, G.B. Nucleotide sequence of bacteriophage lambda DNA. J. Mol. Biol. 162:729-773 (1982).

342. Baer, R., Bankier, A.T., Biggins, M.D., Deininger, P.L., Farrell, P.J., Gibson, T.J., Hatfull, G., Hudson, G.S., Satchwell, S.C., Seguin, C., Tuffnell, P.S., and Barrell, B.G. DNA sequence and expression of the B95-8 Epstein-Barr virus genome. Nature 310:207-211 (1984).

343. Clarke, L., and Carbon, J. A colony bank containing synthetic ColE1 hybrid plasmids representative of the entire *E.coli* genome. Cell 9:91-99 (1978).

344. Edwards, A., Voss, H., Rice, P., Civitello, A., Stegemann, J., Schwager, C., Zimmerman, J., Erfle, H., Caskey, C.T., and Ansorge, W. Automated DNA sequencing of the human HPRT locus. Genomics 6:593-608 (1990)

345. Roe, B.A., Crabtree, J.S., and Khan, A.S. "DNA Isolation and Sequencing." John Wiley and Sons, Ltd. West Sussex, UK (1996).

346. Birnboim, H.C., and Doly, J. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucl. Acids Res. 7:1513-1523 (1979).

347. Pan, H.Q., Chissoe, S.L., Bodenteich, A., Wang, Z., Iyer, K., Clifton, S.W., Crabtree, J.S., and Roe, B.A. The complete nucleotide sequence of the SacB11 kan domain of the P1 pAD10-SacB11 cloning vector and three cosmid cloning vectors: pTCF, svPHEP and LAWRIST 16. Genet. Anal. Tech. Appl. 11:181-186 (1994).

348. Willis, E.H., Mardis, E.R., Jones, W.L., and Little, M.C. Prep-A-Gene: a superior matrix for the purification of DNA and DNA fragments. Biotechniques 9:92-99 (1990).

349. Marko, M.A., Chipperfield, R., and Birnboim, H.C. A procedure for the large scale isolation of highly purified plasmid DNA using alkaline extraction and binding to glass powder. Anal. Biochem. 121:382-387 (1982).

350. Carter, M.J., and Milton, I.D. An inexpensive and simple method for DNA purifications on silica particles. Nucl. Acids Res. 21:1044 (1993).

351. Fry, G., Lachenmeier, E., Mayrand, E., Giusti, B., Fisher, J., Johnston-Dow, L., Cathcart, R., Finne, E., and Kilaas, L. A new approach to template purification for sequencing applications using paramagnetic particles. Biotechniques 13:124-131 (1992).

352. Amersham/Molecular Dynamics - robot users manual.

353. Pogue, R.R., Cook, M.E., Livingstone, L.R., Hunt, S.W., 3d. Preparation of template for automated sequencing using QIAGEN resin. Biotechniques 15:376-378 (1993).

354. Chowdhury, K. One step 'miniprep' method for the isolation of plasmid DNA. Nucl. Acids Res. 19:2792 (1991).

355. Serghini, M.A., Ritzenthaler, C., and Pinck, L. A rapid and efficient 'miniprep' for isolation of plasmid DNA. Nucl. Acids Res. 17:3604 (1989).

356. Mardis, E.R., and Roe, B.A. Automated methods for single-stranded DNA isolation and dideoxynucleotide DNA sequencing reactions on a robotic workstation. Biotechniques 7:736-746 (1989).

357. Kornberg, A. Biological synthesis of deoxyribonucleic acid. Science 131:1503-1508 (1960).

358. Modrich, P., and Richardson, C.C. Bacteriophage T7 deoxyribonucleic acid replication *in vitro*. Bacteriophage T7 DNA polymerase: an enzyme composed of phage- and host-specific subunits. J. Biol. Chem. 250:5515-5522 (1975).

359. Modrich, P. and Richardson, C.C. Bacteriophage T7 deoxyribonucleic acid replication *in vitro*. A protein of *Escherichia coli* required for bacteriophage T7 DNA polymerase activity. J. Biol. Chem. 250:5508-5514 (1975).

360. Tabor, S., and Richardson, C.C. Selective inactivation of the exonuclease activity of bacteriophage T7 DNA polymerase by *in vitro* mutagenesis. J. Biol. Chem. 264:6447-6458 (1989).

361. Stenish, J., and Roe, B.A. DNA polymerase from mesophilic and thermophilic bacteria. I. Purification and properties of DNA polymerase from *Bacillus licheniformis* and *Bacillus stearothermophilus*. Biochem. Biophys. Acta. 272:156-166 (1972).

362. Ye, S.Y., and Hong, G.F. Heat-stable DNA polymerase I large fragment resolves hairpin structure in DNA sequencing. Sci. Sin. [B] 30:503-506 (1987).

363. Mead, D.A., McClary, J.A., Luckey, J.A., Kostichka, A.J., Witney, F.R., and Smith, L.M. Bst DNA polymerase permits rapid sequence analysis from nanogram amounts of template. Biotechniques 11:76-78 (1991).

364. Protocol supplied with ABI PRISM double-stranded cycle sequencing kit for use with Amplitaq DNA polymerase.

365. Protocol supplied with ABI PRISM double-stranded cycle sequencing kit for use with Amplitaq CS+ DNA polymerase.

366. Klenow, H., and Henningsen, J. Selective elimination of the exonuclease activity of the deoxyribonucleic acid polymerase from *Escherichia coli* B by limited proteolysis. Proc. Natl. Acad. Sci., USA 65:168-175 (1970).

367. Setlow, P. DNA polymerase I from *Escherichia coli*. Meth. Enzymol. 29:3-12 (1974).

368. Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., and Erlich, H.A. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239:487-491 (1988).

369. Barnes, W.M. The fidelity of Taq polymerase catalyzing PCR is improved by an N-terminal deletion. Gene 112:29-35 (1992).

370. Protocol supplied with ABI PRISM double-stranded cycle sequencing kit for use with Amplitaq FS DNA polymerase.

371. Ansorge, W., Sproat, B., Stegemann, J., Schwager, C. and Zenke, M. Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. Nucl. Acids Res. 15:4593-4602 (1987).

372. Ansorge, W., Sproat, B., Stegemann, J., and Schwager, C. A non-radioactive automated method for DNA sequence determination. J. Biochem. Biophys. Meth. 13:315-322 (1986).

373. Applied Biosystems Incorporated Users Manual ABI 373 (1994).

374. Applied Biosystems Incorporated Users Manual ABI 377 (1996).

375. Mullis, K.B., and Faloona, F.A. Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction. Meth. Enzymol. 155:335-350 (1987).

376. Applied Biosystems, Inc. Users bulletin No. 11. Synthesis of fluorescent dye-labeled oligonucleotides for use as primers in fluorescent based DNA sequencing. Jan 10, 1989.

377. Smith, L.M., Sanders, J.Z., Kaiser, R.J., Hughes, P., Dodd, C., Connell, C.R., Heiner, C., Kent, S.B.H., and Hood, L.E. Fluorescence detection in automated DNA sequencing analysis. Nature 321:674-679 (1986).

378. Smith, L.M., Kaiser, R.J., Sanders, J.Z., and Hood, L.E. The synthesis and use of fluorescent oligonucleotides in DNA sequence analysis. Meth. Enzymol. 155:260-301 (1987).

379. Kristensen, T., Voss, H., Schwager, C., Stegemann, J., Sproat, B., Ansorge, W. T7 DNA polymerase in automated dideoxy sequencing. Nucl. Acids Res. 16:3487-3496 (1988).

380. Roe, B.A., Johnston-Dow, L., and Mardis, E. Use of a chemically modified T7 DNA polymerase for manual and automated sequencing of supercoiled DNA. Biotechniques 6:520 (1988).

381. Beckman Biomek 1000 Users Manual (1991).

382. Qiagen filtration apparatus, catalog number 19504.

383. Beckman Biomek 2000 Users Manual (1996).

384. Staden, R. Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. Nucl. Acids Res. 12:505-519 (1982).

385. Meyers, E.W. Incremental alignment algorithms and their applications (Technical report 86-2). Department of Computer Science, University of Arizona, Tucson, AZ 85721 (1986).

386. Huang, X. An improved sequence assembly program. Genomics 33:21-31 (1996).

387. Phred/Phrap/Consed developed by Phil Green (http://chimera.biotech.washington.edu/uwgc/).

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

389. Claverie, J.M., and States, D.J. Informtion enhancement methods for large scale sequence analysis. Comp. Chem 17:1919-1921 (1993).

390. Uberbacher, E.C., and Mural, R.J. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. Proc. Natl. Acad. Sci., USA 88:11261-11265 (1991).

391. Maizel, J.V., Jr., and Lenk, R.P. Enhanced graphic matrix analysis of nucleic acid and protein sequences. Proc. Natl. Acad. Sci., USA 78:7665-7669 (1981).

392. Staden, R. The current status and portability of our sequence handling software. Nucl. Acids Res. 14:217-231 (1986).

393. Wilbur, W.J., and Lipman, D.J. Rapid similarity searches of nucleic acid and protein databanks. Proc. Natl. Acad. Sci., USA 80:726-730 (1983).

394. Higgins, D.G. CLUSTAL V: multiple alignment of DNA and protein sequences. Meth. Mol. Biol. 25:307-318 (1994).

395. GCG Sequence analysis software version 7.2.

396. Sonnhamer, E.L., and Durbin, R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:GC1-GC10 (1995).

397. Henikoff, S., and Henikoff, J.G. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad Sci., USA 89:10915-10919 (1992).

398. Bankier, A.T., and Barrell, B.G. Shotgun DNA sequencing in "Techniques in the Life Sciences B5" Flavell, R.A., ed. Cambridge, MA. pp. 1-34 (1983).

399. Schriefer, L.A., Gebauer, B.K., Qui, L.Q.Q., Waterston, R.H., and Wilson, R.K. Low pressure DNA shearing: a method for random NDA sequence analysis. Nucl. Acids Res. 18:7455-7456 (1990).

400. Sharma, R.C., and Schimke, R.T. Preparation of electro-competent *E.coli* using salt-free growth medium. Biotechniques 20:42-44 (1996).

401. Lewin, B. "Genes IV" Oxford University Press, Oxford (1990).

402. Bio-Rad *E. coli* Transformation Apparatus. Operating Instructions and Application Guide. Biorad Laboratories, Richmond, CA (1994).

403. Bird, A.P. CpG-rich islands and the function of DNA methylation. Nature 321:209-213 (1986).

404. Gardiner-Garden, M., and Frommer, M. CpG islands in vertebrate genomes. J.Mol. Biol. 196:261-282 (1987).

405. Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. CpG islands as gene markers in the human genome. Genomics 13:1095-1107 (1992).

406. Cross, S.H., Charlton, J.A., Nan, X., and Bird, A.P. Purification of CpG islands using a methylated DNA binding column. Nature Genet. 6:236-244 (1994).

APPENDIX - Protocols

I. Electroporation Competent Cell Preparation[400]

1. Grow a fresh overnight culture of XL1BlueMRF' cells in liquid YENB medium (7.5 g bacto yeast extract, 8 g bacto nutrient broth, to 1L with double distilled water, autoclave to sterilize).
2. Inoculate 1L of fresh YENB medium with 5-10 µl of the fresh ON culture and grow cells at 37°C with shaking at 250 rpm. Cells should be harvested between an A_{600} of 0.5 to 0.9 (usually 4-5 hours). If cells are overgrown, dilute back to an A_{600} of approximately 0.2 and regrow to desired A_{600}.
3. To harvest cells, pour 250 ml of culture in each of four 500 ml centrifuge bottles, chill on ice for 5 minutes and centrifuge at 4000 X g (approximately 4900 rpm in GS3 rotor) for 10 minutes at 4oC. It is important not to centrifuge at a higher g value than needed to pellet the cells: more centrifugal force applied for a longer time makes the cells difficult to resuspend gently, which, in turn, leads to cell disruption and reduces the electrotransformation efficiency.
4. Pour the media off into another flask and autoclave prior to disposal. Invert each bottle and wope the interior rim with a clean Kimwipe. Add 100 ml sterile cold water to each centrifuge bottle, gently resuspend the cells and centrifuge as in step 3. Discard supernatant as above, and add 100 ml sterile cold water to each bottle, gently resuspending the cells. Transfer 100 ml cell suspension from each of two bottles to the remaining two bottles, leaving 200 ml suspension per bottle, and centrifuge as in step 3. Remove and discard supernatant as above.
5. Resuspend the cells in one bottle in 20 ml cold sterile 10% glycerol and transfer the 20 ml suspension to the other bottle containing a cell pellet and resuspend those cells. Centrifuge as in step 3; remove and discard supernatant.

6. Resuspend the cells (from 1L of culture) in a final volume of 2-3 ml cold sterile 10% glycerol. The cell number in the suspension should be 2-4 x 10^{10} cells/ml. These competent cells can be used fresh or frozen (see step 7) for future use.
7. Aliquot competent cells into 1.5 ml microfuge tubes (40 µl/tube: measure accurately-insufficient volume can cause arcing during electrotransformation) and flash freeze by placing the tubes on dry ice until frozen. Store at -70°C. Competent cells stored in this manner are viable for 1-2 years.

II. Electroporation (to be performed in cold room)[402]

1. Gently thaw competent cells on ice. Remove sterile cuvettes (0.2 cm) from their pouches and place them in the cold room to pre-chill.
2. Add 2-3 µl ligation mix tot he thawed, chilled competent cells and mix by gently pipetting up and down a few times.
3. Let the mix sit on ice for 1 minute (time this).
4. During the incubation on ice, turn on the Gene Pulser apparatus (switch to the bottom right side of the apparatus) and set it to 2.5 kV by simultaneously pressing the "raise" and "lower" pads on the Pulser twice until the LED display reads "2.50."
5. Pipette the chilled cell suspension into the narrow portion of the cuvette between the aluminum plates. Gently shake the suspension to the bottom of the cuvette. Put on the white cuvette cap supplied and place the capped cuvette int he cuvette slide. Push the slide into the shocking chamber (under the plastic shield) until it makes firm contact witht he shocking chamber electrodes (between the red and black electrode leads).
6. To charge the capacitor and deliver a pulse, depress and hold both rectangular red pulse buttons until a continuous tone sounds. The LED display will flash

"Chg" indicating that the capacitor is being charged to the selected voltage. The tone signals that the pulse has been delivered and the pulse buttons may be released. (If arcing occurs this indicates that the sample is too conductive. The cells may be plated, but the yield will be significantly decreased.

7. Remove the cuvette fromt he chamber and IMMEDIATELY add 1 ml of YENB medium. (If YENB is not added immediately after the pulse, the number of transformants is reduced).
8. Check the time constant by simultaneously pressing the the "Actual Volts (kV)" and the "Set Volts (kV)" pads on the front panel of the apparatus. The time constant should be close to 5.0 milliseconds (an experimentally determined optimal value to be used when working with high resistance samples). Process one sample at a time and turn off the pulser after finishing all samples.
9. Transfer the cell suspension to a 17 x 100 mm Falcon tube and incubate at 37°C with shaking at 350 rpm for 1hour, and plate the cells using the standard protocol[345].
10. If additional transformations for the same project are desired, the cuvettes may be re-used for the same ligation mix. Rinse the cuvettes thoroughly with sterile double distilled water and repeat steps 2-9.
11. Rinse cuvettes well with sterile double distilled water, air dry and store for possible re-use ONLY for the same project.

III. Large Scale Double Stranded Isolation for plasmid, cosmid and BAC

1. Pick a smear of bacterial colonies harboring the plasmid, cosmid or BAC DNA of interest into a 12 x 75 mm Falcon tube containing 2 ml LB media (10 g bacto tryptone, 5 g bacto yeast extract, 10 g NaCl, double distilled water to 1L, autoclave to sterilize) supplemented with the appropriate antibiotic and incubate at 37°C for 8 hours with shaking at 250 rpm. Transfer the culture to an

Erlenmeyer flask containing 50 ml similar media, and incubate further for 8 hours. Transfer 12.5 ml of the culture ot each of 4 liters of similar media and incubate for an additional 8 hours.

2. Harvest the cells by centrifugation at 7000 rpm for 20 minutes in 500 ml bottles in the RC5-B using the GS3 rotor. Resuspend the cell pellets in old media and combine cell suspensions and recentrifuge as before such that 2 L initial culture for plasmids and cosmids (1 L initial culture for BAC) per 500 ml bottle. Cell pellets may be frozen at -70°C at this point.
3. Resuspend the cell pellets (either from 2 L plasmid or cosmid culture, or 1 L BAC culture) in 35 ml GET (50 mM glucose, 25 mM Tris/HCl, pH 8.0 and 10 mM EDTA, pH 8.0 in double distilled water) by gently teasing the pellet with a spatula. Do not vortex the lysate at any time because this may shear the host chromosomal DNA. Once cells are resuspended, add 2 mg/ml crystal lysozyme and incubate at room temperature for 10 minutes.
4. Add 70 ml alkaline lysis solution (SDS/NaOH: 1% SDS, 0.2 N NaOH in double distilled water) to each bottle, gently mix, and incubate for 5 minutes in an ice-water bath.
5. Add 52.5 ml 3M NaOAc, pH 4.8 (408.24 g NaOAc-$3H_20$ in 300 ml double distilled water, adjust pH to 4.8 with glacial acetic acid and bring final volume to 1 L with double distilled water), cap tightly and gently mix by inverting the bottle a few times slowly. Incubate 30-60 minutes in an ice-water bath.
6. Clear the lysate of precipitated SDS, proteins, membranes and chromosomal DNA by pouring through a double layer of cheesecloth. Transfer the lysate into a 250 ml centrifuge bottle, centrifuge at 10,000 rpm for 30 minutes at 4°C int he RC5-B using the GSA rotor.
7. Pour the supernatants from step 6 into 500 ml bottles and add DNase-free RNase A (20 mg/ml RNase A in 1 mM NaOAc, pH 4.8) and RNase T1 (100

U/µl in 50 mM Tris/HCl, pH7.6) such that the final concentration is 40 µg/ml and 40 U/ml, respectively. Incubate in a 37oC water bath for 30 minutes.

8. Add an equal volume of isopropanol and precipitate at room temperature for 5 minutes. Centrifuge at 9,000 rpm for 30 minutes in the RC5-B using the GS3 rotor. Decant the supernatant and drain the DNA pellet.
9. Resuspend the DNA pellet in 20 ml 10:1 TE buffer (plasmid and cosmid only) and add 40 ml (add 50 ml defined diatomaceous earth for BAC) defined diatomaceous earth in guanidine-HCl (100 mg/ml: suspend 10 g diatomaceous earth in 100 ml distilled water in a 100 ml graduated cylinder and allow to settle for 3 hours. Decant the supernatant and resuspend the pellet in 100 ml of 6 M guanadine HCl, pH 6.4, 50 mM Tris-HCl, 20 mM EDTA). Allow the DNA to bind at room temperature for 5 minutes with occasional mixing. Centrifuge at 9,000 for 10 minutes in the RC5-B using the GS3 rotor.
10. Decant the supernatant, resuspend the pellet in 40 ml diatomaceous earth wash buffer (10 mM Tris/HCl, pH 8.0, 1 mM EDTA, pH 8.0 and 50% ethanol in double distilled water) and centrifuge as above.
11. Decant the supernatant, resuspend the pellet in 40 ml acetone and centrifuge as above.
12. Decant the supernatant and dry the pellet in a vacuum oven.
13. Resuspend the pellet in 20 ml 10:1 TE buffer (10 mM Tris/HCl, pH 7.6, 1 mM EDTA, pH 8.0 in double distilled water) and elute the bound DNA by incubation at 65°C for 10 minutes with occasional mixing.
14. Remove the diatomaceous earth by centrifugation as above. Repeat if necessary
15. Combine the DNA-containing supernatants and precipitate the DNA in 35 ml Corex glass tubes adding 2.5 volumes of cold 95% ethanol/acetate (95% ethanol, 0.12 M NaOAc, pH 4.8 in double distilled water).

16. Resuspend the dried DNA pellet in 2 ml of 10:0.1 TE buffer (10 mM Tris/HCl, pH 7.6, 0.1 mM EDTA, pH 8.0 in double distilled water) and assay for concentration by absorbance readings at 260 nm or by agarose electrophoresis versus known standards.

IV. Midiprep double stranded DNA isolation for plasmid or cosmid

1. Pick a smear of colonies harboring the plasmid or cosmid of interest into a 12 x 75 mm Falcon tube containing 3 ml 2xTY media (16 g bacto tryptone, 10 g bacto yeast extract, 5 g NaCl in double distilled water to 1 L, autoclave to sterilize) supplemented with the appropriate antibiotic and incubate at 37°C for 8-10 hours with shaking at 250 rpm. Transfer the culture to an Erlenmeyer flask containing 50 ml of similar media, and incubate further 11-14 hours.
2. Harvest the cells by centrifugation at 3000 rpm for 5 minutes in 50 ml conical tubes in the Beckman GPR tabletop centrifuge and decant the supernatant. The cell pellets can be frozen at -70°C at this point.
3. Resuspend the cell pellets in 2 ml GET/lysozyme solution (50 mM glucose, 25 mM Tris/HCl, pH 8.0 and 10 mM EDTA, pH 8.0 in double distilled water, supplement with 2 mg/ml lysozyme before use), add 4 ml alkaline lysis solution (1% SDS, 0.2 N NaOH), gently mix and incubate for 5 minutes in an ice-water bath.
4. Add 4 ml 3 M NaOAc, pH 4.8 (408.24 g NaOAc-$3H_20$ in 300 ml double distilled water, adjust pH to 4.8 with glacial acetic acid and bring final volume to 1 L with double distilled water), gently mix by swirling, and incubate in an ice-water bath for 30-60 minutes.
5. Clear the lysate of precipitated SDS, proteins, membranes and chromosomal DNA by pouring through a double layer of cheesecloth into a clean 50 ml

conical tube. Centrifuge at 3000 rpm for 20 minutes at 4°C in the Beckman GPR tabletop centrifuge.

6. Decant the supernatant to a 50 ml polypropylene centrifuge tube, add 20 µl of 20 mg/ml DNase free-RNase A (20 mg/ml RNase A in 1 mM NaOAc, pH 4.8) and incubate in a 37°C water bath for 30 minutes.
7. Add 7 ml (equal volume) of defined diatomaceous earth in guanadine-HCl (20 mg/ml: suspend 2 g diatomaceous earth in 100 ml distilled water in a 100 ml graduated cylinder and allow to settle for 3 hours. Decant the supernatant and resuspend the pellet in 100 ml of 6 M guanadine HCl, pH 6.4, 50 mM Tris-HCl, 20 mM EDTA) and allow the DNA to bind at room temperature for 5 minutes with occasional mixing. Centrifuge at 3000 rpm for 5 minutes in the Beckman GPR tabletop centrifuge.
8. Decant the supernatant, resuspend in 7 ml diatomaceous earth wash buffer (10 mM Tris/HCl, pH 8.0, 1 mM EDTA, pH 8.0 and 50% ethanol in double distilled water) and centrifuge as before.
9. Decant the supernatant, resuspend in 7 ml acetone, and centrifuge as before.
10. Decant the supernatant and dry in a vacuum oven.
11. Resuspend the pellet in 0.6 ml of 10:1 (10 mM Tris/HCl, pH 7.6, 1 mM EDTA, pH 8.0) and elute the bound DNA by incubation at 65°C for 10 minutes with intermittent mixing.
12. Remove the diatomaceous earth by centrifugation at 3000 rpm for 5 minutes in the Beckman GPR tabletop centrifuge.
13. Transfer the supernatant to a 1.5 ml microcentrifuge tube and centrifuge at 12,000 rpm for 5 minutes in a microcentrifuge at room temperature. Transfer the supernatant to a new 1.5 ml microcentrifuge tube and ethanol precipitate.

14. Resuspend the dried DNA pellet in 40 ml 10:0.1 TE buffer (10 mM Tris/HCl, pH 7.6 , 0.1 mM EDTA, pH 8.0) and assay for concentration by agarose gel electrophoresis versus known standards.

V. Miniprep double stranded DNA isolation

1. Pick a smear of bacteria harboring the plasmid or cosmid of interest into a 17 x 100 mm Falcon tube containing 6 ml TB media (12 g bactotryptone, 24 g bacto yeast extract, 4 ml glycerol to 900 ml with double distilled water, autoclave to sterilize, cool, then add 100 ml 10 x TB salts. 10 x TB salts: 2.31 g KH_2PO_4, 12.54 g K_2HPO_4 in 100 ml double distilled water, autoclave to sterilize) supplemented with the appropriate antibiotic and incubate at 37°C for 16-18 hours with shaking at 250 rpm.
2. Harvest the cells by centrifugation at 3000 rpm for 5 minutes in the Beckman GPR tabletop centrifuge, and decant the supernatant. The cell pellets may be frozen at this point.
3. Resuspend the cell pellets in 0.2 ml of TE-RNase solution (50 mM Tris/HCl, pH 7.6, 10 mM EDTA, pH 8.0, 40 µg/ml RNase A and some add 10 U/µl RNase T1) by gently vortexing, add 0.2 ml of alkaline lysis solution (1% SDS, 0.2 N NaOH) gently mix and incubate for 15 minutes at room temperature.
4. Add 0.2 ml 3M NaOAc, pH 4.8 (408.24 g NaOAc-$3H_20$ in 300 ml double distilled water, adjust pH to 4.8 with glacial acetic acid and bring final volume to 1 L with double distilled water), gently mix by swirling, transfer to 1.5 ml microcentrifuge tubes and incubate in an ice-water bath for 15 minutes.
5. Clear the lysate of precipitatated SDS, proteins, membranes and chromosomal DNA by centrifugation at 12,000 rpm for 15 minutes in a microfuge at 4°C.

6. Transfer the supernatant to a fresh 1.5 ml microcentrifuge tube, incubate in an ice-water bath for 15 minutes and centrifuge as above for an additional 15 minutes.

7a. For standard alkaline lysis purification, precipitate the DNA by adding 1 ml 95% ethanol and resuspend the dried DNA pellet in 200 µl 10:0.1 TE buffer (10 mM Tris/HCl, pH 7.6, 0.1 mM EDTA, pH 8.0). Assay for concentration and purity by agarose gel electrophoresis.

7b. For diatomaceous earth based purification, add 1 ml of defined diatomaceous earth (20 mg/ml: suspend 2 g diatomaceous earth in 100 ml distilled water in a 100 ml graduated cylinder and allow to settle for 3 hours. Decant the supernatant and resuspend the pellet in 100 ml of 6 M guanadine HCl, pH 6.4, 50 mM Tris-HCl, 20 mM EDTA) and allow the DNA to bind at room temperature for 5 minutes with occasional mixing. Meanwhile, soak the Prep-A-Gene nitrocellulose membrane in isopropanol for at least 3 minutes, and assemble the Prep-A-Gene manifold as described[403].

8b. Turn on the vacuum pump and adjust the vacuum level to 8 in. Hg, let the membrane dry for 1 minute, and then release the vacuum.

9b. Pour the well-mixed samples into the wells of the Prep-A-Gene manifold and filter through at 8 in. Hg, until all the liquid is filtered through.

10b. Wash the samples four times with 250 µl diatomaceous earth wash buffer using a repeat pipette, allowing all of the liquid to filter through between washes.

11b. Reduce the vacuum to 5 in. Hg before turning the vacuum off at the stopcock. Without unscrewing the black clamps, release the white clamps and place the collection rack with clean 1.5 ml screwcapped tubes into the manifold. Clamp the manifold with the white clamps, and apply 300 µl of 10:1 TE buffer (10 mM Tris/HCl, pH 7.6, 1 mM EDTA, pH 8.0 in double distilled water) heated

to 65°C and pull the eluted DNA through into the collection tubes at 5 in. Hg. Once the liquid has filtered through, raise the vacuum to 10-12 in. Hg, and let the membrane dry for 1 minute.

12b. Turn off the vacuum at the stopcock and remove the collection rack containing the tubes and eluted DNA. Ethanol precipitate the DNA and resuspend the dried DNA pellets in 30 ml 10:0.1 TE buffer (10 mM Tris/HCl, pH 7.6, 0.1 mM EDTA, pH 8.0).

VI. Manual 96-well double stranded template miniprep isolation

1. Pick colonies of interest using a toothpick into 1.5 ml TB media (12 g bactotryptone, 24 g bacto yeast extract, 4 ml glycerol to 900 ml with double distilled water, autoclave to sterilize, cool, then add 100 ml 10 x TB salts. 10 x TB salts: 2.31 g KH_2PO_4, 12.54 g K_2HPO_4 in 100 ml double distilled water, autoclave to sterilize) supplemented with the appropriate antibiotic in a 96 well block with cover, and incubate at 37°C with shaking at 350 rpm for 24 hours.
2. Harvest the cells by centrifugation at 2500 rpm for 7 minutes in the Beckman GPR benchtop rotor, and decant supernatant. Cell pellets may be stored at -20°C at this point.
3. Add 200 µl TE-RNAse A (50 mM Tris/HCl, pH 7.6, 10 mM EDTA, pH 8.0, 40 µg/ml RNase A and some add 10 U/µl RNase T1) to the pellets and resuspend by pipetting up and down ten times with a 12 channel pipette.
4. Add 200 µl alkaline lysis solution (1% SDS, 0.2 N NaOH in double distilled water) and mix as before.
5. Add 200 µl 3M NaOAc, pH 4.8 (408.24 g NaOAc-$3H_20$ in 300 ml double distilled water, adjust pH to 4.8 with glacial acetic acid and bring final volume to 1 L with double distilled water) to each well, mix as before, and shake at 350 rpm at 37°C for 10 minutes.

6. Incubate at -20°C for 15-30 minutes and centrifuge at 3000 rpm for 45 minutes at 4°C in the Beckman tabletop centrifuge to clear the lysate.
7. Using a 12 channel pipette, remove the top 400 ul of the supernatant to a clean 96 well block. Ethanol precipitate and dry under vacuum. Resuspend DNA pellets in 50 ul 10:0.1 TE buffer (10 mM Tris/HCl, pH 7.6, 0.1 mM EDTA, pH 8.0).

VII. Automated 96 well double-stranded DNA miniprep on the Biomek 1000 laboratory workstation.

1. Pick colonies using a toothpick into 1.8 ml TB (12 g bactotryptone, 24 g bacto yeast extract, 4 ml glycerol to 900 ml with double distilled water, autoclave to sterilize, cool, then add 100 ml 10 x TB salts. 10 x TB salts: 2.31 g KH_2PO_4, 12.54 g K_2HPO_4 in 100 ml double distilled water, autoclave to sterilize) supplemented with the appropriate antibiotic and incubate for 22 hours at 37°C with shaking at 350 rpm in a 96 well block with cover.
2. Harvest cells by centrifugation at 2500 rpm for 7 min. Decant supernatant and allow pellets to drain inverted. Cell pellets may be frozen at this point.
3. Turn on Biomek, begin the program DSISOL2 and set up the biomek as indicated in the configuration function on the screen. Specifically, place TE-RNase A (50:10 TE buffer containing 40 µg/ml DNase-free RNase A) in the first module, alkaline lysis solution (1% SDS, 0.2 M NaOH) in the second reagent module and 3 M NaOAc, pH 4.8 in the third module ((408.24 g NaOAc-$3H_20$ in 300 ml double distilled water, adjust pH to 4.8 with glacial acetic acid and bring final volume to 1 L with double distilled water).
4. Place the 96 well block containing cells onto the biomek tablet at the position labeled "1.0 ml Minitubes". Place a Millipore filter plate in the position labeled "96well flat bottomed microtitre plate".

5. Press ENTER to continue with the program.

6. First the biomek will add 100 µl TE-RNase A to the cell pellets and mix to partially resuspend. Place in 37°C shaker for 5 minutes.

7. Next, the biomek will add 100 µl alkaline lysis solution to the wells of the filter plate.

8. The biomek then will mix the cell suspension again, transfer the entire volume to the filter plate containing alkaline lysis solution, and mix again. Set up the filtration apparatus with a clean 96 well block to collect the filtrate (wash and reuse the block used for growth). Pause the biomek and allow to incubate for 20 minutes at room temperature.

9. The biomek will add 100 ul 3M NaOAc, pH 4.8 to the wells of the filter plate and mix at the sides of the wells. Some choose to place the filter plate at -20°C for 5 minutes at this point. Transfer the filter plate to the QiaVac Vacuum Manifold 96 (Qiagen Cat. No. 19504) and filter using water vacuum only (do not do a harsh filtration as the plates are fragile and will loose their seal). This will typically take less than 20 minutes.

10. The supernatant collected in the 96 well block is the crude DNA and must be ethanol precipitated before use. Dry the pellets under vacuum.

11. Resuspend in 30 ul 10:0.1 TE (10 mM Tris-HCl pH 7.6 and 0.1 mM EDTA pH 8.0) and assay by agarose gel electrophoresis versus known standards.

VIII. Automated 96 well double stranded template miniprep isolation on the Biomek 2000 laboratory workstation

1. Pick colony harboring plasmid of interest into 96 well square well block containing 1.5 ml TB (12 g bactotryptone, 24 g bacto yeast extract, 4 ml glycerol to 900 ml with double distilled water, autoclave to sterilize, cool, then add 100 ml 10 x TB salts. 10 x TB salts: 2.31 g KH_2PO_4, 12.54 g

K_2HPO_4 in 100 ml double distilled water, autoclave to sterilize) with appropriate antibiotic. Grow 22-24 hours at 37°C with shaking at 350 rpm.

2. Harvest cells by centrifugation at 2500 rpm for 7 minutes. Decant supernatant and autoclave before disposal. Cell pellets may be frozen at this point.
3. Place 4 blocks containing cell pellets on the biomek tablet in the configuration indicated in the program "dsisol.v1". Blocks containing cell pellets should be placed at positions A5, A6, B5 and B6; P250 biomek tips should be placed at positions A2, A3, B2 and B3; reagent half modules should be placed at positions A4 and B4; tool rack at position A1 with MP200 tool in the right hand rack; position B1 is empty. TE-RNase A+T1 should be placed in the left side of A4, alkaline lysis solution should be placed in the right side of A4, and 3 M Na/K OAc should be placed in the left side of B4. The right module of B4 is empty.
4. Start the program. The biomek will add 200 µl TE-RNase A+T1 (50 mM Tris/HCl, pH 7.6, 10 mM EDTA, pH 8.0, 40 µg/ml RNase A and some add 10 U/µl RNase T1) to each well of the four blocks and mix 20 times to resuspend. This will use all four boxes of tips. After addition and mixing (1 hour, 6 minutes), the biomek will pause and request tips to be changed at position A2. At this time, *all four tip boxes should be changed and "Replace All" should be clicked.*
5. Next, the biomek will add 200 µl alkaline lysis solution (1% SDS, 0.2 N NaOH in double distilled water) and will mix an additional 10 times. This will use all four boxes of tips. The biomek will pause and request tips to be changed at position A2. *Place clean tips at A2 only and click "OK."*
6. The biomek adds 200 µl 3M NaOAc, pH 4.8 (or 3M KOAc, pH 4.8) to the blocks. In doing this, the biomek only uses one row of tips. Leave this box of tips on the biomek...it will reuse the row of tips for the first supernatant

transfer. You then remove the four blocks, cover with plate sealers, vortex briefly and place in 350 rpm shaker for 10 minutes.

7. Incubate blocks in -20°C for 15-30 minutes and centrifuge for 45 minutes at 3000 rpm to pellet the precipitate.
8. Place the blocks back on the biomek tablet in the new configuration as shown on the screen. Centrifuged blocks are placed in positions A5, A6, B5 and B6; clean blocks are placed at positions A3, A4, B3 and B4; P250 tips are placed at positions A2, B1 and B2. *(NOTE: The graphical display of the new configuration does not show tips at B1, but transfer requires tips at this position. Just a programming glitch).* The biomek will transfer the upper 400 µl to clean blocks starting with A5 to A3 using tips at A2. The second transfer is blocks at A6 to A4 using tips at position B1. Then the biomek will ask for.new tips to be replaced at position A2. Replace the tips at A2 and click "OK." The third transfer is blocks at B5 to B3 using tips at A2 and the fourth transfer is B6 to B4 using tips at position B2.
9. Ethanol precipitate by adding 1 ml 95% ethanol to each well of the block and cover with plate sealer (do not invert...the ethanol will degrade the glue of the sealer). Incubate at -20°C for 30 minutes to overnight, centrifuge 3000 rpm for 30 minutes, decant. Add 500 µl 70% ethanol, centrifuge an additional 15 minutes at 3000 rpm. decant and dry under vacuum. Resuspend in 50 µl 10:0.1 TE (10 mM Tris/HCl, pH 7.6, 0.1 mM EDTA, pH 8.0) and assay 2 ul by agarose gel electrophoresis.