INFORMATION TO USERS

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

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

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

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

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

UMI

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

UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

SEQUENCE AND ANALYSIS OF APPROXIMATELY 0.5 MEGA-BASEPAIR OF THE DIGEORGE SYNDROME CRITICAL REGION IN HUMAN CHROMOSOME 22 BAND Q11 AND A SYNTENIC MOUSE BAC

A Dissertation

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

FENG CHEN

Norman, Oklahoma

1997

SEQUENCE AND ANALYSIS OF APPROXIMATELY 0.5 MEGA-BASEPAIR OF THE DIGEORGE SYNDROME CRITICAL REGION IN HUMAN CHROMOSOME 22 BAND Q11 AND A SYNTENIC MOUSE BAC

A Dissertation APPROVED FOR THE
DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

BY

Acknowledgements

I would like to express my sincere gratitude to my wife Jing Li, and my parents for their support and love. I also would like to thank my brother and sisters for their encouragement. I would like to thank all the teachers I had in the way of my growth for everything they taught me. A very special acknowledgment goes to my major professor, Dr. Bruce A. Roe, for his guidance, knowledge, and support. I also extend my appreciation to the other members of my graduate advisory committee, Drs. Leroy Blank, Phil E. Klebba, David R. McCarthy, and Ann West.

A big thanks goes to everyone in Dr. Roe's Lab, both past and present, for their positive support, mutual help, team effort, and inspiring discussion: Mueed Ahmad, Cathy Anadu, Dennis Burian, Linda Cantu, Lingzhi Chu, Sandy Clifton, Arlena Coulberson, Judy Crabtree, Chesca Craig, Angela Dorman, Whitney Elkins, Jennifer Gray, Glenda Hall, Karen Hartman, Jennifer Hausner, Andrew Horning, Naiqing Hu, Emily Huang, Axin Hua, Nora Ivanoff, Steve Kenton, Akbar Khan, Doris Kupfer, Hongshing Lai, Lisa Lane, HioIeong Lao, Michele Lasley, Sharon Lewis, Shaoping Lin, Eda Malaj, Anthony McDaniel, Sean Meadows, Fares Najar, Nientung Ma, Yichen Ma, Thuan Nguyen, Huaqin Pan, Murli Rao, Adonis Reece, Qun Ren, Ging Sobhraksha, Lin Song, Rhys Strasia, Will Tankersley, Steve Toth, Yonathan Tilahun, Jeffery Viken, Ha Vu, YingPing Wang, Yungfang Wang, Zhili Wang, Rusty Wayt, Heather Wright, Ziyun Yao, Sola Yu, Xiling Yuan, Min Zhan, Guozhong Zhang, and Hua Zhu. I would like to especially thank Drs. Nientung Ma, Steve Toth, and Zhili Wang for their help on the project.

I thank our collaborators, Drs. Marcia L. Budarf and Beverly S. Emanuel from the Children's Hospital of Philadelphia, PA and member of their research groups for

providing mapped cosmids and BACs and their collaboration. This work was supported by grants from the National Institute of Health, the Oklahoma Center for the Advancement of Science and Technology and the Oklahoma Department of Commerce Center of Excellence in Molecular Medicine Program. Financial supports to the author provided by the Department of Chemistry and Biochemistry and the Graduate College of the University of Oklahoma are deeply appreciated.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

2xTY - medium containing tryptone, yeast extract and salt

ABI, ABD, or PE-ABD - Applied Biosystems Incorporated (a division of Perkin Elmer)

ADU - a female patient with DiGeorge syndrome

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

AP-2 - retinoic acid-inducible activator protein 2

APS - ammonium persulfate

α-S-dNTP - alpha-thio-deoxynucleotides

BAC - bacterial artificial chromosome

BCR - breakpoint cluster region

BCRL2 - breakpoint cluster like region 2

BLAST - basic local alignment search tool

bp - base pair

BRCA1 - breast-ovarian cancer gene

Bst - *Bacillus stereothermophilus*

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

CATCH22 - Cardiac Abnormality/abnormal facies, T cell deficit due to thymic hypoplasia, Cleft palate, Hypocalcemia due to hypoparathyroidism resulting from 22q11 deletion

CCD - charge coupled device

cDNA - complementary deoxyribonucleic acid

CLTC or CHC - clathrin heavy chain

CLTCL or CHC2 - clathrin heavy chain-like

COMT - catechol-O-methyltransferase

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

CREB - cyclic AMP regulatory element binding protein

CTP - citrate transport protein

DGCR - DiGeorge sydrome critical region

DGS - DiGeorge syndrome

DMSO - dimethyl sulfoxide

dATP - 2'-deoxy-adenosine triphosphate

dCTP - 2'-deoxy-cytidine triphosphate

dc^{7}GTP - 2'-deoxy-7-deaza guanosine triphosphate

ddATP - 2',3'-deoxy-adenosine triphosphate

ddCTP - 2',3'-deoxy-cytidine triphosphate

ddGTP - 2',3'-deoxy-guanosine triphosphate

ddNTP - dideoxynucleotide

ddTTP - 2',3'-deoxy-thymidine triphosphate

dGTP - 2'-deoxy-guanosine triphosphate

dITP - 2'-deoxy-inosine triphosphate

DNA - deoxyribonucleic acid

DNase - deoxyribonuclease

dNTP - deoxynucleotide

dTTP - 2'-deoxy-thymidine triphosphate

DVL-22 - human homolog to *Drosophila* dishevelled segment-polarity gene

EDTA - disodium ethylenediamine tetraacetate

EMBL - European Molecular Biology Laboratory

EST - expressed sequence tag

F34D10.2 - a *C. elegens* hypothetical protein

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

FISH - fluorescence *in situ* hybridization

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

GR - glucocorticoid receptor

GTE - buffer containing glucose, Tris-HCl and EDTA

H4TF-2 - human histone H4 gene-specific transcription factor

HIR1 - repressor 1 of histone gene transcription

HIR2 - repressor 2 of histone gene transcription

HIRA - homolog of repressors 1 and 2 of histone gene transcription

hox-1.5 - a homeobox gene

IPTG - isopropyl-β-D-thiogalactopyranoside

indels - insertions/deletions

kb - kilobase

lacZ - gene coding for β-galactosidase

lacZ' - gene coding for first 146 amino acids of β-galactosidase

LB - medium containing tryptone, yeast extract and salt

LF-A1 - liver specific trans-acting factor

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

LZTR-1 - a gene coding for a putative DNA-binding protein

mb - megabase

MDGCR - minimum DiGeorge syndrome critical region

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

MIR - mammalian-wide interspersed repeats

MLT - a human repeat element

MOPS - 4-morpholinepropanesulfonic acid

MRC - Medical Research Council

mRNA - messenger ribonucleic acid

MST - human repeat element

NCBI - National Center for Biotechnology Information

NF-1 - nuclear factor 1

NF-1/L - nuclear factor 1-like

NF-E - erythroid specific nuclear factor

NF-kappaE2 - nuclear factor kappa enhancer 2

ORF - open reading frame

OSP - oligonucleotide selection program

PCR - polymerase chain reaction

PEG - poly(ethylene glycol)

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

RFLP - restriction fragment length polymorphism

RNA - ribonucleic acid

RNase - ribonuclease

rRNA - ribosomal ribonucleic acid

SDS - sodium dodecyl sulfate

snRNPs - small nuclear ribonucleoprotein particles

SP1 - a transcription factor

SRF - serum response factor

STS - sequence tagged site

SV40 - simian virus 40

T10 - a gene coding a serine/threonine rich protein

Taq - *Thermus aquaticus*

TB - terrific broth

TBE - buffer containing Tris-HCL, boric acid and EDTA

TCAS - terminator ammonium cycle sequencing buffer containing Tris-HCl, $MgCl_2$ and $(NH_4)_2SO_4$

TE - buffer containing Tris-HCl and EDTA

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

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

TFIID - transcription factor IID

THE1b - transposon like human repeat element

TM - melting temperature or buffer containing Tris-HCl and magnesium

TOH - a male patient with DGS

tRNA - transfer ribonucleic acid

TTE - buffer containing Tris-HCl, Triton X-100, and EDTA

TUPLE1 - tup-like/enhancer of split gene 1

UFD - ubiquitin fusion-degradation

UFD1L - ubiquitin fusion-degradation protein 1-like

UTR - untranslated region

VCFS - velo-cardio-facial syndrome

VDU - a female patient with DGS, mother of ADU

WD40 - a number of highly conserved repeating units ending with trp-asp (W-D)

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.

YAC - yeast artificial chromosome

ZNF74 - a zinc finger gene

ABSTRACT

The complete nucleotide sequence of four regions of DiGeorge Sydrome Critical Region in human chromosome 22, band q11, totally 482 kb, has been determined by sequencing an overlapping set of cosmid genomic clones that have been mapped to these regions. The sequencing strategy involved a modified diatomaceous earth-based large scale DNA isolation procedure to obtain the genomic cosmid clone, physical shearing of the clones by nebulization to generate randomly distributed DNA fragments, subcloning the end-repaired, kinased and size-selected DNA fragments into pUC-based vectors to get chimeric subclones. The chimeric subclones then were isolated using a modified alkaline lysis procedure manually or automatically on a Beckman® Biomek 2000 robotic workstation. Cycling Sequencing reactions were performed using dye-labeled primer or terminator chemistry, and loaded on an ABI® 377 DNA sequencer for automated sample electrophoresis, detection, data collection and base calling. The ABI-called shotgun sequences were transferred to Sun® SPARCstation and then edited, assembled and proofread by using several available editing and assembly programs. The final contiguous sequence in each regain was analyzed for features such as potential promoters, potential coding regions, repeated elements and database homology using XGrail and Powerblast.

From the data collected, part of the genomic sequence of the human clathrin heavy chain like gene (CLTCL), entire genomic sequence of the human homolog of yeast repressors of histone gene transcription HIR1 and HIR2 (HIRA), the human ubiquitin-fusion degradation protein 1-like gene (UFD1L), the human homolog of mouse T10 gene, and a putative transmembrane gene were revealed. CLTCL may be involved in endocytosis. HIRA may be a negative regulator of core histone gene transcription. UFD1L may be involved in protein degradation. The T10 gene and the putative

transmembrane gene have no homolog to any known functional genes in the database. The sequence of this approximate 0.5 mb of human genomic sequence reveals a gene density of slightly more than one gene per 100 kb with numerous ESTs and an extensive amount of repeat sequences. In addition, a ribosomal protein L34 pseudogene and a human Tigger 1-like transposon were reported. Different families of interspersed repetitive elements and their contents also were determined. The total content of these repetitive elements in these region is about 32%. The comparison of two 36 kb fragments of human genomic sequences from two individuals revealed a overall difference of 0.2%. A fragment of 90 kb of mouse genomic sequence syntenic to human HIRA gene also was sequenced and compared with human sequence. Although the coding regions of the human and mouse HIRA gene are almost identical and the genomic sequence level similarity between these two genes is low, their genomic organization is similar.

Chapter I

Introduction

1.1 DiGeorge syndrome

1.1.1 History and nomenclature

The original description of DiGeorge syndrome was presented in a published discussion at an immunology meeting [1] and DiGeorge [2] published a formal report three years later. A report by Strong [3] predated this formal report and probably represented the same variable disorder and since DiGeorge syndrome overlaps clinically with the disorder described by the Japanese as "conotruncal anomaly face syndrome" [4,5,6], it would be appropriate to use Takao syndrome for those cases with cardiac defects in contrast to the low T cell and hypocalcemic levels in DiGeorge syndrome and the craniofacial and palatal abnormalities typical of Shprintzen syndrome. Because all three phenotypes have been shown to have a 22q11 deletion, Wilson et al. [7] proposed the acronym CATCH22 (Cardiac Abnormality/abnormal facies, T cell deficit due to thymic hypoplasia, Cleft palate, Hypocalcemia due to hypoparathyroidism resulting from 22q11 deletion) as a collective acronym for those with this common genetic etiology. Shprintzen [8] objected to lumping velocardiofacial syndrome with the DiGeorge anomaly, arguing that there was no valid evidence to suggest that velocardiofacial syndrome was etiologically heterogeneous whereas the DiGeorge anomaly was known to be so, but Hall [9] cited data of Driscoll et al. [10] indicating that velocardiofacial syndrome was etiologically heterogeneous.

1.1.2 Description and clinical features

The major symptoms of DiGeorge syndrome are hypocalcemia arising from parathyroid hypoplasia, thymic hypoplasia, and outflow tract defects of the heart. [2] Disturbance of cervical neural crest migration into the derivatives of the pharyngeal arches and pouches has been proposed to account for the phenotype, which in most cases, results from a deletion of chromosome 22q11. [7, 10, 14, 15, 22, 24, 25] This deletion also may have a variety of phenotypes: Shprintzen syndrome, conotruncal anomaly face (or Takao) syndrome, and isolated outflow tract defects of the heart. [5-8] A small number of cases of DiGeorge syndrome have defects in other chromosome, notably 10p13 and in the mouse, a transgenic hox-1.5 knockout produces a phenotype similar to DiGeorge syndrome as do the tetratogens retinoic acid and alcohol. [9]

DiGeorge syndrome is clinically characterized by neonatal hypocalcemia, which may result in tetany or seisures, due to hypoplasia of the parathyroid glands, and susceptibility to infection due to a deficit of T cells. The immune deficit is caused by hypoplasia or aplasia of the thymus gland. A variety of cardiac malformations also are seen, in particular, those affecting the outflow tract. These include tetralogy of Fallot, type B interrupted aortic arch, truncus arteriosus, right aortic arch and aberrant right subclavian artery. In infancy, micrognathia may be present and the ears are typically low set and deficient in the vertical diameter with abnormal folding of the pinna. Telecanthus with short palpebral fissures also is seen and both upward and downward slanting eyes have been described. The philtrum is short and the mouth relatively small. In older children the features overlap Shprintzen syndrome with a rather bulbous nose and square nasal tip and hypernasal speech associated with submocous or overt palatal clefting. Cases presenting later tend to have a milder spectrum of cardiac defect with ventricular septal defect being common. [2, 7, 11-25]

Short stature and variable mild-to-moderate learning difficulties are common in patients with DiGeorge syndrome and a variety of psychiatric disorders have been described in a small proportion of adult cases of velocardiofacial syndrome. These have included paranoid schizophrenia and major depressive illness. Clinical features seen more rarely include hypothyroidism, cleft lip, and deafness. [2, 7, 23]

1.1.3 Genetics of DiGeorge syndrome

1.1.3.1 Inheritance

DiGeorge Syndrome is usually sporadic and results mainly from *de novo* 22 deletion. A long series of reports has recognized the variable features resulting from this deletion in multiple family members with the variable phenotype behaving as an autosomal dominant trait [17, 22-26]. Stevens et al. [26] suggested that such familial cases should be regarded as being velocardiofacial syndrome. The variable phenotype was described by Strong [3] prior to the recognition of DiGeorge syndrome. The mother of that family developed a psychotic illness. The first dominant pedigree in which marked clinical variability was associated with dominant transmission of a 22q11 deletion was reported by Wilson et al. [27]. The mother had the typical dysmorphic features and of her three affected children, one had coarctation of the aorta, one a ventricular septal defect, and one DiGeorge syndrome. Wilson et al. [27] also found five of nine families with familial outflow tract defects to have 22q11 deletion. Subtle dysmorphic features typical of those seen in DiGeorge syndrome were apparent in several of these affected family members.

1.1.3.2 Cytogenetics

De la Chapelle et al. [28] suggested that DiGeorge syndrome may be due to a deletion within chromosome 22 or partial duplication of chromosome 20 p arm, based on finding the syndrome in members of a family with a 20;22 translocation. Specifically, they observed DiGeorge syndrome in four members of one family and demonstrated monosomy of 22pter-q11 and 20p duplication. They also mentioned a patient previously reported briefly by Rosenthal et al. [29] with complete monosomy 22 and DiGeorge syndrome. Their interpretation that DiGeorge syndrome might result from monosomy for 22q11 was confirmed by Kelley et al. [30] in three patients with translocation of 22q11-qter to other chromosomes.

Greenberg et al. [31] observed partial monosomy due to an unbalanced 4;22 translocation in a two-month-old male with type 1 truncus arteriosus and features of DiGeorge syndrome. The asymptomatic mother showed partial T-cell deficiency and the same unbalanced translocation with deletion of proximal 22q11. Augusseau et al. [32] observed telecanthus, microretrognathia, severe aortic coarctation with hypoplastic left aortic arch, decreased E rosettes, and mild neonatal hypocalcemia in a patient with balanced translocation involved chromosome 2 and 22: t(2:22)(q14:q11). The same translocation was present in the clinically normal mother and maternal aunt who had her fourth pregnancy aborted because of cardiac and other malformations detected on ultrasound and this translocation has proved important in analysis of the expressed sequences in the deleted segment.[28, 30, 31, 32]

The recognition of the importance of 22q11 deletion grew with improving techniques. Greenberg et al. [33] found chromosome abnormalities in 5 of 27 DiGeorge syndrome cases, 3 with 22q11 deletion though only one of these was an interstitial

deletion. Wilson et al. [34] reported high resolution banding in 30 of 36 cases of DiGeorge syndrome and demonstrated 9 cases of interstitial deletion. All other cases were apparently normal. Use of molecular dosage analysis and fluorescence *in situ* hybridization with probe isolated from within the deleted area revealed deletion in 21 of 22 cases with normal karyotypes giving pooled results of 33 deleted among the consecutive series of 35 cases [35]. Driscoll et al. [36] also found deletions at the molecular level in all 14 cases studied.

Whereas 90% of cases of DiGeorge syndrome may now be attributed to a 22q11 deletion, other chromosome defects have been identified. In the report of Greenberg et al. [33], there was one case of DiGeorge syndrome with a deletion in 10q13 and one with a 18q21.33 deletion. Other chromosome defects also were reported to be associated with DiGeorge syndrome. A deletion of 10p was reported by Monaco et al. [37], Shapira et al. [38], and Schuffenhauer et al. [39]. Lindgren et al. reported that a chromosome duplication of 9q(21.12-q22.1) was related to DiGeorge syndrome [40]. Greenberg et al. reported that deletion 17p13 was associated with DGS [41] and van Essen et al. reported a case with isochromosome 18q(46, XX, i(18q)) who had combined manifestations of monosomy 18p and trisomy 18q and with DiGeorge syndrome [42]. Lai et al. reported a DiGeorge syndrome case with simultaneous partial monosomy 10p and trisomy 5q [43]. Fukushima et al. found a female infant with a deletion of 4q21.3-q25 and with DiGeorge syndrome [44]. DiGeorge syndrome implicated with either of the two sex chromosomes, X or Y, has been reported. Most of them involved in a chromosomal translocation of 22, with either X or Y, resulting in a monosomy of 22q. The possibility of an unrecognized submicroscopic deletion of 22q11 also was considered in these cases, although it is clear that the disturbance of neural crest migration presumed to underlie DiGeorge syndrome may be caused by several distinct defects at the molecular level. [45-49]

1.1.3.3 Mapping

Cytogenetic studies support the presence of a DGS critical region in band 22q11 but a small number of cases with microdeletions were undetectable by cytogenetic analysis even with high-resolution banding techniques. Molecular analysis methods can detect deletions in those patients with normal karyotypes. RFLP (restriction fragment length polymorphism), gene dosage analysis and FISH (fluorescence *in situ* hybridization) have been used to study in mapping the critical region of DGS. [50-53] In RFLP analysis, genomic DNA is digested by certain restriction enzymes and the resulting fragments are separated by electrophoresis. Radiolabeled probes are hybridized to appropriate bands by Southern blot. Normally, there are two bands for each probe, one is from maternal and the other is from paternal. If one of the probe-binding region is deleted, only one band can be seen. Gene dosage analysis is to determine the copy number of the DNA by quantitative analysis of Southern blots. A control probe binding to a region not involved in DGS is used simultaneously as a comparison. FISH analysis with a fluorescence labeled probe that can hybridize to metaphase chromosomes can be used to visualize the location of a large deletion. Since 1989, the laboratories of Drs. Emanuel in Philadelphia, Desmaze in Paris and Scambler in London have been isolating probes from 22q11 to determine the smallest region of overlap in DGS patients. These different groups used different probes on different patients to locate the DiGeorge critical region [50-53].

Studies [36,50,54] conducted by Dr. Emanuel's group showed that loci D22S9 (p22/34) and D22S43 (pH32) flank the critical region proximally. Loss of an allele at more distal locus, BCRL2, in one of the patient suggested that the distal boundary for DGCR might be in proximity to the BCRL2 locus. Other loci used as probes include

pH98 (D22S57), pH11 (D22S36), pR32 (D22S259), pH160b (D22S66), pH162 (D22S68), N25 (D22S75), W21G (D22S24), and 22C1-18 (D22s10). Probes N25, H160b, pR32 and translocation cell line GM5878 (a human -hamster hybrid carrying t(10;22)(q26.3;q11.2) translocation and retaining the der(10) as the only relevant human chromosome) were detected in all patients studied. Desmaze et al. [52] used many of same probes as well as additional probes such as 33-11Tg (carrying a reciprocal t(x;22) translocation), 48F8, 100C10, NOT54, and ICB100 (carrying a recurrent t(11;22)(q23.2;q11.2) translocation). They found the longest deletion region is flanked by probe D22S9 and ICB100 and the smallest deletion region is between pH11 and GM5878. Scambler et al. [55] on the other hand, detected shortest deletion region which is between pH11 and GM5878. Maps of DGCR from these three groups are provided in Figure 1. The overlapping region of the three maps occurs from pH11 to GM5878. Because it is still possible that the disease may be caused by contiguous gene deletion, it is important to study the largest deleted region to understand the variable spectrum of phenotypes.

1.1.3.4 Molecular genetics

Augusseau et al. [32] described a female patient (ADU) with DGS (see 1.3.2 Cytogenetics) that had a balanced translocation involved chromosome 2 and 22: t(2;22)(q14;q11). The same translocation also was present in her mother (VDU). The original paper reported that VDU had no features of DGS, but Budarf et al. [56] later observed that subsequent publications cited VDU as mildly affected with VCFS and DGS. The DGS phenotype in ADU, the VCFS phenotype in VDU and a balanced translocation of chromosome in both led Budarf et al. [56] to clone the translocation, sequence the region containing the breakpoint, and analyze the DNA sequence for

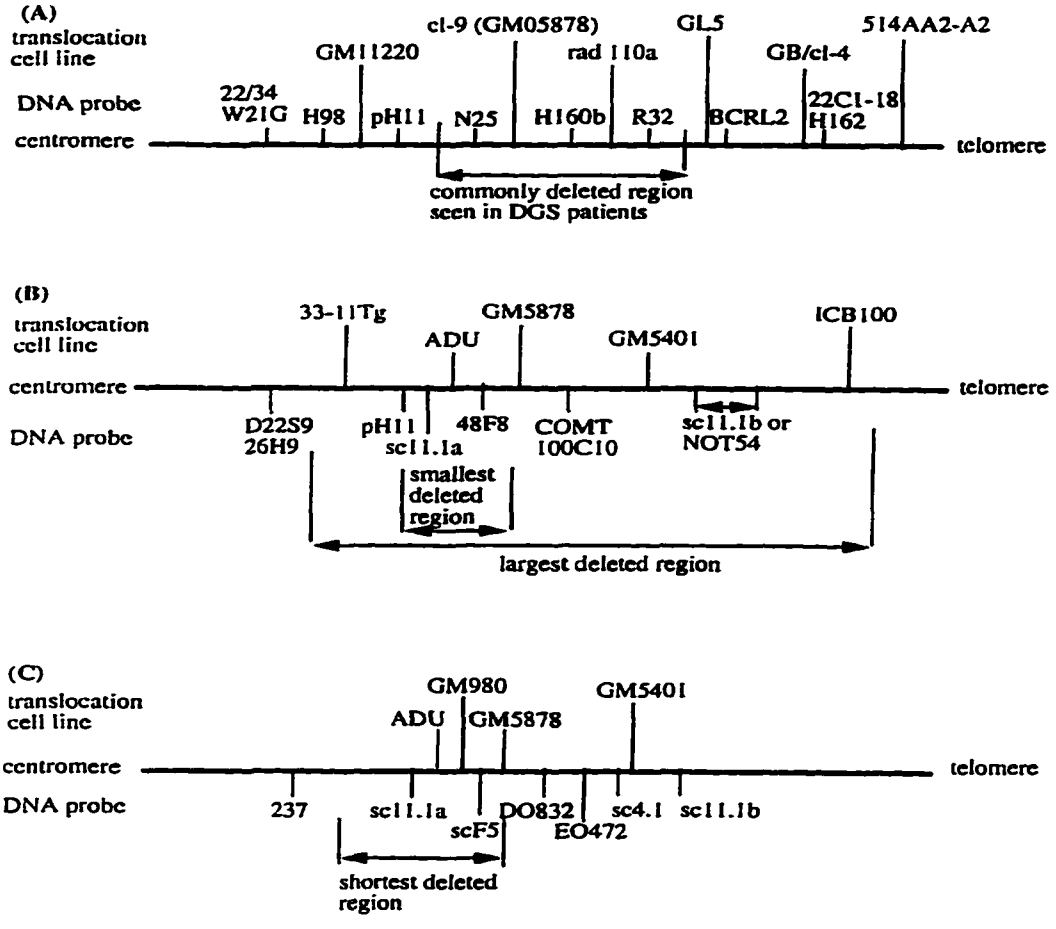


Figure 1: Map of DiGeorge syndrome critical region

transcript identification. A gene (DGCR3) disrupted by the rearrangement was identified. Their analysis suggested that there are at least two transcripts (DGCR3 and DGS-B) on opposite strands in the region of the t(2;22) breakpoint. The breakpoint disrupted a predicted open reading frame of one of these genes (DGS-B), deleting 11 nucleotides at the translocation junction. Additional fluorescence *in situ* hybridization studies and Southern blot analysis demonstrated that the deletions in chromosome 22 deletion-positive patients with DGS/VCFS include both of the transcripts at the t(2;22) breakpoint. The answer to whether either of these two putative genes may be of significance in the etiology of DGS will lie in determining whether all deleted patients are hemizygous for these loci and whether mutations in these genes are detectable in non-deleted patients with DGS. Lacking such evidence, the possibility that the translocation separates a locus control region from its target gene or produces a position effect will still remain. Most recent research suggested that a common polymorphism in DGS-B may lead to a frameshift in this open reading frame (Holmes et al., unpublished). A transcription map of the DGS and VCFS minimal critical region on 22q11 was published in 1996 by Gong et al. [57]. This study found transcription units such as DGS-A, LAN (DGS-C), DGS-D, DGS-E, DGS-F, DGS-G, DGS-H, DGS-I, CTP (DGS-J), CLTCL (DGS-K) in the minimal critical region. LAN gene, also named IDD gene, which is close to ADU breakpoint, was first isolated by Demczuk et al. [58] and Wadey et al. [59]. This is an integral membrane protein and is a putative adhesion receptor protein, which could play a important role in neural crest migration to the third and fourth pouch in embryogenesis, a process which has been proposed to be improper in the development of DGS. DGS-G is 94% identical in amino acid level to *Mus musculus* serine/threonine kinase gene [57]. CTP (citrate transport protein) is 98% similar in amino acid level to rat tricarboxylate transport protein gene [57].

A patient named TOH has several features which also have been reported in individuals with DGS/VCFS. This patient carries a *de novo* constitutional balanced translocation (21;22)(p12;q11) which has been identified within minimal DGCR [60]. Budarf et al. [61] cloned and sequenced the breakpoint of translocation and showed that it occurred in the CLTCL gene (result partially from the research of this dissertation). Because the breakpoint of the translocation on 21p12 is in a repeated region telomeric to the rDNA cluster, it is unlikely that the patient's features are caused by interruption of sequences on 21p12. The chromosome 22 breakpoint disrupts the 3' coding region of the CLTCL gene and leads to a truncated transcript.

Halford et al. [62] analyzed the extent of deletion in patients with DGS/VCFS phenotypes and refined the molecular cytogenetic map of the region. They found in no case was a deletion smaller than 300 kb and the boundary of the smallest deletion region is between sc11.1a and C1/79. C1/79 is a probe about 275 kb telomeric to scF5. (see Figure 1). Probe scF5 was used to screen a human fetal brain cDNA library and a mouse embryo cDNA library and located a cDNA for a transcriptional regulator, TUPLE1, which is similar to yeast TUP1 gene. This gene contained WD40 repeat motifs, which were first described in the beta subunit of transducin, the GTP-binding protein. Two years later, the study of Lamour et al. [63] showed that this gene is more similar to two yeast genes, HIR1 and HIR2, that are repressors of histone gene transcription. Therefore the name of TUPLE1 was changed to HIRA. This dissertation will reveal the whole context of the genomic organization of this gene.

Because the majority of the patients have larger deletions than the minimal DGCR, studies on neighboring regions also were carried out. Aubry et al. [64] identified a zinc finger gene ZNF74 which has an open reading frame containing 12 zinc finger motifs as well as N-terminal and C-terminal non-zinc finger domains. This

gene was found to be hemizygously deleted in 23 out of 24 DGS patients and expressed in human and mouse embryos but not in adult tissues. Hemizygous deletion of this gene may result in changes of the dosages of putative transcription factor. Halford et al. [65] isolated a cDNA at D22S134, which is distal to minimal DGCR. This gene, named T10, is expressed during early embryogenesis in mouse and encodes a serine/threonine rich protein. The genomic organization of this gene was determined during part of this dissertation research and will be discussed later. Kurahashi et al. [66] reported a gene, designated LZTR-1, which has an open reading frame encoding 552 amino acids and has several characteristics of DNA-binding proteins. This gene is transcribed in several essential fetal organs. Their study showed that this gene is hemizygously deleted in 7 out of 8 DGS patients. Its structural characteristics identifying it as a transcriptional regulator suggest that it may play a role in embryogenesis and that haploinsufficiency of this gene may be partly related to the development of DGS. Pizzuti et al. [67] found a human cDNA homologous to the *Drosophila*'s dishevelled (dsh) segment-polarity gene, a gene required for establishment of fly embryonic segments. The 3' UTR (untranslated region) of this transcript (DVL-22) was located in DGCR and was deleted in a DGS patient they studied. The pivotal role of dsh in fly development suggests an analogous key function in vertebrate embryogenesis of its homologue gene. Since DGS may be due to perturbation of differentiation mechanisms at decisive embryological stages, this gene also might be a candidate for the pathogenesis of DGS. Demczuk et al. [68] also found a gene, DGCR6 in DGS critical region. The putative gene encodes a protein similar to *Drosophila*'s gonadal protein, which participates in gonadal and germ-line cells development, and the γ-1 chain of human laminin, which upon polymerization with α- and β- chains forms the laminin. Laminin binds to cells through interaction with a receptor and has functions in cell attachment, migration and tissue organization during development. COMT gene was first mapped to chromosome 22 by Brahe et al. [69] and

further mapped to 22q11 by Winqvist et al. [70]. Lachman et al. associated this gene to DGS/VCFS in 1996 [71]. COMT (catechol-O-methyltransferase) has two forms. One is soluble and the other is membrane bound. Lachman et al. identified a polymorphism in membrane-bound form leads to a valine to methionine substitution at amino acid 158. Homozygotes for COMT158met has a 3-4 fold of reduction in enzymatic activity compared with homozygotes for COMT158val [71].

Demczuk et al. [72] pointed to the existence of a strong tendency for 22q11.2 deletions in DGS, VCFS and isolated connotruncal cardiac disease to be of maternal origin. With their experience of 22 cases in which parental origin could be determined, combined with recent results from literature, 24 cases were found to be of maternal origin and 8 of paternal origin, yielding a probability level lower than 0.01.

1.1.3.5 Heterogeneity and population genetics

The association of the DiGeorge syndrome with at least 2 and possible more chromosomal locations suggests strongly that several genes are involved in control of migration of neural crest cells and their subsequent fixation and differentiation at different sites. In the mouse, Chisaka and Capecchi [73] described a knockout of hox-1.5 which produced a recessive phenocopy of DGS. Mice heterozygous at the hox-1.5 locus appear normal, whereas hox-1.5^{-}/hox-1.5^{-} mice die at or shortly after birth. The homozygous mice have deficiencies remarkably similar to the pathology of the human congenital disorder DGS but this gene maps to human chromosome 7, in an area not yet implicated in the cause of the human DiGeorge syndrome.

One explanation for the wide variation in phenotype would be the need of more than one gene defect to produce the severe version. Thus, for example, multiple

signals and receptors should be effected to produce the full phenotype. Environmental factors also could play an additional role as features of DGS have been described in children with clinical evidence of fetal alcohol syndrome [74]. The alcohol may have directly disrupted neural crest migration or have exposed a genetic predisposition. In addition Lammer et al. [75] studied a series of pregnancies exposed to the teratogen isotretinin (vitamin A) and several of the infants satisfied the diagnostic criteria of DGS. Thus, it is likely that these environmental changes are exposing the same susceptible pathways of development as are impaired by the 22q11 deletion although the possibility of an interaction between the insult and genotype remains open.

No formal population study has yet been completed, although a preliminary study in the Northern region of England, which has a birth population of 40,000 per annum, has revealed 9 cases born in 1993 with 22q11 deletion who presented with neonatal DGS features. The overall birth prevalence of DGS therefore would appear to be approximately 1 in 4,000 [76].

1.1.4. Sequencing analysis

The majority of patients with DGS/VCFS (about 90%) have a common large deletion, and a minority have smaller deletions, unbalanced translocations or no detectable abnormality of 22q11.2 [61]. Though the phenotype is highly variable, there is no apparent correlation between deletion size and phenotype [36]. The minimal DGS/VCFS critical region (MDGCR), defined as the region most likely to contain the gene or genes for DGS/VCFS, has been studied extensively. Different groups have their own definitions of MDGCR because of the different probands they used (see 1.1.3.3. Mapping for detail). With the help of DNA sequencing, a transcription map of MDGCR has been published [57]. Combined with this transcription map and studying

of balanced translocation, three disrupted transcription units were found. One of them, DGCR3 [57], seems to be within another newly found transcription unit, DGCR5 [77]. The other transcription unit recently was found to be the functional gene for CLTCL, clathrin heavy chain-like gene [61]. The sequence data of the genomic context of this gene revealed that the breakpoint is within intron 27 [61]. DGCR5 and CLTCL genes are the only two genes in MDGCR which span the balanced translocations. However, the relationship between genes and the features of DGS/VCFS still remains unknown. It is possible that the translocation breakpoints are exerting position effects on a gene or genes in the region or the neighboring region. To further understand the deletion mechanism and the genomic organization of the critical region of DGS/VCFS, Dr. Roe's research group has been sequencing the entire 1.4 megabase DGCR. This sequence data has been proven to be a great help in establishing the transcription map, localizing known genes in the region, studying the genomic organization of genes, revealing the details of translocation breakpoint, and discovering new genes within the region [56, 57, 61, 78]. This sequence data in the DGCR is from numbers of overlapping cosmids, fosmids, BACs, and PACs. A map of these cosmids, fosmids, BACs and PACs is provided in Figure 2. Database searches reveal the locations of several known genes and human homologs of known genes of other species. The presence of numerous STS's and EST's also has aided the discovery of new genes in this region. Computational approaches, such as Xgrail, promoter scan, and signal scan programs, are used to predict the potential exons and promoter regions. Combining the locations of predicted exon cluster, potential promoter region and EST's together can be very useful to discover new potential genes. A large number of interspersed repeat elements and single repeats also have been identified. The distribution of these repeats throughout the region can help to understand the clue of how the chromosome DNA is organized. After all known and potential genes are localized in the region and the genomic organization of all these genes are revealed, one can possibly answer the

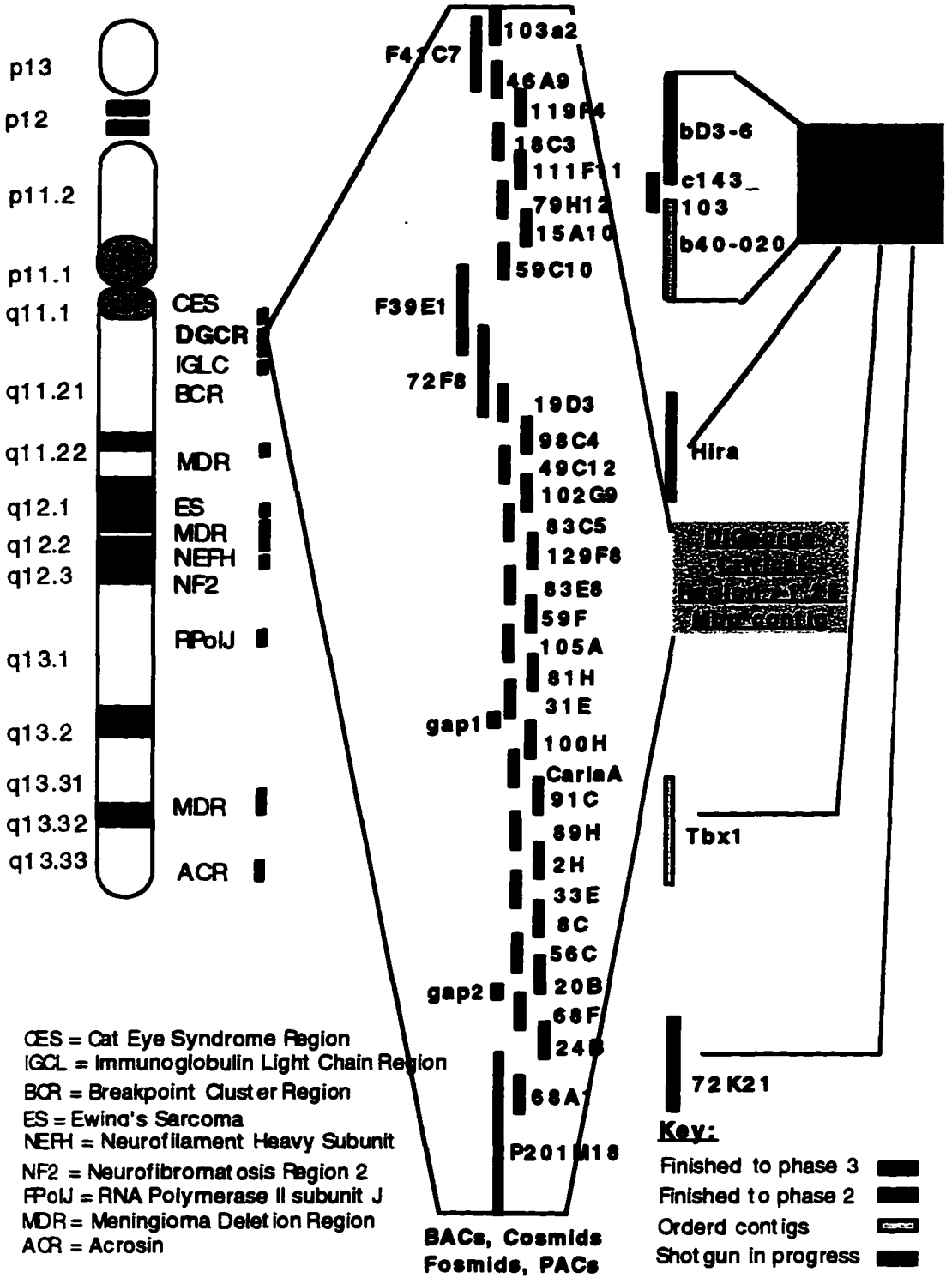


Figure 2: Map of cosmids, fosmid, BACs and PAC in DGCR

questions: Which gene or genes are responsible for DGS/VCFS or which regulatory factor or factors are responsible for DGS/VCFS? What causes the broad spectrum of severity observed in DGS/VCFS patients? How are the products of potentially responsible genes involved in early embryonic development which causes DGS/VCFS? What is the cause of deletions? What is the deletion mechanism?

1.2 DNA

1.2.1 Inheritance and DNA

Far before DNA was discovered, the phenomenon of inheritance was observed and attempts were made to understand its mechanism. As ancient Chinese put it: "Dragon breeds dragonet, a phoenix breeds a little phoenix, and mice's son can surely dig holes". Inheritance was defined as the ability of passing parental characteristics to next generation. However, its physical basis was not understood until the first year of the twentieth century, when, during a remarkable period of creative activity, the chromosomal theory of heredity was established. Hereditary transmission through the sperm and egg became known by 1860, and in 1868 Haeckel, noting that sperm consisted largely of nuclear material, postulated that the nucleus was responsible for heredity. When the theories of mitosis, meiosis and fertilization were accomplished, it could be seen that unlike other cell constituents, the chromosomes were equally divided between daughter cells. Moreover, the complicated chromosomal changes in meiosis which reduced the number of chromosomes of sperm and egg in half became understandable as necessary to keep the chromosome number consistent after fertilization. These facts suggested that chromosomes carry the heredity information [79, 80, 81].

It was not until 1900 that Gregor Mendel's basic rules of heredity were accepted by the scientific community. Although Mendel proposed his basic rules of heredity in 1865, they were completely ignored by prominent biologists of that time. After de Vries, Correns and Tschermak, all working independently, reached similar conclusions before they knew of Mendel's work, the great importance of Mendel's forgotten work was realized. Mendel's work traced the results of breeding experiments between strains of garden peas (*Pisum sativum*) differing in well defined characteristics. Mendel first related the physical appearance (phenotype) with its genetic composition (genotype), and proposed that a pair of "particulate factors", one from each parent, carried the information of heredity, passed unchanged from parent to progeny, and control the offspring's appearance. He also developed the concepts of dominance, recessivity, segregation and independent assortment. The dominance/recessivity concept discriminated two particulate factors carried by one individual. The dominant one will control the phenotype while the recessive one is in effect irrelevant. These two particulate factors have no permanent effect on one another when present in the same individual but segregate unchanged by passing into different gametes. When two different characteristics were reviewed, he concluded that the factor-pairs controlling different phenotypes will segregate independently, in modern word, nonhomologous chromosomes undergo independent segregation [79, 80, 81].

Walter Sutton, in 1903, emphasized the importance of the factor that the diploid chromosome group consists of two morphologically similar sets and that, during meiosis, every gamete receives only one chromosome of each homologous pair. He used this fact to explain Mendel's results by relating "particulate factors" to chromosomes. Though Sutton's work didn't prove the chromosomal theory of heredity, it was very important in bringing together the independent disciplines of genetics and cytology [79, 80, 82]. In 1910, Johannsen called these "particulate factors"

genes [79, 80, 81]. In 1910, Morgan's work on *Drosophila melanogaster* proved that a specific gene always lies on a certain chromosome. It also was observed that if two genes are on the same chromosome (linked genes), they tended to be inherited together (linkage). Linkage is seldom complete because homologous chromosomes attach to each other during meiosis and often break at identical spots and rejoin crossways [79, 80, 81, 83].

The relationship between the structures of the genes and the biochemical mechanism by which they control the cellular characteristics was the next question. Each gene exists in a variety of different forms called alleles. Different alleles arise from inheritable changes called mutations. Studies on large amount of spontaneous mutations in *Drosophila* led to the conclusion that many genes worked together to control one character such as eye color and wing shape [79, 80, 81, 84]. Garrod was the first to propose the general hypothesis of a gene-enzyme relationship in 1909. In his studies of an hereditary disease affecting amino acid metabolism, he suggested that the illness was caused by a lack of a particular enzyme, which was required in the amino acid metabolism pathway, and proposed that the gene controlled the synthesis of the enzyme [79, 80, 85]. In 1941, Beadle and Tatum introduced a red bread mold (*Neurospora*) as a genetic model system to study metabolism. Different mutants were generated by irradiation with X-ray and selected for their inability to grow on a medium that could allow the wild-type cell to grow. By adding additional nutritional compounds to the medium, the mutant's ability to grow was restored. They concluded that each mutation disrupted a certain gene whose product is involved in one of the metabolic steps and made the cell unable to use the nutrition the wild-type is able to use. Because each mutation blocked a particular step of metabolic pathway, adding the nutrients which appeared downstream of this step restored the cell growth. Based on this study, they proposed the "one gene-one enzyme" theory and confirmed Garrod's

hypothesis [81,86]. In more modern terms, considering some proteins consist of more than one kind of subunit, it is more precise to express "one gene-one enzyme" as "one gene-one polypeptide chain" [81].

Studies by Fred Griffith in 1928 on bacterium *Pneumococcus* strains brought the first idea of DNA as the genetic material. *Pneumococcus* kills mice by causing pneumonia. The virulence of the bacterium is determined by the capsular polysaccharide, a component of the surface, which gives smooth appearance and protects the bacterium from destruction by the host. A wildtype strain of *Pneumococcus* (I S), which is able to synthesize a capsular polysaccharide and has the ability to kill the mouse, and a mutant strain of *Pneumococcus* (II R), which has lost the ability to synthesize the capsular polysaccharide, has rough surface and can not kill the mouse, were used. Their results showed that coinjection of heat-killed smooth bacteria (I S) and live rough bacteria (II R) into mice resulted in pneumonic infection and death, whereas injection of only Heat-killed I S or live II R did not. Furthermore, the virulent bacteria recovered from the mixed infection had the smooth coat of type I. Based on these observations, Griffith presented his transformation hypothesis that some "transforming principle" from dead type I S bacteria can transform the live avirulent II R bacteria so that they make the type I capsular polysaccharide, and as a result become virulent [81, 87]. Griffith's "transforming principle" was chemically characterized by Avery, MacLeod and McCarty in 1944 after purifying the "transforming principle" from extracts. When the "transforming principle" was treated with deoxyribonuclease, an enzyme that specifically degrades DNA molecules to their nucleotides and has no effect on the integrity of protein molecules or ribonucleic acid (RNA), transforming activity of the extract was destroyed. The addition of either enzyme ribonuclease, which degrades RNA, or various proteolytic enzymes, which destroy protein, had no influence on transforming activity. They then concluded that

the "transforming principle" to be deoxyribonucleic acid (DNA) [81, 88]. At that time, DNA was not even known to be a component of bacterium. However, it had been recognized for many decades that DNA was a major component of eukaryotic chromosomes. The results that showed that the genetic material of a prokaryote also is DNA indicated DNA is probably the genetic material in both prokaryote and eukaryote [81].

In 1952, Hershey and Chase confirmed that DNA also was the genetic material of a quite different system, T2 bacteriophage. When phages were labeled with ^{32}P in DNA or with ^{35}S in protein of phage coats, the ^{32}P label entered the infected bacteria and can be recovered from progeny phages. The ^{35}S label was found to be in empty phage coats but not inherited by the progeny [81,89]. The similar observation also was obtained from studies in eukaryotes. When purified DNA containing information of particular phenotype mixed with single eukaryote cells which lack this particular phenotype, some of the cells could uptake the DNA and obtain this phenotype [81].

These experiments showed directly that DNA is the genetic material. Now we know that all known organisms and many viruses use DNA as genetic material. However, some viruses use ribonucleic acid (RNA) as genetic material. In these circumstances, RNA serves the same role as DNA.

1.2.2 DNA components

DNA is a polymer consisting of a chemically linked sequence of nucleotides. Each nucleotide contains three major components: a nitrogenous base, a five carbon sugar in ring form, and a phosphate group. DNA has four kinds of nitrogenous bases (Figure 3). They fall into two groups, pyrimidines and purines. DNA has adenine and

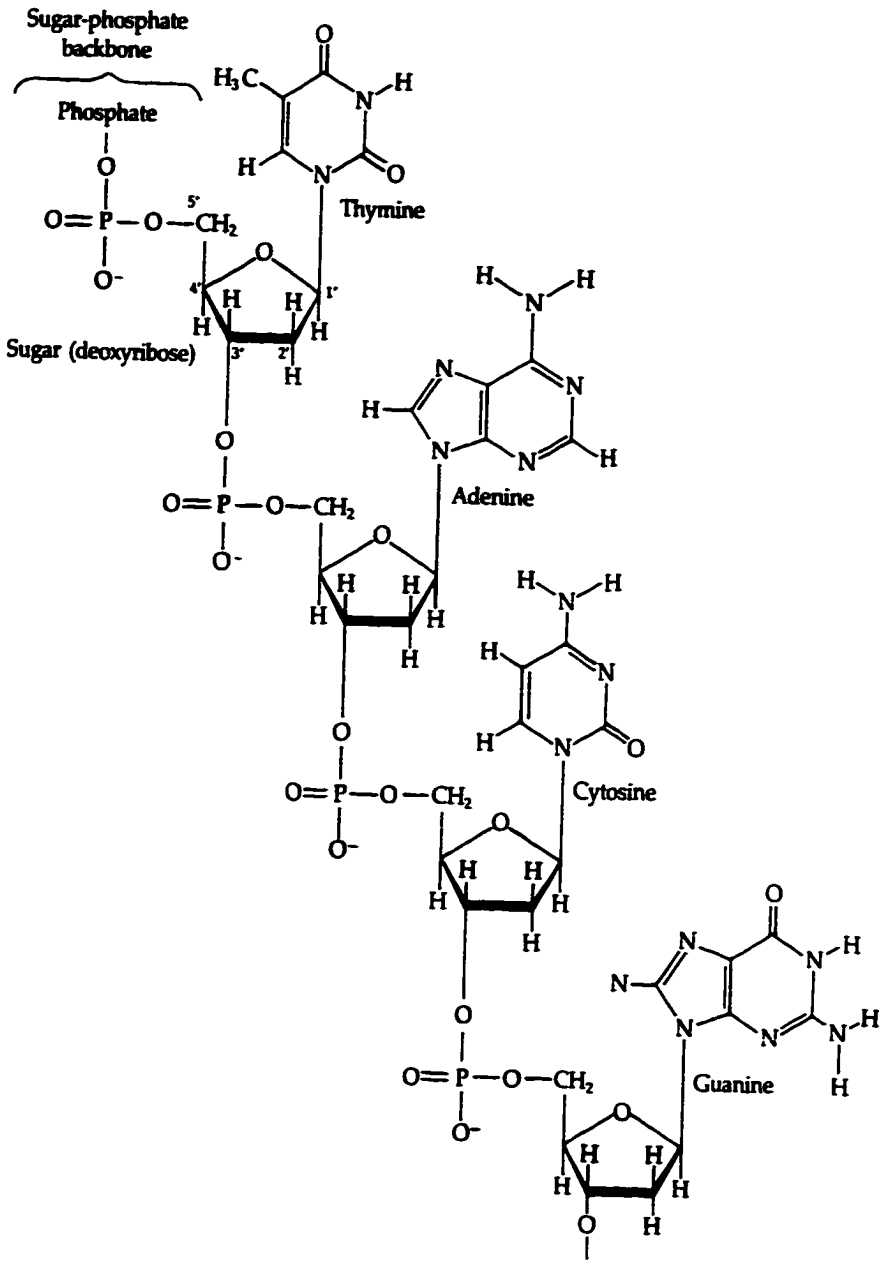


Figure 3: Structures of bases, pentose and phosphate in DNA.

guanine as purines and cytosine and thymine as pyrimidines. The five carbon sugar in DNA is 2-deoxyribose (Figure 3). The nitrogenous base is linked to position 1 on the pentose ring by a N-glycosidic bond from N_9 of purines or N_1 of pyrimidines. The phosphate group is linked to 5' position of the pentose via phosphodiester bond. Nucleotides are building blocks from which DNA is constructed. The DNA polynucleotide is an unbranched polymer whose backbone consists of an alternating series of sugar and phosphate residues. The 5' phosphate of one nucleotide is bound covalently to the 3' hydroxyl of the sugar on the adjacent nucleotide. The terminal nucleotide at one end of the chain has a free 5' phosphate group and the other end has a free 3' hydroxyl group. Thus, the DNA chain has a direction. In 1951, Chargaff, et al. found out that double stranded DNA consists of equal amounts of adenosine (A) and thymidine (T) and equal amounts of guanosine (G) and cytidine (C). The molar ratios of ([A]+[T]/[G]+[C]) differed among organisms, a finding which suggested that DNA molecules might be more complicated 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 [79, 80, 90].

1.2.3 DNA structure

Three major pieces of evidence helped Watson and Crick in 1953 conclude that DNA usually consists of two complementary polymeric strands that form a double helix. First, X-ray diffraction data showed that DNA has the form of a regular helix, making a complete turn every 34 Angstroms (3.4 nm), and with a constant diameter of 20 angstroms (2 nm). Because the distance between adjacent nucleotides is 3.4 angstroms, there are 10 bases per turn of the helix. Second, the density of DNA suggests that the helix must contain two polynucleotide chains. The constant diameter

of the helix can be explained only if the bases in each chain face inward and a purine is always opposite a pyrimidine, because purine-purine pairs are too large and pyrimidine-pyrimidine pairs are too small. The hydrophobicity of bases also suggests that bases face inward, making a hydrophilic backbone on the surface of the helix. Third, the proportion of G and C is always the same in DNA (G = C), and it also is true for the proportion of A and T (A = T). This indicated that A might pair with T and G might pair with C [79, 80, 81, 91, 92, 93].

Watson and Crick proposed that most of the time, hydrogen atoms in the bases are found at precise locations instead of as was previously assumed, that they are highly mobile, randomly moving from one ring nitrogen or oxygen to another. The nitrogen atoms attached to the bases are almost always in the amino (NH_2) form rather than imino (NH) form and the C-6 atoms are keto (C=O) rather than enol (COH). These structure features of the bases allow hydrogen bonds only between A and T, and G and C (Figure 4). Thus, the DNA molecule has two polynucleotide chains coiled together in opposite direction (antiparallel) to form a right-handed double-helix. The nitrogenous bases face inward and two hydrogen bonds between A and T in different chains and three hydrogen bonds between G and C keep these two chains together. Bases are flat structures lying in pairs perpendicular to the axis of the helix (base-stacking). This base-stacking allows interaction of the aromatic π electron clouds and diminishes the electrostatic repulsion expected between the negatively charged phosphate backbones of two chain. Besides the hydrogen bonds, hydrophobic base-stacking also keeps the DNA structure stable. The twisting of two strands around one another gives two grooves on the surface of the helix, narrow and wide grooves. The narrow groove is about 12 angstroms and the wide groove is about 22 angstroms [79, 80, 81].

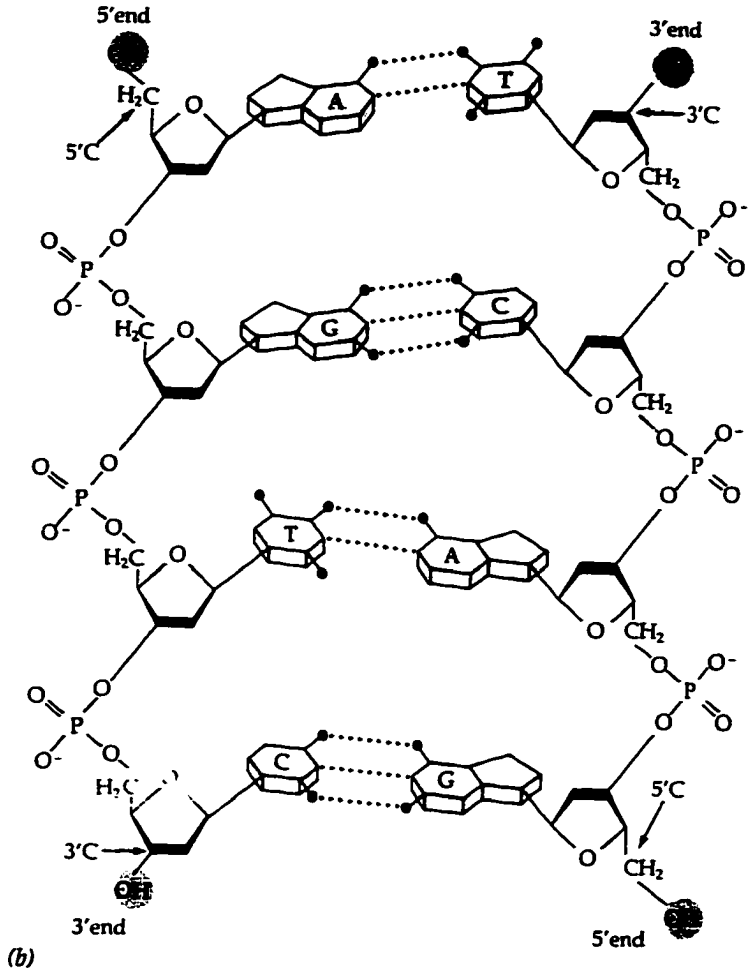


Figure 4: Hydrogen bonds between bases and antiparallel backbones.

This structure represents a common form of DNA, B-form, existing in solution under physiological conditions, but there are other forms of DNA. The A-form DNA is very close to the conformation of double-stranded regions of RNA, where the presence of 2'-OH group prevents adoption of the B-form. In A-form DNA, there are 11 bases per turn and the diameter of the helix is about 23 angstroms and the bases are not lying perpendicular to the axis but tilted [80, 94, 95, 96, 97, 98]. The Z-form DNA, is a left handed double helix having only one deep groove with a greater density of negative charges than either groove in B-form (See Figure 5). Z-DNA contains 12 base pairs per turn with a diameter of approximately 18 angstroms, making this structure much more compact than B-form DNA [80, 99, 100, 101]. The Z-form DNA double helix occurs in polymers with a sequence of alternating purines and pyrimidines as in CpG island which is seen in the promoter region of most house keeping genes. Besides the sequence of the nucleotides, negative supercoiling is another factor that determines which form DNA adopts. Negatively supercoiled DNA seems to have greater possibility of exiting in a Z-form than relaxed DNA . Since Z-form DNA has been reported in regions of SV40 virus that regulate transcription, Z-DNA is thought to have an effect on transcription [102]. Proteins bound to DNA also may affect the ability of DNA to adopt different forms [103]. Methylation of key cytosine residues also may influence gene functioning through its potentiation of the transition from B-form to Z-form [104,105]. Miller, et al. reported that the conversion from B-form DNA to Z-form in nucleosomes could change the superhelical density of a DNA segment, and therefore alter gene expression [106].

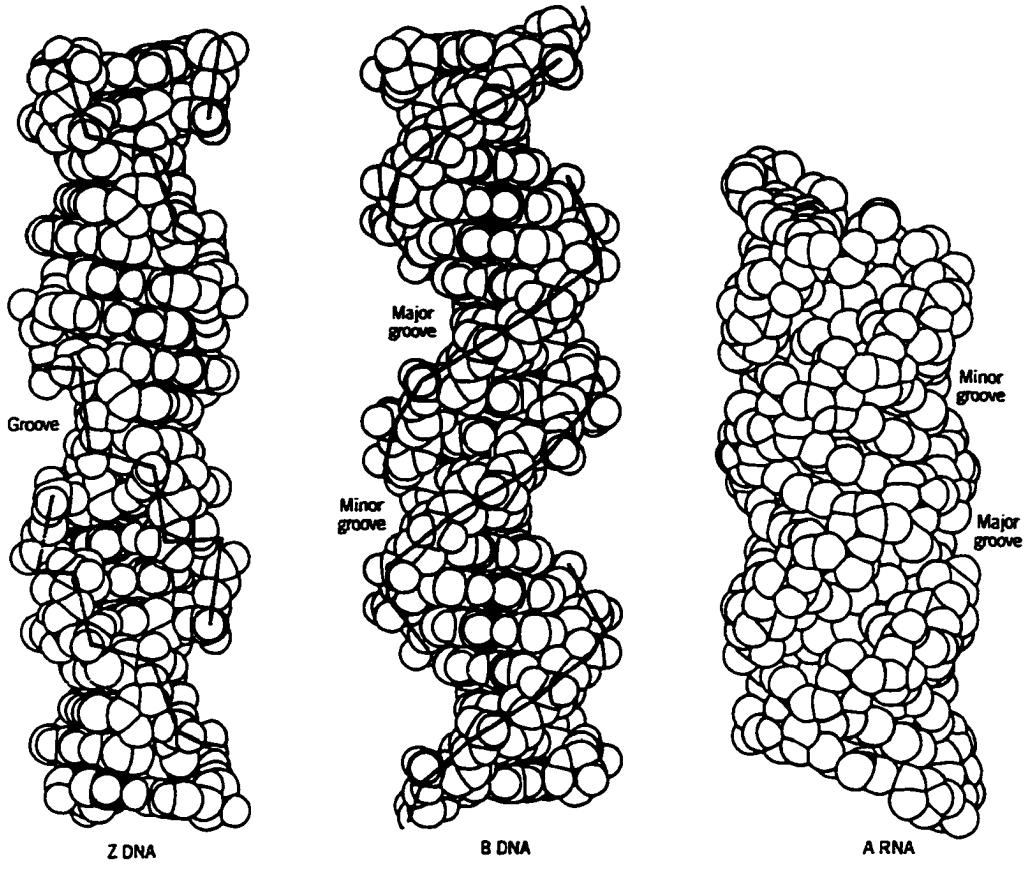


Figure 5: Comparison of A-form, B-form, and Z-form DNA.

1.2.4 DNA and genetic information

How does the DNA structure relate to the genetic information it carries? How is this information coverted to biochemical metabolic functions? The central dogma proposed by Francis Crick may answer these questions [107]. The central dogma defined the paradigm of the molecular biology: genes are perpetuated as sequences of DNA (or in some cases, RNA), but the functions are carried out by being expressed in the form of proteins.

Central dogma contains two components (Figure 6): The first is DNA replication. Duplication of DNA to make an identical copy passes the genetic information from one generation to next. The second is gene expression which contains two stages. DNA replication provides a way for genetic information to pass from parent to offspring. Two major features should be mentioned here for DNA replication. First, DNA can only be synthesized from 5'-end toward 3'-end. Second, no known DNA polymerase activity can initiate a DNA chain *de novo*, instead, it is extended from a pre-existing chain (primer), DNA or RNA, with a free 3'-OH end [80, 81, 110, 111, 112, 113]. In gene expression, transcription generates a single-stranded RNA identical in the sequence with one of the strands of the duplex DNA. Translation converts the sequence of RNA into the sequence of amino acids comprising a protein. In eukaryotes, DNA is in the nucleus and protein synthesis occurs in the cytoplasm. RNA acts as a messenger to transfer the information DNA carries outside the nucleus into the cytoplasm. This RNA is called mRNA [108,109]. Only one DNA strand is the template for mRNA synthesis. The template DNA strand which is copied and thus complementary to the mRNA is called the anticoding or negative strand. The other one which has the same sequence as mRNA (except in RNA a U is used instead of a T) is called the coding or positive strand [79, 80, 81].

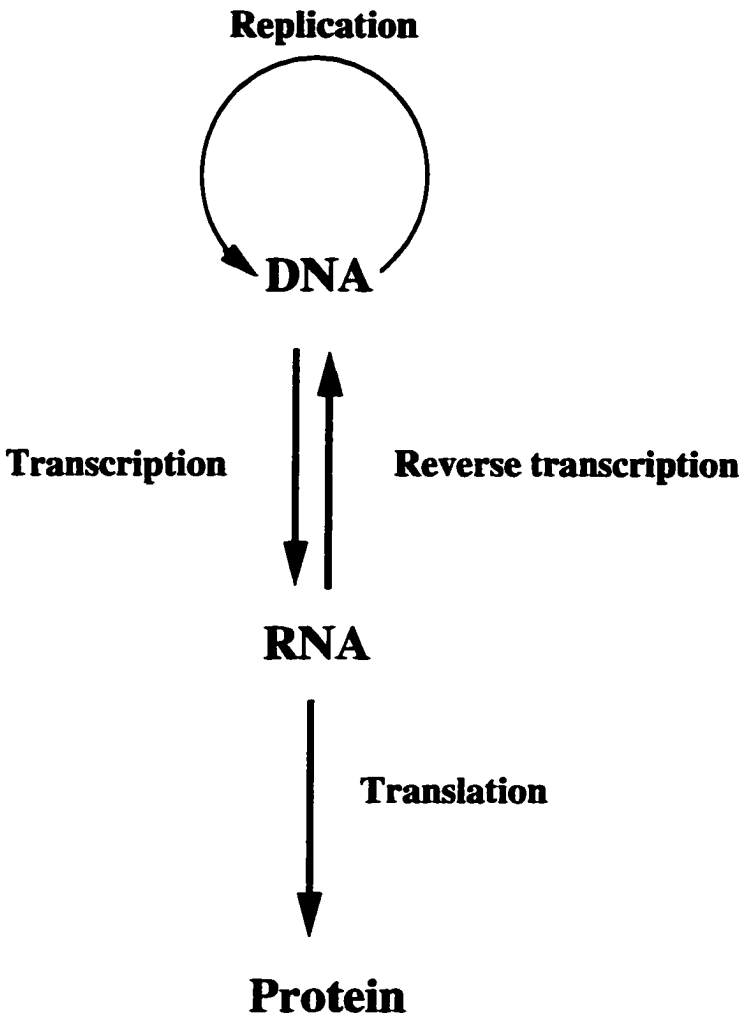


Figure 6: Central dogma of molecular biology

Transcription utilizes RNA polymerase. There are three RNA polymerases in eukaryotes. RNA polymerase I is responsible for synthesis of ribosomal RNA (rRNA) which is a constituent of ribosome [114], the factory of protein synthesis. RNA polymerase II is responsible for mRNA synthesis. RNA polymerase III is responsible for synthesis of transfer RNA (tRNA), which recognizes a free amino acid and transfers it to a ribosome for protein synthesis [115]. Transcription is facilitated by the presence of specific promoters, containing specific consensus sequences, which are recognized by RNA polymerase. Promoters for RNA polymerase I, and most promoters for RNA polymerase II are located upstream of the transcription startpoint [116, 117, 118, 119], and in most cases promoters for RNA polymerase III lie downstream of the startpoint [120]. Because RNA polymerase II transcription results in mRNA, it will be emphasized here. RNA polymerase II usually requires specific consensus sequences in promoter region, such as the TATA box, with a consensus sequence of TATAA, which is located 25-30 base pairs upstream from the transcription start site. The TATA box is recognized by TFIID (transcription factor IID) and delineates the correct positioning of the RNA polymerase II for transcription initiation and specifies the 5' cap termini in the mature mRNA [121, 122]. Further upstream from TATA box is a region which can contain one or more upstream regulatory elements such as the CAAT box and the Sp1 (a transcription factor containing zinc-fingers) consensus binding site GGGCGG. The position of CAAT box can vary considerably from the transcription startpoint. It may also be present in either orientations. CAAT box may play a role in controlling the efficiency of the promoter. GGGCGG sequence (GC box) usually has multiple copies in different locations and can occur in either orientations [123, 124, 125, 126, 127]. Even though these regulatory elements are often described as promoters, none of them are essential, and a gene promoter region may lack any one of them and still be transcribed. In addition, several eukaryotic promoters

have been shown to not necessarily function alone. Enhancers, which may be located at a variable distances from promoter, upstream or downstream, can increase the activity of promoter enormously. Enhancers also may not be specific for certain promoters, and can stimulate any promoter placed in its vicinity [128, 129, 130]. Response elements are found either in promoters or enhancers, and are recognized by factors that coordinate the transcription of particular groups of genes. Response elements are switches that can turn particular gene(s) on in the presence of certain factors, such as some metals, hormones (e.g. glucocorticoid) and serum. Such a factor binds to a receptor in cytoplasm, the complex then goes into the nucleus and binds to response elements that regulate the transcription [131, 132].

Little is known about the mechanism of termination of transcription of RNA polymerase II. A consensus sequence, known as poly(A) signal, is usually found near the end of transcription stop point. This sequence may not signal the termination of the transcription but rather signal the addition of poly(A) tail to mRNA. Actually, RNA polymerase II may continue elongating the RNA chain far beyond the polyadenylation signal site, but after it is somehow terminated, the nascent pre-mRNA then is cut by a specific endonuclease, which creates the correct 3' terminus for polyadenylation and releases downstream sequences for rapid degradation. The poly(A) tail is believed to control the stability and the lifetime of a mRNA [133, 134, 135].

The coding regions, termed exons, usually are not continuous in eukaryote genes as they are separated by non-coding regions, termed introns. The size of the introns usually is much larger than the size of the exons [80, 81, 136, 137, 138, 139]. The immediate product of RNA polymerase II is only a precursor of mRNA because it contains large amount of non-coding regions. Before the mRNA precursor can be used as a template of protein synthesis, the introns must be removed by a process called

splicing, which occurs in the nucleus [138, 140]. During splicing the pre-mRNA is cleaved after a G residue (underlined in 3'-end of exon and 5'-end of intron consensus junction: (C/A)A**G**/**gt**(a/g)agt, the bold is most conserved [141, 142]) at the end of the upstream exon. The resulting free guanosine at the end of the intron reacts with a residue (usually an A) in the middle of the intron (usually near the right end of the intron) to form a 5' to 2' phosphodiester bond [143, 144, 145]. After this lariat structure is formed, the 3'-end of intron and 5'-end of the down stream exon consensus junction $((t/c)_n n(c/t)$**ag**/G) [141, 142] is cleaved between two G's and the two exons are ligated. This process has been studied in detail and shown to require small nuclear ribonucleoprotein particles, termed snRNPs, which contain small nuclear RNAs and special proteins [146, 147, 148].

1.2.5 DNA packaging: genome and chromosome

The three billion base pairs of the human genomic nuclear DNA in a single cell is roughly one meter long and mainly B-form double helix. Human genomic DNA is organized and packaged into a set of 46 chromosomes in a haploid cell, two copies of chromosomes 1-22 and the sex chromosomes XX or XY. The shape of a typical chromosome in metaphase is provided in Figure 7. A large number of proteins is bound to DNA. Some of these proteins, histone and non-histone proteins, help package the DNA into very dense form. The basic structural unit of a chromosome is chromatin. Chromatin consists of densely packed nucleosomes. A nucleosome contains a characteristic length of DNA, usually about 200 bp, wrapped two times around an histone octamer containing two of each histone: H2A, H2B, H3, and H4. Histone H1 binds to the DNA between nucleosomes and facilitates chromatin packaging. Nucleosomes are organized into chromatin which is a helical fiber of 30 nm

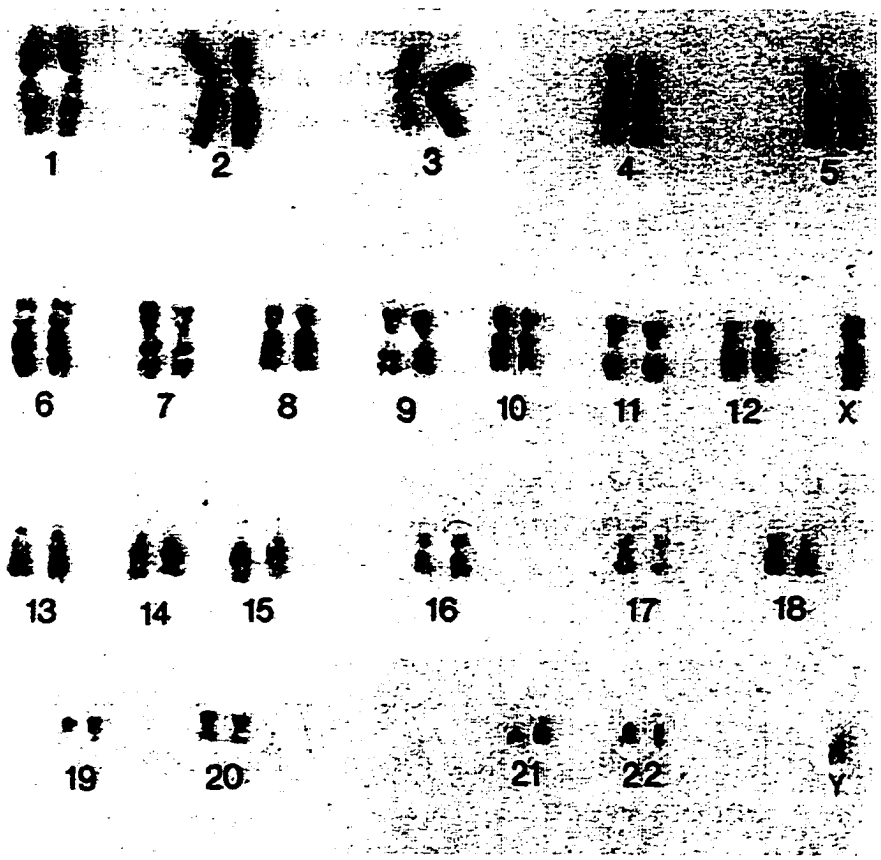


Figure 7: Typical human chromosomes in metaphase.

diameter having six nucleosomes per turn. Without H1, the chromatin is 10 nm in diameter, in which nucleosomes are organized edge to edge and stacked along. Non-histone proteins are responsible for further packaging [149, 150, 151, 152, 153]. The chromatin packaging is regulated by phosphorylation of histone H1 prior to mitosis, and dephosphorylation following mitosis. As a result of this covalent modification, chromatin has different densities. The very densely packed region, in a condition comparable to that of the chromosome at mitosis, is known as heterochromatin. Heterochromatin is a region of relatively inactive transcription [154]. Two kinds of heterochromatin exists. Constitutive heterochromatin consists of particular regions that are not expressed, such as repetitive sequences. Facultative heterochromatin is regions inactive in one cell lineage but which may be expressed in other cell lineages [81]. In most regions however, the chromatin is less densely packed than in the mitotic chromosome. These regions are called euchromatin. Euchromatin is actively transcribed [155].

1.3 Large scale DNA sequencing

1.3.1 History of DNA sequencing

The first approach to nucleic acid sequencing followed the same strategy of protein sequencing. The molecule was first broken into small overlapping pieces with different reagents and the composition of each small fragment were determined. By overlapping the small pieces, the exact sequence of the molecule was deduced. This method is practical for protein because twenty amino acids provide a large variety, but for DNA, there are only four different nucleotides and the composition hardly reflects the sequence. It was only possible for this initial approach to sequence very short DNA [156, 157].

In 1978, two different methods of DNA sequencing were introduced. The first one, developed by Maxam and Gilbert in Cambridge, MA, is a chemical cleavage based approach [158] and the second, developed by Sanger and Coulson in Cambridge, England, is an enzymatic synthesis based approach [159]. The two methods share the same general principle. Each method generates a series of single-stranded DNA molecules, each molecule is one nucleotide longer than the last. This set of DNA molecules is separated on an acrylamide gel, because the acrylamide gel has the ability to separate DNA molecules only one nucleotide different in size. If there are four different reactions which can generate a series of DNA molecules with one of their ends known, upon separation on the gel in four parallel lanes with each lane containing only one specific ending, one can deduce the sequence according the ladder of bands on the gel. The difference between these two methods lies in the methods to generate the series of DNA molecules.

The Maxam and Gilbert technique involved chemical cleavage of a DNA strand by using different base-specific modifying chemicals. Two aspects must be emphasized. First, the cleavage must be specific. Because four distinct chemical agents were not available, combinations of chemicals were used. G cleavage (dimethylsulfate and heat) is very effective with guanine but it also has a weak effect on adenine. The difference of the reactivity is sufficient to distinguish on a gel. The reaction, dimethylsulfate and acid, cleaves at A, but also cleaves weakly at G. Hydrazine and salt together will cleave only at C, but hydrazine alone will react with C and T. Second, the cleavage should be partial at any susceptible position. Any site with appropriate base should only have a small possibility to be cleaved so that every sites will have the same opportunity. With only the 5' end labeled, if the cleavage is

complete, only one band can been seen on the gel and the sizes of resulting fragments will not reflect the DNA sequence [158].

Sanger and Coulson developed a sequencing method which utilized the polymerase activity of the Klenow fragment of *E. coli* DNA polymerase I. With the presence of small amount of 2',3'dideoxynucleotides, the chain elongation will be terminated if dideoxynucleotide is randomly incorporated because the ending dideoxynucleotide lacks the 3' hydroxyl group required for polymerization. If four dideoxynucleotides are used in four separated reaction, each reaction will give a set of DNA molecules in different length with the end according to the dideoxynucleotide used. DNA polymerization is from a defined site beginning with an oligonucleotide primer which is required for the enzyme and different labeling systems can be used for the detection of the resulting nested fragments. Primer, Dideoxynucleotides or building block deoxynucleotides can be labeled. Because initially, Sanger and Coulson used ^{32}P-labeled deoxynucleotides and four separated terminator reactions, the resulting sets of DNA molecule fragments were loaded into separated parallel lanes on the gel and the DNA sequence could be deduced by the pattern of bands ladder on the gel. For example, in the A-specific lane, a band terminated with a dideoxyATP presents a T that occurred in the complementary template sequence. This method, dideoxy chain termination sequencing, is very simple and extremely accurate and today is preferred over the Maxam/Gilbert chemical method [159].

1.3.2 Improvement of DNA sequencing

Since the original Sanger sequencing method was introduced, efforts have emerged that result in improvement in both its sensitivity and efficiency. Among these modifications, several are worthy of mention here.

1.3.2.1 Enzymes in DNA sequencing

Different enzymes have been used in DNA sequencing. In Sanger's original method, the Klenow fragment of *E. coli* DNA polymerase I was used. The Klenow fragment is the large C-terminal fragment of the enzyme resulting from proteolysis of the enzyme. The Klenow fragment has both the DNA polymerization and 3' to 5' exonuclease activities but has the 5' to 3' exonuclease activity domain removed because it will remove the common 5' end [162]. Modified bacteriophage T7 DNA polymerase (Sequenase) is another enzyme often used in DNA sequencing. This enzyme's 3' to 5' exonuclease activity is eliminated by *in vitro* mutagenesis in addition to already lacking 5' to 3' exonuclease activity [163]. It has higher efficiency of polymerization and wider tolerance for nucleotide analogs than Klenow fragment. Sequenase also tolerates the fluorescent dye-labeled dideoxynucleotide terminators. More recently two thermal stable enzymes have been introduced: Bst (*Bacillus stearothermophilus*) DNA polymerase [164] and Taq (*Thermus aquaticus*) DNA polymerase [165]. The optimum reaction temperature for Bst DNA polymerase is 65°C. This feature of the enzyme can help to reduce the DNA secondary structure and GC compression problem. The ability of polymerization even under low concentration of template of this enzyme because of high polymerization efficiency can significantly reduce the cost of sequencing.

Taq DNA polymerase has an even higher temperature tolerance and thus has an advantage over Bst DNA polymerase. The optimum temperature of Taq DNA polymerase is 72°C. Taq DNA polymerase as well as several mutant forms of this enzyme are the most widely used enzyme used in DNA sequencing today. The first mutant of original Taq DNA polymerase is AmpliTaq form ABI (Applied Biosystems Division of Perkin Elmer) [166]. AmpliTaq is formed by removal of 3' to 5'

exonuclease activity domain from original Taq DNA polymerase. It can tolerate dNTP analogs and DMSO can be included in the reaction to reduce the secondary structure. It also can be used in both dye labeled primer and terminator chemistries. AmpliTaq exhibits a strong discrimination against incorporation of ddNTPs and requires high concentrations of expensive dye labeled terminators. [166, 169]

DNA polymerases differ in their ability to distinguish between dNTP and ddNTP. [169] T7 DNA polymerase prefers dNTP by only several folds [163, 167]. Others prefer dNTP by several hundreds to several thousands folds [165, 168]. The molecular basis of this lower discrimination against ddNTPS of T7 DNA polymerase was first recognized by Tabor and Richardson, and resided in a single hydroxyl group on the benzene ring of the residue tyrosine at position 526 [169]. At corresponding position of Taq DNA polymerase, position 667, it is a phenylalanine. When Phe667 is mutated to Tyr, the discrimination against ddNTPs of Taq DNA polymerase reduces 250 to 8000-fold [169]. The decreased discrimination against ddNTP not only reduces the cost of DNA sequencing but also improves the quality and read length of sequence data by giving more even signals and less background. KlenTaq-TR, developed by Barnes et. al., is a mutant form of Taq DNA polymerase I [170, 171]. Actually, the precursor of KlenTaq-TR is KlenTaq, a Taq DNA polymerase I mutant without 5' to 3' exonuclease activity in addition to its already absence of 3' to 5' exonuclease activity (wildtype Taq DNA polymerase I lacks the 3' to 5' exonuclease activity already). KlenTaq DNA polymerase first was developed to increase the fidelity of PCR reaction [172]. KlenTaq-TR, also named as Taquenase, contains the 3' to 5' exonuclease KlenTaq deletion but also carries the Phe667Tyr mutation. This mutation reduces the discrimination against incorporation of ddNTPS. ThermoSequenase, developed by Tabor and Richardson and marketed by Amersham, is similar to KlenTaq-TR but the deletion in N-terminal (responsible for removing exonuclease activity) in

ThermoSequenase is shorter. AmpliTaq-FS [173] also has a similar Phe667Tyr mutation and is marketed by ABI, but the N-terminal mutation of this enzyme has not been revealed.

1.3.2.2 Fluorescence labeling and cycle sequencing

DNA sequencing was improved dramatically by the introduction of fluorescent labeling methods to replace radioisotopic detection. Fluorescent 5'-labeled primers were first used instead of radioisotope ^{32}P labeled primers [174]. One year later, fluorescent labeled ddNTP terminators were introduced [175]. These two labeling system were commercialized by ABI. In ABI's sequencing kits, fluorescein and rhodamine derivatives, FAM, JOE, TAMRA and ROX, each with a slightly different maximum emission wavelength, are used in dye-primer reactions. Four rhodamine derivative dyes, R110, R6G, TAMRA and ROX are used in dye-terminator reactions. Four different dyes in sequencing reaction allow loading four base-specific reactions into one lane and still can be distinguished. In dye-terminator chemistry, four base-specific reactions even can be done in the same single tube. Fluorescent labeled dideoxynucleotides also reduce compressions because the large fluorescent group attached to the ddNTP does not allow base stacking on the end of newly synthesized chain. Background also is reduced because abortive stops are undetected if they are not terminated with a labeled ddNTP. [160, 161]

With thermostable enzymes available, cycle sequencing was introduced to increase the amount of reaction product and thus improve its detection on fluorescence [176, 177]. In cycle sequencing, all the reaction components are combined and incubated for several repeated temperature cycles of denaturation, annealing and extension. Cycle sequencing can yield a large amount of fluorescent labeled DNA extension products by

cycling linear amplification. Furthermore, the use of thermostable enzyme elevates the extension temperature and reduces the secondary structure which interferes with the DNA sequencing and allows for lower template concentration. Cycle sequencing also is convenient for sequencing reactions using double-stranded template because no initial denaturing step is needed. For custom primer reactions in gap closure stage, dye labeled primers are impractical because the labeling of each primer increases the cost and time dramatically. In contrast, a dye labeled terminator reaction can use any primer, and therefore this approach is used exclusively in the gap closure stage.

1.3.2.3 Nucleotide analogs

Deoxyguanosine triphosphate has been replaced by deoxyinosine triphosphate, α-thionucleotide triphosphate or deoxy-7-deaza guanosine triphosphate to eliminate the gel artifacts caused by short stretches of dyad symmetry, especially those regions containing a high proportion of G/C, that result in intrastrand secondary structure. Although dITP can pair with dCTP, only two hydrogen bonds are formed and it, like α-thionucleotide, can reduce the possibility of secondary structure. 7-deaza-dGTP does not participate in base stacking thus also reduces the possibility of secondary structure [160, 161].

1.3.3 Large scale DNA sequencing strategy

Strategies for large DNA sequencing basically fall into two major categories: directed sequencing and random shotgun sequencing.

1.3.3.1 Directed sequencing strategy

Directed sequencing entails sequencing the target DNA directly from both ends using a series of primers. Primers for the first set of sequencing reactions are based on the known sequencing data in the cloning vector. The primers for the next set of sequencing reactions are synthesized based on the data of previously sequenced regions. Every step uses the newly synthesized primers and the same target DNA as template. The sequence data can be extended into the unknown region by this stepwise pattern. By sequencing 300-400 bp each time, all of the unknown region can be covered. This strategy is very straightforward and does not result in any assembling problems, because the exact locations of all newly sequenced data are known. This strategy is very useful and has been widely used to sequence relatively short cloned DNA fragments such as eukaryotic cDNAs. If the target DNA fragment is large, prior digestion with restriction enzymes is necessary. The resulting shorter DNA fragments then are cloned and directly sequenced. This strategy has shortcomings. First, it is slow as only two sequencing reactions can be performed at the same time and the next step can not be carried out until the sequence data is available. Second, it is expensive where every sequencing reaction requires a newly synthesized primer. Furthermore, dye labeled primer reactions are impractical for this strategy because of the expensive labeling process.

Another directed sequencing strategy eliminated the requirement of primer synthesis. Instead of walking along the target DNA into the unknown region, in this method the target cloned DNA is degraded unidirectionally by exonuclease III, which digests the DNA from the 5' protruding or blunt end but the sequencing primer binding site is protected by having a 3' overhang. The nested set of deletions from one end of

the target DNA then are ligated back to the end with sequencing primer binding site. In this way, the unknown region of the target DNA is effectively moved closer to the sequencing primer site. Universal sequencing primer on the vector then is used to sequence the target DNA and the assembly is easy because the order of the nested set of DNA fragments is known based on the time of the exonuclease digestion [178, 179].

1.3.3.2 Random shotgun sequencing strategy

In the random shotgun sequencing strategy, the target DNA is not sequenced directly, but instead, various techniques are employed to generate a library of subclones which contain randomly distributed DNA fragments from the initial target DNA. These resulting subcloned DNAs then serve as templates for sequencing. Techniques involved in generation of randomly distributed DNA fragments are sonication [180, 181], nebulization [182, 183], transposon insertion [184], and partial restriction digestion [185, 186]. The random shotgun sequencing strategy can deal with a very large DNA project, and with new automation techniques, is efficiently applied to large sequencing projects such as large insert clones from the human genome project. Because a large amount of sequence data from randomly distributed DNA fragments is generated, computer programs are needed to assemble the overlapping fragments into a single contiguous sequence.

The number of random subclones required depends on the length of the target DNA and accuracy demanded. The accuracy of the sequence depends largely on the redundancy, i. e., how many times a base has been sequenced with high redundancy yielding high accuracy. The accuracy also is improved when both strands of the DNA are sequenced. For example, a compression problem observed in one sequencing direction may not occur in the opposite direction. Because a very large number of

subclones is needed to cover 100% of the target DNA by a totally random shotgun approach, after data from a certain number of sequencing reactions has been collected and the sequence of the most of the target DNA is known, a more directed approach then is used. This is because most of the gel readings collected after that time will fall into the known region and will not help to cover the gaps between known regions. When the redundancy reaches a certain level, which is practically determined to be enough to maintain the accuracy, random sequencing will be stopped. Directed sequencing then is engaged to close the remaining gaps. The redundancy only reflects the average number of times a base has been sequenced. If the subclone generation were absolutely random, every base would have same coverage, but in reality this process is not random. However, the randomness of subclone generation is sufficient if highly random cleavage approaches are employed and these more randomly generated subclones will require fewer gel readings to reach the same degree of accuracy. The question of how many gel readings needed to finish a project turns out in more of a question of when the shotgun sequencing should stop and the directed sequencing phase should start. A project can be finished by using a very large number of shotgun sequencing gel readings with only a few short gaps left, therefore using only a few directed sequencing gel readings, or it can be finished by using a minimum number of random sequencing gel readings and a large amount of directed sequencing data to (1) close large numbers of remaining gaps and (2) resequence large numbers of regions with less accuracy. The answer to this question lies in the answers to the following questions: How expensive is a shotgun random sequencing reaction? How expensive is a directed sequencing reaction? What accuracy is required for a project? Which, shotgun sequencing or directed sequencing, needs less time?

When random shotgun sequencing is stopped and all gel readings are assembled, we will get a series of contigs. The orientation of these contigs are

important to know in the gap closure stage if PCR is used to generate the templates over the gap regions. In M13-based single-stranded sequencing, the order of the contigs and their relative positions in the genomic target sequence are difficult to determine, although restriction map and some sequenced tags may be helpful. In plasmid-based double-stranded sequencing, because both ends of the insert are sequenced, the orientation of the matched forward and reverse sequence pair is known. This information is extremely useful to determine the orientation of the contigs as well as to assure the correctness of final assembly.

Gap closure stage is one of the most challenging and time consuming steps in large scale DNA sequencing because various difficulties may be encountered. In this dissertation, a strategy of gap closure has been worked out. This strategy involved resequencing the ends with different chemistries, primer walking, PCR, and PCR product and subclone fragmentation (see Methods and Materials for detail).

1.3.3.3 Cloning vectors in large scale DNA sequencing

DNA cloning is an extremely powerful way to amplify individual DNA sequences. Here, the target DNA is inserted into a carrier, called a vector, which then is transformed into the host cell where it can be replicated by host enzymes. The vector has the ability to replicate both itself and inserted DNA. The vectors used in this study include plasmids, bacteriophages, cosmids, forsmids, BACs and PACs.

Plasmids are double stranded, extrachromosomal DNA often naturally found in several bacterial strains. Naturally occurring plasmids typically contain important genes such as antibiotic resistance genes, bacterial conjugation gene for its transferring from one cell to another, and replication regulatory genes. [187] Plasmids also have sequence

representing the region for origin of replication, but the replication of plasmids usually relies on entirely on enzymes supplied by the host. [188] In some instances, plasmid replication also is regulated by several proteins encoded by the plasmid itself. [189] Under normal growth condition, a few copies of the plasmids are maintained in a host cell. [190] To be used as a cloning vector, the negative regulation system of the replication is weakened by introducing mutation into the relevant gene, [188] therefore a larger number of plasmids can be obtained from a single cell. pUC vectors are such kind of vector genetically engineered to have the copy number as high as 500-700. [191] The number and position of restriction enzyme cleavage sites where the foreign DNA insertion will take place are engineered. pUC vectors now have been engineered to contain over a dozen such sites. [192] The capacity to accept fragments of foreign DNA for a plasmid is limited in the transformation step, because the efficiency of transformation is inversely related to the size of the plasmid. Selectable markers are important characteristics of a plasmid vector. The most commonly used selectable markers are genes that confer resistance to antibiotics. [193] When plated on antibiotic-containing media, only those cells which uptake the plasmid upon transformation can survive. Other selectable markers are to distinguish host cells carrying nonrecombinant plasmids from those carrying an interesting DNA fragment. Here, a segment of DNA derived from the lac operon of *E. coli* was included in the original pUC vector system because it contains the regulatory sequences and the coding information for the first 146 amino acids (Lac Z') of the enzyme β-galactosidase which cleaves lactose. [192] A polycloning site is inside the β-galactosidase gene but does not disrupt the function of the gene. This kind of vector is used with host cells that code for the carboxyl-terminal portion of β-galactosidase. Although neither the host-encoded nor plasmid-encoded fragments are active, they can associate to form an enzymatically active protein. [194] Thus, when foreign DNA is inserted into the polycloning site, the insert causes expression of a dysfunctional lac Z' protein. In the presence of isopropyl-β-D-

thiogalactopyranoside (IPTG), an inducer of the synthesis of Lac Z', and 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (Xgal), a chromogenic substrate for β-galactosidase, cells carry a nonrecombinant plasmid will produce a functional β-galactosidase protein which cleaves Xgal resulting in blue colonies, whereas those with inserted DNA will produce dysfunctional β-galactosidase and resulting colorless colonies. This color selection mechanism allows for differentiation between cells with only vector pUC DNA and those cells with pUC vector containing inserts.

A bacteriophage or phage is a virus that infects a bacterium. The infection of the host cell by phage involves injection of phage DNA into host cell. After replication and expression, a large amount of phage progeny will be released into the medium. This feature of phage's life cycle makes the phage a suitable cloning vector. [187] Major families of bacteriophage vectors are λ phage derivatives and M13-alike vectors. The λ phage derivative vectors have linear double-stranded genome. Soon after entering the host cell, the cohesive termini of the linear DNA associate to form a circular molecule. Foreign DNA is inserted by cutting the linear phage genome into two pieces and integrating the foreign DNA into the λ phage genome. The life cycle of this vector is relatively complex. [195, 196] The length of foreign DNA that can be cloned were limited because the capacity of the phage head prevents genomes that are too long from being packaged. Thus, the nonessential portions of the phage genome were removed. However, the resulting genome itself is too short to be packaged into the phage head. Thus only those genomes with approximately 20 kb of foreign DNA fragments can generate phage progeny. The size of insert DNA depends on how much of the λ phage genome is removed, because the λ phage can pack the DNA from 78% to 105% of the size of wild-type λ phage genome. [187] M13 bacteriophage is a quite different bacterial virus. It is one of the members of filamentous bacteriophage. [197, 198, 199] It contains a single-stranded closed circular molecule of DNA and after entering the host cell, the

single-stranded DNA (+ strand) is converted to double-stranded RF DNA where the (-) strand is used as templates for expression and replication. The replication mechanism is rolling-circle replication. The resulting (+) strand progenies are cleaved and circularized. [200-203] Another feature of M13 is the morphogenesis of phage particles. Unlike most other bacterial viruses, they do not assemble phage particles inside the cell. Instead, the DNA, bound with single-stranded DNA binding protein, moves to the host membrane where the binding protein is stripped and the DNA is extruded through the membrane and becomes covered with capsid. [204, 205] This important feature gives this vector no strict limit to the size of foreign DNA to be cloned. [206] For M13-cloning, the RF DNA of the M13 phage is isolated and used as vector because it is double-stranded. Like plasmid vectors, M13 vectors also contain genetically engineered polycloning sites, antibiotic resistance gene(s), and other selectable markers.

The attempt to combine some of the advantages of plasmids and phages led to the construction of cosmids. Cosmid vectors are modified plasmids that carry a copy of the DNA sequences, *cos* sequences, required for packaging DNA into bacteriophage λ particles. Because cosmids carry an origin of replication and selectable markers as plasmids, cosmid vectors can be propagated as plasmids. Because they contain *cos* sequences, they can be packaged *in vitro* into λ phages. By eliminating the large-sized phage genome, they can carry foreign DNA fragments up to 40 kb. [187, 207-210] YAC (yeast artificial chromosome) and recently developed BAC (bacterial artificial chromosome) fosmid and PAC vectors also are useful for cloning even larger DNA fragments. [211-214]

1.3.3.4 Automated instruments for large scale DNA sequencing

The development of cycle sequence and fluorescent labeling systems opened the door for automated large scale DNA sequencing. Cycle sequencing was developed based on the same strategy as PCR but with only one primer used. [176, 177] Cycles of denaturation, primer annealing and extension allow the repeated use of the template resulting in very high signal intensity from a small amount of template. These two advantages of cycle sequencing made the automation feasible. The high signal intensity made the signal detection easier and with less template being required, template isolation was less expensive. The fluorescent labeling system brought a safer, quicker and more efficient way of signal detection. Several fluorescent DNA sequencing instruments were developed in about a decade. One DNA sequencer was developed by Ansorge et al. [215, 216] at the European Molecular Biology Laboratory (EMBL) and is marketed by Pharmacia. This instrument employs only a single dye, with the four base-specific reactions performed in four separate tubes and electrophoresed in four different wells of a denaturing polyacrylamide gel. The fluorescent dyes are excited by a laser beam directed into the gel and the detector is position perpendicular to the direction of laser excitation. Hitachi developed a similar fluorescent sequencer. 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. Li-Cor introduced an automated DNA sequencer in 1992, which differs from Pharmacia and Hitachi as it uses infrared fluorescence technology. It also uses single dye labeling system and it can simultaneously detect two reactions in one four-lane set by using two different dyes and loaded together. [217, 218]

The DNA sequencer developed by Perkin Elmer's Applied Biosystems was the first commercially available fluorescent sequencer and today is most widely used. [219, 220] All data in this dissertation research was collected from either of the two types of ABI's instruments, ABI DNA sequencer 373A or the updated model ABI 377. 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, thereby increasing the number of samples which can be processed per machine run by a factor of two or four with respect to the other instruments available. Both ABI instruments have a scanning argon laser, as in Pharmacia and Hitachi's machines, for excitation, but the ABI 377 uses a CCD (charged coupled device) camera instead of a PMT (photomultiplier tube), which is used in ABI 373, for detection. In the 373, the laser 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 PMT 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 but a series of lenses collect and focus the emitted light onto a spectrograph diffraction grating which separates the light based on wavelength into a predictably spaced pattern across a 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. Over the past few years, the sample loading capacity of ABI 373A and ABI 377 has been increased from 24, to 36, to 48, and now to 64 samples per gel. The polyacrylamide gel thickness used in ABI 377 is 0.2 mm and 0.4 mm in the ABI 373A.

Chapter II

Methods and Materials

2.1 Random shotgun sequencing strategy

Large scale DNA sequencing can be achieved by two strategies: one is directed sequencing and the other is random shotgun sequencing. In directed sequencing, the target DNA initially is sequenced directly from both ends to obtain the sequence at the vector-insert junction. From this data, sequence walking primers are synthesized and sequencing reactions once again are carried out using these newly synthesized primers and the target DNA as template. Thus, the sequencing data can be extended from the known into the unknown region. This procedure is repeated until all of the unknown region is sequenced. This method has several shortcomings. First, it is slow. Every time sequencing reactions are performed, only two ends of the target DNA are sequenced. Second, it is expensive. New unique primers are needed for every sequencing reaction. Random shotgun sequencing strategy in the other hand, does not use the target DNA directly as a template. Instead, the target DNA first is broken into small pieces and after sub-cloning, these small pieces are used as templates for the sequencing reaction. Because these small pieces are subcloned into the same kind of vector, the same universal primer, whose complementary sequence resides inside the vector, can be used to sequence different inserts. This approach solves the problems directed sequencing has. However, random shotgun sequencing also has its disadvantages. Because the random breaks in the target DNA are not truly random and because of the possible bias of subcloning, every base (or region) in the target DNA may not have the same chance to be sequenced. Some regions may be sequenced excessively while some regions may not be sequenced even once. The problem can be

solved by applying the directed sequencing strategy once the initial data is collected by random shotgun sequencing.

Random shotgun sequencing strategy can be divided into two stages. The first stage is break down. Target DNA such as cosmid, PAC, or BAC usually is large. It has to be broken into small pieces about 3 kb in size. Such small pieces of DNA are subcloned into vectors such as M13 or pUC18. The end sequences of these small pieces are obtained and entered into a database. The second stage is assembling. The end sequences of these small pieces are assembled together into large contigs by the means of computer programs which find overlaps of small sequences and align them. Finally, all the sequences are joined to form a full length contig which represents the sequence of the original target DNA (Figure 8). However, even after several-fold coverage of shotgun clones have been sequenced, there still are regions for which sequence data has not been obtained and directed sequencing should be performed. Therefore the key consideration in any strategy is when to switch from random sequencing to directed sequencing. This depends on two things. One is how much coverage is needed for every base of the sequence and the second is the cost of directed sequencing. Usually, directed sequencing will cost more than shotgun sequencing, and if shotgun sequencing is stopped too early, it is difficult to reach the required coverage for every base without an excessive number of directed sequencing reactions, and the cost of a project will be extremely high. On the other hand, if too many shotgun sequencing reactions are performed, large amounts of unnecessary coverage occurs at many areas and this also increases the cost.

To obtain a final sequence accuracy of less than one error or uncertain base per 10,000 bases sequenced, it has been established that every base must have at least three-fold coverage and that the sequence from at least one opposite directioned clone

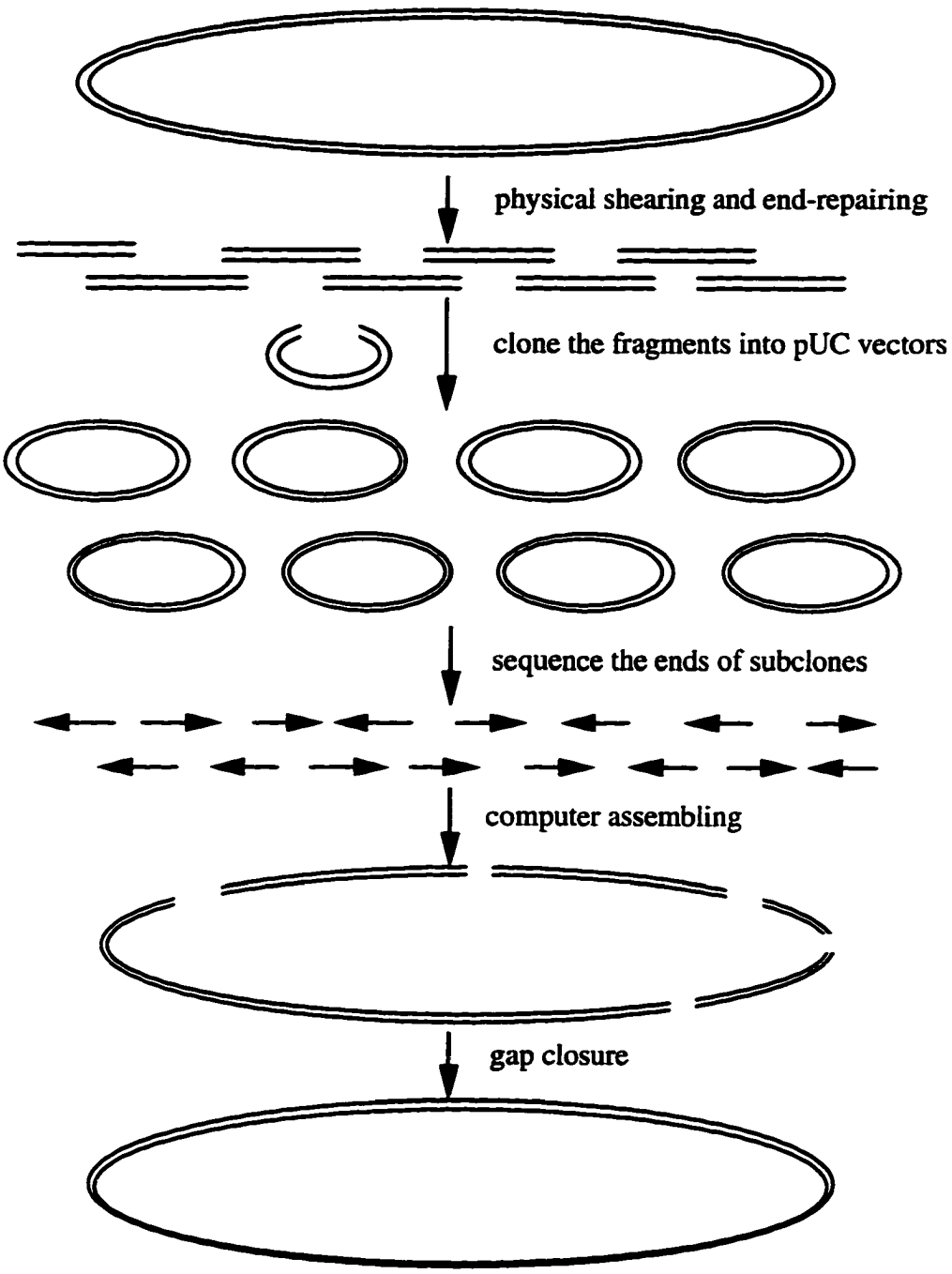
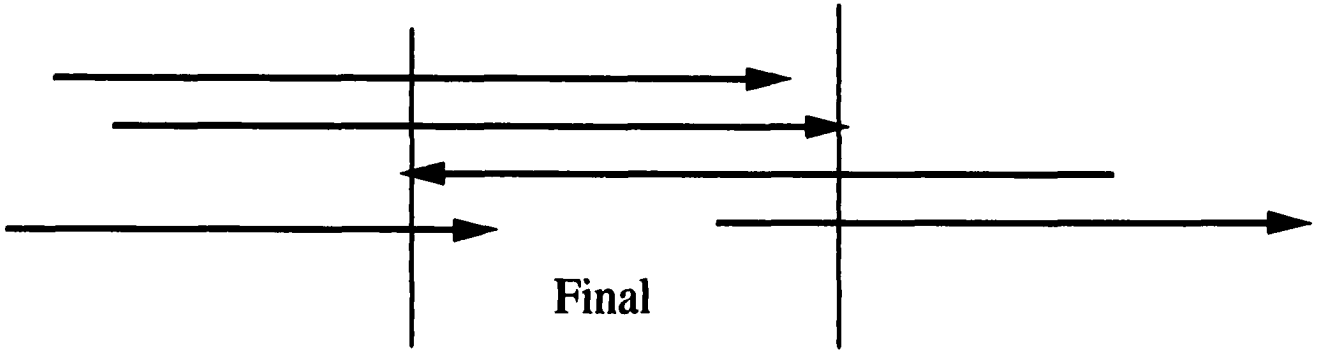


Figure 8: Shotgun sequencing strategy

Figure 9: Rule of three coverage:



Note: The region labeled with "final" meets the criteria of rule of three. All other regions are considered as gaps.

spans each base. A diagram is provided to visualize this requirement (Figure 9). Thus, it has been established that, based on cost and efficiency, large numbers of shotgun clones should be sequenced until the average coverage reaches five to six fold and then, the directed sequencing phase should begin. Directed sequencing in this stage usually entails primer walking on existing subclones covering the region, or if there is no subclones covering the region, a PCR product covering the region is generated from the original target DNA. Because the shotgun stage is usually easy and routine, most difficulties will be encountered in this gap closure stage. A gap closure strategy will be discussed below.

At the early stage of this dissertation research, sonication was used as the physical force to shear the large-sized target DNA. Because of the possible unrandomness of sonication shearing and the difficulty of controlling of the sonic strength needed to produce desired size of products, nebulization was employed to overcome the shortcomings of sonication.

Single-stranded M13mp18 vector also was used at the initial stage of this work because the single-stranded DNA sequencing method was sophisticated at that time. It was believed that single-stranded template was easier to sequence than the double-stranded ones. With the development of thermo-stable DNA polymerases, the cycling sequencing reaction with high denaturing temperature, and improved double-stranded template preparation, the double-stranded DNA sequencing became practical. Double-stranded DNA sequencing has several advantages in large scale DNA sequencing. Obviously, template number required for finishing a project can be reduced by one-half by using double-stranded template comparing with using single-stranded template because double-stranded template allows one to sequence both ends of the insert with universal forward and reverse primers. Double-stranded DNA sequencing also has the

advantage of being helpful in assembling the gel readings and organizing the database. When two sequences from both ends of one template are present in the database, the information about the orientation of these two sequences is known. This information can be used as quality assurance that the final assembly is correct and thus a misjoint would be observed if the orientation of two sequences from one template is not correct. It also is very useful to obtain the orientation information about two unjoined contigs if there are enough information of pairs of sequences from same templates. When a large-sized repeat region is encountered, known orientation of two sequences from one template also can be very useful to determine where one sequence should be oriented when the position of the other one is known.

2.2 Material preparation

2.2.1 Large scale DNA isolation

The method used for large scale cosmid DNA isolation is based on a modification of alkaline lysis procedure. [221] After crude DNA was isolated, different methods were tested to further purify the DNA. At the early stage of the dissertation research, equilibrium ultracentrifugation in cesium chloride-ethidium bromide gradients was used. [222] This method gave very pure DNA but the dealing with the carcinogen ethidium bromide was dangerous and the procedure was time consuming. A more recent method using diatomaceous earth to bind-elute DNA was less hazard and easy to do. [223] The disadvantage of diatomaceous earth based method was that in some cases the yield of DNA was reduced. Recent results indicated that the diatomaceous earth purification step could be omitted if extreme care was taken during the cleared lysis phase, and the resulting DNA quality was sufficient to allow it to be used as a template for shotgun sequencing.

2.2.1.1 Large scale DNA isolation by diatomaceous earth

A smear of colonies of cosmid were picked and transferred into a 12X75 mm Falcon tube containing 3 ml of LB medium (10 g Bacto-Tryptone, 5 g Bacto-yeast extract, and 10 g NaCl in 1 L H_2O, autoclaved) with appropriate antibiotic. After incubating at 37°C for 8 hours with 250 rpm shaking, the culture was transferred to a 250 ml flask containing 50 ml of the same medium and incubated for 8 hour under the same conditions. Then, all 53 ml of the culture was transferred into a 2 liter flask containing 1 liter same medium and incubated for additional 8 hours. After harvesting the cells by centrifugation at 7000 rpm for 20 minutes in 500 ml bottles in the RC5-B using the GS3 rotor, cell pellets were frozen and stored at -70°C. Prior to use, the cells were thawed and resuspended in 35 ml of GET (50 mM glucose, 25 mM Tris-HCl, pH 8.0, and 10 mM of EDTA, pH 8.0) by gently teasing the pellet with a spatula and then 70 mg of lysozyme was added to a final lysozyme concentration of 2 mg/ml. After mixing gently, to make sure lysozyme was dissolved completely, the solution was incubated at room temperature for ten minutes. Then 70 ml of alkaline lysis solution (0.2 N NaOH and 1% SDS) was added and after gently mixing until the solution was homogenous, it was incubated for 5 minutes in an ice-water bath. Then 70 ml of 3 M NaOAc, pH 4.8 was added and mixed gently by inverting the bottle several times and then the bottle was placed an in ice-water bath for 45-60 minutes. The lysate was cleared from precipitated SDS, proteins, membranes, and chromosomal DNA by pouring through a double-layer of cheesecloth, transferring the lysate into 250 ml centrifuge bottles and centrifuging at 10,000 rpm for 30 minutes at 4°C in the RC5-B using the GSA rotor. If necessary, an additional centrifugation was performed to ensure that all insolubles were removed. The supernatant then was transferred into a 500 ml bottle and DNase-free RNase A (stock solution of 20 mg/ml of RNase A in 1

mM NaOAc, pH 4.5 is boiled for 10 minutes and cooled down in room temperature before using) was added to the final concentration of 100 ug/ml. After incubating in 37°C water bath for 30 minutes, an equal volume of isopropanol was added and mixed well, the solution was held at room temperature for 5 minutes prior to centrifugation at 9,000 rpm for 30 minutes in RC5-B using GS3 rotor. The supernatant was decanted and the pellet drained. The DNA pellet then was dissolved in 20 ml of 10:1 TE (10 mM Tris-HCl, pH 7.6, 1 mM of EDTA, pH 8.0), divided into two 50 ml Corning centrifuge tubes and 40 ml of the defined diatomaceous earth (suspend 10 g of diatomaceous earth in 100 ml distilled water and let it settle down for 3 hours, decant the supernatant) in 6M Guanidine-HCl (100 mg defined diatomaceous earth/ml, 50 mM Tris-HCl, 20 mM EDTA, final pH 6.4) was added into each tube. After mixing well and setting at room temperature for 5 minutes with occasional mixing, and centrifugation in Beckman GS-6R centrifuge at 3000 rpm for 10 minutes, the supernatant was decanted and 40 ml of wash buffer (equal volume of 10:1 TE and ethanol) was added to each tube to resuspend the diatomaceous earth. After centrifugation as before, and decanting the supernatant, the diatomaceous earth was suspended in 40 ml of acetone and centrifuged for 10 minutes, following by decanting the supernatant and drying in a vacuum oven. After the diatomaceous earth was dry, 20 ml of 10:1 TE buffer was added and after incubation at 65°C water bath for 10 minutes, and centrifugation in the GS-6R for 15 minutes at 3000 rpm, the supernatant was collected. To improve the recovery, another 20 ml of 10:1 TE buffer was added to the diatomaceous earth, and after resuspending and incubating at 65°C, the supernatant was collected as before and pooled with the first supernatant prior to standard ethanol precipitation (2.5 volume of 95% ethanol with 0.12 M NaOAc, incubate in ice-water bath for 30-45 minute and precipitate). The final pellet was washed with 1 volume of 70% ethanol and dried. The dried pellet usually was dissolved in 2 ml of ddH_2O and the DNA concentration was estimated by agarose gel electrophoresis.

2.2.1.2 Large scale DNA isolation by CsCl gradient purification

The cell growth and cell lysis procedure were the same as in the procedure given above, but once lysed and centrifuged, the clear lysate was mixed with one-fourth volume of 50% PEG/0.5 M NaCl by swirling and incubated in an ice-water bath for 1-2 hour. After collecting the PEG-precipitated DNA by centrifugation at 7000 rpm for 20 minutes at 4°C in the RC5-B using the GSA rotor, the pellets were dissolved in a combined total of 32 ml solution containing 100:10 TE buffer, 5 ml of 5 mg/ml ethidium bromide, and 37 g cesium chloride. After transferring the sample into 35 ml polyallomer centrifuge tubes, removing all of the air bubbles, and sealing, the tubes were centrifuged at 60,000 rpm for 16-20 hours at 15-20°C in Sorvall OTD-75B ultracentrifuge using the T-865 rotor. The DNA bands were visualized under long-wave UV light and the lower DNA band representing the supercoiled DNA was collected by puncturing the tubes with a 25 gauge needle on a 5 ml syringe. The ethidium bromide was removed from the DNA sample by passing through a previously equilibrated 1.5 ml Dowex AG column (the Dowex AG resin is treated with 1 M NaOH, water and then 1 M Tris-HCl, pH 7.6, until the Dowex solution has a pH of 7.6) and 0.5 ml fractions were collected. Fractions with an A260 of 1.0 or higher were pooled into 35 ml Corex glass tubes and the DNA was precipitated, washed and quantitated as described above.

2.2.2 Mini-prep isolation of subcloned DNA

The mini-prep DNA isolation methods for both single-stranded and double-stranded subclones are discussed below because even though the doubled-stranded

pUC vector was used throughout most of this research, m13 single-stranded vector was used at the early stage.

2.2.2.1 Double-stranded plasmid isolation

Double-stranded plasmid DNA was isolated in 1.5 ml eppendorf tubes during the initial stage of this research but more recently in a 96 well block format employing the same general strategy as the large scale DNA isolation described above. Here, 1.75 ml of TB medium (12 g Bacto-Tryptone, 24 g Bacto-yeast extract, and 4 ml glycerol in 900 ml ddH_2O, autoclaved) with TB salt (10x TB salt stock solution: 2.31 g KH_2PO_4, and 12.54 g K_2HPO_4 in 100 ml of ddH_2O) and 100 ug/ml of ampicilin was added to each well of a Beckman 96 well block. Individual white pUC plasmid colonies were picked with a toothpick from an LB plate and inoculated into separate wells on the block. The cells were incubated in a 37°C shaker at 350 rpm for 22 hours and then harvested by centrifugation in Beckman GS-6R tabletop centrifuge at 1800 rpm for 5 minutes. The supernatant was decanted and the cell pellets were drained by inverting on a paper towel. Cell pellets then either were frozen at -70°C for at least one hour before use or were stored at -70°C for later use. The following isolation steps were done either manually or on a Beckman's Biomek 2000 workstation robot. The cells were resuspended in 200 ul of TE-RNase solution (50 mM Tris-HCl, pH 7.6, 10 mM EDTA, pH 8.0 and 100 ug/ml of DNase-free RNase A) by pipetting up and down several times with 12-channel P-200 pipette until all the cells were resuspended. Then, an equal volume (200 ul) of alkaline lysis buffer (0.2 N NaOH, 1% SDS) was added. The block was placed in a 350 rpm shaker for about 15 minutes to mix and then an equal volume (200 ul) of 3 M NaOAc (pH 4.8) was added to every well and the block was shaken again in a 350 rpm shaker for 10 minutes. The block then was sealed and set in a -20°C or -70°C freezer for at least half hour. The block then was removed

from the freezer and the solution was thawed. After centrifugation at 3000 rpm in Beckman GS-6R for 25 minutes to pellet the precipitated SDS, proteins membranes and *E. coli* genomic DNA, 400 ul of the clear supernatant from each well was transferred to another block containing 1 ml of 100% ethanol in each well. The block was incubated in an ice-water bath for 30-45 minutes, centrifuged at 3000 rpm for 25 minutes at 4°C in Beckman GS-6R centrifuge and the solution was decanted. The pellet was washed with 1 ml 70% ethanol and after 5 minutes spinning at 3000 rpm, the solution was drained and the pelleted DNA was dried in a vacuum oven. Resulting pellet was resuspended in 50 ul of ddH_2O.

2.2.2.2 Single-stranded M13 DNA isolation

An early log phase culture of JM101 was prepared by inoculating a glycerol stock of JM101 into 50 ml of 2xTY medium (16 g Bacto-tryptone, 10 g Bacto-yeast extract, and 5 g NaCl in 1 L ddH_2O, autoclaved) and pre-incubating the medium in a 37°C water bath for one hour without shaking. Then 1 ml of early log phase JM101 culture was aliquoted into 12x75 mm Falcon culture tubes and individual M13 plaques were picked into these tubes with sterile toothpicks. The infected cultures then were incubated in 37°C 250 rpm shaker for 4-6 hours as incubation for more than 6 hours may result in cell lysis and contamination of the phage DNA with host genome DNA. The culture then was transferred to 1.5 ml microcentrifuge tubes and centrifuged for 15 minutes at 12,000 rpm at 4°C to separate bacterial cells from the viral-containing supernatant. The microcentrifuge tubes then were placed in a special rack which fits into the Biomek 1000 workstation and the following pipetting and transferring steps were performed by Biomek 1000 robot. Two 250 ul of the supernatant of each sample were transferred into wells of 96-well flat-bottomed microtiter plate (Dynatech). Fifty ul of 20% PEG in 2.5 M NaCl solution was added to each well and mixed by pipetting

up and down. The plate was covered with plate sealer and incubated at room temperature for 15 minutes. After centrifugation in the GS-6R at 2400 rpm for 20 minutes to pellet the precipitated phage, the PEG solution was decanted by gently draining the plate upside down on a Kimwipe. The plate was returned to Biomek 1000 workstation and 200 ul of PEG:TE (1:3) was added to each well. After centrifugation and draining as above, 70 ul of TTE (10 mM Tris-HCl, pH 8.0, 0.5% Triton X-100, and 0.1 mM EDTA, pH 8.0) were added to each well and gently tapped to resuspend. The solution from two paired wells were pooled into 1.5 ml microcentrifuge tubes and incubated in 80°C water bath for 10 minutes to denature the viral protein coat. The DNA was precipitated by standard ethanol precipitation and after drying was resuspend in 20 ul of ddH_2O.

2.3 Random shotgun DNA subclone library construction

Large-sized cosmid DNA was sheared initially by sonication and then at later stage of this research by nebulization. The size of the resulted DNA was controlled by adjusting the strength of the physical force and the viscosity of the solution. The resulting DNA fragments have different types of the ends, i. e., blunt, 3' protruding or 5' protruding and in the later two instances, to subclone them into blunted-end vectors, the ends have to be repaired to become blunted-end. After end-repair and phosphorylation, the DNA fragments were size-selected by electrophoresis and elution from a low-melting temperature agarose gel with appropriate molecular weight marker. The excised DNA band was extracted with phenol-chloroform method and after ligation into blunted-end vector, was plated out. A diagram is provided to show the general procedure for the library construction (Figure 10).

2.3.1 DNA shearing

2.3.1.1 DNA shearing by sonication

The generation of random DNA fragments by sonication was performed by placing a Microcentrifuge tube containing the buffered DNA sample into an ice-water bath in a cup horn sonicator and sonicating for a varying number of 10 second bursts using maximum output and continuous power essentially as described by Bankier and Barrell. [181] The optimal sonication conditions were determined for each given DNA sample as they varied from sample to sample. Approximately 100 ug of DNA sample was dissolved in 350 ul of 1x TM buffer (50 mM Tris-HCl, pH 8.0, 15 mM $MgCl_2$). This DNA solution was distributed into ten 1.5 ml microcentrifuge tubes. Five tubes were used to test and determine the optimal sonication conditions. Each of these five was placed in a Heat Systems Ultrasonics W-375 cup horn sonicator set on 'HOLD', 'CONTINUOUS', and Maximum 'OUTPUT CONTROL' =10. The DNA sample then was subjected to sonication for 10, 20, 30, 40, and 50 seconds with at least one minute pause between every 10 second bursts to cool down the sample. This cooling down period was important because temperature increase will result in uneven fragment distribution patterns. After briefly centrifuging to collect the samples, 10 ul of each was loaded onto an agarose gel along with the appropriate molecular weight markers to determine which condition produced the desired DNA size distribution. The remaining 5 tubes then were sonicated under the chosen conditions. All of the fragmented DNA samples were pooled and precipitated by adding 2.5 volumes of 95% ethanol containing 0.12 N NaOAc. After washing the sample with one volume of cold 70% ethanol, and drying in a speedy-vacuum, the DNA sample was dissolved in 35 ul of 1x TM buffer and ready for end-repair.

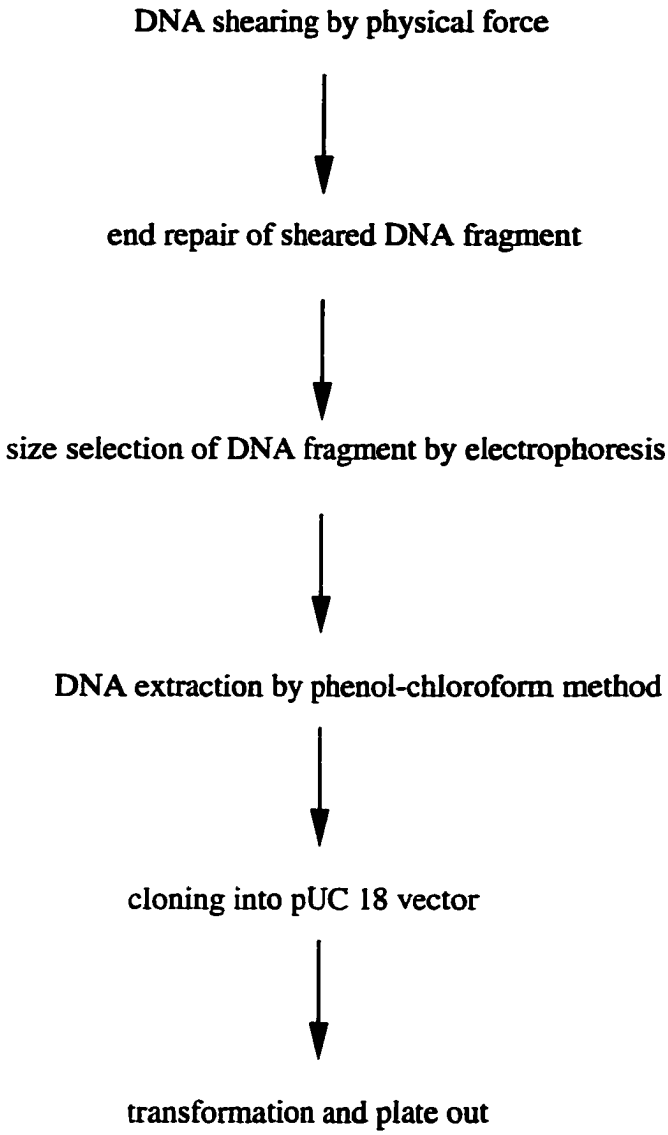


Figure 10: Shotgun DNA subclone library construction

2.3.1.2 DNA shearing by nebulization

A nebulizer is a small device which originally was designed for clinical use to break liquid samples containing either medication or nutrition, into fine droplets. A picture of the nebulizer used in this research is shown in Figure 11. After the DNA liquid sample was added to the bottom of the bowl, pressured nitrogen gas was introduced into the nebulizer from the top. The nitrogen gas was guided into a small chamber that was connected to a tube that extended to the bottom into the liquid DNA sample. The flowing nitrogen gas created a vacuum in the chamber and the DNA sample traveled up the tube and out the hole in the bottom of the small upper chamber, where it formed small droplets after colliding with the protruding semi-spherical shaped hard plastic surface. The size of product DNA fragment is inversely proportional to the pressure of the nitrogen (actually the flow rate of the nitrogen gas), and also is related to the temperature and viscosity of the liquid DNA sample. Glycerol was added to the DNA solution to increase the viscosity of the solution and to prevent the freezing of the DNA solution. [182, 183]

Approximately 50 ug of DNA was added to the nebulizer to a final volume of 2.0 ml containing 500 ul of glycerol, 200 ul 10X TM buffer. The nebulization was carried out in a -20 °C alcohol-dry-ice bath. The nitrogen pressure was adjusted to 8 to 12 psi and the nebulization time was 2.5 minutes. After nebulization, the sample was collected by briefly centrifuging the device in a Beckman tabletop centrifuge at 1500 rpm. DNA sample then was divided into four microcentrifuge tubes and ethanol precipitated. The dried DNA was dissolved into 35 ul of 1X TM buffer and ready for end-repair.

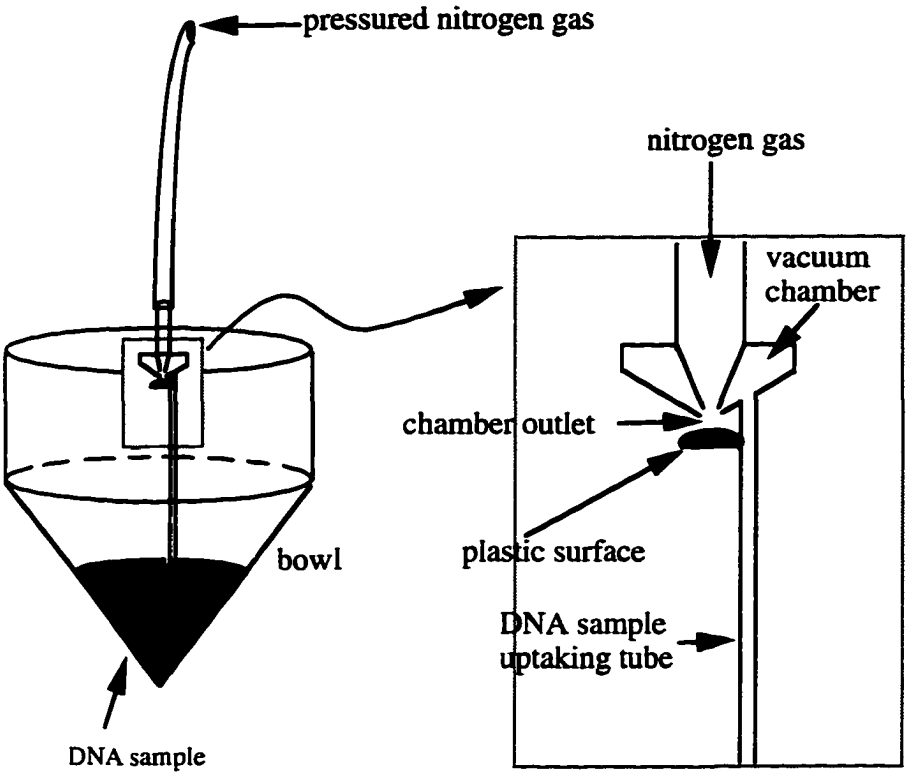


Figure 11: A nebulizer (left) and a closer look of DNA shearing parts (right).

2.3.2 End-repair and phosphorylation

Since both sonication and nebulization generated DNA fragments contain a certain percentage of single-stranded ends, the fragments were end-repaired prior to ligation into blunt-ended vectors. *E. coli* DNA polymerase I, large (Klenow) fragment was used in end-repair. Klenow fragment is a proteolytic product of *E. coli* polymerase I which retains 5' to 3' polymerization activity and 3' to 5' exonuclease activity, but has lost 5' to 3' exonuclease activity. The 5' overhang of the DNA fragment was filled-in by the enzyme's 5' to 3' polymerization activity and the 3' overhang of the DNA fragment was degraded by enzyme's 3' to 5' exonuclease activity. Because T4 DNA polymerase has a 3' to 5' exonuclease activity which is much more active than that found in the DNA polymerase I Klenow fragment, it also was useful to add T4 DNA polymerase as well. After the ends were repaired, T4 polynucleotide kinase was used to add a phosphate group to 5' hydroxyl terminus of the DNA fragment. T4 polynucleotide kinase also catalyzed the removal of 3'-phosphoryl group if any were present.

Here, 35 ul of DNA solution in 1X TM buffer, 6 ul of 10X kinase buffer, 6 ul of 10 mM rATP, 10 ul of 0.25 mM dNTPs, 2 ul of T4 DNA polynucleotide kinase (New England Biolabs), 2 ul of T4 DNA polymerase, and 3 ul of Klenow DNA polymerase I fragment (New England Biolabs) were added and the reaction was incubated at 37°C water bath for 30 minutes. After stopping the reaction by incubating at 70°C water bath for 10 minutes and then cooling the sample on ice, 6 ul of agarose loading buffer was added to the solution, which then was loaded onto a low-melting temperature agarose gel for size selection.

2.3.3 DNA size selection

DNA solutions from end-repair and phosphorylation reactions with about 10% of agarose loading buffer were loaded into separate lanes of a 0.8% low-melting temperature agarose gel with Hind III digested λ DNA and Hae III digested ϕX-174 DNA size markers run in parallel. After electrophoresing the gel at 120 mA for 60 minutes, DNA bands in the size range from 1.3 to 4 kb were quickly excised from the gel with sterile blazer while being visualized on a long wave UV light box in dark room because the duration of exposure of the DNA sample to the UV light should be minimized to prevent DNA damage. The gel slices then were placed into 1.5 ml microcentrifuge tubes and frozen in -20°C for at least 30 minutes. After melting the sample by incubating in a 70°C water bath, an equal volume of TE (10 mM Tris-HCl, pH 7.6, 1 mM EDTA, pH 8.0) saturated phenol was added and after a brief vortexing, was centrifuged for 5 minutes at room temperature to separate the phases. The upper aqueous layer was transferred to a clean tube and the lower organic phase was discarded. Half volume of TE saturated phenol and half volume of chloroform then were added to the aqueous phase, and after vortexing and centrifugation as above, the upper layer was transferred to a clean tube and an equal volume of chloroform was added. After again vortexing and centrifugation as above, the upper layer was transferred to a clean tube. An equal volume of water saturated ether was added, vortexed briefly and centrifuged for 5 minutes. The upper ether layer was removed and discarded and the DNA in the aqueous phase was concentrated by ethanol precipitation. After drying, each tube of DNA was resuspended into 20 ul of ddH_2O and ready for ligation.

2.3.4 DNA ligation and transformation

The phosphorylated blunt-ended DNA was ligated into SmaI linearized and dephosphorylated double stranded either M13 replicative form or pUC18 vector. The reaction, catalyzed by T4 DNA ligase (New England Biolabs), results in the formation of a phosphodiester bond between juxtaposed 5' phosphate and 3' hydroxyl termini in duplex DNA. It is also able to repair single stranded nicks in duplex DNA. For every set of DNA from different nebulizations, different amounts (usually from 0.1 to 1 ug) of DNA were used to determine the optimal ligation condition. The ligation reaction always was carried out with positive and negative controls. The positive control was the ligation reaction using the target DNA previously digested with Alu restriction enzyme. The purpose of this control was to insure the reaction was carried out correctly and that every ingredient functioned as they should. The negative control was the ligation reaction containing everything but insert DNA fragment. This was to determine the extent of any vector self-ligation background.

A typical ligation reaction was as follows: in a microcentrifuge tube, an optimal amount of DNA fragment (usually from 0.1 to 1 ug), 20 ng of vector (pUC18 SmaI-linearized, Calf Intestinal Alkaline Phosphatase-dephosphorylated vector), 1 ul of 10X reaction buffer, 1 ul of T4 DNA ligase (400 U/ul), and ddH_2O to bring the final reaction volume to 10 ul were mixed and the reaction was incubated at 4 ^{o}C over night. It then was ready for transformation.

DNA transformation is the process by which bacterial cells take up DNA molecules, in this case, pUC18 plasmid containing a ligated insert. Because the plasmid DNA has an origin of replication recognized by the host cell DNA polymerase, the bacterial will replicate the plasmid DNA along with its genomic DNA. The plasmid

also encodes the ampicilin (or any other antibiotic) resistant gene and the host bacterial genome does not. Thus when the host is subjected to an environment containing ampicilin (or other antibiotic), only those host containing such plasmid will survive. Because the object of transformation is to get the host cells containing foreign plasmids which carry insert DNA of interest, another selection is needed to discriminate host cells containing plasmid with and without an insert of interest. Plasmid such as pUC18 has lac Z' spanning the polylinker region. lac Z' operon, when induced by IPTG and when not interrupted, will express lac Z' gene and the product of this gene is galactosidase. Galactosidase will metabolize X-Gal, a modified galactose, to produce blue-colored products. When the plasmid carries an insert at polylinker site, lac Z' gene is interrupted and no blue color will be seen. It should be noted that both the M13 bacteriophage and pUC18 used in this study contain similar polylinker cloning sites and the lac Z' region.

The host bacterial cells usually have limited ability to take up foreign DNA. Thus the cells need to be treated to increase this ability. The cells with increased efficiency of transformation are called competent cells. [224] To prepare competent cells, a frozen glycerol stock of the appropriate strain of *E. coli*, JM101 for M13 based vector and XL1 Blue for pUC based vector was thawed, transferred to 50 ml of 2xTY in 250 ml flask and incubated at 37°C water bath for 1 hour without shaking. Then, the cells were further incubated in a 37°C shaker at 250 rpm for about 2.5 hours to reach early log phase of growth. The cells then were transferred to a sterile 50 ml polypropylene centrifuge tube and collected by centrifugation at 2000 rpm in a Beckman GPR centrifuge for 5 minutes. After decanting the supernatant, the cells were suspended in 25 ml of cold 50 mM $CaCl_2$ and incubated in an ice-water bath for 20 minutes. For M13 transfection, 10 ml of original cells culture was saved in an ice-

water bath for later use. The cells were collected by centrifugation as above, resuspended in 5 ml cold 50 mM $CaCl_2$ and then were ready for transformation.

For transformation, the ligation reaction was mixed with 200 ul of competent cell in a 12x75 Falcon tube. After incubating in an ice-water bath for 45-60 minutes, the cells were heat shocked by incubation at 42°C water bath for two minutes without any shaking. The cells then were chilled in ice-water bath for 5 minutes. For M13 based transfection, 25 ul of 25 mg/ml IPTG in ddH2O, 25 ul of 20 mg/ml x-gal in dimethylformamide, 0.2 ml of non-competent cells, and 2.5 ml of 55°C lambda top agar (10 g Bacto-tryptone, 10 g Bacto-agar, and 5 g NaCl in 1 L of ddH2O, autoclaved and kept warm) were added, briefly mixed and quickly poured onto a lambda agar plate (10 g Bacto-tryptone, 15 g Bacto-agar, and 2.5 g NaCl in 1 L ddH2O, autoclaved and poured into petri dishes). After the top agar solidified, the plates were inverted and incubated at 37°C overnight. For pUC based transformation, 1 ml of 2xTY was added and the samples were incubated in a 37°C shaker for 25 minutes. The cells were collected by centrifugation at 1500 rpm for 2 minute, the supernatant was decanted and then the cells were suspended in 200 ul of 2xTY, to which 25 ul of IPTG and 25 ul of X-Gal were added. The samples then were plated onto an LB plate (10 g Bacto-tryptone, 5 g Bacto-yeast extract, 10 g NaCl, and 15 g Bacto-agar, autoclaved, cooled to 55°C to add antibiotic and poured to petri dishes) with an ampicilin concentration of 100 ng/ml. The plates were inverted and incubated at 37°C overnight. White colonies containing DNA fragments of interest were picked and grown in TB medium as described above for plasmid DNA miniprep isolation. Positive and negative controls were performed in parallel where the positive control contained circular intact pUC plasmid to determine if the competent cells were viable and the transformation was carried out correctly, while for the negative control, only competent cells were plated to

determine if the competent cells were contaminated by other antibiotic-resistant plasmids.

2.4 Dideoxynucleotide DNA sequencing

Sequencing methods used in this research all were based on the Sanger's dideoxynucleotide terminator sequencing procedure with detection exclusively using fluorescent dyes. The labeling methods included dye-labeled primers and dye-labeled dideoxynucleotide terminators and different DNA polymerases were used to achieve the optimal results for different sequencing templates.

Fluorescent dye labels instead of radioactive labels in the sequencing reaction are safer to handle, easier to automate and overall yield better quality sequencing results. The dye primer sequencing reaction, as will be discussed later, requires four separate reactions for four specific bases. In gap closure stage, dye primer chemistry is rarely used because every custom synthesized primer needs to be labeled. The dye terminator chemistry does not have these disadvantages but the difficulty with this chemistry is that the originally available DNA polymerases have very low affinity for then-existing dye labeled terminators. Dye terminator chemistry now is widely used since new enzymes with high affinity for dye terminators, such as KlenTaq-TR and AmpliTaq-FS DNA polymerase, are available. In the very early stage of this research, dye labeled primer was used, but for the majority of the research in this dissertation, dye labeled terminators were used. AmpliTaq DNA polymerase was the enzyme of choice for most of dye primer reactions and more recently was replaced by KlenTaq-TR or AmpliTaq-FS DNA polymerase in the dye terminator reactions. Sequenase was used for dye terminator reaction in gap closure stage before ABI introduced new dyes for terminator labeling. AmpliTaq-FS DNA polymerase was largely used in gap

closure stage to resequence the contig ends and also used to resequence the regions difficult to sequence by other DNA polymerases. Automated sequencing instruments from Perkin Elmer's Applied Biosystem Division were used throughout this work. The ABI 373A initially was used and then was replaced by the more advanced ABI 377. Both instruments use a laser to excite the fluorescent dye but the emission detection are different, as discussed above, and the ABI 373A uses a PMT but the ABI 377 uses a CCD camera for detection.

2.4.1 AmpliTaq catalyzed fluorescent labeled primer cycle sequencing

The fluorescent dyes used to label primers from PE's ABD were JOE (4',5'-dichloro-2'7'-dimethoxy-6-carboxyfluorescein), FAM (5'-carboxyfluorescein), TAMRA (6-carboxytetramethylrhodamine), and ROX (6-carboxy-X-rhodamine). The dyes were covalently attached to the 5' end of the primer by an amide linkage. Each base-specific fluorescent-labeled cycle sequencing reaction routinely included approximately 100 or 200 ng of single-stranded DNA or 200 or 400 ng of double-stranded DNA for A and C or G and T reactions respectively, buffer, four deoxy- and one dideoxy-nucleotides, AmpliTaq DNA polymerase and primer labeled with one of four fluorescent dyes. Sequencing reaction was carried out in four separate tubes, with the terminating dideoxynucleotide corresponding to the fluorescent labeled primer for that particular terminator. The final reaction volumes for A and C were 5 ul and G and T were 10 ul. Each ingredient in G and T tubes was doubled over that in A and C tubes. The final reaction contained 80 mM Tris-HCl, pH 9.0, 20 mM $(NH_4)_2SO_4$, pH 9.0, 5 mM of $MgCl_2$, pH 7.0, and 1% of DMSO as reaction buffer. The concentrations for dNTPs and ddNTP were as follows:

For A reaction: 12.5 uM dATP, 50 uM dCTP, 75 uM of c^7dGTP, 50 uM dTTP, and 300 uM ddATP,

For C reaction: 50 uM dATP, 12.5 uM dCTP, 75 uM c^7dGTP, 50 uM dTTP, and 150 uM ddCTP,

For G reaction: 50 uM dATP, 50 uM dCTP, 19 uM c^7dGTP, 50 uM dTTP, and 25 uM ddGTP, and

For T reaction: 50 uM dATP, 50 uM dCTP, 75 uM c^7dGTP, 12.5 uM dTTP, and 250 uM ddTTP.

The cycling conditions were 30 cycles of a seven temperature profile of 95°C for 4 seconds, 55°C 10 seconds, 72°C for 1 minutes, 95°C for 4 seconds, 72°C for 1 minute, 95°C for 4 seconds, 72°C for 1 minute. This cycling profile was linked to a 4°C soak file for storage. After cycling, four reactions for each template were pooled together and precipitated with 250 ul 95% ethanol. Dried samples were stored at -20°C prior to loading.

2.4.2 Sequenase catalyzed fluorescent dye terminator sequencing

Single-stranded template Sequenase dye-terminator reaction requires about 2 ug of template DNA. DNA template and 3.2 umol of primer in 12 ul of 40 mM MOPS, pH 7.5, 50 mM NaCl, 10 mM $MgCl_2$, 5 mM $MnCl_2$, 15 mM isocitrate and 2.5% glycerol solution were denatured at 65-70°C for 5 minutes. The primer was allowed to bind to the template by allowing the reaction to cool at room temperature for 15 minutes. To each reaction, 7 ul of ABI terminator mix, 6.5 unit of Sequenase and 1 ul of 2 mM of α-S-dNTPs were added and the reaction was incubated for 10 minutes at 37°C water bath. Then 20 ul 9.5 M ammonium acetate and 100 ul 95% of ethanol were added to the mixture to stop the reaction and precipitate the DNA, which then was washed with 300 ul of 70% ethanol twice to remove any unincorporated dye terminators. The sample was dried and stored at -20°C prior to loading.

For double-stranded template, 5 ug of double-stranded template was denatured in 12 ul of 0.33 N NaOH solution at 65-70°C for 5 minutes. After denaturing, 3 ul of 8 uM primer, 4 ul of MOPS-acid buffer (1.35 M acid form of MOPS and 100 mM $MgCl_2$) and 8 ul ddH_2O was added and after brief mixing, 4 ul of 10X Mn^{2+}/isocitrate buffer (50 mM $MnCl_2$, 150 mM isocitrate and 25% glycerol), 6 ul of ABI terminator mix, 2 ul (6.5 units) of Sequenase, and 1 ul of 2 mM α-S-dNTPs then were added. After incubating for 10 minutes at 37°C water bath, the reactions were stopped and precipitated as above mentioned.

2.4.3 AmpliTaq catalyzed fluorescent labeled terminator cycle sequencing

About 0.5 ug of single-stranded or 1 ug of double stranded template, 0.8 or 3.2 pmol of primer for single-stranded or double-stranded template respectively, and 8 ul of the reaction pre-mix were mixed in a total volume of 20 ul. Reaction pre-mix for ten reactions was pre-made by adding 40 ul of 5X TCAS buffer (400 mM Tris-HCl, 10 mM $MgCl_2$, and 100 mM $(NH_4)_2SO_4$ pH 9.0), 10 ul of dNTPs mix (750 uM dITP, 150 uM dATP, 150 uM dTTP, and 150 uM dCTP), 10 ul of each dye-labeled ddNTP terminator (concentration of ddNTPs was not provided by the manufacturer), 5 ul of AmpliTaq DNA polymerase (8 U/ul) together. The reaction was cycled for 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 cycling profile was linked to a 4°C soak file. The reaction then was applied to a Sephadex G-50 spin column to remove any unincorporated dye terminators, dried and ready to load.

2.4.4 KlenTaq-TR (Taquenase) catalyzed fluorescent labeled terminator cycle sequencing

KlenTaq-TR DNA polymerase is a mutant form of Taq DNA polymerase, in which the N-terminal 280 amino acids has been deleted to eliminate the activity of 5' to 3' exonuclease, and a mutation of F667Y reduces the discrimination against dideoxynucleotides that is observed with the wild-type Taq DNA polymerase. About 0.5 ug of plasmid DNA was used for sequencing reaction and the reaction buffer contained 50 mM Tris-HCl, pH 9.2, 16 mM $(NH_4)_2SO_4$, 3.5 mM $MgCl_2$, 1 mM $MnSO_4$. The concentrations of the dNTPs were 150 uM for each of dATP, dCTP, and dTTP, and 450 uM of dITP. The primer amount was 6.5 pmol. The dye labeled dideoxynucleotides were used as 0.01 ul of ddA, 0.02 ul of ddG, 0.004 ul of ddT and ddC from ABI (initial concentrations were not given by ABI and thus were unknown). If the pure dye labeled ddNTPs were not available, 0.01 ul of regular dideoxy terminator premix from ABI of unknown concentration were used. The final reaction volume was 20 ul. The reaction was cycled for 25 cycles of a three temperature profile of 96-98°C for 15 seconds, 50°C for 1 second, and 60°C for 4 minutes. This cycling profile was linked to a 4°C soak file. The post cycling treatment was the same as in the AmpliTaq catalyzed dye terminator sequencing.

2.4.5 AmpliTaq-FS catalyzed fluorescent labeled terminator cycle sequencing

AmpliTaq-FS DNA polymerase is similar to KlenTaq-TR and is commercially available from PE-ABD. As recommended by ABI, the amount of single-stranded template was from 50 to 100 ng and the amount of double-stranded template was from 300 to 500 ng. For purified PCR product, about 30 to 180 ng was recommended. The

template was mixed with 3.2 pmol of primer, 8 ul of reaction pre-mix (ddNTPS, dITP, dATP, dTTP, dCTP, Tris-HCl pH 9.0, $MgCl_2$, thermal stable pyrophosphatase, and AmpliTaq-FS DNA polymerase, the concentrations of above mentioned reagents were not available from the supplier and the information of other probably used ingredients also was not provided by ABI). The cycle condition was 25 cycles of 96°C for 10 seconds, 50°C for 5 seconds, 60°C for 4 minutes. The post cycling treatment was the same as in the AmpliTaq catalyzed dye terminator sequencing.

2.4.6 Unincorporated dye terminator removal by Sephadex G-50 spin column

Sephadex G-50 was prepared by adding 20 g of Sephadex G-50 powder to 250 ml ddH_2O and allowed the G-50 to hydrate for several hours before using. The column was packed by adding 400 ul of hydrated Sephadex G-50 to each well of Millipore 96 well filter plate. The filter plate was placed on the top of a microtiter plate and centrifuged at 1600 rpm for 2 minutes in a GS-6R centrifuge. The water that was collected in the microtiter plate was discarded and the plate was placed back onto the bottom of the filter plate. Then, 200 ul of 1:5 diluted G-50 was added to each well of the filter plate and centrifuged for another 2 minutes at 1600 rpm as before. The water collected in the lower microtiter plate was discarded as before and a new clean microtiter plate was placed below to collect the sample. All 20 ul of terminator reaction was added to the center of the top surface of G-50 in each well of the filter plate. After centrifugation at 1600 rpm for 2.5 minutes to collect the DNA sample, the microtiter plate was covered with Kimwipe, the samples were dried under vacuum and stored in -20°C before loading.

2.4.7 Sequencing gel preparation, sample loading and electrophoresis on ABI 377

The polyacrylamide gel used in automated DNA sequencing on the ABI 377 instrument usually was 4.25% polyacrylamide gel containing 8 M of urea in TBE (10 X TBE buffer was made by dissolving 108 g Tris, 55 g boric acid, and 8.4 g EDTA in 1 L solution) buffer. A stock of gel mix was prepared by adding appropriate amount of urea (360 g in 1 L solution) to 1 X TBE buffer and dissolving at 55°C with stirring. After the urea dissolved completely, the solution was cooled and the appropriate amount acrylamide and bisacrylamide (40% stock solution is made by dissolving 380 g of acrylamide and 20 g bisacrylamide in 1 L solution) were added. The solution was filtered to remove any undissolved particle and degassed by vacuum to remove any air dissolved in the solution. Before gel casting, 625 ul of 10% ammonium persulfate (APS) and 25 ul of TEMED were added into 50 ml gel mix to catalyze the polymerization. After mixing by swirling gently, the solution was immediately poured into a preassembled plate set. The thickness of the gel was 0.2 mm. The sharktooth comb was inserted backward and the glass plates were clamped tightly. The polymerization was at room temperature for usually more than 2 hours. If long time storage was needed, both ends of the plate are sealed with plastic wrap covering a moist paper towel to avoid gel drying.

After the gel had been polymerized, the glass plates were washed with warm water, rinsed with ddH_2O and ethanol, and dried in a hood. The sharktooth comb was removed, inverted and inserted back to form the sample wells. The plates were assembled into the ABI 377 machine using the gel cassette, and a plate check using the PLATE CHECK option in the ABI data collection software was performed to assure the cleanness of plates. Then, 1 X TBE buffer was added to the upper and lower

buffer chambers and the gel was pre-run by choosing the PRERUN option in the ABI data collection software until the gel temperature reached 51°C.

Then, 1 ul of loading dye (10 mg/ml blue dextran and 5 mM EDTA in deionized formamide) was added to each dried sample, which was heated to 95°C for 4 minutes to denature. The samples were ready to load. However, before loading, the sharktooth well was cleaned with 1 X TBE buffer using a syringe. The odd number lanes were loaded first and after a 2-5 minutes prerun, the even number lanes were loaded and RUN option of the ABI data collection software was chosen to electrophoresis the sample and collect the data while for 48 lane loading, every third lane were loaded prior to the prerun followed by additional two cycles of loading and prerun. A typical ABI 377 run took 3.5 hours at 3 kV (limiting parameter) using a 4.25% polyacrylamide, 8 M urea gel and a Long Ranger run on the ABI 377 took 7 hours at 1.68 kV (limiting parameter) using 5.3% Long Ranger polyacrylamide, 8 M urea gel.

The collected data was analyzed by the ABI analysis software. This analysis software required manual tracking because the tracking lines created by the software did not represent the actual lanes. Once the gel had been tracked, the software would extract the data from each lane, correct the baseline, enhance the signal, correct the mobility shift from the fluorescent dyes, calculate base spacing, and finally, call the bases. The collected sequence data then was manually transferred to SUN SPARCstation for further analysis.

2.5 Shotgun sequence data editing and assembling

In contrast to the directed sequencing approach, each gel reading of random shotgun sequencing has an unknown position relative to others. Thus the gel readings

must be assembled by aligning them based on overlapping and for large scale random shotgun sequencing data assembling, various computer programs can be used. These include XGAP (X-window genome assembly program) [225], FAKII (fragment assembly kernel), CAP2 (contig assembly program) [226], and PHRAP (Phil's revised assembly program) [227]. XGAP, CAP2 and FAKII require pre-edited sequence data as input and TED (trace editing program) [228] is used for editing purpose. PHRAP has it own editing program called PHRED [227] and these two programs are linked together as a package. All four of these assembly programs are based on Smith-Waterman's algorithm [229].

XGAP was written by Roger Staden at the MRC in Cambridge, England and is based on basic Smith-Waterman algorithm. With this user friendly program, the user can view and edit the aligned sequence data at all the stages of assembly easily. The sample electrophoresis chromatogram can be easily viewed from within XGAP to verify the quality of the base calling. The parameters for the alignment can be set by the user and the assembled contig can be broken if the user considers the assembly has any problem. A "Find internal joins" option lets the user view the alignment before assembly under a looser stringency, and a "Join contigs" option lets the user join two contigs. XGAP also can calculate the consensus of a contig and put the resulting output into different formats. It also allows the comparison between the incoming sequence data and pre-assembled data already in the database and assembles the incoming data into the pre-existing database. Because XGAP does not have a base calling quality check function, the sequence data must be manually edited by TED before assembly. TED lets the user view the electrophoresis chromatogram and mask the 5' end cloning vector sequence and 3' end ambiguous sequence. The sequences containing vector only or very short insert are screened out before the remaining data is entered into the database. The shortcomings of this program are that it is time

consuming, requires significant manual interaction, and does not provide a measure of sequence data quality.

FAKII was written by Gene Myers and Susan Larson at the University of Arizona and based on an extended Smith-Waterman algorithm. This program also needs pre-edited (by TED) sequence data as input. It assembles all of the sequence data each time when a new data set is entered and the results can only be viewed and edited after converting the FAKII database into XGAP format. No quality measurements are provided with FAKII but it has the advantage of speed and is capable of handling large amounts of data.

CAP2 was written by Dr. Huang at the Michigan Technical University and used a further modified Smith-Waterman algorithm. As with FAKII, CAP2 needs pre-edited (by TED) sequence data as input and assembles all of the sequence data when new data is added. It has some degree of quality checking as certain regions arbitrarily defined by the user to have higher quality will be weighted heavier during assembling. The program does not have the capability to let the user define the better quality region for each individual gel reading. Instead, these numbers are pre-set for each project and the program considers that all of the gel readings have the same quality distribution pattern. The main advantage of CAP2 is that the initial data included in the assembly is weighted for accuracy and thus usually produces fewer contigs than FAKII. Like FAKII, the CAP2 database must be converted to XGAP format to be viewed and edited.

PHRED/PHRAP is a program package written by Phils Green at the University of Washington. It has a quality check method which gives a quality score to each base and thus is more precise than CAP2. PHRED, the editing part of the package, calculates the ratio of signal to noise for each base and the peak to peak spacing to

produce the quality value which then is used in assembling. Because every base in the gel reading has a quality score, the 3' end ambiguous sequence does not need to be masked and can be used as low quality information in assembling. The result of PHRED/PHRAP can be viewed and edited by CONSED (consensus sequence editing program) or it can be converted to an XGAP viewable and editable database.

Because each assembly program has it own advantages and disadvantages, the research in this dissertation was assembled by different assembly programs and the results of each assembly program compared. Any discrepancies were resolved by additional sequencing.

2.6 Final sequence data analysis

After the sequence project was finished, the consensus of the sequence was converted to fasta format for analysis. Generally speaking, the analysis contains two major parts. First, the sequence was used as a query to search all major DNA databases to identify any similarities of known features, which include known genes, EST's (expressed sequence tags), STS's (sequenced tag sites), repeats and other sequences already in the databases. This part of the analysis was done using a computer program called BLAST (basic local alignment search tool). [230] Several modified BLAST programs can be used. [231] Basically, the repeat elements in the query sequence were filtered by searching the human repeat database. The unique nonrepeat regions then were compared with known sequences in the databases and the similarities were aligned with query sequences for reference. Second, any potential features of the sequence were predicted. These features include potential exons, promoters, CpG islands, and poly(A) signals. The structure feature prediction was done using GRAIL (gene recognition and analysis internet link) program. [232] This

program also gives a visualized pattern of C/G content, full length repeat elements and single base repeats. Potential exon prediction was based on codon usage in open reading frame regions, hexanucleotide distribution and the occurrence of a consensus exon-intron splice site.

In addition to these two basic analysis tools, other programs were used to study selected regions of interest further. The following GCG (Genetic Computer Group) programs were used in this research [233] : the BESTFIT program for DNA alignment, the TRANSLATE program to translate DNA sequence to amino acid sequence, the COMPOSITION program to calculate the G/C content of the DNA sequence, the STEMLOOP program to view inverted repeats, the MAP, MAPSORT and MAPPLOT programs to display the position of restriction sites, the COMPARE and DOTPLOT programs to graphically display the homology comparison for two sequences. Des Higgin's Multiple Sequence Alignment program ClustalW [234] also was used for amino acid sequences alignment and the Cross-match program (developed by P. Green at Washington University, unpublished) also was used at the later phase of this study for homology comparison. In some cases, Staden's SIP (Sequence Identification Program) and NIP (Nucleotide Interpretation Program) were used as alternative analysis methods. [235] Dr. Dan Prestridge's SIGSCAN program [236] was used to find and list homologies of cis-acting regulatory elements and trans-acting factor binding sites in DNA sequence. Finally, an on-line protein secondary structure prediction program, TMpred [237] , provided by Human Genome Center at Baylor College of Medicine (http://kiwi.imgen.bcm.tmc.edu: 8088/search-launcher/) was used to study protein secondary structure.

2.7 Gap closure

After 5-6 fold coverage of shotgun sequence data has been collected and assembled into a database, collecting more random shotgun sequences is not economical and thus a directed sequencing method was applied to close the remaining gaps. Gap closure stage is one of the most challenging stages in large scale sequencing because various difficulties may be encountered. In this dissertation research, the following gap closure strategy was developed (Figure 12):

I. All the contig ends were resequenced using the AmpliTaq-FS kit from ABI with 5-10% DMSO on a 48 cm, 5.3% Long Ranger gel with 8 M urea. Sometime the denature temperature of cycling reaction also was increased to eliminate the secondary structure of the templates. Usually, AmpliTaq-FS polymerase reaction products resolved on a Long Ranger gel will give longer reading in a single run than KlenTaq-TR. However, because AmpliTaq-FS kit is more expensive than KlenTaq-TR which was used in shotgun sequencing stage, a study performed as part of this dissertation research revealed that the sequencing reaction volume could be reduced from 20 ul recommended by ABI to 5 ul without any significant loss of signal. Typically, 70-80% of initial gaps and ambiguous region could be closed by this method.

II. Primer walking then was performed using custom synthesized primers on the template pUC-based clones covering the remaining gap. Before primer walking, the gap size was determined by PCR using universal forward and reverse primers. If the gap was very large (>2 kb), either the PCR product would be nebulized directly, or the clone would be digested by appropriate restriction endonuclease to extract the insert and the insert would be nebulized. The nebulized DNA fragment then were subcloned

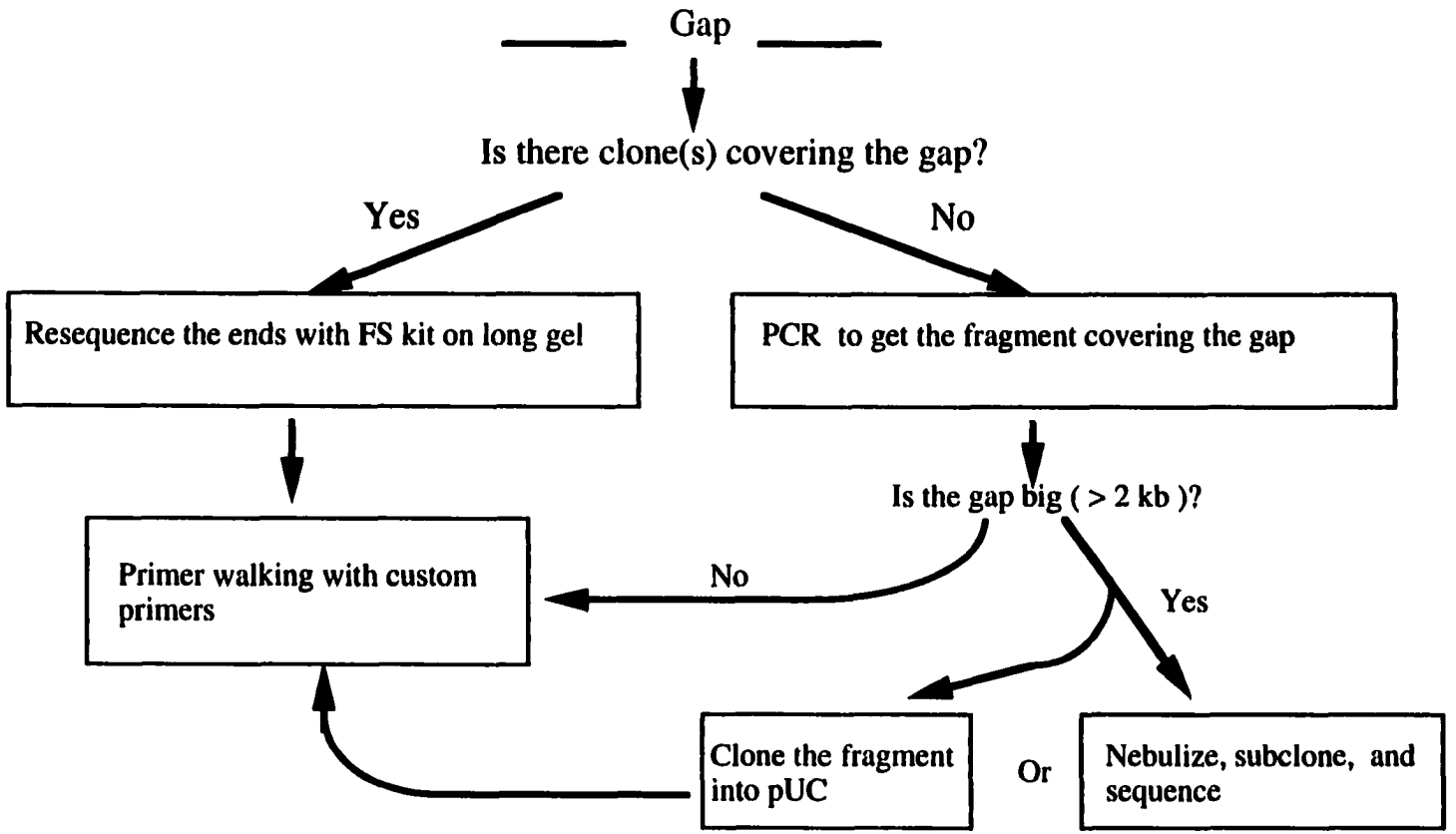


Figure 12: Gap closure strategy

and sequenced. If there was no clone covering the gap, PCR reaction would be performed as described below.

III. For any gaps with no clone coverage, the gap usually could be obtained by PCR using synthesized flanking primers and cosmid DNA as template. The size of the PCR product could be estimated by determining the size of the PCR product on agarose gels and then, either primer walking or nebulization would be used to close the gap. Usually, all the gaps could be closed by one of the above mentioned three alternative methods.

IV. In the rare cases that there was no clone covering the gap and attempts to PCR the gap failed, the first choice was to re-nebulize the entire cosmid using a low pressure (4-6 psi) and low temperature (-20°C) to generate larger DNA fragment, i.e. approximately 4-8 kb. These fragments were subcloned and end-sequenced to find a clone covering the gap. In other instance, larger DNA fragments could be generated by appropriate restriction endonucleases. The resulting restriction digested fragments then were gel purified, cloned and end-sequenced to find the clone covering the gap. The restriction endonucleases chosen were selected by mapping the available sequence of contigs of the cosmid with GCG's MAPPLOT program.

2.7.1 Oligonucleotide primer synthesis and PCR

Oligonucleotide primers, either used in primer walking or PCR, were chosen by computer programs using the following criteria. First, the primer must be unique i.e., it must bind only to the one desired site. Second, it must have about 40-50% of G/C with no obvious inverted repeat motifs and the melting temperature (Tm) of the primer should be between 50-55°C. Third, when possible, the 3' end of the primer should be

either a G or a C. The computer programs used in this research to select primers were the primer selection function in XGAP, the OSP (oligonucleotide selection program) from Washington University, the PRIMER program from Whitehead Institute for Biomedical Research, and in the later stages of this research, the Primo program from Southwest Medical Center, Dallas.

Oligonucleotides were synthesized according to manufacturer's procedures on either the Beckman Oligo 1000M DNA synthesizer or ABI 392 DNA/RNA synthesizer using the phosphoramidite chemistry. Post-synthesis treatment of oligonucleotide entailed ammonium hydroxide cleavage of oligonucleotide from the solid support, removal of protection group by heating the ammonium hydroxide solution, concentration by butanol precipitation and measurement of the final primer yield. Here one ml of ammonium hydroxide was used to cleave the oligonucleotide from the column at room temperature for 1.5 hours. The solution then was incubated in 70°C water bath for at least 2 hours. Four 100 ul aliquots were removed and each was mixed with 1.25 ml of butanol. After vortexing to mix and centrifugation at 4°C for 10 minutes at 13,000 rpm in a microcentrifuge, the oligonucleotides were dried in vacuum, and dissolved in 100 ul of ddH_2O. The amount was estimated in a spectrometer by reading the 260 nm absorption and assuming one A_{260} unit in 1 ml solution in a 1 cm path is 37 ug of single-stranded DNA.

PCR (Polymerase chain reaction) [238] typically was performed by mixing 10-20 ng of template DNA, 100 pmole of each primer, 20 nmole of each deoxynucleotide, 5 units of AmpliTaq DNA polymerase and 10 ul of 10X PCR buffer (500 mM KCl, 100 mM Tris-HCl, pH 8.5, 15 mM $MgCl_2$) in 100 ul reaction volume. The solution was denatured in 95°C for 5 minutes before thermal-cycling begins. The reaction was cycled for 25 cycles of a three temperature profile of 96°C for 1 minute, 55°C for 1

minute, and 72°C for 2 minutes. This cycling profile was linked to a 4°C soak file. After the cycling, 10 ul of the reaction was loaded on a 0.8% agarose gel to estimate the size and amount of PCR product. The PCR product then was purified by low-melting agarose gel electrophoresis followed by phenol-chloroform-ether extraction.

For the templates difficult to PCR, 5% to 10% of DMSO [239, 240] was added, the denaturation temperature was increased, or a new primer corresponding to a nearby region was synthesized. For long PCR (5 to more than 40 kb), the GeneAmp XL PCR kit from Perkin Elmer was used according to the protocol provided by the manufacturer.

Chapter III

Results and Discussion

3.1 Summary

Totally 13 cosmids and one PCR product that form four contigs from the DGCR region of human chromosome 22 and one BAC clone syntenic to this region have been sequenced. Cosmid 59c10 makes one contig in the proximal portion of the DiGeorge Critical Region. Cosmid 19d3, 98c4, 49c12, 102g9, 83c5, and 129f8 form the second contig more distal to the first with the mouse b264n1 BAC syntenic to a portion of this second contig. Cosmid 105a, 81h, 31e, Gap1 (the PCR product) and 100h form the third more distal to the DiGeorge Critical Region contig. Cosmid 68f and 24b form the fourth contig in the most distal portion of the DiGeorge Critical Region. A cosmid map is provided in Figure 13. The size, the number of gel readings used, and the redundancy of the coverage of each cosmid are listed in Table 1.

The accuracy of the sequence data was estimated by three methods: observing the discrepancies in overlapping regions between cosmids, observing the discrepancies between genomic sequence data and published cDNA sequences in the region, and comparing the sequenced cosmid vector sequence with the known sequence of the same vector. In this dissertation research, there was no discrepancy between any two overlapping cosmids and the sequence of the cosmid vector agreed exactly with known sequence of the cloning vector. There were some discrepancies between the genomic sequence data and published cDNA sequences in the same region and all these discrepancies were examined carefully and evaluated. Because of polymorphism

Table 1: Summary of cosmids sequenced

Contig name	cosmid name	size of insert	# of gel readings	coverage
contig1				
	59c10	43995	682	3.35
contig2				
	19d3	39706	795	4.74
	98c4	38712	851	4.05
	49c12	39546	949	4.47
	102g9	43945	979	6.27
	83c5	42074	780	5.36
	129f8	43699	492	2.15
contig3				
	105a	38478	1285	6.13
	81h	43481	1011	5.39
	31e	35711	591	4.57
	gap1	4717	164	11.14
	100h	40646	727	5.56
contig4				
	68f	39573	748	5.69
	24b	38987	1085	3.91
		Total=533276		
mouse BAC	b264n1	89500	1905	5.46
		Grand total=622776		

Note: Gap1 is a PCR product covering the physical gap between cosmid 31e and 100h

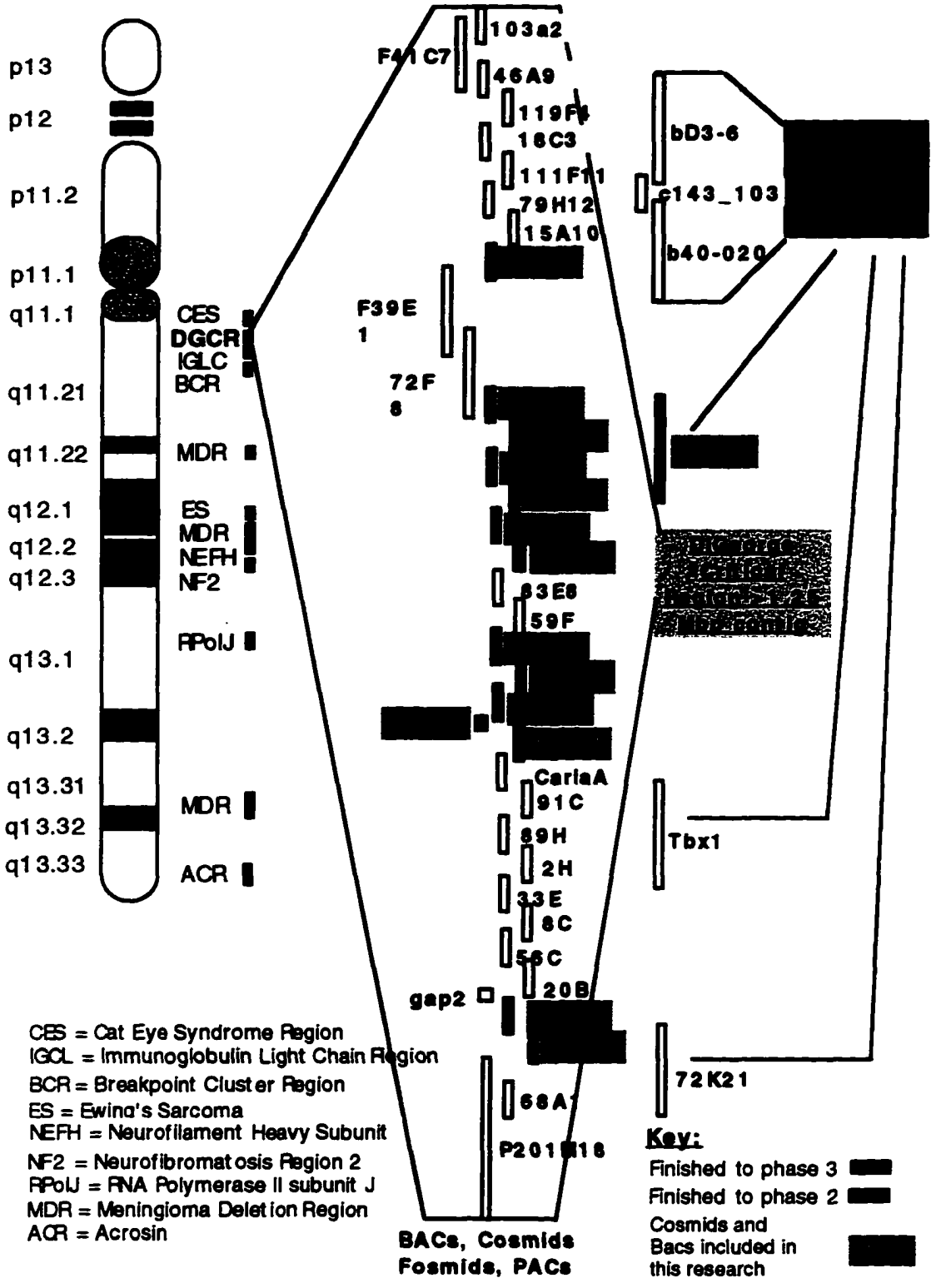


Figure 13: Map of clones completed in this research.

between different individuals, these discrepancies may not reflect sequencing errors in either the genomic or cDNA sequence.

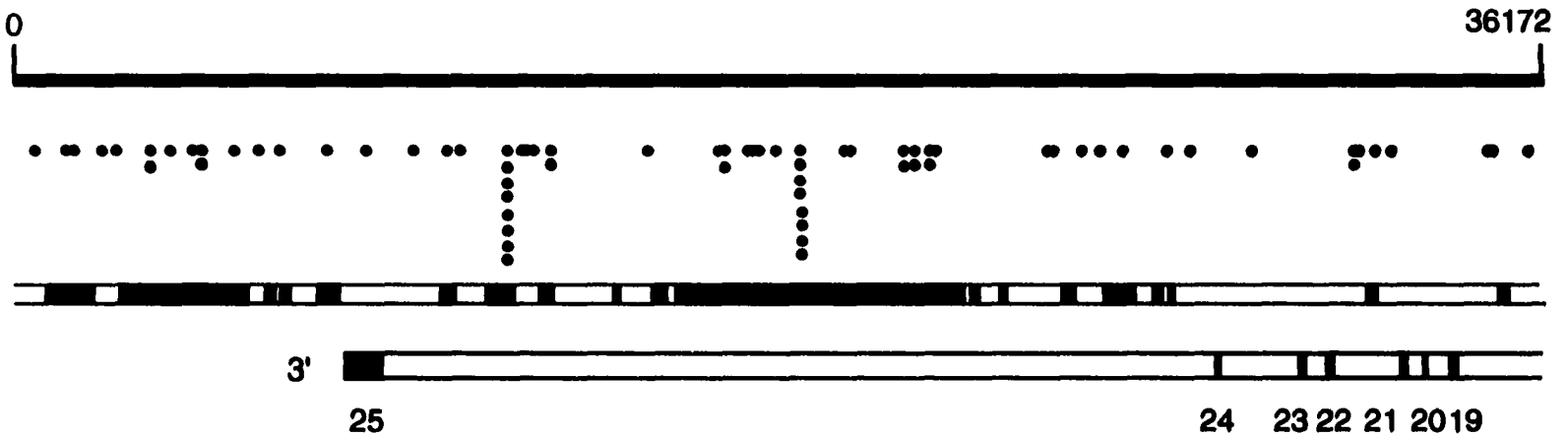
Ambiguity of the sequence data was calculated by dividing the total number of bases in all contigs by the total numbers of N in the sequence. Overall, the ambiguity was less than 0.0027% in the four human genomic regions of over 533.27 kb.

3.2 Comparison of 36 kb of genomic sequences from two individuals

Cosmid 19d3 contains 36172 bp and overlaps with BAC 72f which was sequenced by Axin Hua in the lab. Because this cosmid and BAC come from two different individuals, comparison of these two sequences revealed the allelic variations in this region. After aligning the two sequences with the GCG BESTFIT program, the results shown in Figure 14 were obtained. As also shown in Figure 14, this region contains the 3' end of human HIRA gene and the adjacent 3' region of the gene. The position of exons and introns of the human HIRA gene in this region also are given in this figure.

There are 73 (0.20%) differences between these two sequences. Among the 73 allelic variations, 34 (46.58%) are transitions (nucleotide substitutions of a purine by a purine, or a pyrimidine by a pyrimidine), 15 (20.55%) are transversions (nucleotide substitutions of a purine by a pyrimidine, or a pyrimidine by a purine), and the remaining 24 (32.88%) are nucleotide insertions or deletions. The frequencies of transitions, transversions, and insertions/deletions observed in this study are different from previous report. [241, 242] In the factor IX gene, [241] the frequencies of transition, transversion and indels (insertions/deletions) are 62%, 23% and 14.6% respectively. In human immunoglobulin λ gene, [242] the three frequencies are 58%,

Figure 14: Allelic variation map



Note: Solid black line is the sequence. Blue dots are allelic variations. Filled red boxes are interspersed repetitive elements. Filled green boxes are exons of human HIRA gene. The numbers under the filled green boxes are exon numbers.

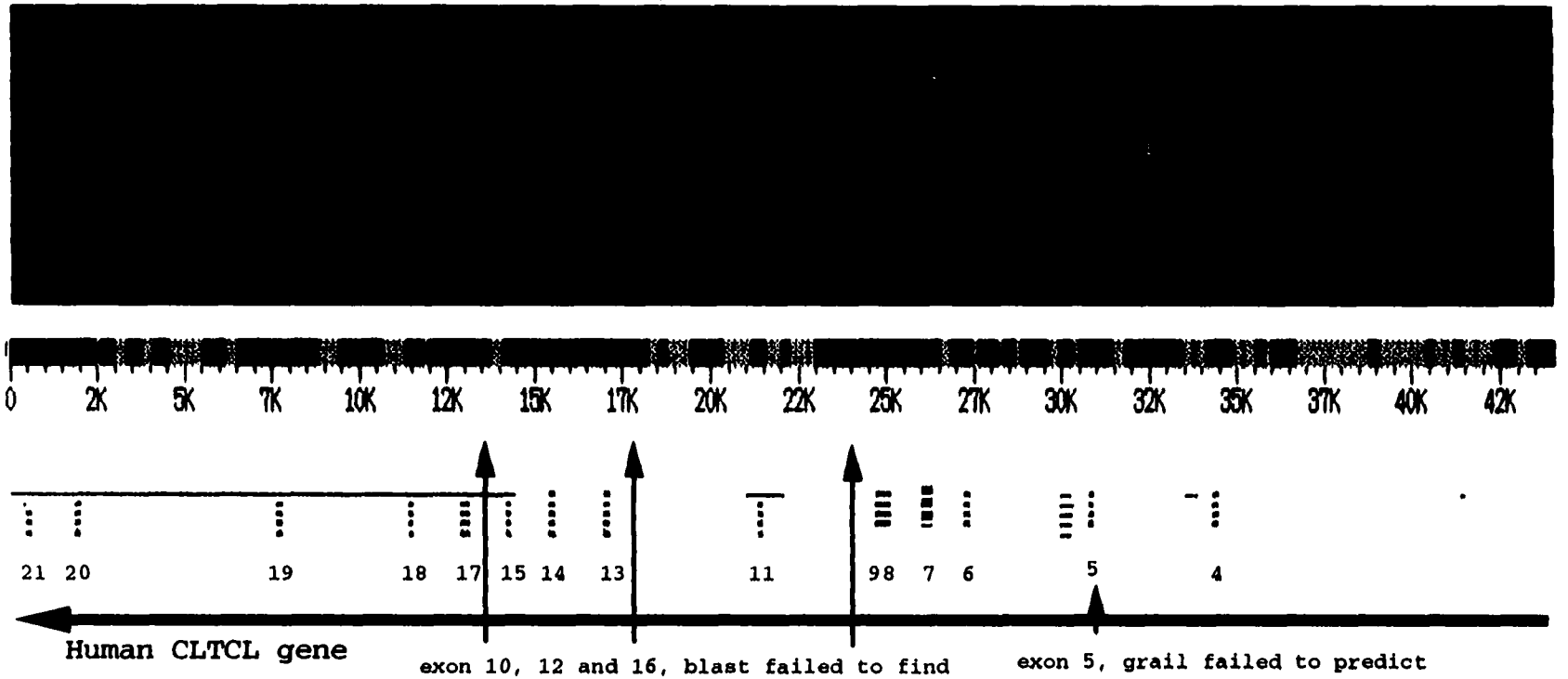
30% and 12% respectively. In this region, transitions occur more frequently than transversions which agrees with those reported previously. However, the insertions/deletions in this region is much higher than those observed before. Most of the insertions/deletions (22 out of 24) occur inside repetitive elements. Considering all the variations, more than half (49, 67.12%) occur inside repetitive elements which include Alu, L1 and MLT. Only one variation occurs inside a human HIRA gene exon and that is in the last exon, in the 3' untranslated region.

The variation is not equally distributed throughout the region. There are two "clusters" of variations in this region. In position 11529-11597, there are 8 variations and in position 18533-18647, there are 8 variations. A map of all variations is provided in Figure 14.

3.3 Analysis of contig 1

Contig 1 consists only one cosmid, 59c10, which is located within the 250 kb region of DiGeorge minimal critical region (MDGCR). [57] It is 43996 bp and overlaps the proximal cosmid 15c10 and the distal BAC 72f. The database similarity search was performed with Powerblast program and the functional regions were predicted by Xgrail program. Figure 15 shows the results of Powerblast and Xgrail analysis. The sequence shown is in the direction from centromere to telomere and reveals 18 exons of the clathrin heavy chain-like gene (CLTCL). In position 29955 to 30250, which is within intron 5 of CLTCL, there is an EST with homology to 60S ribosomal protein L34 that probably represents a non-functional pseudogene fragment.

Figure 15: Result of Powerblast and Xgrail analysis of contig 1



Note: Xgrail result is on the top. The gray line is DNA sequence. The peaks are predicted exons. The small color bars above or below the peaks indicate the quality of exons, Green for excellent, blue for good and red for marginal. Peak above the gray line means the exon is on forward strand while peak under the line, reverse. The intensity of the gray line means the GC contents. The lighter the region is, the higher the GC contents.
Powerblast result is on the bottom. The black-gray line is DNA sequence. The gray regions are repeat sequence. The small colored bars are matches found in GenBank. The numbers under red bars are exon numbers of CLTCL.

3.3.1 Clathrin heavy chain-like gene

Exons 4 to 21 of the clathrin heavy chain-like gene are within this contig. The alignment of the genomic sequence and the cDNA sequence of CLTCL is provided in Figure 16. Clathrin heavy chain-like gene is present in various species. [61] The function of CLTCL is unknown except its homology to clathrin heavy chain (CLTC). [243, 244] The similarity of CLTCL and CLTC in human is 85% at amino acid level but the expression of CLTCL is more restricted to adult. It was revealed that CLTCL is expressed highly in human skeleton muscle but lowly in heart and other adult tissues. [61] Because velopharyngeal insufficiency which presents in more than 90% of VCFS and DGS patients results from hypoplasia of the lateral pharyngeal wall musculature, it suggests CLTCL may play a role in the etiology of these syndromes [61].

A DGS patient (TOH) who shows some but not all of the common features associated with DGS carries a *de novo* t(21;22)(p12;q11) balanced translocation which disrupts the CLTCL. [61] Clathrin is known to form triskelion structures composed of three heavy chains and three light chains. The triskelions form a lattice that line coated pits and vesicles and the carboxy terminus of the heavy chain is believed to important for trimerization. [244, 245] The translocation in patient TOH disrupts CLTCL product at C-terminus. [61] If the product of CLTCL is like CLTC to form triskelion, the disruption of the gene in TOH will clearly disrupt endocytosis and membrane receptor trafficking. Further more, it was detected that CLTCL is expressed in early (9-12 week) fetal tissues. If CLTCL is involved in receptor-mediated endocytosis like CLTC does, it may play an important role in intercellular signaling processes in early developmental stage when DGS evokes. [61]

In another report, Northern and Southern blot analysis of meningiomas suggested that CLTCL also may be involved in the tumorigenesis and can be considered as a candidate tumor suppresser gene. [246]

Comparing the CLTCL exons with both Xgrail predicted and Powerblast found exons, one can see that Xgrail missed exon 5 and Powerblast failed to find exons 10, 12 and 16. In addition, four exons predicted by Xgrail in this region are not present in CLTCL mRNA (see Figure 15 for detail).

The CLTCL DNA sequence from genomic clone agrees perfectly with the mRNA sequences in GenBank, except for two bases. The first discrepancy locates at base 1907 of contig1 which is a "T" but in mRNA sequences it is a "C". This discrepancy results in a change of codon from GCG to GCA both of which are for alanine. The genomic sequence of this site has four-fold coverage with one forward and three reverse gel readings. It is possible that this discrepancy is a result of sequence polymorphism. The second disagreement is located at base 17026 which is a "C". The five mRNA sequences in GenBank differ in this position too, namely HSCLH22AS (X95486), HSU411763 (U41763) and HUMEX23G (L42356) agree with contig1 but HSU60803 (U60803) and HSU60802 (U60802) have a "T" at this position. The "T" to "C" change results in a difference of codon from AAG to GAG which code for lysine and glutamic acid respectively. The genomic sequence of this base has seven coverage with three forward and four reverse gel readings and clearly is a "C" in this cosmid.

Figure 16: Alignment of genomic sequence and CLTCL

```
59C10       > 401       atcagtgagatgcagagccctagggagaggggactcactt

59C10       > 441       ac|tgctcctgctggctgactgaacaacttccaggtaagag
U60803      < 3486         ....................................
clathrin h  < 1136           S  R  S  A  S  Q  V  V  E  L  Y  S

59C10       > 481       gaagggtcgtcccctctgatataggagttgatggcttcct
U60803      < 3448      ........................................
clathrin h  < 1122      S  P  D  D  G  R  I  Y  S  N  I  A  E  K

59C10       > 521       tcaccaaatctttctggagctgggcttgggccagctgact
U60803      < 3408      ........................................
clathrin h  < 1109        V  L  D  K  Q  L  Q  A  Q  A  L  Q  S



59C10       > 561       ccacacagcaggctcattgcatctctccgcaaactcatat
U60803      < 3368      ........................................
clathrin h  < 1096       W  V  A  P  E  N  C  R  E  A  F  E  Y

59C10       > 601       gcccggtccaggtttccaatgtgctcgatcaggac|ctagg
U60803      < 3328      ..................................... ..
clathrin h  < 1083      A  R  D  L  N  G  I  H  E  I  L  V   Q

59C10       > 641       ggttatgagaggacttccattctctgcaacatctagtcag

                        ------------------------------------
                        intron 20, sequence 681-1800 not shown
                        ------------------------------------
59C10       > 1801      gtggcagaaggtgctaggtccgacatgcttacactcagca

59C10       > 1841      gcccac|ctggattgctgaggcattcatatcaaacttgtgg
U60803      < 3295          .. ................................
clathrin h  < 1073          V   Q  I  A  S  A  N  M  D  F  K  H

59C10       > 1881      aaaacggtgaaggcctcctcatacagtgcgctgctgacag
U60803      < 3259      ..........................c.............
clathrin h  < 1059      F  V  T  F  A  E  E  Y  L  A  S  S  V  A

59C10       > 1921      cgatgctcgcgatgtccagtgcgtcatagttgtccaggcg
U60803      < 3219      ........................................
clathrin h  < 1046        I  S  A  I  D  L  A  D  Y  N  D  L  R

59C10       > 1961      gctgatgtactccatgacccgtgtgcggtctgccttgatg
U60803      < 3179      ........................................
clathrin h  < 1033       S  I  Y  E  M  V  R  T  R  D  A  K  I

59C10       > 2001      gcagtcaggatcaacagattctgtagattc|ctgaggagag
U60803      < 3139      ............................. ...
clathrin h  < 1022      A  T  L  I  L  L  N  Q  L  N   R

59C10       > 2041      agggtggtcagcacggcatcccaggaacgaaggctgcaca

                        ----------------------------------------
                        intron 19, sequence 2081-7560 not shown
                        ----------------------------------------
59C10       > 7561      cagaagatgagtatcaagcagatgttttgaatctgcctgt
```

```
59C10       > 7601      gtccacagctacccctggcatagcctcac|ctgtgctcgct
U60803      < 3110                                  .. ..........
clathrin h  < 1019                                   R  H  E  S

59C10       > 7641      gaagacagagttatccagaactatcttctccagcagttca
U60803      < 3098      ........................................
clathrin h  < 1006       F  V  S  N  D  L  V  I  K  E  L  L  E

59C10       > 7681      atcagttcattaggcaggtcggctgtcataaaggctttga
U60803      < 3058      ........................................
clathrin h  < 992       I  L  E  N  P  L  D  A  T  M  F  A  K  V

59C10       > 7721      cagtgaccgaaatctcttcaggatcccgtgtttctgacaa
U60803      < 3018      ........................................
clathrin h  < 979         T  V  S  I  E  E  P  D  R  T  E  S  L

59C10       > 7761      tgctgtctgtaccac|ctggttttaaaaagcatttgaaaga
U60803      < 2978      ............... .....
clathrin h  < 972        A  T  Q  V  V   Q  D

59C10       > 7801      acttgattaaagtcaattcaagactgtatctttgaagcta

                        ----------------------------------------
                        intron 18. sequence 7841-11360 not shown
                        ----------------------------------------
59C10       > 11361     ggcaaaatgcaggtgtgcagacatacccatctctgtgcct

59C10       > 11401     cac|ctggtcaattagctgtctcctggatgggttggtctcc
U60803      < 2966      ... ....................................
clathrin h  < 962        V   Q  D  I  L  Q  R  R  S  P  N  T  E

59C10       > 11441     tcaaggacgtgagcccagagctccggatcctttctgcata
U60803      < 2926      ........................................
clathrin h  < 948       E  L  V  H  A  W  L  E  P  D  K  R  C  V

59C10       > 11481     ccaggtagcgggcctcgcttttgaacagagaattctcatt
U60803      < 2886      ........................................
clathrin h  < 935         L  Y  R  A  E  S  K  F  L  S  N  E  N

59C10       > 11521     gcacac|ctgaaatgagcacactcatgtgtgtgacacagca
U60803      < 2846      ...... ..
clathrin h  < 932        C  V   K

59C10       > 11561     gccaaccagcggaagctgcttgggaggacacaccccaatt

                        ----------------------------------------
                        intron 17, sequence 11601-12840 not shown
                        ----------------------------------------
59C10       > 12841     aagcttccctgggagcccagggttacatgggcagctcagc

59C10       > 12881     acatctgctacccac|cttgatgagctcaaggtcacactgc
U60803      < 2843                  ... ......................
clathrin h  < 925                    V   K  I  L  E  L  D  C  Q

59C10       > 12921     ccccgctcataggcaacacaggccagatgggggtctcgct
U60803      < 2815      ........................................
clathrin h  < 911       G  R  E  Y  A  V  C  A  L  H  P  D  R  K

59C10       > 12961     tctcacagtagcggcccaccacgctgctgtcatagtaggc
U60803      < 2775      ........................................
clathrin h  < 898         E  C  Y  R  G  V  V  S  S  D  Y  Y  A
```

```
59C10       > 13001     attctctctcaggaagcactcggggctgttgttgctgtcg
U60803      < 2735      ....................................
clathrin h  < 885        N  E  R  L  F  C  E  P  S  N  N  S  D

59C10       > 13041     atgtagattttagccagtgcattgtgagtggcaggctcct
U60803      < 2695      ....................................
clathrin h  < 871       I  Y  I  K  A  L  A  N  H  T  A  P  E  E

59C10       > 13081     cacagccttcctgaatctgggactccagccagggaagcag
U60803      < 2655      ....................................
clathrin h  < 858         C  G  E  Q  I  Q  S  E  L  W  P  L  L

59C10       > 13121     cagcttgagc|ctttgaaagaaggaaggtgctgtaaagtct
U60803      < 2615      .......... ..
clathrin h  < 854        L  K  L   R

59C10       > 13161     caaggctaccactggaagtcttgagggcaactccagggct

                        ----------------------------------------
                        intron 16, sequence 13201-13400 not shown
                        ----------------------------------------

59C10       > 13401     attagacagaaacaacaacagacactagaatccaggggta

59C10       > 13441     agagaaaacaccaagtaaggacaccttac|CTATTTCTTTT
                                                      R  N  R  K

59C10       > 13481     TTCTACTTCAGCCACCAACTCATCAGTAGAGAACTGTCCT
                         E  V  E  A  V  L  E  D  T  S  F  Q  G

59C10       > 13521     CTCACTGCCATGATTAAGTGTTTAATCACTTCCTCAGAAC
                        R  V  A  M  I  L  H  K  I  V  E  E  S  C


59C10       > 13561     AATCCACATCAAGCAGCCCTCCAATCACAGCTGGGGTCCG
                          D  V  D  L  L  G  G  I  V  A  P  T  R

59C10       > 13601     GCTAGGGTTGAC|ctagggtagtcaaggtcaagtaacttca
                         S  P  N  V

59C10       > 13641     gtgttgcaaacacaagttattgttagaagcctcctacgac

                        ----------------------------------------
                        intron 15, sequence 13681-14160 not shown
                        ----------------------------------------
59C10       > 14161     gcccagaggcagccgggtgtcactccagggagccagcagt

59C10       > 14201     ac|cttctgcacgtagatctcaatgtacctctgcaggttgt
U60803      < 2464      .. ....................................
clathrin h  < 794       V   K  Q  V  Y  I  E  I  Y  R  Q  L  N  N

59C10       > 14241     tgcggtataaatataggacaaggtcatggacaaagccaaa
U60803      < 2424      ....................................
clathrin h  < 781         R  Y  L  Y  L  V  L  D  H  V  F  G  F

59C10       > 14281     acgatcacacacgatgatgaggggaagctggtctgtgagc
U60803      < 2384      ....................................
clathrin h  < 768        R  D  C  V  I  I  L  P  L  Q  D  T  L

59C10       > 14321     ttggcctc|ctgaggattagaatgcacagaacaagtcaggg
U60803      < 2344      ........ ..
clathrin h  < 764       K  A  E   K

59C10       > 14361     agtggagaaaccagacagccagacctcggactcattgcac
```

```
                ----------------------------------------
                intron 14, sequence 14401-15360 not shown
                ----------------------------------------
59C10       > 15361     ccatcagcaagggcagctcatggaaggaggcagcactgga

59C10       > 15401     acacac|cttcaggaagttcttcacacgctctgggttgtag
U60803      < 2337            . ................................
clathrin h  < 754               K  L  F  N  K  V  R  E  P  N  Y

59C10       > 15441     cagctgctctctcggcatatcctctccacctccttgatct
U60803      < 2302      ........................................
clathrin h  < 740       C  S  S  E  R  C  I  R  E  V  E  K  I  Q

59C10       > 15481     gccctgtcttacaggcagcctgaatgtatttcagatgcac
U60803      < 2262      ........................................
clathrin h  < 727         G  T  K  C  A  A  Q  I  Y  K  L  H  V

59C10       > 15521     atctgggtcttggctgaagttcacgattgagcccaggaag
U60803      < 2222      ........................................
clathrin h  < 714        D  P  D  Q  S  F  N  V  I  S  G  L  F


59C10       > 15561     tagaagaggc|ctatgaagagagaccattccatttttgtcc
U60803      < 2182      .......... ..
clathrin h  < 710       Y  F  L  G

59C10       > 15601     ttgtaacacctgcctgtctcctccgtagctgctgcccacc

                ----------------------------------------
                intron 13, sequence 15641-16920 not shown
                ----------------------------------------
59C10       > 16921     aggaggtgtagtcacatggtggggtcggcgagaggctgca


59C10       > 16961     tggtttac|ctttgtaactcttgaaggattcaaagagctcc
U60803      < 2173            . ................................
clathrin h  < 700               G  K  Y  S  K  F  S  E  F  L  E

59C10       > 17001     accagggcctgcgtgcccagctgctcgtggtacttagagg
U60803      < 2140      ........................t...............
clathrin h  < 686       V  L  A  Q  T  G  L  Q  K  H  Y  K  S  A

59C10       > 17041     ccacctgcacacacagctgaaggttctgtctgatgttagc
U60803      < 2100      ........................................
clathrin h  < 673         V  Q  V  C  L  Q  L  N  Q  R  I  N  A

59C10       > 17081     agacagcatggcatgcagacactccacagaatcctccacc
U60803      < 2060      ........................................
clathrin h  < 660        S  L  M  A  H  L  C  E  V  S  D  E  V

59C10       > 17121     gataaggagccaaagaaattgacaagcca|ctgggaacaag
U60803      < 2020      ............................ ......
clathrin h  < 649       S  L  S  G  F  F  N  V  L  W   E P
                                                      \
                                                      |
                                                      c
clathrin h  < 649

59C10       > 17161     gaagagtctgtccacaacgatgcaccactctaaagattgc

                ----------------------------------------
                intron 12, sequence 17201-17680 not shown
                ----------------------------------------
```

```
59C10       > 17681     acattaattgcatagacaacagccataataggagtgccca

59C10       > 17721     ggggtacacagtac|CTCGGGATTGAGGAGGTGAGTGTGGA
                                        E  P  N  L  L  H  T  H  V

59C10       > 17761     CCACAGCCCTCTTGATGTCATAGAGGTCGGTGTAGTGCTC
                          V  A  R  K  I  D  Y  L  D  T  Y  H  E

59C10       > 17801     CAGTGCTTGCTGCAGGAGGCCTGCCTTCTCACAGAGCTGG
                         L  A  Q  Q  L  L  G  A  K  E  C  L  Q

59C10       > 17841     GCAATGTGGGCCCGGTCGTAATGAGTAAACATTTTATTTC
                        A  I  H  A  R  D  Y  H  T  F  M  K  N  G

59C10       > 17881     CAAGGATGGCATCTGCAAC|ctatgaaacagggagttcggt
                          L  I  A  D  A  V

59C10       > 17921     gagaagccctgatcaatggcaataacttttaagaccagc

                        ----------------------------------------
                        intron 11, sequence 17961-21280 not shown
                        ----------------------------------------
59C10       > 21281     gctcaagtctgcaaacccagagaggtgccacaggtgctct

59C10       > 21321     gcctctcggtgttagcctacaagtcactgac|ctggggtgc
U60803      < 1828                                  .. .........
clathrin h  < 592                                   V  Q  P  A

59C10       > 21361     atgaacaaggttcatctccaacagccatgtctgcaggagt
U60803      < 1817      ........................................
clathrin h  < 579        H  V  L  N  M  E  L  L  W  T  Q  L  L

59C10       > 21401     ccctcagctgggcgattattcttcaaggcatccaataaga
U60803      < 1777      ........................................
clathrin h  < 565       G  E  A  P  R  N  N  K  L  A  D  L  L  F

59C10       > 21441     aggaagtacactgctgaattaaactgttttccatgaaaat
U60803      < 1737      ........................................
clathrin h  < 552         S  T  C  Q  Q  I  L  S  N  E  M  F  I

59C10       > 21481     gtccacaat|ctgaaacatatccaagccacattttcactca
U60803      < 1697      ......... ...
clathrin h  < 548        D  V  I  Q

59C10       > 21521     gtatgttttcattcagtgtttcattttcctattgaaatag

                        ----------------------------------------
                        intron 10, sequence 21561-23920 not shown
                        ----------------------------------------
59C10       > 23921     cgaatcaaagaaatcttctagatggctctaaaaagccaca

59C10       > 23961     aaaagttgtccaaaactgcctacgctcac|CTGGCTAATGT
                                                    Q  S  I  N

59C10       > 24001     TGGCCAGCGGCTCCTCGTCCTGCACTAGCATTCGAGAAAA
                          A  L  P  E  E  D  Q  V  L  M  R  S  F

59C10       > 24041     CTGCAGGCCCTGTTCCGGACTGATCTTCATTACACCCCTC
                         Q  L  G  Q  E  P  S  I  K  M  V  G  R

59C10       > 24081     AGCAGAAAGATCCAGTCTGGGGTGTACCCAAC|ctagaagc
                        L  L  F  I  W  D  P  T  Y  G  V

59C10       > 24121     aagggagcaccaatcaggaaaatcaatgaaaaaccctggg
```

```
                ----------------------------------------
                intron 9, sequence 24161-24640 not shown
                ----------------------------------------
59C10       > 24641     gagtaatctgtgagatgccagtggttcaacacacgttac|
U60803      < 1566                                             .
clathrin h  < 508

59C10       > 24681     ctttttggcatagagcacaattttctggaattggcctgtt
U60803      < 1565      ........................................
clathrin h  < 495        K  K  A  Y  L  V  I  K  Q  F  Q  G  T

59C10       > 24721     tctgcaaaacactggatcactttgcttggcacatttgccc
U60803      < 1525      ........................................
clathrin h  < 481       E  A  F  C  Q  I  V  K  S  P  V  N  A  R

59C10       > 24761     gaaggtacacactcagagcgagcatggggtcagtggtttt
U60803      < 1485      ........................................
clathrin h  < 468         L  Y  V  S  L  A  L  M  P  D  T  T  K

59C10       > 24801     aaccaagtctccgagctcctctgagcactccag|ctgtggt
U60803      < 1445      g.............................
clathrin h  < 457        V  L  D  G  L  E  E  S  C  E  L

59C10       > 24841     atgccaagggacaaggcaaagttaggaggcaggtagggag

                      ---------------------
                            intron 8
                      ---------------------
59C10       > 24881     cctggaccaagttttcaagctgtgcgggggggctacgagc

59C10       > 24921     tcacaagtgtttctac|cttatcttctttcagccacttctc
U60803      < 1412                     ......................
clathrin h  < 449                       K  D  E  K  L  W  K  E

59C10       > 24961     taggagttgcttacgcccctgctgaagaaccagatggcaa
U60803      < 1388      ........................................
clathrin h  < 436        L  L  Q  K  R  G  Q  Q  L  V  L  H  C

59C10       > 25001     agttctaaggattcaagtttattgagctgaccctggtcga
U60803      < 1348      ........................................
clathrin h  < 422       L  E  L  S  E  L  K  N  L  Q  G  Q  D  L

59C10       > 25041     gcaggattccgaagtactgcagcaatggagaagcctggcc
U60803      < 1308      ........................................
clathrin h  < 409         L  I  G  F  Y  Q  L  L  P  S  A  Q  G

59C10       > 25081     agactgagcgggtatactctggaatttctggaccgtctct
U60803      < 1268      ........................................
clathrin h  < 396        S  Q  A  P  I  S  Q  F  K  Q  V  T  E

59C10       > 25121     ctggtacgcaggattcc|ctaataataaatggaaaaatagg
U60803      < 1228      ................. ..
clathrin h  < 389       R  T  R  L  I  G   K

59C10       > 25161     tcattgtgttgtaatcacaggggatcccttcactcaactt

                ----------------------------------------
                intron 7, sequence 25201-25960 not shown
                ----------------------------------------
59C10       > 25961     attttcctctcaaggttggttcttaacactcagaacactt
```

```
59C10       > 26001     gaacagtattcaaggtttatcac|ctttggtgcagacgctg
U60803      < 1212                           . ................
clathrin h  < 384                              K  P  A  S  A  A

59C10       > 26041     caactttggcggcttcagcatagctgccctgtgcaaagag
U60803      < 1194      ........................................
clathrin h  < 371        V  K  A  A  E  A  Y  S  G  Q  A  F  L

59C10       > 26081     ggtattgaattttctcacaaacaacttctctgccccagcc
U60803      < 1154      ........................................
clathrin h  < 358        T  N  F  K  R  V  F  L  K  E  A  G  A

59C10       > 26121     aggttactacgaacggccaaacgcagaccaaggtctggat
U60803      < 1114      ........................................
clathrin h  < 344       L  N  S  R  V  A  L  R  L  G  L  D  P  N

59C10       > 26161     tctgaagcacgttggttgcataattcacaatgttatcttc
U60803      < 1074      ........................................
clathrin h  < 331        Q  L  V  N  T  A  Y  N  V  I  N  D  E

59C10       > 26201     ctcaacacaaactgacagtac|ctgtaaggacacaacaagt
U60803      < 1034      ..................... ....
clathrin h  < 323        E  V  C  V  S  L  V   Q

59C10       > 26241     gagagcagcccggcctagaagagtgctccaccatccctct

                        ----------------------------------------
                        intron 6, sequence 26281-27160 not shown
                        ----------------------------------------
59C10       > 27161     agtccctaatgaatggaaaaggttcgcaaatgaatgtcag

59C10       > 27201     agttacacac|ctggccctttttgttgacaccaataattcc
U60803      < 1006                  ..............................
clathrin h  < 314                    Q  G  K  K  N  V  G  I  I  G

59C10       > 27241     agaggttggtttgtgtggagcagtgacaaatattgtgtca
U60803      < 983       ........................................
clathrin h  < 301        S  T  P  K  H  P  A  T  V  F  I  T  D

59C10       > 27281     gcactaatacggttcatgcagatgcacacgccagactcta
U60803      < 943       ........................................
clathrin h  < 287       A  S  I  R  N  M  C  I  C  V  G  S  E  L

59C10       > 27321     ggtcgtacagatgaagatagccatactttgtgatcaagta
U60803      < 903       ........................................
clathrin h  < 274         D  Y  L  H  L  Y  G  Y  K  T  I  L  Y

59C10       > 27361     aataacaccatgtttagctccaat|ctataagaagacagag
U60803      < 863       ......................... ..
clathrin h  < 265        I  V  G  H  K  A  G  I   Q

59C10       > 27401     agcaagaggttggaaagtaaggaggagaagccatgcgggt

                        ----------------------------------------
                        intron 5, sequence 27441-30720 not shown
                        ----------------------------------------
59C10       > 30721     cctgtatctggggaggtgctagagtagatttaggacatga

59C10       > 30761     agatgtttagagagcagcacattcgtac|ctgcatagccac
U60803      < 854         ........c-----------..ca.t ............
clathrin h  < 268           H  K  A            G  I   Q  M  A  V
```

```
59C10       > 30801     tggaaaatcattctgtgcctctggaggaaaaaacacatct
U60803      < 827       .....................................
clathrin h  < 249        P  F  D  N  Q  A  E  P  P  F  F  V  D

59C10       > 30841     actgctttctttacaaaaggttggtttcccgctgcaggct
U60803      < 787       .....................................
clathrin h  < 235       V  A  K  K  V  F  P  Q  N  G  A  A  P  Q

59C10       > 30881     gtccaacttcaatgatgtgcaa|ctagaagagagattttag
U60803      < 747       ..................... ..
clathrin h  < 227         G  V  E  I  I  H  L  K

59C10       > 30921     gtcaatcaaggggaccgcctgtggtggtgggcaagtgcta

                ---------------------------------------
                intron 4, sequence 30961-34240 not shown
                ---------------------------------------
59C10       > 34241     ccagctcatgacacttttctcagggagaagacagtagagt

59C10       > 34281     gggatcttac|cttgcctcctgtgggattacgtacagcaaa
U60803      < 725                 ...........................
clathrin h  < 218                  K  G  G  T  P  N  R  V  A  F

59C10       > 34321     gcagaaaagggtggcaggcttggcattcccctccatcttg
U60803      < 695       .....................................
clathrin h  < 205        C  F  L  T  A  P  K  A  N  G  E  M  K

59C10       > 34361     aactctgcaaaagccgcagcatggccttctatgggttgtg
U60803      < 655       .....................................
clathrin h  < 191       F  E  A  F  A  A  A  H  G  E  I  P  Q  S

59C10       > 34401     aaaccttcctatccacagagtagagctgcattgctccaac
U60803      < 615       .....................................
clathrin h  < 178         V  K  R  D  V  S  Y  L  Q  M  A  G  V

59C10       > 34441     cacacggttttg|ctaagaaaagatattatgcaatgaaagg
U60803      < 575       ............ ..
clathrin h  < 173        V  R  N  Q   Q

59C10       > 34481     gagagaaaagagaggaatataaagaaacaattcttaaaat
```

Note: Not all sequences are shown. The transcription direction of CLTCL is from telomere to centromere on the complementary strand of 59c10. "." indicates the identity. Discrepancies are bold. "|" indicates the splice site. Capitalized sequences represent the exons 10, 12 and 16 which the Powerblast program failed to find.

Table 2: Splice sites of CLTCL comparing with the consensus sequence.

	exon		intron		exon	
			3	...cttttcttag|**C**AAAACCGTG...		4
4	...AGGAGGCAAG|gtaagatccc...		4	...tctcttctag|**T**TGCACATCA...		5
5	...GGCTATGCAG|gtacgaatgt...		5	...cttcttatag|**A**TTGGAGCTA...		6
6	...AAAGGGCCAG|gtgtgtaact...		6	...gtccttacag|GTACTGTCAG...		7
7	...TGCACCAAAG|gtgataaacc...		7	...ttattattag|GGAATCCTGC...		8
8	...AGAAGATAAG|gtagaaacac...		8	...cataccacag|**C**TGGAGTGCT...		9
9	...TGCCAAAAAG|gtaacgtgtg...		9	...ttgcttctag|GTTGGGTACA...		10
10	...CATTAGCCAG|gtgagcgtag...		10	...tatgtttcag|**A**TTGTGGACA...		11
11	...TGCACCCCAG|gtcagtgact...		11	...tgtttcatag|GTTGCAGATG...		12
12	...CAATCCCGAG|gtactgtgta...		12	...ttgttcccag|**T**GGCTTGTCA...		13
13	...AGTTACAAAG|gtaaaccatg...		13	...ctcttcatag|GCCTCTTCTA...		14
14	...CTTCCTGAAG|gtgtgttcca...		14	...taatcctcag|GAGGCCAAGC...		15
15	...CGTGCAGAAG|gtactgctgg...		15	...actaccctag|GTCAACCCTA...		16
16	...AAAGAAATAG|gtaaggtgtc...		16	...tctttcaaag|GCTCAAGCTG...		17
17	...GCTCATCAAG|gtgggtagca...		17	...ctcatttcag|GTGTGCAATG...		18
18	...AATTGACCAG|gtgaggcaca...		18	...ttaaaaccag|GTGGTACAGA...		19
19	...GCGAGCACAG|gtgaggctat...		19	...ctctcctcag|GAATCTACAG...		20
20	...AGCAATCCAG|gtgggctgct...		20	...taacccctag|GTCCTGATCG...		21
21	...AGCAGGAGC**A**|gtaagtgagt...		21			
consensus	(C/A)AG|gt(a/g)agt.....$(t/c)_n$n(c/t)ag|G					

Note: Upper case letters are exons while lower case letters, introns. The bold letters are those do not match with the consensus.

The splice sites for exons 4 through 21 are listed in Table 2. The sequences of these sites were located by comparison with the consensus splice site sequences and analyzing the amino acid codons at the terminals of each exon. It is shown that most of the splice sites are conservative with one exon end and several exon beginnings as exceptions. The study of the splice sites of the gene is important because it is possible that there exists alternatively spliced transcript forms of CLTCL gene. [243] Determining the entire genomic context of this gene now makes it possible to understand the proteins produced by these splice variants that may reflect differences in the resulting protein function.

3.3.2 Ribosomal protein L34 pseudogene

In region 29881 to 30320 of this contig, Powerblast found a homology to putative human ribosomal protein L34. However, it is more likely that this region contains a non-functional pseudogene similar to ribosomal protein L34. This region contains almost all the mRNA sequence without any intron. It lacks an ATG start codon and there is no putative promoter within 7 kb of the 5' end based on Xgrail prediction. Previous studies showed that it is likely that there are several copies of L34 genes in the mammalian genome, but in no instance has it been shown that more than one of the genes is functional [247-250]. The presumption is that for each ribosomal protein the genome contains only one gene that is expressed, and the other copies are non-functional pseudogenes [251]. There is a putative ribosomal protein L34 gene proximal to BRCA1 at 17q21. [252] Alignment of the cDNA sequence and genomic sequence shows only 82% identity. In position 30229 and 30230, an "A" to "T" and a "G" to "A" changes occur respectively, making a stop codon UAG instead of CUG for leucine in the mRNA sequence. The alignment is shown in Figure 17.

Figure 17: Alignment of region 29881 to 30320 of contig1 and human ribosomal protein L34 cDNA.

```
59C10       > 29881     aggtgtgagccaacgcgcccggccaagatttcttatatat

59C10       > 29921     tttacatccaactggtttatctacagaggaatgttttctg
L38941      < 365                                          ......
ribosomal   < 114                                           K  Q

59C10       > 29961     actctgtgcctttgacacttccacaacaatattctgctcc

L38941      < 359       .........t.g........t.....tg..t.........
ribosomal   < 109        S  Q  A  Q  A  V  K  V  I  I  K  Q  E
                                    \
                                     |
                                  ccttca
ribosomal   < 107                   K  L

59C10       > 30001     tcaataaggaaagcacacttgatcctgtcacgaacacact
L38941      < 313       ..g.............g.......................t.
ribosomal   < 85        E  I  L  F  A  R  K  I  R  D  R  V  C  K

59C10       > 30041     tagcacacacggaaccaccacaggccctgctgacatgtgt
L38941      < 273       .........t..........t..................t.
ribosomal   < 72          A  C  M  S  G  G  Y  A  R  S  V  H  K

59C10       > 30081     ttctgttttagggaacctcacaagaacttcaggtctcata
L38941      < 233       c.t......g.ac..t....t........t......t.c.
ribosomal   < 59         K  T  K  S  L  R  M  L  V  K  P  R  V

59C10       > 30121     gcacgaactcctcgaagtctggctaggcacatgccacatg
L38941      < 193       .g......c.........t..c..g......ca.......
ribosomal   < 45        P  R  V  G  R  L  K  G  P  C  V  G  C  A

59C10       > 30161     caggttcaggtgttttcccaaccttcttgatgtaaaggtg
L38941      < 153       ...a..tt....c.................g.a.......a
ribosomal   < 32          S  K  P  A  K  G  V  K  K  T  Y  L  Y

59C10       > 30201     aataattctcccaccaggggcttgggactacctagtttca
L38941      < 113       ..c......att.........t.c......ag........tg
ribosomal   < 19         V  I  R  N  G  P  T  R  S  L  R  T  K
                                                     *
59C10       > 30241     ctagaggctgcgtgtcaaaggctgggccattctgaatgcc
L38941      < 73        t.........
ribosomal   < 15        N  S  A  T

59C10       > 30281     agctcaaaggactgtccctgggagctctaggctccctcaa
```

Note: A "*" indicates where a stop codon occurs instead of leucine.

3.3.3 Repetitive elements in contig 1

DNA sequences containing interspersed repetitive elements and simple repeating units were identified using both Xgrail and Powerblast programs with a library of human repeat sequences. Alu is the most commonly occurring repetitive element found in the human genome. It is believed that Alu is a pseudogene derived from 7SL RNA. Previous studies estimated that one Alu sequence appeared in about every 2 kb of genomic DNA. A typical Alu is about 300 bp in size and consists two homologous halves, each of which has an A-rich tail. Alu sequences found in the human genome can be whole length, half length or quarter length. [253-256] There are other interspersed repetitive elements in contig1, L1, MER and MLT, which occur at much lower frequencies. Powerblast identified 33 individual Alu elements, 11 L1's, 1 MLT and 1 MER. Besides these interspersed repeat elements, Xgrail identified 21 single repeat units within this contig, all present in intron regions. The interspersed repetitive elements are summarized in Table 3.

Table 3: Interspersed repetitive elements in contig 1.

Family	number	bases	% of region
Alu	**33**	**8579**	**19.5**
L1	11	4150	9.4
Mer	3	297	0.7
Mlt	2	809	1.8
Total non-Alu	**16**	**5256**	**11.9**

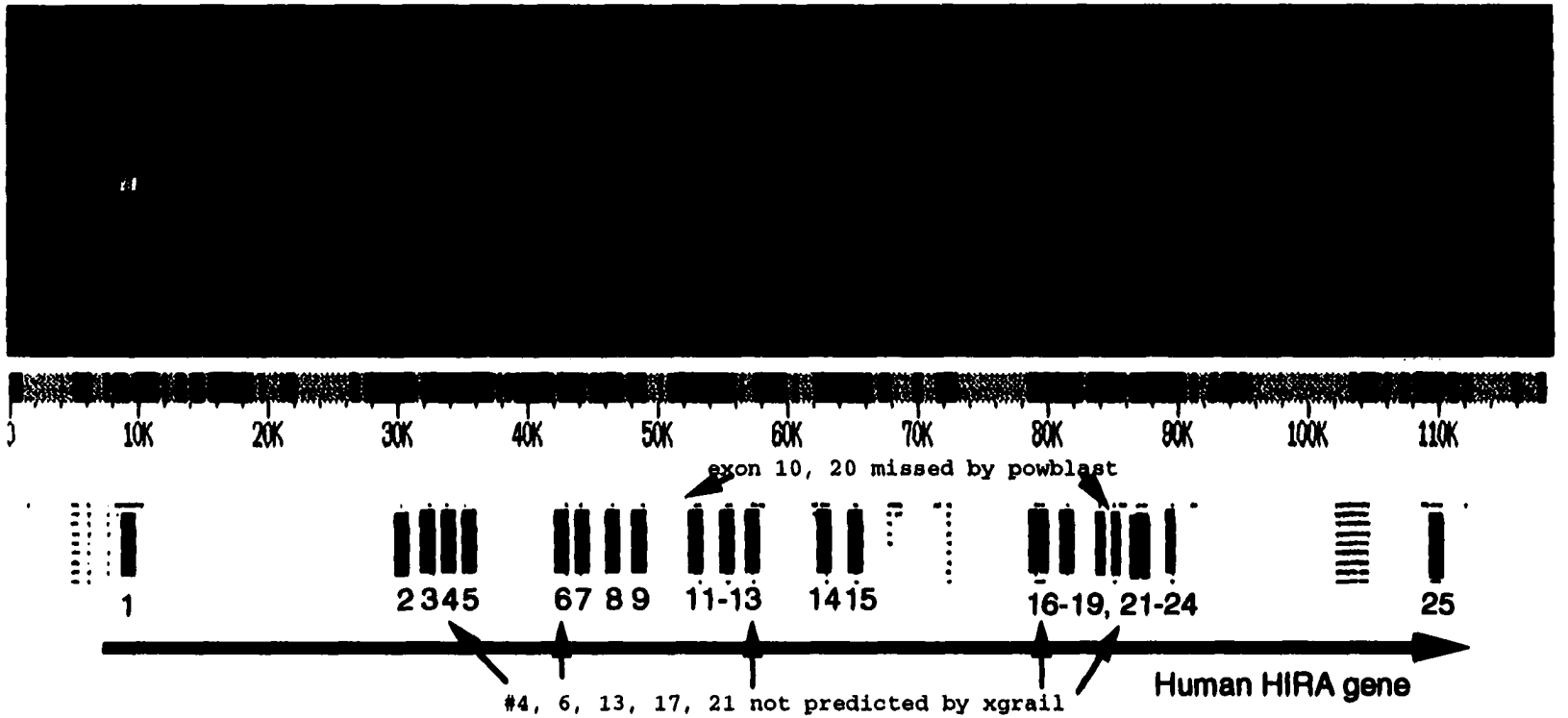
3.4 Analysis of contig 2

Contig 2 contains 6 cosmids, namely from centromere to telomere, they are 19d3, 98c4, 49c12, 102g9, 83c5 and 129f8. They form a contig of 211,386 bp in length. For the convenience of analysis with Xgrail, this contig has been divided into two contigs. The first contig, called contig2.1, has 118,000 bp and contains the entire human HIRA gene. This contig is presented as the reverse and complement here because human HIRA gene is expressed in the telomere to centromere direction. The second contig, named contig2.2, has the remainder of this 211,386 bp sequence.

3.4.1 Contig 2.1

The total length of contig2.1 is 118,000 bp. Database homology search reveals that this region contains the human HIRA gene. The results of Xgrail and Powerblast are provided in Figure 18. Comparing the results of Xgrail and Powerblast, Powerblast failed to identify only two of the known exons of the human HIRA gene. Both of the exons are relatively small, 71 bp and 59 bp, and are very close to neighboring exons. In contrast, Xgrail failed to predict five of the 25 HIRA exons and instead predicted several exons which are not found in the human HIRA cDNA sequence. Most of the additional predicted exons lie in three large introns of the HIRA gene, namely intron 1, 21 kb, intron 15, 14 kb and intron 24, 20 kb (see below for detail). Powerblast results showed that those additional exons were homologues to widely expressed EST's in the GenBank but the similarities were low. These additional exons may belong to a family of pseudogenes that appear to be the result of tandem reiterations of DNA segments during evolution. The structural similarities of

Figure 18: Result of Powerblast and Xgrail analysis on contig 2.1



Note: Xgrail result is on the top. The gray line is DNA sequence. The peaks are predicted exons. The small color bars above or below the peaks indicate the quality of exons, Green for excellent, blue for good and red for marginal. Peak above the gray line means the exon is on forward strand while peak under the line, reverse. The intensity of the gray line means the GC contents. The lighter the region is, the higher the GC contents. The pink boxes on the gray line are CpG islands.
Powerblast result is on the bottom. The black-gray line is DNA sequence. The gray regions are repeat sequences. The small colored bars are matches found in GenBank. The numbers under green bars are exon numbers of human HIRA gene.

these DNA segments to specific functional genes are the reason why Xgrail predicted these regions as exons. [80-81]

3.4.1.1 Human HIRA gene

Analysis of contig2.1 reveals the genomic organization of the human HIRA gene. The transcription direction of this gene is from telomere to centromere. The human HIRA gene contains 25 exons. Comparative analysis of the genomic sequence and the cDNA sequence reveals the position and the size of all 25 exons (Table 4). The HIRA cDNA sequence originally reported contained only 61 nucleotides upstream of the initiation codon [63] which later was shown not to be correct since RNase protection experiments indicated that the 5' end of the cDNA sequence should extend further 159 nucleotides. [257] The full-length HIRA transcription therefore consists of 4018 nucleotides and a poly(A) tail. The boundaries of exon/intron are determined with the available genomic data to preserve the nearly invariant consensus 5' GT and 3' AG dinucleotides of intron boundaries. The comparison of all splice sites and the consensus is provided in Table 5.

The alignment of genomic DNA and HIRA cDNA is provided in Figure 19. A poly(A) signal at position 110194-110199 indicates the end of the last exon. These two sequences match perfectly except in their 3' UTR as here there are five mismatches (including three "n"s in cDNA sequence) and five indels. All discrepancies were checked carefully. In all cases, there was more than six-fold redundancy in genomic DNA sequence. The alignment of human (X89887) and mouse (X92590) cDNA shows great similarity (87.53% identity by cross_match). Approximately 40% of the discrepancies are observed in the 5' and 3' UTR. The identity at amino acid level is about 97%.

Table 4: Position and size of exons and size of introns of human HIRA gene in contig 2.1

Exons				Introns	
Exon	Start	End	Size	Intron	Size
1	9232	9488	257	1	20662
2	30150	30212	63	2	2122
3	32335	32445	111	3	1208
4	33654	33744	91	4	1303
5	35048	35142	95	5	7699
6	42842	42937	96	6	1044
7	43982	44142	161	7	2277
8	46420	46587	168	8	2127
9	48715	48828	114	9	3545
10	52374	52444	71	10	666
11	53111	53216	106	11	1974
12	55191	55406	216	12	1814
13	57221	57306	86	13	5552
14	62859	63056	198	14	2075
15	65132	65293	162	15	13697
16	78991	79195	205	16	385
17	79581	79685	105	17	1752
18	81438	81586	149	18	2284
19	83871	84032	162	19	601
20	84634	84692	59	20	364
21	85055	85160	106	21	1640
22	86801	86923	123	22	476
23	87400	87563	164	23	1909
24	89473	89561	89	24	19799
25	109361	110221	861		

Note: Exon 1 contains 5' untranslated region. Start codon ATG begins at position 231 of the first exon. Exon 25 contains 3' untranslated region. Stop codon TAG begins at position 115 of the last exon.

Table 5: Exon/intron boundary of human HIRA gene

	exon	intron	exon	
1	...AACCACAA**T**G|gtgagtgcgc...	1	...ttgtttccag|GCAAGCCGAT...	2
2	...GGAGGACAAG|gtatgtttat...	2	...tttgttccag|GGCAGGATTC...	3
3	...AATCACTTAG|gtattacaga...	3	...ttcttcccag|CATGTGTGAA...	4
4	...AGCGGGCTA**C**|gtaagcatca...	4	...tctcacacag|GTACATCGGC...	5
5	...CATTCAGGC**G**|gtgagtggga...	5	...cttctttcag|**A**TGTGATGGA...	6
6	...AAGTTCCCAG|gtctggggcc...	6	...ttcttcacag|**A**AATTCTAGC...	7
7	...TTTTGATGAG|gtaagtgggg...	7	...cttctttcag|**T**GTGGAGGAA...	8
8	...GACTGTCG**T**G|gtgagtgccc...	8	...ttggctttag|**A**AATTCAACC...	9
9	...TTCTGTCT**G**G|gtgagcctgc...	9	...ttccttccag|CTCACATGTC...	10
10	...ATATTTCC**T**G|gtaaggtgga...	10	...tgtctttcag|GACTCTGAAT...	11
11	...GGAGGAGAAG|gtaggctgac...	11	...gcttccccag|**A**GCCGCATTC...	12
12	...TATCAGGAAG|gtgagagccc...	12	...gctttgccag|**A**ATCTTTTGA...	13
13	...TGGACACT**G**G|gtactggttc...	13	...ttgatgacag|GGACTTCTCC...	14
14	...ATAAAGACAG|gttagtgggt...	14	...ttcaccccag|**T**ATGAATGCT...	15
15	...CTGTGGAAAG|gtaggtggta...	15	...ttttgtgcag|GTTAAAAGAG...	16
16	...GTCTGTCCAG|gtgaggccag...	16	...gctcttccag|**T**CTCCAGCTG...	17
17	...CACCCTCCAG|gtaagttcta...	17	...cttcatccag|GTCAGCTCCG...	18
18	...CGGGCAGC**T**G|gtaagggtgg...	18	...gccttctcag|**T**GACGTGGTG...	19
19	...TCTCTGTC**T**G|gtgagaggca...	19	...tccctcctag|GGATGTTCAC...	20
20	...ATCCTGGCAG|gtaatgcagg...	20	...ctgtcctcag|GAAGTGATAT...	21
21	...TTTCCACAT**G**|gtaagcaagc...	21	...gcctccccag|GAACCTGGTT...	22
22	...GCACCTCCA**A**|gtacgtgacc...	22	...gtctctgcag|CTCGGGAAGG...	23
23	...GTAAACGAAG|gtgaggctcc...	23	...ctgtaattag|GGTTTGAATA...	24
24	...AACAGTAG**T**G|gtaggtgctt...	24	...gtgtccacag|GGTCTGCGGA...	25

consensus: (C/A)AG|gt(a/g)agt.....$(t/c)_n$n(c/t)ag|G

Note: Upper case letters are exons and lower case letters, introns. The bold letters are those do not match with the consensus.

Figure 19: Alignment of genomic DNA and human HIRA cDNA

```
contig2.1  > 9213     gaggaggagcggcgcaggggatgcggctgtggtggcggcg
 human     > 1                              ..................

contig2.1  > 9253     gcggcggccgagcgcgggtggcggctgtggcggcggaggg
 human     > 22       ....................................

contig2.1  > 9293     gggcgcgggccggcgatggcgcggcggccctgagggcgcg
 human     > 62       ....................................

contig2.1  > 9333     gggcgggcggcggccggagggcgggcggcgcgggaggaag
 human     > 102      ....................................

contig2.1  > 9373     cggcggcggtggctccatggcccgggcgcgctgagggacc
 human     > 142      ....................................

contig2.1  > 9413     cggcgctcgcctcagcccggcggcggcggcggccgaacaa
 human     > 182      ....................................

contig2.1  > 9453     tgaagctcctgaagccgacctgggtcaaccacaatg|gtga
 human     > 222      ..................................
HIRA       > 1        M  K  L  L  K  P  T  W  V  N  H  N  G

contig2.1  > 9493     gtgcgcgcaggggtcgcgggaggccgagcccggagtcggg

----------------intron 1, sequence 9533-30092 not show----------------

contig2.1  > 30093    acacttcaatgaataaaagccaatttagatactgtatttt

contig2.1  > 30133    tttttctttgtttccag|gcaagccgatttttcagttgat
 human     > 257                      .. ....................
HIRA       > 13                         G  K  P  I  F  S  V  D

contig2.1  > 30173    attcaccctgacgggaccaagttcgcaactggaggacaag|
 human     > 281      ....................................
HIRA       > 21        I  H  P  D  G  T  K  F  A  T  G  G  Q

contig2.1  > 30213    gtatgtttatggaataaagaaagggaaagcttaacagccc
 human     > 321      .
HIRA       > 34       G

contig2.1  > 30253    tgccaccttgctcttctcctgctgtccatttagctgtaga

----------------intron 2, sequence 30293-32292 not shown----------------

contig2.1  > 32293    ctactccttgcacattcagccatttctgctccttttgttcc

contig2.1  > 32333    ag|ggcaggattctgggaaggttgtgatctggaatatgtct
 human     > 319      ....................................
HIRA       > 34         G  Q  D  S  G  K  V  V  I  W  N  M  S

contig2.1  > 32373    ccagtcctccaggaggatgacgagaaggatgaaaatattc
 human     > 359      ....................................
HIRA       > 47        P  V  L  Q  E  D  D  E  K  D  E  N  I

contig2.1  > 32413    ccaagatgctttgccagatggacaatcacttag|gtattac
 human     > 399      ..............................
HIRA       > 60       P  K  M  L  C  Q  M  D  N  H  L

contig2.1  > 32453    agagtggccctttgcaggctttgcgtgttggcagagtgcc

----------------intron 3, sequence 32493-33612 not shown----------------
```

```
contig2.1  > 33613     gactactggattttcatcaccttatttttttttcttccca
 human     > 430                                              .
HIRA       > 70

contig2.1  > 33653     g|catgtgtgaactgtgtgcggtggtcaaacagtgggatgt
 human     > 431       . ....................................
HIRA       > 71          A  C  V  N  C  V  R  W  S  N  S  G  M

contig2.1  > 33693     atttagcttctgggggagatgacaaactgattatggtgtg
 human     > 471       ........................................
HIRA       > 84        Y  L  A  S  G  G  D  D  K  L  I  M  V  W

contig2.1  > 33733     gaagcgggctac|gtaagcatcattttccttttcttcacatt
 human     > 511       ............ ...
HIRA       > 98          K  R  A  T  Y

contig2.1  > 33773     ttatttttagaagctgctttgctggtctttatcaaagtca

~~~~~~~~~~~~~~~intron 4, sequence 33813-34972 not shown~~~~~~~~~~~~~~~~~

contig2.1  > 34973     gctggctgcagggagcatttctatgcagtgcccagagcac

contig2.1  > 35013     tccatgagtgccttttttggccccttttctcacacag|gtaca
 human     > 523                                            .....
HIRA       > 102                                             Y

contig2.1  > 35053     tcggccccagcaccgtgttcggctccagtggtaagcttgc
 human     > 528       ........................................
HIRA       > 103       I  G  P  S  T  V  F  G  S  S  G  K  L  A


contig2.1  > 35093     caatgtggagcagtggcggtgtgtctctatcctccggaat
 human     > 568       ........................................
HIRA       > 117         N  V  E  Q  W  R  C  V  S  I  L  R  N

contig2.1  > 35133     cattcaggcg|gtgagtgggactgcctgttggggtgggtgg
 human     > 608       ..........
HIRA       > 130        H  S  G

contig2.1  > 35173     gcccctgacacctggctcagatcagcagcaccgacacgtc

~~~~~~~~~~~~~~~intron 5, sequence 35213-42772 not shown~~~~~~~~~~~~~~~~~

contig2.1  > 42773     ctggccaagctaatagacaagggcatcaccaggtaagaga

contig2.1  > 42813     tctcaccagaaccctctgtcttctttcag|atgtgatggat
 human     > 617                                  . ..........
HIRA       > 133                                    D  V  M  D

contig2.1  > 42853     gtagcatggtctccccacgatgcctggctagcctcatgca
 human     > 629       ........................................
HIRA       > 137        V  A  W  S  P  H  D  A  W  L  A  S  C

contig2.1  > 42893     gcgtggataacactgtcgtcatctggaatgctgtaaagtt
 human     > 669       ........................................
HIRA       > 150       S  V  D  N  T  V  V  I  W  N  A  V  K  F

contig2.1  > 42933     cccag|gtctggggcctcctctgactggttaacagcaagca
 human     > 709       .....
HIRA       > 164         P

contig2.1  > 42973     gtactgcctgtcagtttctggagcaggggttgggctgtct
```

```
----------intron 6, sequence 43013-43932 not shown----------

contig2.1  > 43933     tcatgtttccctcagacaacacctctccacattccttttt

contig2.1  > 43973     tcttcacag|aaattctagctactctgagaggtcattctgg
 human     > 711           ... ............................
HIRA       > 164           P  E I L A T L R G H S G

contig2.1  > 44013     cttggtcaaagggttgacatgggaccctgttggtaaatac
 human     > 745     ........................................
HIRA       > 176       L V K G L T W D P V G K Y

contig2.1  > 44053     atagcttctcaagctgatgaccgcagcctaaaggtgtgga
 human     > 785     ........................................
HIRA       > 189      I A S Q A D D R S L K V W

contig2.1  > 44093     ggacgctggactggcagttggagaccagcatcaccaagcc
 human     > 825     ........................................
HIRA       > 202     R T L D W Q L E T S I T K P

contig2.1  > 44133     ttttgatgag|gtaagtggggacagctcagggagctgagcc
 human     > 865     ..........
HIRA       > 216       F D E

contig2.1  > 44173     cccgcagctgccccagcctcttcctgacttctgtgtacag

----------intron 7, sequence 44213-46372 not shown----------

contig2.1  > 46373     cacaaaggtggccttgagctcctttcaagtgcagatgctt

contig2.1  > 46413     ctttcag|tgtggaggaacgacccatgtgttgcggctcagc
 human     > 873          .. ................................
HIRA       > 218          E  C G G T T H V L R L S

contig2.1  > 46453     tggtcacctgatgggcattacctggtgtctgcccatgcca
 human     > 908     ........................................
HIRA       > 230      W S P D G H Y L V S A H A

contig2.1  > 46493     tgaacaactcaggccccactgcccagatcatcgaacggga
 human     > 948     ........................................
HIRA       > 243     M N N S G P T A Q I I E R E

contig2.1  > 46533     gggatggaagaccaacatggactttgttgggcaccggaaa
 human     > 988     ........................................
HIRA       > 257       G W K T N M D F V G H R K

contig2.1  > 46573     gctgtgactgtcgtg|gtgagtgcccagggaactaaaggga
 human     > 1028    ..............
HIRA       > 270      A V T V V

contig2.1  > 46613     gatgggctccaagattcagcaaatattaattcttggttta

----------intron 8, sequence 46653-48652 not shown----------

contig2.1  > 48653     gggcacaggggacatgcctgatgcttgattaattattttt

contig2.1  > 48693     aacatctgttatttggctttag|aaattcaacccaaaaatc
 human     > 1042                          . ................
HIRA       > 275                             K F N P K I

contig2.1  > 48733     ttcaaaaagaagcagaagaatgggagttctgcgaagccta
 human     > 1061    ........................................
HIRA       > 281      F K K K Q K N G S S A K P
```

```
contig2.1  > 48773     gctgcccgtactgctgctgtgctgttggcagcaaggaccg
 human     > 1101      ........................................
HIPA       > 294       S  C  P  Y  C  C  C  A  V  G  S  K  D  R

contig2.1  > 48813     ctcgctttctgtctgg|gtgagcctgctatcctgagctttt
 human     > 1141      ................
HIRA       > 308         S  L  S  V  W

contig2.1  > 48853     gttgcaggacgggaggtgtaggttggtaaagtagcaagtg

~~~~~~~~~~~~~~~~intron 9, sequence 48893-52332 not shown~~~~~~~~~~~~~~~~

contig2.1  > 52333     ttcagtaatttcattaaccttgactgtgtctttccttcca

contig2.1  > 52373     g|CTCACATGTCTGAAACGGCCGCTGGTGGTCATCCATGAA
                          L  T  C  L  K  R  P  L  V  V  I  H  E

contig2.1  > 52413     CTGTTTGACAAATCCATCATGGATATTTCCTG|gtaaggtg
                         L  F  D  K  S  I  M  D  I  S  W

contig2.1  > 52453     gattctacttacaggaataggttaagaaaggagaagccac

~~~~~~~~~~~~~~~~intron 10, sequence 52493-53052 not shown~~~~~~~~~~~~~~~~

contig2.1  > 53053     ggcacatgtgtctgcctggggctcttgctaccaggctgct

contig2.1  > 53093     gatggccatgtctttcag|gactctgaatgggctgggcatc
 human     > 1227                        . ......................
HIRA       > 336                        W  T  L  N  G  L  G  I

contig2.1  > 53133     ttggtatgctctatggacggctctgtggcattcctcgact
 human     > 1250      ........................................
HIRA       > 344       L  V  C  S  M  D  G  S  V  A  F  L  D

contig2.1  > 53173     tctcccaggatgagcttggcgatcccctgagcgaggagga
 human     > 1290      ........................................
HIRA       > 357       F  S  Q  D  E  L  G  D  P  L  S  E  E  E

contig2.1  > 53213     gaag|gtaggctgacggctcttgtgctgctgggtgctctgt
 human     > 1330      ....
HIRA       > 371         K

contig2.1  > 53253     ggccattgccacaatagctctgagagacccctttgactac

~~~~~~~~~~~~~~~~intron 11, sequence 53293-55132 not shown~~~~~~~~~~~~~~~~

contig2.1  > 55133     cagcactggccacactggcactcatgctggcatgtcatgc

contig2.1  > 55173     ctctccttgcttccccag|agccgcattcaccagtccacct
 human     > 1332                        .. ......................
HIRA       > 371                        K  S  R  I  H  Q  S  T

contig2.1  > 55213     atggcaagagcctagccatcatgaccgaggcccagctctc
 human     > 1356      ........................................
HIRA       > 379       Y  G  K  S  L  A  I  M  T  E  A  Q  L  S

contig2.1  > 55253     cacagccgtcattgagaaccctgagatgctcaagtaccag
 human     > 1396      ........................................
HIRA       > 393         T  A  V  I  E  N  P  E  M  L  K  Y  Q

contig2.1  > 55293     cgaaggcagcagcagcagctggaccagaagagtgctg
 human     > 1436      ........................................
HIRA       > 406       R  R  Q  Q  Q  Q  Q  L  D  Q  K  S  A
```

```
contig2.1  > 55333     cgaccagggagatgggctcagccacctcagtcgcaggcgt
 human     > 1476      ........................................
HIRA       > 419       A  T  R  E  M  G  S  A  T  S  V  A  G  V


contig2.1  > 55373     tgtcaacggggagagtcttgaagatatcaggaag|gtgaga
 human     > 1516      ................................
HIRA       > 433         V  N  G  E  S  L  E  D  I  R  K

contig2.1  > 55413     gccctgcctggtggcgctaaagacatggacatatgcccag

----------------intron 12, sequence 55453-57172 not shown----------------

contig2.1  > 57173     tagttggccaaaacttttgcattcatttatgtaactctgc

contig2.1  > 57213     tttgccag|aatcttttgaagaaacaagttgagactcggac
 human     > 1548           .. ............................
HIRA       > 443            K   N  L  L  K  K  Q  V  E  T  R  T

contig2.1  > 57253     agcagatggccggagaagaatcacgcctctctgcatagca
 human     > 1582      ........................................
HIRA       > 455         A  D  G  R  R  R  I  T  P  L  C  I  A

contig2.1  > 57293     cagctggacactgg|gtactggttcattccccttttcatggt
 human     > 1622      ............. .
HIRA       > 468        Q  L  D  T  G

contig2.1  > 57333     cttatctggcgagatcttcctgtcttctttcagcggttaa

----------------intron 13, sequence 57373-62812 not shown----------------

contig2.1  > 62813     tggcacctgactttcaaggctgacttttgttcttacttga

contig2.1  > 62853     tgacag|ggacttctccacggcattctttaacagcatcccc
 human     > 1635           . ..............................
HIRA       > 472            G   D  F  S  T  A  F  F  N  S  I  P

contig2.1  > 62893     ctctcgggctccctggcgggcaccatgctctcttctcata
 human     > 1670      ........................................
HIRA       > 484        L  S  G  S  L  A  G  T  M  L  S  S  H

contig2.1  > 62933     gcagtccacagctactgccactggactccagtacccctaa
 human     > 1710      ........................................
HIRA       > 497       S  S  P  Q  L  L  P  L  D  S  S  T  P  N

contig2.1  > 62973     ctccttcggcgcctcgaagccttgcacagagcctgtggtg
 human     > 1750      ........................................
HIRA       > 511         S  F  G  A  S  K  P  C  T  E  P  V  V

contig2.1  > 63013     gctgccagtgccagacctgcaggcgattctgtcaataaag
 human     > 1790      ........................................
HIRA       > 524        A  A  S  A  R  P  A  G  D  S  V  N  K

contig2.1  > 63053     acag|gttagtgggtgccaaggcttgttcctgtccctgggg
 human     > 1830      ....
HIRA       > 537       D  S

contig2.1  > 63093     aaagttcaggtttgcacatagttttgaggacagagaattt

----------------intron 14, sequence 63133-65052 not shown----------------

contig2.1  > 65053     gctccaaataatgcaaatactagacctccagatcctttca
```

```
contig2.1  > 65093     ggattgcgcttgacatgcctcccctcttcttcaccccag|t
 human     > 1831                                          ... .
HIRA       > 538                                            S

contig2.1  > 65133     atgaatgctacctctactcctgctgcattgtcaccttctg
 human     > 1835      ....................................
HIRA       > 539        M  N  A  T  S  T  P  A  A  L  S  P  S

contig2.1  > 65173     tgttaacgaccccgtccaagatcgaacccatgaaagcgtt
 human     > 1875      ....................................
HIRA       > 552       V  L  T  T  P  S  K  I  E  P  M  K  A  F

contig2.1  > 65213     tgactcccggttcacagagcggtccaaagccacaccaggt
 human     > 1915      ....................................
HIRA       > 566         D  S  R  F  T  E  R  S  K  A  T  P  G

contig2.1  > 65253     gctcctgccctgaccagcatgactccgacagctgtggaaa
 human     > 1955      ....................................
HIRA       > 579        A  P  A  L  T  S  M  T  P  T  A  V  E

contig2.1  > 65293     g|gtaggtggtagctgcaggaccctgaaggaccaggcaggg
 human     > 1995      . ..
HIRA       > 592       R

contig2.1  > 65333     tgcaggcatgggtcaacagcacagtggcagtgttcccacc

----------------intron 15, sequence 65373-78932 not shown----------------

contig2.1  > 78933     cactctttcttgcatgaatggtttctgaagggaagtgtga

contig2.1  > 78973     tgtaatgtttttgtgcag|gttaaaagagcagaaccttgtg
 human     > 1994                       .. ..................
HIRA       > 592                         R  L  K  E  Q  N  L  V

contig2.1  > 79013     aaagagctgaggccccgagacctcctggagagcagcagtg
 human     > 2018      ....................................
HIRA       > 600        K  E  L  R  P  R  D  L  L  E  S  S  S

contig2.1  > 79053     acagcgatgagaaagtccctttggctaaggcttcctcact
 human     > 2058      ....................................
HIRA       > 613       D  S  D  E  K  V  P  L  A  K  A  S  S  L

contig2.1  > 79093     gtccaagcgaaaacttgagcttgaggtagagacagtagag
 human     > 2098      ....................................
HIRA       > 627         S  K  R  K  L  E  L  E  V  E  T  V  E

contig2.1  > 79133     aagaagaagaaagggcggcctcggaaggactctcgtctca
 human     > 2138      ....................................
HIRA       > 640        K  K  K  K  G  R  P  R  K  D  S  R  L

contig2.1  > 79173     tgcctgtgtctctgtctgtccag|gtgaggccagtgggccc
 human     > 2178      ....................
HIRA       > 653       M  P  V  S  L  S  V  Q

contig2.1  > 79213     atcctgccttctttagcccacctttctgtacttagagctt

----------------intron 16, sequence 79253-79532 not shown----------------

contig2.1  > 79533     accccgtagtgttagagcaaggctcaggaacgtttctggc

contig2.1  > 79573     tcttccag|tctccagctgccctaaccgcagagaaggaggc
 mouse     > 2195                .............gt.ta.............
HIRA prote > 659                  S  P  A  A  L  S  T  E  K  E  A
```

```
contig2.1  > 79613     catgtgtctgtctgcaccagcacttgcactgaagctgcca
 mouse     > 2227      ........................................
HIRA prote > 670        M  C  L  S  A  P  A  L  A  L  K  L  P

contig2.1  > 79653     attccaagcccccagagagcattcaccctccag|gtaagtt
 mouse     > 2267      ......g....a........g.......... ..
HIRA prote > 683        I  P  G  P  Q  R  A  F  T  L  Q   V

contig2.1  > 79693     ctaggaccctgccagttctccctcaggcacacccagacat

~~~~~~~~~~~~~~~intron 17, sequence 79733-81372 not shown~~~~~~~~~~~~~~~

contig2.1  > 81373     ccttggagcatgtcacctgggctgcggcccttgtctgagc

contig2.1  > 81413     agacccgtgtgtttacttcatccag|gtcagctccgatcct
 human     > 2301                          ..... ..............
HIRA       > 694                           L  Q   V  S  S  D  P

contig2.1  > 81453     tccatgtacattgaggtggagaatgaagtgacagtggtgg
 human     > 2321      ........................................
HIRA       > 701        S  M  Y  I  E  V  E  N  E  V  T  V  V

contig2.1  > 81493     ggggcgtgaagctgagccgcctgaagtgcaaccgggaagg
 human     > 2361      ........................................
HIRA       > 714       G  G  V  K  L  S  R  L  K  C  N  R  E  G

contig2.1  > 81533     gaaggagtgggagacggtactcaccagccggatcctcact
 human     > 2401      ........................................
HIRA       > 728         K  E  W  E  T  V  L  T  S  R  I  L  T

contig2.1  > 81573     gctgcgggcagctg|gtaagggtgggcaggtccttagccgg
 human     > 2441      ............
HIRA       > 741        A  A  G  S  C

contig2.1  > 81613     cccaggtgacacaggcacacatgccatcattctgggcctg

~~~~~~~~~~~~~~~intron 18, sequence 81653-83812 not shown~~~~~~~~~~~~~~~

contig2.1  > 83813     actgtcttggctgcagcactcagcccccatgacacctgcc

contig2.1  > 83853     cccactctgccttctcag|tgacgtggtgtgtgtcgcctgt
 human     > 2454                          . ....................
HIRA       > 745                          C   D  V  V  C  V  A  C

contig2.1  > 83893     gaaaaaaggatgctgtcagtgttctccacctgtggtcgcc
 human     > 2477      ........................................
HIRA       > 753        E  K  R  M  L  S  V  F  S  T  C  G  R

contig2.1  > 83933     gtctcctctctcccatcctcctgccatccccgatctctac
 human     > 2517      ........................................
HIRA       > 766       R  L  L  S  P  I  L  L  P  S  P  I  S  T

contig2.1  > 83973     tttgcattgcacaggctcctacgtcatggcgctcaccgct
 human     > 2557      ........................................
HIRA       > 780         L  H  C  T  G  S  Y  V  M  A  L  T  A

contig2.1  > 84013     gcagccacactctctgtctg|gtgagaggcaggcacagggt
 human     > 2597      ..................... .
HIRA       > 793        A  A  T  L  S  V  W

contig2.1  > 84053     ggggtgggcttcagccagggcaggacagggctgtgtaggc

~~~~~~~~~~~~~~~intron 19, sequence 84093-84572 not shown~~~~~~~~~~~~~~~
```

```
contig2.1  > 84573     aggcacactctgatgtgtttacctgggttggtgagtaaac

contig2.1  > 84613     tgttccctgtctccctcctag|GGATGTTCACAGACAGGTG
                                             D  V  H  R  Q  V

contig2.1  > 84653     GTTGTGGTGAAAGAAGAGTCTCTACACTCCATCCTGGCAG|
                        V  V  V  K  E  E  S  L  H  S  I  L  A

contig2.1  > 84693     gtaatgcaggcaacaggctgtccctactcttttgggggcc

~~~~~~~~~~~~~~~~intron 20, sequence 84733-85012 not shown~~~~~~~~~~~~~~~~

contig2.1  > 85013     gctgatcagagcacacccagaactcattcccatgctgtcc

contig2.1  > 85053     tcag|gaagtgatatgacggtatcacagatcttgctgacgc
 human     > 2673      ... ....................................
HIRA       > 818       A   G  S  D  M  T  V  S  Q  I  L  L  T

contig2.1  > 85093     agcatggaatcccagtaatgaacctgtccgatgggaaggc
 human     > 2712      ....................................
HIRA       > 831       Q  H  G  I  P  V  M  N  L  S  D  G  K  A

contig2.1  > 85133     gtactgctttaatccgtcactttccacatg|gtaagcaagc
 human     > 2752      ............................. .
HIRA       > 845         Y  C  F  N  P  S  L  S  T  W

contig2.1  > 85173     tgacccccagtcctattccctggaagagtgtctgtctgct

~~~~~~~~~~~~~~~~intron 21, sequence 85213-86732 not shown~~~~~~~~~~~~~~~~

contig2.1  > 86733     tctggttggcaaagaggccaaccagatctcttggttttcc

contig2.1  > 86773     tgagctctgctcctgcctttgcctccccag|gaacctggtt
 human     > 2781                                  . ..........
HIRA       > 854                                   W   N  L  V

contig2.1  > 86813     tctgacaagcaggactcactggctcagtgtgcagacttta
 human     > 2792      ....................................
HIRA       > 858       S  D  K  Q  D  S  L  A  Q  C  A  D  F

contig2.1  > 86853     ggagcagcctgccatcccaggacgccatgctgtgctcagg
 human     > 2832      ....................................
HIRA       > 871       R  S  S  L  P  S  Q  D  A  M  L  C  S  G

contig2.1  > 86893     accgttagccataatccagggccgcacctccaa|gtacgtg
 human     > 2872      ..............a.................
HIRA       > 885         P  L  A  I  N  Q  G  R  T  S  N

contig2.1  > 86933     acctgaatgccatggccactgtccttagtcctgccctggg

~~~~~~~~~~~~~~~~intron 22, sequence 86973-87332 not shown~~~~~~~~~~~~~~~~

contig2.1  > 87333     cacgggggcagttcctgactctgtgtagtgcctgctgccc

contig2.1  > 87373     catcatggaaactctcctggtctctgcag|ctcgggaaggc
 human     > 2905                                  ...........
HIRA       > 896                                     S  G  R

contig2.1  > 87413     aggctgcccggctcttctccgtgcctcatgtggtgcagca
 human     > 2916      ....................................
HIRA       > 899       Q  A  A  R  L  F  S  V  P  H  V  V  Q  Q
```

```
contig2.1  > 87453     agagaccaccctggcctacctagagaaccaggtggcagca
 human     > 2956      ........................................
HIRA       > 913         E  T  T  L  A  Y  L  E  N  Q  V  A  A

contig2.1  > 87493     gcactcaccctgcagtccagccacgagtaccgccattggc
 human     > 2996      ........................................
HIRA       > 926        A  L  T  L  Q  S  S  H  E  Y  R  H  W

contig2.1  > 87533     tcctcgtctacgcacggtacctcgtaaacgaag|gtgaggc
 human     > 3036      ................................... .
HIRA       > 939       L  L  V  Y  A  R  Y  L  V  N  E  G

contig2.1  > 87573     tcctggcagccctgagcagccttctccttctgcccctcca

----------------intron 23, sequence 87613-89412 not shown----------------

contig2.1  > 89413     aatgtgttcctatttctgttttgtttttgttgttgttgt

contig2.1  > 89453     ttgtttgtttttctgtaattag|ggtttgaataccgacttc
 human     > 3067                          .. ...............
HIRA       > 950                             G  F  E  Y  R  L

contig2.1  > 89493     gagaaatatgcaaggacttactgggtccggttcactactc
 human     > 3087      ........................................
HIRA       > 956       R  E  I  C  K  D  L  L  G  P  V  H  Y  S

contig2.1  > 89533     cactggaagccagtgggagtcaacagtagtg|gtaggtgct
 human     > 3127      ................................. .
HIRA       > 970         T  G  S  Q  W  E  S  T  V  V

contig2.1  > 89573     tctttcattgttgctgcaagaattctttcttgaaacatga

----------------intron 24, sequence 89613-109292 not shown----------------

contig2.1  > 109293    gccaggaaagtccccacaatctaggtctcttcactgacac

contig2.1  > 109333    ttgcactcagctgtcactctgtgtccacag|ggtctgcgga
 human     > 3157                                  . .........
HIRA       > 980                                     G  L  R

contig2.1  > 109373    agagggagctgctgaaggagctgctaccagtcatcgggca
 human     > 3168      ........................................
HIRA       > 983       K  R  E  L  L  K  E  L  L  P  V  I  G  Q

contig2.1  > 109413    gaacctccgattccagcgcctcttcaccgagtgtcaggaa
 human     > 3208      ........................................
HIRA       > 997         N  L  R  F  Q  R  L  F  T  E  C  Q  E

contig2.1  > 109453    cagctcgacatcctgagggacaagtagcctgccccagcct
 human     > 3248      ........................................
HIRA       > 1010       Q  L  D  I  L  R  D  K ^^^

contig2.1  > 109493    gccctggctgcagcaagggcagggccacactctcgccgct
 human     > 3288      ........................................

contig2.1  > 109533    gatgacatgcaggaccgcctctcacctgaccaggctgtag
 human     > 3328      ........................................

contig2.1  > 109573    ggggaggagacactggcaggagatgtgctgtcctgcacca
 human     > 3368      ........................................

contig2.1  > 109613    gcgccagcccagctccctgggcagatgtgccctgtgtcct
 human     > 3408      ........................................
```

```
contig2.1  > 109653    gggtctgacattgcctcaggaggggaagctcatccctcc
 human     > 3448      ....................................

contig2.1  > 109693    ctccaagcccccgatgcggagctagggctggagctctgcc
 human     > 3488      ....................................ng.

contig2.1  > 109733    caggtgccctggggccagcaaggcagtcccaggcctgccg
 human     > 3528      ....................................

contig2.1  > 109773    tctccagcacggccccaaggtggacactagcccctgctgc
 human     > 3568      ....................................

contig2.1  > 109813    tgcaggcgccatgctgcttcagcagtgacgattgagccat
 human     > 3608      ....................................

contig2.1  > 109853    ttgtgagagagaatccggaagcgttaggtattacagcaga
 human     > 3648      ........c..........n..........n.......
                                          \      \
                                          |      |
                                          a      a

contig2.1  > 109893    ctgacctaagacctatgtaaaaggaaagctgttccaacac
 human     > 3690      ....................................
                                    \      \
                                    |      |
                                    c      a

contig2.1  > 109933    acggactatttttgtactagaatttgctaacttgtaatat
 human     > 3732      ....................................
                              \
                              |
                              a

contig2.1  > 109973    tgaatttccttttggccgataatcaggatttccctataagt
 human     > 3773      ....................................

contig2.1  > 110013    cacttggacattggtcacttgtaggaaatttaaactctaa
 human     > 3813      ....................................

contig2.1  > 110053    ttatgacagctacactgaaaaataattgtactgaaattaa
 human     > 3853      ....................................

contig2.1  > 110093    cttgtctatcttcatttggttttaatttttaaatgtttgt
 human     > 3893      ....................................

contig2.1  > 110133    aaaaagagactgttttgggggaatggggcaaaggggtggg
 human     > 3933      ....................................

contig2.1  > 110173    cgatttcttttgtaagtgtaaaataaatgaacacgcattg
 human     > 3973      ....................................

contig2.1  > 110213    aataccaaaccctcctgatttctcttgccttttccttcta
 human     > 4013      ......

contig2.1  > 110253    actttcatcctccccacacagtgcatctctctcctgaagt
```

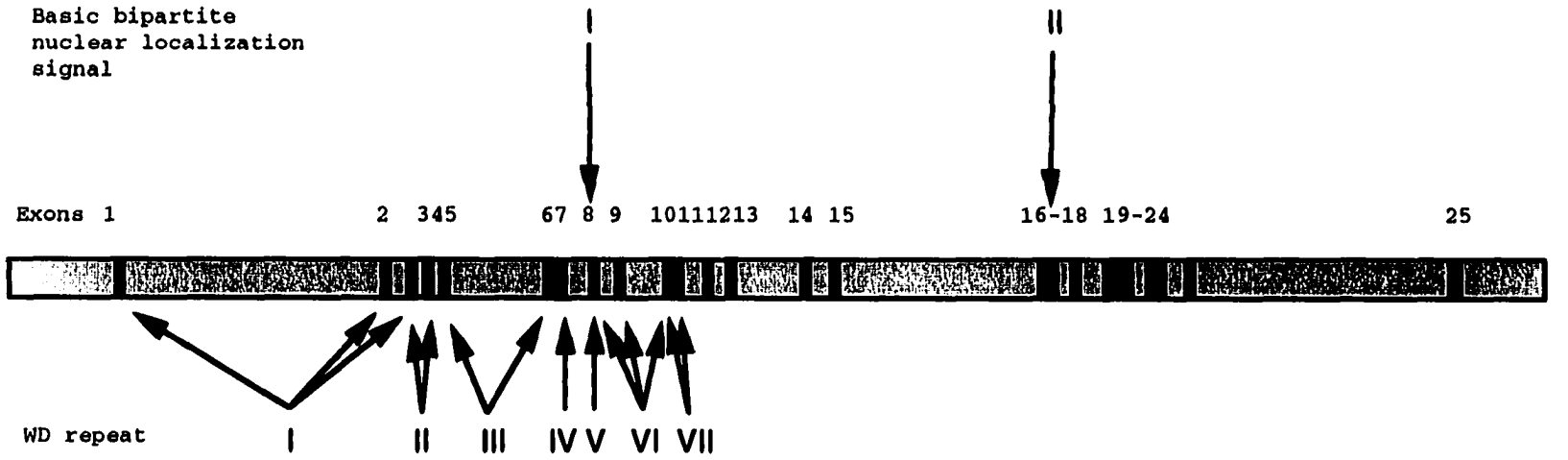
Note: Not all sequences are shown. "." indicates the identity. "|" indicates the splice site. Capitalized sequences are exon 10 and 20 which Powerblast program failed to find. For exon 17, Powerblast only found mouse cDNA. Poly(A) signal is bold.

The N-terminal part of HIRA protein containing exons 1 to 11 is highly similar to a corresponding region in the yeast HIR1 protein. [63, 258] They both contain seven WD [259] repeats and have a similar arrangement. No relationship is observed between the structure of exon/intron and the distribution of the WD repeats. WD repeat I spans exons 1, 2 and 3. II spans exons 3 and 4. III is in exons 5 and 6. IV and V are in exons 7 and 8 respectively. VI is in exons 8, 9 and 10. Finally, VII is in exons 10 and 11 (Figure 20). The C-terminal part of HIRA protein (exons 13 - 25) encodes a region of significant similarity to the yeast HIR2 protein. [63, 258] Exon 12 encodes a glutamine-rich region unique to HIRA.

The yeast HIR1 and HIR2 gene products have been identified as a negative regulator of core histone gene transcription. [258] They are required for the periodic repression of three of four histone gene loci during the cell cycle. Not surprisingly, HIRA transcription product contains basic dipartite nuclear localization signals (one in exons 8 and 9, one in exon 16) which direct it to the nucleus. [260, 261] As with HIR1 and HIR2, HIRA does not contain any of the known DNA or RNA binding motifs. [63]

Part of the 3' end of the HIRA gene lies within the smallest DGS critical region. *In situ* hybridization experiments performed on mouse and chick embryo sections showed that the murine and avian homologs of HIRA were expressed in the neural crest region. [262, 263] This observation is compatible with the implication of the gene product in the early development of the tissues affected in DGS. If like HIR1 and HIR2 in yeast, HIRA protein is expected to interact with several other proteins to preserve the function of the multicomponent regulatory complex, it will be necessary in future studies to find the human counterparts of those proteins to study the function of HIRA and the involvement of this gene in DGS.

Figure 20: Exon/intron organization and features of human HIRA gene in contig2.1



Note: Exon (green) sizes are not shown in scale. Introns are shown in scale approximately.

3.4.1.2 Nucleotide composition of the 5' region of the HIRA gene

The overall G/C content of contig2.1 is 46%. The G/C content of the noncoding region of exon 1 is as high as 86%. The G/C content of the 5' region of the HIRA gene from -1020 to 220 (see Figure 21), which is from 8212 to 9452 in contig2.1, is 74%. As predicted by Xgrail program, there is a CpG island from position 7705 to 9721 which includes about 1.6 kb of 5' region, exon 1 and about 0.23 kb of intron 1. This large CpG island categorizes HIRA gene to the family of widely transcribed house keeping genes containing a CpG island in their promoter region.

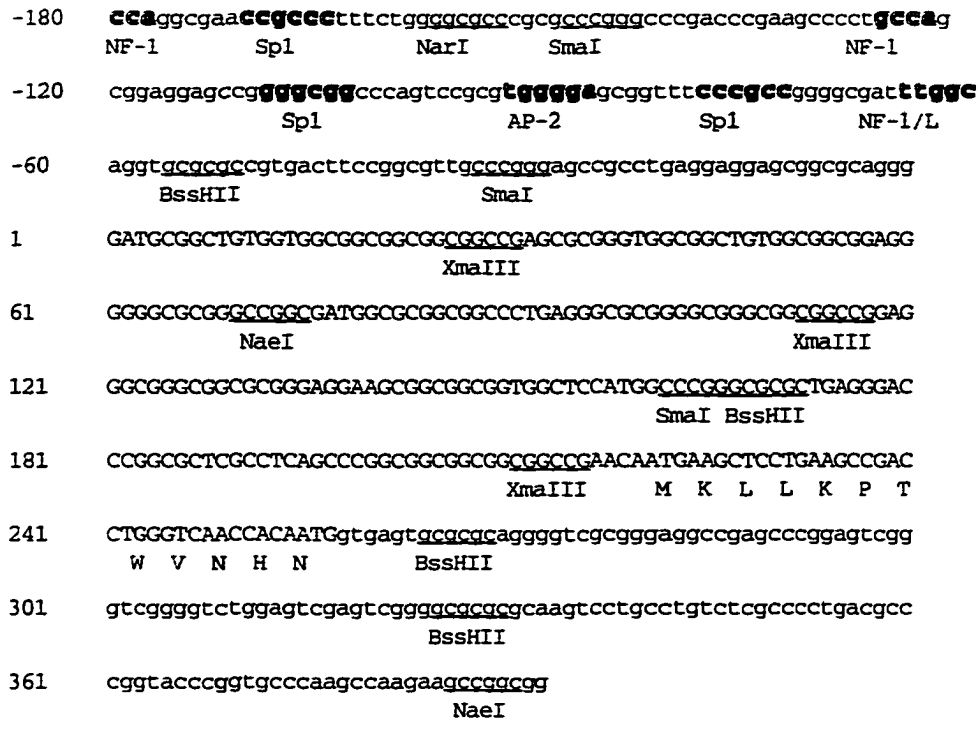
Detailed analysis of the 5' region of the HIRA gene is visualized in Figure 21. The studied region is from -1200 to 393 (defining the start position of exon 1 as position 1) which is from 8123 to 9716 in contig2.1. This region consists of 1.2 kb of proximal HIRA promoter, followed by exon 1 and 228 bp of the 5' end of intron 1. This region has an unusual cluster of 19 rare G/C-containing restriction sites including five BssHII (gcgcgc), three NaeI (gccggc), one NarI (ggcgcc), one SacII (ccgcgg), five SmaI (cccggg) and four XmaIII (cggccg).

A TATA box is not evident in the sequence of this region. However, it is normal for a G/C rich promoter not to have a TATA box. [264, 265] Furthermore, several GC boxes (GGGCGG), which bind the Sp1 transcription factor and have been shown to be required for interaction of transcription factor IID (TFIID) with a TATA-less promoter, [264, 266] are present. There also is a putative TFIID binding site. Other putative regulatory elements in the region were identified using online SIGNALSCAN against the TRANSFAC database to identify transcription factor binding sites. These identified potential sites are: AP-2 (retinoic acid-inducible activator protein 2), CCAAT box, CREB (cyclic AMP regulatory element binding protein), GR (glucocorticoid

receptor), H4TF-2 (human histone H4 gene-specific transcription factor), LF-A1 (liver specific trans-acting factor), NF-1 (nuclear factor 1), NF-1/L (nuclear factor 1-like), NF-E (erythroid specific nuclear factor), NF-kappaE2 (nuclear factor kappa enhancer 2), SRF (serum response factor). The positions of these potential elements are shown in Figure 21.

Figure 21: G/C-rich region of 5' of HIRA gene

```
-1200  attgcaatgctcccaaacccggggcacttctaactacttgatatgacagcaaggtcaacg
                          SmaI LF-A1

-1140  tggaaggtttttccaaaatctttcgttcagaccacagaggacttaagcaggaccgctgctc
                 TFIID                    GR

-1080  gccagggtgctcagattttcatgcataagagtcacgaccagatggctcatccctgagata
       NF-1

-1020  ttcgtggactgggaggcgagcagggagcacaggcgagtcctggcggacagtcgcgaagac

-960   cccgcacgcgccctgggcactccgctatccggccagcgtccgcactcacccgctagtcgg
                     LF-A1      NF-E     NF-1

-900   gcgcagggctagcgagatacttcgcagcacggaggccgtcatggctaccctcgccgcttc
-840   cgggcgtgcgccccteeacagtccggccccggtgccccctgctgaggtccggcagacca
                 Sp1             SacII             H4TF-2

-780   ggcgcccagcgctcaggcgttccggctgcgtgggctgggacgcgcctgtccccgcccccg
                                                    GR   Sp1

-720   acgcagccccgcctagggtgtcagctgcgcacacacgagggccttgcgtttcacgcggct
                                 Sp1

-660   cggggtcctcccttccgtccccgcccctgtgctgggagcgcgccctctgcagcgggcacagt
         H4TF-2             Sp1             BssHII          LF-A1

-600   ctagccagtcaggccctagccggttggtcggtctgcgccctcgcgcgggcctcggctccg
         NF-1                              Sp1

-540   cagtgcagccggagcggccgagggggcgtccctctgctcctccgagtccaaatgttcatg
                   XmaIII    Sp1                       NF-1

-480   tctaggcaaacggccccgatcttctaataaacgcgtccgcccagcctcccccaccaggg
                            GR              Sp1

-420   ttggctcgtcccccgcattcgttcgagtccgaagcccggatcgcattccggcgccggcgc
                                                          NaeI

-360   ggagcccggcgctcgcaagatgccccaatagcgcgtgttgatgaatgacggaagctagtta
                          SRF/CCAAT box          CREB

-300   ataagtcggactcagactggcgaggcagagctggtgacgggcggtccacgccacctgggc
                                GR      CREB Sp1 H4TF-2   NF-kappaE2

-240   cccgcctgcctctgcgggcggcccgggttgcaactgcaggccccgctcgctgtccgccgt
                      Sp1   SmaI                          NF-E
```



Note: Position 1 corresponds to the start site of exon 1 of the complete cDNA sequence. The sequence of exon 1 is capitalized, with the deduced amino acid sequence below the coding region. The G/C containing restriction sites are underlined and the putative regulatory elements are outlined.

It also is interesting to note that there are three H4TF-2 (human histone H4 gene-specific transcription factor) binding sites present in the promoter region of the HIRA gene since the putative function of the human HIRA gene is to negatively regulate core histone gene transcription and H4TF-2 is required for maximal transcription of histone H4 gene. Thus it is possible that there is a mechanism controlling the competitive binding of H4TF-2 to the histone H4 gene and the HIRA gene to maintain the appropriate amount of H4 in the cell, or H4TF-2 binds to the HIRA gene as an inhibitor of HIRA gene transcription to achieve the high transcription rate of the H4 gene in the S phase of the cell cycle. [258]

3.4.1.3 Repetitive elements in the HIRA gene region

Alu, L1, Mer, Mir and Mlt are interspersed repetitive elements found in contig2.1. Total repeat sequence content of contig2.1 is 32.0% (37737/118000). The individual Alu elements found in this region tend not to be full length and are small in size and the Alu percentage (total Alu bases divided by total contig length) is only 15.7%. A summary of all repetitive elements is listed in Table 6. Comparing the repetitive elements in the HIRA region and those in other regions, one can tell that there are big differences. For Alu elements, the percentage in other regions is almost two-fold higher than in the HIRA region (31.2% and 15.7% respectively). The percentage of non-Alu elements in other regions also is significantly higher (18.9% Vs 13.3%). It is worth noting that none of the interspersed repetitive elements appears to be in translated regions within the contig.

Table 6: Interspersed repetitive elements in contig 2.1

	HIRA gene region			other regions			total
Family	number	bases	% region	number	bases	% region	% region
Alu	**62**	**15824**	**15.7**	**19**	**5312**	**31.2**	**17.9**
L1	15	9818	9.7	4	2421	14.2	10.4
Mer/Mir/Mlt	9	3562	3.5	2	800	4.7	3.7
Total non-Alu	**24**	**13380**	**13.3**	**6**	**3221**	**18.9**	**14.1**

The human HIRA gene is a relative large gene which covers approximately 101.1 kb. The percentage of intronic sequence of this gene is 95.9%, which is higher than the average percentage of intronic sequence of most large genes (89.8%), while the percentage of repeat elements of this gene is 29.0% which is almost the same as the average (30.4%). [267]

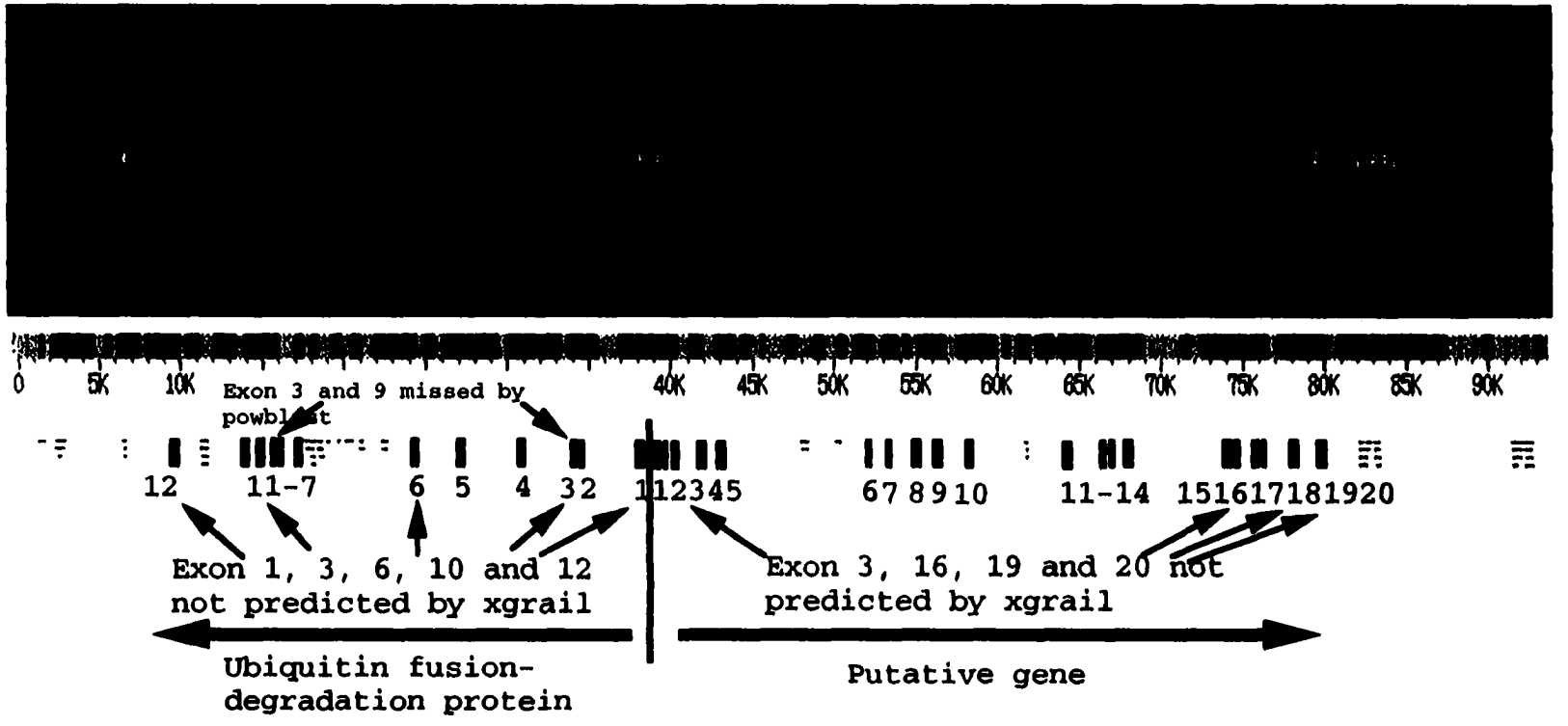
3.4.2 Contig 2.2

Contig2.2 is 93,386 bp in length. It contains one known gene, the gene for human ubiquitin fusion-degradation protein (UFD1L). The entire UFD1L gene lies in the first one third of contig2.2 and the transcription direction is from telomere to centromere. The results of Xgrail and Powerblast are provided in Figure 22. The Xgrail result showed there is a exon cluster region right after the human UFD1L gene and Powerblast found that the first three exons belong to an EST which is the 5' end of a cDNA similar to heparan sulfate proteoglycan. This adjacent exon cluster region, therefore may represent a second putative gene in this region.

3.4.2.1 Human ubiquitin fusion-degradation protein gene

Human ubiquitin fusion-degradation protein gene is found in contig 2.2 from position 38227 to 9699 with coding on the negative strand. Previous study revealed that this gene encodes a product containing 343 amino acids, [268] but this does not agree with the genomic DNA sequence. However, mouse UFD1L cDNA agrees with human genomic DNA very well. Carefully examining these three DNA sequences revealed that the previously reported human UFD1L cDNA has an extra exon which is almost identical to an adjacent exon. This extra exon, termed exon 5', encodes for thirty-six amino acids. The first twenty-seven amino acids can be found at the

Figure 22: Result of Powerblast and Xgrail analysis of contig 2.2



Note: Xgrail result is on the top. The gray line is DNA sequence. The peaks are predicted exons. The small color bars above or below the peaks indicate the quality of exons, Green for excellent, blue for good and red for marginal. Peak above the gray line means the exon is on forward strand while peak under the line, reverse. The intensity of the gray line means the GC contents. The lighter the region is, the higher the GC contents. The pink boxes on the gray line are CpG islands. Powerblast result is on the bottom. The black-gray line is DNA sequence. The gray regions are repeat sequence. The small colored bars are matches found in GenBank. The numbers under green bars are exon numbers of human UFDL1 gene and putative gene.

beginning of exon 5 which is right after exon 5' according to previous report. The last six amino acids can be found at the end of exon 4 which proceeds exon 5'. [268] A dotplot showing the comparison of the previous reported cDNA and human genomic DNA is in Figure 23. The DNA sequence of human genomic DNA (contig 2.2) has been carefully examined, and no suspect region was found. Based on the accuracy of genomic DNA sequence data and the exon structure of mouse UFD1L cDNA, it might be concluded that the previous report of human UFD1L cDNA contained an error. In addition, another discrepancy observed is that the previous reported cDNA has an additional ten bases at the beginning, and this also may represent a sequencing artifact in the cDNA sequence. Alignment of genomic sequence, human UFD1L cDNA and mouse UFD1L cDNA is provided in Figure 24.

Despite the extra exon and the extra ten bases at the beginning of exon one, the similarity of human cDNA and genomic DNA is very high, 99%. The deduced human UFD1L gene coding region sequence and the amino acid sequence from genomic DNA are listed in Figure 25.

The size of each of the exons and the size of the introns as well as the splice site of the exon/intron boundaries are listed in Table 7.

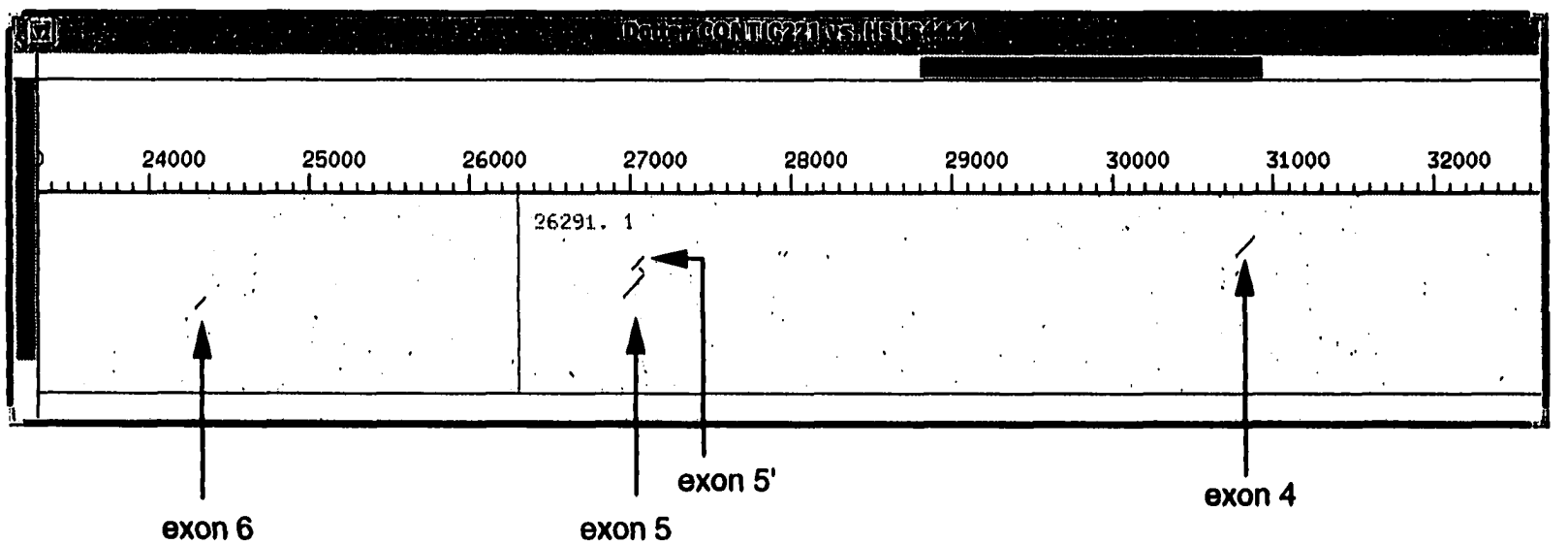


Figure 23: Dotplot comparison of human UFD1L cDNA (U64444) and human genomic DNA. Genomic DNA is horizontal and part of UFD1L cDNA is vertical.

Figure 24: Alignment of genomic sequence, human UFD1L cDNA and mouse UFD1L cDNA

```
contig2.2  > 9405      tttagtcaccttcataggttttgacagaattagggatgca

contig2.2  > 9445      gaattgctcctgctgaggcagtgggagctcccctcaagct
U64445     < 1294          ....-.........t...a.g..---...g.....

contig2.2  > 9485      gaaagcgtgaagaggtgaggaggcagctatctacaggccc
U64445     < 1262      .g.---..t.g.ga.........caac.a...g...a..
                                  \
                                  |
                                  g

contig2.2  > 9525      tcagggacagctctttgtctttccccaaatcaagcataca
U64445     < 1224      c...c.------.....---.....g........c...

contig2.2  > 9565      actacaaagtcctgctgctccacttccaagtcaaacttaa
U64445     < 1193      ..ct.c....t........ttgta....g.a...g...g.

contig2.2  > 9605      taattcttaggtaactctaaataaatcttaagtaacagag
U64445     < 1153      ..g....c........t....c...c.a...a........

contig2.2  > 9645      tattctctgatgaggctcgtccctgtcagtgccagtaact
U64445     < 1113      ...-......a----------.t......a..........

                                                poly(A) signal
contig2.2  > 9685      aaaagccaagatgttgcaaatgattctttttattattttcc
U64445     < 1084      gt......g.----.....tc.ga.........---....
U64444     < 1156                .........aa..............

contig2.2  > 9725      aatcagccaacagtcctcacttagggctttcttccctttt
U64445     < 1051      .g...........tt.a.....................c.
ubiquitin  < 302                       ^^^ P  K  R  G  K  K
                                  \
                                  |
                                  g
U64444     < 1130      ........................................
ubiquitin  < 338                       ^^^ P  K  R  G  K  K

contig2.2  > 9765      tacgcaatgactgtccttctccagagaaagcgacgaatct
U64445     < 1010      .c....gc.....................a........
ubiquitin  < 289         R  L  S  Q  G  E  G  S  F  A  V  F  R
U64444     < 1090      ........................................
ubiquitin  < 325         R  L  S  Q  G  E  G  S  F  A  V  F  R

contig2.2  > 9805      gcctccagcttcatc|ctaggtttgaagagaagatactatt
U64445     < 970       ............... ..
ubiquitin  < 283        G  G  A  E  D   E
U64444     < 1050      ............... ..
ubiquitin  < 319        G  G  A  E  D   E
```

```
contig2.2  > 9845      ttcagaaactccctggaggtttacaataaatgccttactc

              --------intron 11, sequence 9885-13764 not shown--------

contig2.2  > 13765     atcactagcaaataaatctcctcattaccaaccaaggaaa

contig2.2  > 13805     tacaaggaaatacactcac|ctcttcaacctttttgacaag
U64445     < 956                         . ......................
ubiquitin  < 277                           E  E  V  K  K  V  L

contig2.2  > 13845     gggacgtgaatttctgatgaaagttatcttaccaagttta
U64445     < 934       ........................g...........c...
ubiquitin  < 264        P  R  S  N  R  I  F  T  I  K  G  L  K

contig2.2  > 13885     aattcataattgggaattcct|ctgtaagaagagacagaga
U64445     < 894       .....g.....a......... ..
ubiquitin  < 256       F  E  Y  N  P  I  G   R

contig2.2  > 13925     cagagaaatgttatttccaggaaaaaaggactaaaattca

              --------intron 10, sequence 13965-14684 not shown-------

contig2.2  > 14685     aggctgcccgaccccagcgggctggcacagatcctcctcc

contig2.2  > 14725     tgtcctggaggagagccaggacagacctac|cttttaatat
U64444     < 953                                     ..........
ubiquitin  < 289                                    R  K  I  D
U64445     < 864                                              .
ubiquitin  < 253                                              D

contig2.2  > 14765     ctccaggcttgattggggaggggctgggctctacccctttt
U64444     < 943       ........................................
ubiquitin  < 276         G  P  K  I  P  S  P  S  P  E  V  G  K
U64445     < 863       ......................a.....t..a........
ubiquitin  < 240         G  P  K  I  P  S  P  S  P  E  V  G  K

contig2.2  > 14805     cttctttccatccagtctattgccagatccagagaaagc|c
U64444     < 903       ....................................
ubiquitin  < 263        K  K  G  D  L  R  N  G  S  G  S  F  A
U64445     < 823       t.....c..............c.....a........
ubiquitin  < 228        K  K  G  D  L  R  N  G  S  G  S  F

contig2.2  > 14845     tagacaggaagtaggtgagtcatataataagcccaagtgt

              --------intron 9, sequence 14885-15604 not shown--------

contig2.2  > 15605     ctctaatttcaaggagcagagggcacagcaatataagtca

contig2.2  > 15645     agtggtctggtactcac|GCGGAAGCCCAGCTCTCCAGCAT
                                        R  F  G  L  E  G  A  Y
```

```
contig2.2  > 15685     AGCCACTGTGGTCGGCTTCACCTTC|ctgacagggaaaaag
                         G  S  H  D  A  E  G  E

contig2.2  > 15725     aaaaacaggtgagctcctcagcggggacacccagcttccc

              ~~~~~~~intron 8, sequence 15765-15884 not shown~~~~~~~

contig2.2  > 15885     agcatgcagcaggacagcaaaggacactgaggataacact

contig2.2  > 15925     tac|tgtcgactcctcatgctggacttgtctttcgggttct
U64444     < 817         . ..................................
ubiquitin  < 235            T  S  E  E  H  Q  V  Q  R  E  P  E

contig2.2  > 15965     ttgtagcccaggggagcatcaaagtccac|ctgtgtttcca
U64444     < 779       ............................
ubiquitin  < 225       K  Y  G  L  P  A  D  F  D  V

contig2.2  > 16005     ggattttaaaagtaaaatattgagacatgaccacccttgg

              ~~~~~~~intron 7, sequence 16045-17084 not shown~~~~~~~

contig2.2  > 17085     ggtcaccaactgcccatcctcccagggcccaggctcactg

contig2.2  > 17125     taactggacagagccactcac|gttcatgtcacactcaatg
U64444     < 753                          ... ..............tga
ubiquitin  < 219                           V   N  M  D  C  E  H

contig2.2  > 17165     atggacactgccttgtcgggtttggtctccatcacacgca
U64444     < 731       ....................................
ubiquitin  < 205       I  S  V  A  K  D  P  K  T  E  M  V  R  L

contig2.2  > 17205     gttcgtagat|ctgtggggcaaacacagaaatcagttggat
U64444     < 691       .......... ..
ubiquitin  < 201         E  Y  I   K

contig2.2  > 17245     ggtttcctggcagcactggagctgtcgctgtccatgttac

              ~~~~~~~intron 6, sequence 17285-24204 not shown~~~~~~~

contig2.2  > 24205     ccccattcaaacattagaatggacactgcaggtggctctt

contig2.2  > 24245     gtcaggaatgacctgcatgaactgggctcac|cttttcatt
U64444     < 717                                      .........
ubiquitin  < 207                                       K  E  N

contig2.2  > 24285     atagttgatggcaatcacatccccggtggtcagacaggca
U64444     < 672       g.....a.....................................
ubiquitin  < 186        Y  N  I  A  I  V  D  G  T  T  L  C  A
```

```
contig2.2  > 24325     aagttcctaagtgcgttttctaat|ctgcagacacattctg
U64444     < 632       ........................
ubiquitin  < 178       F  N  R  L  A  N  E  L


contig2.2  > 24365     tcaaggcaacatggcaagatggataccttaatctgaaaca

              -------intron 5, sequence 24405-26884 not shown-------


contig2.2  > 26885     tttaacaaccgccagcacactctcctttgtctagaagaac


contig2.2  > 26925     tccgcttatttgaaaaagatac|acggctttggggttggtg
U64444     < 608                             ...................
ubiquitin  < 172                             V  A  K  P  N  T
U64445     < 519                                   ........a
ubiquitin  < 136                                   P  N  T



contig2.2  > 26965     atgtccaggaagtcagggctctgaggttggaatttggagt
U64444     < 590       ................c.......................
ubiquitin  < 158       I  D  L  F  D  A  S  Q  P  Q  F  K  S  Y
U64445     < 510       ..a..........t..........c.....c..a....
ubiquitin  < 122       I  D  L  F  D  P  S  Q  P  Q  F  K  S  Y


contig2.2  > 27005     aggtggccacttgaaggttgacgctctccacctggaccag
U64444     < 550       ........................................
ubiquitin  < 145         T  A  V  Q  L  N  V  S  E  V  Q  V  L
U64445     < 470       ....c........g........a..t........a.....
ubiquitin  < 109         T  A  V  Q  L  N  V  S  E  V  Q  V  L


contig2.2  > 27045     gccgccttcttccaagagtaagttctgcatcat|ctgaaag
U64444     < 510       ................................
ubiquitin  < 134        G  G  E  E  L  L  L  N  Q  M  M
U64445     < 430       ...c.....c.....c..c..a............
ubiquitin  < 98         G  G  E  E  L  L  L  N  Q  M  M


contig2.2  > 27085     gaagaagaggctacatgagactcctagagatgaagcagcc

              -------intron 4, sequence 27225-30685 not shown-------


contig2.2  > 30685     gggccaggactctcgagtgcaggagatgtctctgcctaag


contig2.2  > 30725     tggctgctctctagagaccctggcagccccactcac|ccag
U64444     < 369                                          ....
ubiquitin  < 97                                            W


contig2.2  > 30765     tgtgggaggtagcagatgccctcatcagccacaaactcca
U64444     < 365       ....................t..............t....
ubiquitin  < 83        H  P  L  Y  C  I  G  E  D  A  V  F  E  L
U64445     < 390          ........a........t........a..........
ubiquitin  < 83           P  L  Y  C  I  G  E  D  A  V  F  E  L
```

```
contig2.2  > 30805    gcacgccacaatgcgtcatgcggtccgaattcttattggt
U64444     < 325      .....................................
ubiquitin  < 70         V  G  C  H  T  M  R  D  S  N  K  N  T
U64445     < 353      .t..a..g..g..t.....c..a..t.............
ubiquitin  < 70         V  G  C  H  T  M  R  D  S  N  K  N  T


contig2.2  > 30845    cagtttgaacagcatgggataggtaatgttaagtcggc|ct
U64444     < 285      .....................................
ubiquitin  < 57         L  K  F  L  M  P  Y  T  I  N  L  R  S
U64445     < 313      ..a...a........a.............
ubiquitin  < 60         L  K  F  L  M  P  Y  T  I  N


contig2.2  > 30885    gaaaataagagagagactatgaggcacagctcctatgaag

            -------intron 3, sequence 30925-34084 not shown-------


contig2.2  > 34085    actcttggtgtccatatgtacattatctaagtccttaagt


contig2.2  > 34125    cactcagaagatacttac|TGAGTTGGTCCAGGGCCGAGGG
                                        L  Q  D  L  A  S  P

contig2.2  > 34165    TGGCATAATTA|ctgtggaaagaacatttaagaaatattaa
                       P  M  I  I


contig2.2  > 34205    aaaataaaaataaaaaggacagaccaaaggcaagatgcag

            -------intron 2, sequence 34245-34484 not shown-------


contig2.2  > 34485    ctacagcattgcagacttagggcaggaaaagttactattc


contig2.2  > 34525    aaataaaaacaagcacatac|tcttccctcctttctccaca
U64444     < 214                           .....................
ubiquitin  < 40                              K  G  G  K  E  V
U64445     < 242                           .....................
ubiquitin  < 40                              K  G  G  K  E  V


contig2.2  > 34565    tctgacctgtcattaggccctgctagcatggacacagaga
U64444     < 194      ..................a.....................
ubiquitin  < 26       D  S  R  D  N  P  W  A  L  M  S  V  S  F
U64445     < 222      ....................................g....
ubiquitin  < 26       D  S  R  D  N  P  G  A  L  M  S  V  S  F


contig2.2  > 34605    agcagcggtactgtgtggagaagcggttttggaagaccct
U64444     < 154      .....................................
ubiquitin  < 13         C  R  Y  Q  T  S  F  R  N  Q  F  V  R
U64445     < 182      ............c.............c.........g
ubiquitin  < 13         C  R  Y  Q  T  S  F  R  N  Q  F  V  R


contig2.2  > 34645    gggaatagggtggtcgaacatgttgaaagagaa|ctagaag
U64444     < 114      ................................... .
ubiquitin  < 2          P  I  P  H  D  F  M  N  F  S  F
U64445     < 142      ......c........a.............
ubiquitin  < 3          P  I  P  H  D  F  M  N  F  S
```

```
contig2.2  > 34685     gaggaaagagaaacaagtattaaaacaaaggtacatttct

              --------intron 1, sequence 34725-38084 not shown--------

contig2.2  > 38085     cctgccccgtcgggtcctacctcaggccccgggccaaggc

contig2.2  > 38125     ccagcccccgtccgctgtccgtcagcgcttac|catgatgg
U64444     < 81                                        ........
ubiquitin  < 1                                          M

contig2.2  > 38165     acaccacctggcagactccgctcctctcaggcaatgcaac
U64444     < 73        ....................................

contig2.2  > 38205     gaagaaaccccgccgaccgctctcccagccgccgctgccg
U64444     < 33        .....................

contig2.2  > 38245     ctgccgccgcgccaagccggtacgccccagaggctcaccg
```

Note: Not all sequences are shown. The transcription direction of UFD1L is from telomere to centromere while contig2.2 is opposite. "." indicates the identity. "|" indicates the splice site. Poly(A) signal is bold. "^^^" indicates stop codon. Capitalized sequences are exon 3 and 9 which Powerblast program failed to find. Exon 5' is not shown. U64444 is human cDNA and U64445 is mouse cDNA.

Figure 25: The human UFD1L gene coding region deduced from genomic DNA

```
1       atgttctctttcaacatgttcgaccaccctattcccagggtcttccaaaa
         M  F  S  F  N  M  F  D  H  P  I  P  R  V  F  Q  N
51      ccgcttctccacacagtaccgctgcttctctgtgtccatgctagcagggc
          R  F  S  T  Q  Y  R  C  F  S  V  S  M  L  A  G
101     ctaatgacaggtcagatgtggagaaaggagggaagataattatgccaccc
         P  N  D  R  S  D  V  E  K  G  G  K  I  I  M  P  P
151     tcggccctggaccaactcagccgacttaacattacctatcccatgctgtt
         S  A  L  D  Q  L  S  R  L  N  I  T  Y  P  M  L  F
201     caaactgaccaataagaattcggaccgcatgacgcattgtggcgtgctgg
          K  L  T  N  K  N  S  D  R  M  T  H  C  G  V  L
251     agtttgtggctgatgagggcatctgctacctcccacactggatgatgcag
         E  F  V  A  D  E  G  I  C  Y  L  P  H  W  M  M  Q
301     aacttactcttggaagaaggcggcctggtccaggtggagagcgtcaacct
         N  L  L  L  E  E  G  G  L  V  Q  V  E  S  V  N  L
351     tcaagtggccacctactccaaattccaacctcagagcctgacttcctgg
          Q  V  A  T  Y  S  K  F  Q  P  Q  S  P  D  F  L
401     acatcaccaaccccaaagccgtattagaaaacgcacttaggaactttgcc
         D  I  T  N  P  K  A  V  L  E  N  A  L  R  N  F  A
451     tgtctgaccaccggggatgtgattgccatcaactataatgaaaagatcta
         C  L  T  T  G  D  V  I  A  I  N  Y  N  E  K  I  Y
501     cgaactgcgtgtgatggagaccaaacccgacaaggcagtgtccatcattg
          E  L  R  V  M  E  T  K  P  D  K  A  V  S  I  I
551     agtgtgacatgaacgtggactttgatgctcccctgggctacaaagaaccc
         E  C  D  M  N  V  D  F  D  A  P  L  G  Y  K  E  P
601     gaaagacaagtccagcatgaggagtcgacagaaggtgaagccgaccacag
          E  R  Q  V  Q  H  E  E  S  T  E  G  E  A  D  H  S
651     tggctatgctggagagctgggcttccgcgctttctctggatctggcaata
          G  Y  A  G  E  L  G  F  R  A  F  S  G  S  G  N
701     gactggatggaaagaagaaaggggtagagcccagcccctccccaatcaag
         R  L  D  G  K  K  K  G  V  E  P  S  P  S  P  I  K
```

```
751     cctggagatattaaaagaggaattcccaattatgaatttaaacttggtaa
         P  G  D  I  K  R  G  I  P  N  Y  E  F  K  L  G  K
801     gataactttcatcagaaattcacgtccccttgtcaaaaaggttgaagagg
          I  T  F  I  R  N  S  R  P  L  V  K  K  V  E  E
851     atgaagctggaggcagattcgtcgctttctctggagaaggacagtcattg
        D  E  A  G  G  R  F  V  A  F  S  G  E  G  Q  S  L
901     cgtaaaaagggaagaaagccctaa
         R  K  K  G  R  K  P  ^^^
```

Note: Disagreements with human cDNA are shown in bold. "^^^" indicates stop codon.

Table 7: Size and boundary of exons and introns of the human UFD1L gene

exon size	exon #		intron #	intron size		exon #
91bp	1	...GTCCATCA**TG**|gtaagcgctg...	1,	3480bp	...ctccttctag|**T**TCTCTTTCA...	2
133bp	2	...GGAGGGAA**GA**|gtatgtgctt...	2,	369bp	...ctttccacag|**T**AATTATGCC...	3
33bp	3	...GACCAACT**CA**|gtaagtatct...	3,	3299bp	...ttattttcag|GCCGACTTAA...	4
122bp	4	...CCCACACT**GG**|gtgagtgggg...	4,	3683bp	...ttcctttcag|**A**TGATGCAGA...	5
131bp	5	...CCAAAGCC**GT**|gtatcttttt...	5,	2598bp	...gtgtctgcag|**A**TTAGAAAAC...	6
73bp	6	...TAATGAAAAG|gtgagcccag...	6,	7062bp	...tgccccacag|**A**TCTACGAAC...	7
69bp	7	...TGACATGAA**C**|gtgagtggct...	7,	1151bp	...ggaaacacag|GTGGACTTTG...	8
66bp	8	...GGAGTCGA**CA**|gtaagtgtta...	8,	217bp	...tccctgtcag|GAAGGTGAAG...	9
48bp	9	...GGGCTTCC**GC**|gtgagtacca...	9,	857bp	...tcctgtctag|GCTTTCTCTG...	10
89bp	10	...ATATTAAAAG|gtaggtctgt...10,		850bp	...cttcttacag|**A**GGAATTCCC...	11
82bp	11	...GGTTGAAGAG|gtgagtgtat...11,		4004bp	...tcaaacctag|GATGAAGCTG...	12
121bp	12	.....ATCATTTGCA				

consensus: (C/A)AG|gt(a/g)agt.....$(t/c)_n$n(c/t)ag|G

Note: Upper case letters are exons while lower case letters, introns. The bold letters are those do not match with the consensus.

Although this protein lacks significant sequence similarity to any human proteins of known function, Blastp program reveals that it has a high similarity to the known yeast protein ubiquitin fusion-degradation protein 1 (UFD1P), therefore this gene most likely represents the human equivalent of ubiquitin fusion-degradation protein. UFD1P protein is believed to be involved in ubiquitin fusion degradation proteolytic pathway (UFD pathway). The common feature of UFD pathway is the covalent conjugation of ubiquitin to short-lived proteins prior to their degradation by the proteasome, a multisubunit, multicatalytic protease. The actual function of this protein still remains unknown although it is known that UFD1P does function at a post-ubiquitination step of UFD pathway. [269, 270]

A multiple sequence alignment of the human, mouse and yeast ubiquitin fusion-degradation proteins using the ClustalW multiple sequence alignment program is provided in Figure 26. The identity between the human and mouse UFD1P is 98.7% while the identity between human and yeast is only 36.6%. The major areas of sequence conservation occur at the middle of this protein, most notably between residues 43 to 200 and the C-termini diverge greatly because of large insertions in yeast UFD1P.

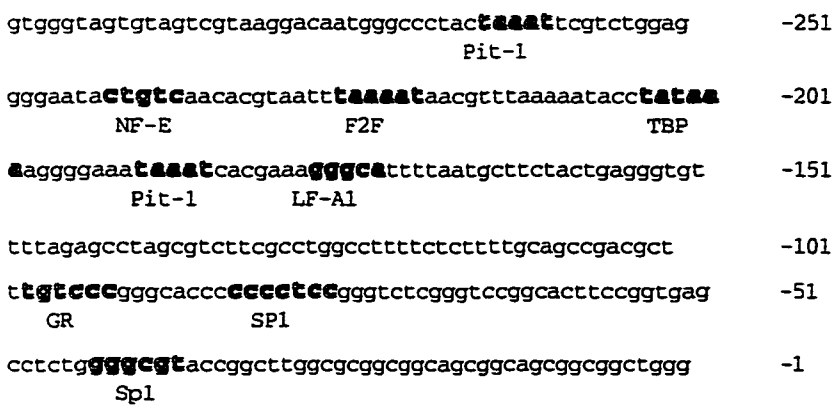
Figure 26: CLUSTAL W(1.60) multiple sequence alignment among yeast, human and mouse ubiquitin fusion degradation protein 1.

```
CLUSTAL W(1.60) multiple sequence alignment
h. vs. m.       ***_***************_************** ************************
humanufd1l      MFS-FNMFDHPIPRVFQNR-FSTQYRCFSVSMLAWPNDRSDVEKGGKIIMPPSALDQLSR
mouseufd1l      MFS-FNMFDHPIPRVFQNR-FSTQYRCFSVSMLAGPNDRSDVEKGGKIIMPPSALDQLSR
yeastufd1p      MFSGFSSFGGGNGFVNMPQTFEEFFRCYPIAMMNDRIRKDDANFGGKIFLPPSALSKLSM
h/m/y           *** *  *      *   . *   .**. ..*.     . *   **** .***** .**


h. vs. m.       ************************************************************
humanufd1l      LNITYPMLFKLTNKNSDRMTHCGVLEFVADEGICYLPHWMMQNLLLEEGGLVQVESVNLQ
mouseufd1l      LNITYPMLFKLTNKNSDRMTHCGVLEFVADEGICYLPHWMMQNLLLEEGGLVQVESVNLQ
yeastufd1p      LNIRYPMLFKLTANETGRVTHGGVLEFIAEEGRVYLPQWMMETLGIQPGSLLQISSTDVP
h/m/y           *** ********   . *.** *****.*.**  ***.***..* .. * *.*. *  .


h. vs. m.       *********** ********************.************************_
humanufd1l      VATYSKFQPQSADFLDITNPKAVLENALRNFACLTTGDVIAINYNEKIYELRVMETKPD-
mouseufd1l      VATYSKFQPQSPDFLDITNPKAVLENALRNFACMTTGDVIAINYNEKIYELRVMETKPD-
yeastufd1p      LGQFVKLEPQSVDFLDISDPKAVLENVLRNFSTLTVDDVIEISYNGKTFKIKILEVKPES
h/m/y           .  . * .*** *****. ******* ****...*  *** * ** * . ....* **.


h. vs. m.       --***** ***************** ____* ******_____________ ********
humanufd1l      --KAVSIHECDMNVDFDAPLGYKEPE----RQVQHEES-------------TEGEADHSG
mouseufd1l      --KAVSIIECDMNVDFDAPLGYKEPK----RPVQHEES-------------IEGEADHSG
yeastufd1p      SSKSICVIETDLVTDFAPPVGYVEPDYKALKAQQDKEKKNSFGKGQVLDPSVLGQGSMST
h/m/y             *.... *.*.  **  *.** **     .  *  *                  *.   *


h. vs. m.       ---****.*_____**********************************************
humanufd1l      ---YAGELG-----FRAFSGSGNRLDGKKKGVEPSPSPIKPGDIKRGIPNYEFKLGKITF
mouseufd1l      ---YAGEVG-----FRAFSGSGNRLDGKKKGVEPSPSPIKPGDIKRGIPNYEFKLGKITF
yeastufd1p      RIDYAGIANSSRNKLSKFVGQGQNISGKAPKAEPK-QDIKDMKITFDGEPAKLDLPEGQL
h/m/y             ***  .        * * *. . **    **    **   *          *


h. vs. m.       ******************___****************_***
humanufd1l      IRNSRPLVKKVEEDEAGG---RFVAFSGEGQSLRKKG-RKP------------------
mouseufd1l      IRNSRPLVKKVEEDEAGG---RFVAFSGEGQSLRKKG-RKP------------------
yeastufd1p      FFGFPMVLPKEDEESAAGSKSSEQNFQGQGISLRKSNKRKTKSDHDSSKSKAPKSPEVIE
h/m/y             .   .. * .*. * *       * *.* **** . **


h. vs. m.
humanufd1l      --
mouseufd1l      --
yeastufd1p      ID
h/m/y
```

Note: "*" indicates identity, "." indicates similarity, " " (space) indicates dissimilarity, and "-" indicates deletion/insertion. Similarity parameters above amino acid sequences indicate the comparison between human and mouse and those below indicate similarity among all three.

The 140 bp in 5' region, the first exon and 276 bp of the first intron of UFD1L gene are located in a CpG island. The CpG island is 488 bp in size and has a G/C content of 67.8 while the overall G/C content of contig2.2 is 48.9%. The 300 bp of 5' region (from 38527 to 38228) was analyzed by the SIGNALSCAN program against the TRANSFAC database and the results are shown in Figure 27. A potential TATA box is present in position -200 and two potential SP1 binding sites are within the -100 region. The position of the TATA box seems too far away from transcription start point indicating that this TATA box might be false. Other putative regulatory elements that were identified include the GR (glucocorticoid receptor), F2F (a transcriptional repressor in nonpituitary cells), LF-A1 (liver specific trans-acting factor), NF-E (erythroid specific nuclear factor), and Pit-1 (pituitary transcription factor-1).

Figure 27: 5' region of UFD1L gene



Note: The start point of transcription is defined as position 1. Outlined regions are putative regulatory element or transcription factor binding sites. See text for detail.

3.4.2.2 A Putative gene in the distal region of contig 2.2

Powerblast program found three regions of homology to the EST database in contig2.2 at the region from about 38700 to 41900 which were homologous to EST64671 (T34235). Actual alignment of this EST and genomic sequence showed that Powerblast failed to find a match between positions 76 to 132 of this EST. T34235 is the 5' end of a human cDNA whose origin is not clear. There is another EST, EST35508 (T31599), which is identical to T34235 but is 62 bases shorter at the 3' end. This EST is a human cDNA's 5' end which is similar to heparan sulfate proteoglycan from a human embryo cDNA library. There are eight other regions with EST homology found, and the Xgrail program predicted a cluster of exons at the region from about 38700 to 79800. Since the direction of EST hits and the cluster of exons are the same, when the results of Powerblast and Xgrail are taken together, a putative gene is expected in this region.

In the first EST hit, there is an ATG start codon at position 39039 followed by an open reading frame. Assuming this is the first exon of the putative gene, the first exon/intron boundary could be located by comparing the genomic DNA sequence and the conservative exon/intron splice site. The second exon is found by comparing the EST sequence (from positions 76 to 135 of the EST) and the genomic sequence. The third and fourth exons also were found by comparing the genomic sequence with the second and third EST hits and the consensus exon/intron splice site. Among the remaining exons, exons 16, 18, 19 and 20 also have homology to ESTs in the data base. The remainder of the exons were deduced by studying the Xgrail results and exon/intron boundary consensus sequence using Xgrail version 1.3 and both exon predicting models, 2 and 1a. Exons 5, 7-15, and 17 were predicted by exon model 2 and exon 6 was predicted by exon model 1a. The other exons were predicted by EST

homology. Exon 20 does not contain any coding region but has homology to the 3' end of an EST whose 5' end is contained within exon 19. Thus, the coding region of this putative gene ends at exon 19, and exon 20 contains the poly(A) signal sequence.

Throughout this dissertation research, the main criterion for determining exons is that the quality of Xgrail predicted exons was high and that the exon/intron boundary agreed well with consensus. If the boundary matched well but this boundary introduced a stop codon into the open reading frame by causing a shift in the translation frame, this boundary would not be used. The open reading frames predicted by this method were translated into amino acid sequence and then used as a query to search the protein database. Any high score (> 100) hit was aligned with the open reading frame using Clustal program and compared. By this comparison, any missed and extra exons predicted would be revealed and the context region of the genomic sequence would be examined to resolve the problem.

The predicted coding region is hypothetical and even arbitrary and may not reflect the true coding region of the gene because of following reasons: 1) the exon/intron boundaries were not always conserved, 2) In some instance, Xgrail did not predict all the exons, and in other, it predicted extra exon(s), 3) there also were more than one "good" exon/intron boundary near each other, 4) this gene predicting method cannot predict the 5' and 3' untranscribed regions without additional direct evidences for promoters and poly(A) signal regions, respectively. Further study is needed to define the true coding region of the gene. Similarity analysis (Powerblast) is the most successful method for accurately predicting coding regions, especially exons with UTRs. However full-length cDNA data sometimes are not available or incomplete. Xgrail predicts exons very well but fails to identify 5' and 3' UTRs. It is

possibly because UTRs lack splice acceptor (5' UTR) or splice donor (3' UTR) and do not start or end with the coding regions.

The alignment of the genomic sequence with the putative gene exons is provided in Figure 28. One can see that the similarity between T34235 and genomic sequence is perfect with only two "n"s in T34235 sequence. The identity between these two sequences is 99.4% (320/322) and the exon/intron boundaries of this putative gene match consensus as shown in Table 8.

Table 8: Comparison of the exon/intron boundary of putative gene and the consensus sequence

	exon		intron		exon	
1	...CCAGAGCCAG	gtgacgccca...	1	...tccccgatag	AGGGTCCTTC...	2
2	...GATCCTTCAG	gtgagttctg...	2	...ctctctccag	GCCTTGTTCC...	3
3	...TAAAGAACAG	gtattgaaga...	3	...ttttttcag	TTTCATTATT...	4
4	...TGTCGTCAAT	gtatacaacg...	4	...ttgaacttag	ATCAAATTAC...	5
5	...GTTAGAAGAG	gtgagtttgg...	5	...cttgctacag	CTTCGCAAGG...	6
6	...GATGTTCTTG	gtggtgttgg...	6	...cctgtatcag	GAGATAGTGG...	7
7	...AGGCCCGGAG	gtgagtctgt...	7	...tcaagtttag	CATTTTTCCA...	8
8	...TGGGACATCG	gtaagtatga...	8	...ttccttacag	TCAGCCATGG...	9
9	...TGTTGCCTTG	gtcagtgcca...	9	...cttcttccag	GTGGGCCATC...	10
10	...AGATCACTCA	gtaaggacac...	10	...ttttttccag	AATGAAATAC...	11
11	...TCTCCTTTGA	gtatgagtat...	11	...ttgactgcag	GCCCAGCCTC...	12
12	...CAGACATGGG	gtgagtgact...	12	...ctctttccag	TCTTCCCCTG...	13
13	...ATAAATTTGG	gtaaacacac...	13	...cccatgacag	GATGAAGGAC...	14
14	...GCCTCTCCAG	gtagcaggag...	14	...cctgttacag	GAGTAACCTG...	15
15	...TCTCATGGAG	gtcaggcttc...	15	...gcctccgcag	GGCACTCCAG...	16
16	...TGTGTGTTCG	gtgaggggcc...	16	...gactccatag	ACAAAGAACC...	17
17	...ACAGGAAGAA	gtgagcagct...	17	...gtgtcagcag	CTTTTTTGGG...	18
18	...GACCTCTCAG	gtgagagtct...	18	...attacttcag	TAATTGAGCT...	19
19	...CTGTCCTAGG	gtgagttaca...	19	...tgttttctag	AATTTGATTC...	20
consensus	(C/A)AG	gt(a/g)agt.....(t/c)nn(c/t)ag			G	

Note: Upper case letters are exons while lower case letters, introns.

Figure 28: Alignment of genomic sequence and exons of putative gene in contig 2.2

```
contig2.2  > 38725     ccaaagtaaagatgtccgcccagcgtcgcttttctgggta
                                    Sp1

contig2.2  > 38765     aaagcgagtcagaggttgaaggggggcacagtgtttgcgtt
                              GR          LF-A1

contig2.2  > 38805     cggcctcagaggtgacgcttctttgggggcgggctggagt
                             GR     CREB           Sp1

contig2.2  > 38845     tttctggccgtgaatggcagagcgctaatggccgcctcca
                                    GR

contig2.2  > 38885     atgtggccccaatcagaaagaaaggaaggctgggaactaag
                            SRF/CCAAT box

contig2.2  > 38925     aagctattggttggtgatccctggaccaatcggaggagcc
                          CCAAT box          SRF/CCAAT box

contig2.2  > 38965     gtgatttggcgggagtcttgaccgccgccgggctcttggt
                               Sp1

contig2.2  > 39005     acctcagcgcgagcgccaggcgtccggccgccgtggctat
T34235      > 1                      ........................
                                                          M

contig2.2  > 39045     gttcgtgtccgatttccgcaaagagttctacgaggtggtc
T34235      > 27       ......n.............................
                         F  V  S  D  F  R  K  E  F  Y  E  V  V

contig2.2  > 39085     cagagccag|gtgacgcccagtccgggacccccgccgaggc
T34235      > 67       .........
                        Q  S  Q

contig2.2  > 39125     cttgccggtggggaagggatgaggggagcgaccgtgggt

~~~~~~~~~~~~~~~intron 1, sequence 39145-39164 not shown~~~~~~~~~~~~~~~

contig2.2  > 39165     gaccgggaggcaggggagggccggggttcgggtcgccgtg

contig2.2  > 39205     ttcagccggtctgctcttccccgatag|agggtccttctct
                                                  R  V  L  L

contig2.2  > 39245     tcgtggcctcggacgtggatgctctgtgtgcgtacaagat
                       F  V  A  S  D  V  D  A  L  C  A  Y  K  I

contig2.2  > 39285     ccttcag|gtgagttctgcggaccctaggagggcggggccg
                         L  Q
```

```
contig2.2  > 39325     gcgcgcgaggtgagggtgctgcgtgggggcgcagggcgga

~~~~~~~~~~~~~~~intron 2, sequence 39365-39964 not shown~~~~~~~~~~~~~~~


contig2.2  > 39965     cttttcaaggtaaattgctaatgtcctaaataagataggc


contig2.2  > 40005     tttgaaaacactctctctccag|gccttgttccagtgtgac
T34235     > 133                          ......................
                                             A  L  F  Q  C  D

contig2.2  > 40045     cacgtgcaatatacgctggttccagtttctgggtggcaag
T34235     > 154       ....................................
                        H  V  Q  Y  T  L  V  P  V  S  G  W  Q


contig2.2  > 40085     aacttgaaactgcatttcttgagcataaagaacag|gtatt
T34235     > 194       ................n.................
                       E  L  E  T  A  F  L  E  H  K  E  Q

contig2.2  > 40125     gaagatgcgttttagaataacgtggctttttactcaattg

~~~~~~~~~~~~~~~intron 3, sequence 40165-41684 not shown~~~~~~~~~~~~~~~


contig2.2  > 41685     gaacaactcagaggcttagtgatgaaggaaaaggggcctc


contig2.2  > 41725     ctcgatatagctactttacaatatcttcctttttttcag|t
T34235     > 226                                            ....

contig2.2  > 41765     ttcattattttattctcataaactgtggagctaatgtaga
T34235     > 230       ....................................
                       F  H  Y  F  I  L  I  N  C  G  A  N  V  D

contig2.2  > 41805     cctattggatattcttcaacctgatgaagacactatattc
T34235     > 270       ....................................
                         L  L  D  I  L  Q  P  D  E  D  T  I  F

contig2.2  > 41845     tttgtgtgtgacacccataggccagtcaatgtcgtcaat|g
T34235     > 310       ............
                        F  V  C  D  T  H  R  P  V  N  V  V  N

contig2.2  > 41885     tatacaacgatacccaggtacttttttgtgctatgccctca

~~~~~~~~~~~~~~~intron 4, sequence 41925-42884 not shown~~~~~~~~~~~~~~~


contig2.2  > 42885     agttttccgtagaaagaggagaggcttaattccttttttgt


contig2.2  > 42925     tttgaacttag|atcaaattactcattaaacaagatgatga
                                  I  K  L  L  I  K  Q  D  D  D

contig2.2  > 42965     ccttgaagttcccgcctatgaagacatcttcagggatgaa
                         L  E  V  P  A  Y  E  D  I  F  R  D  E

contig2.2  > 43005     gaggaggatgaagagcattcaggaaatgacagtgatgggt
                        E  E  D  E  E  H  S  G  N  D  S  D  G

contig2.2  > 43045     cagagccttctgagaagcgcacacggttagaagag|gtgag
                       S  E  P  S  E  K  R  T  R  L  E  E
```

```
contig2.2  > 43085     tttgggtctctcacagctatcccagaggaacttgcactcc

----------intron 5, sequence 43125-52084 not shown----------

contig2.2  > 52085     tccttggcctctctgataccaccagacagagggcttgggg

contig2.2  > 52125     tgccttgctacag|cttcgcaaggatgggagcctaggctcc
                                    L  R  K  D  G  S  L  G  S

contig2.2  > 52165     cactgggcctttgctggggtggatatggatgttcttg|gtg
                        H  W  A  F  A  G  V  D  M  D  V  L

contig2.2  > 52205     gtgttggctagagtagatgggctactgcctaaaagctttc

----------intron 6, sequence 52245-53324 not shown----------

contig2.2  > 53325     aaccttcttgggctgtggggataaattcctggtgcatttg

contig2.2  > 53365     ctccacctttgttctctttgtccctgtatcag|gagatag
                                                      G  D  S

contig2.2  > 53405     tggagcaaaccatgcggaggaggcagcggcgagagtggga
                         G  A  N  H  A  E  E  A  A  A  R  V  G

contig2.2  > 53445     ggcccggag|gtgagtctgtgcttccagctgctcccagcac
                        G  P  E

contig2.2  > 53485     cagaagcccacttgcctgggggtctgtggtgccacccata

----------intron 7, sequence 53525-54024 not shown----------

contig2.2  > 54925     gaaaacaaatagtaccatgatttaatttgcatttcattga

contig2.2  > 54965     atactgtcaagtttag|catttttccaaaatatttggcttc
                                          H  F  S  K  I  F  G  F

contig2.2  > 55005     ctgtactgctacttcctgttgaagacctgcattttatgtc
                        L  Y  C  Y  F  L  L  K  T  C  I  L  C

contig2.2  > 55045     attacagaagagacatcctctttgactacgagcagtatga
                       H  Y  R  R  D  I  L  F  D  Y  E  Q  Y  E

contig2.2  > 55085     atatcatgggacatcg|gtaagtatgaataggtggaactca
                         Y  H  G  T  S

contig2.2  > 55125     ctataaagttctgactccaggggtcagtgtcctcaaatgt

----------intron 8, sequence 55145-56404 not shown----------

contig2.2  > 56405     gccccagcacatgcccctcccatgagccttagacttctct

contig2.2  > 56445     gcttccttacag|tcagccatggtgatgtttgagctggctt
                                      S  A  M  V  M  F  E  L  A

contig2.2  > 56485     ggatgctgtccaaggacctgaatgacatgctgtggtacgt
                       W  M  L  S  K  D  L  N  D  M  L  W  Y  V
```

```
contig2.2  > 56525     agcccctgcggcagctgtgtgggagtatttgttgccttg|g
                         A  P  A  A  A  V  W  E  Y  L  L  P  W

contig2.2  > 56565     tcagtgccacataagcagctctgtcctcccacagggaggg

~~~~~~~~~~~~~~~~intron 9, sequence 56605-58124 not shown~~~~~~~~~~~~~~~~

contig2.2  > 58125     agcccaggtgggctataggccggctccactgccttctctt

contig2.2  > 58165     cttccag|gtgggccatcgttggactaacagaccagtgggt
                                  W  A  I  V  G  L  T  D  Q  W  V

contig2.2  > 58205     gcaagacaagatcactca|gtaaggacacactcccttgcct
                         Q  D  K  I  L  Q

contig2.2  > 58245     tgcagggtcagccctgtgctgtggccaggtgcacagcctg

~~~~~~~~~~~~~~~~intron 10, sequence 58285-64364 not shown~~~~~~~~~~~~~~~~

contig2.2  > 64365     cctaggcttaggctgcctggtaagagctggagggaaccag

contig2.2  > 64405     ccgaggcttgtgctacttttttttccag|aatgaaatacgt
                                                    M  K  Y  V

contig2.2  > 64445     gactgatgttggtgtcctgcagcgccacgtttcccgccac
                         T  D  V  G  V  L  Q  R  H  V  S  R  H

contig2.2  > 64485     aaccaccggaacgaggatgaggagaacacactctccgtgg
                          N  H  R  N  E  D  E  E  N  T  L  S  V

contig2.2  > 64525     actgcacacggatctcctttga|gtatgagtatccttgtgg
                        D  C  T  R  I  S  F  E

contig2.2  > 64565     cccagcctgaggggcacaggcagcacccctgctcagcagg

~~~~~~~~~~~~~~~~intron 11, sequence 64605-66404 not shown~~~~~~~~~~~~~~~~

contig2.2  > 66405     agtgctcagggcggccagtgggtgcttgcatgctccttga

contig2.2  > 66445     ctgcag|gcccagcctccgcctggtgctctaccagcactgg
                                 P  S  L  R  L  V  L  Y  Q  H  W

contig2.2  > 66485     tccctccatgacagcctgtgcaacaccagctataccgcag
                         S  L  H  D  S  L  C  N  T  S  Y  T  A

contig2.2  > 66525     ccaggttcaagctgtggtctgtgcatggacagaagcggct
                        A  R  F  K  L  W  S  V  H  G  Q  K  R  L

contig2.2  > 66565     ccaggagttccttgcagacatggg|gtgagtgactgcctgg
                          Q  E  F  L  A  D  M  G

contig2.2  > 66605     gcctctgcagtgccagccctggccacccccaagggaaaag

~~~~~~~~~~~~~~~~intron 12, sequence 66645-66764 not shown~~~~~~~~~~~~~~~~

contig2.2  > 66765     tggtttgggtttctttccaccctgtcggtgtgtggtgacc
```

```
contig2.2  > 66805     tcactcatgtggcttgggcttgctctttccag|tcttcccc
                                                        L  P

contig2.2  > 66845     tgaagcaggtgaagcagaagttccaggccatggacatctc
                       L  K  Q  V  K  Q  K  F  Q  A  M  D  I  S

contig2.2  > 66885     cttgaaggagaatttgcgggaaatgattgaagagtctgca
                         L  K  E  N  L  R  E  M  I  E  E  S  A

contig2.2  > 66925     aataaatttgg|gtaaacacacatttttctggatttatctt
                        N  K  F  G

contig2.2  > 66965     cattacatccaggttcattaaagtgaagggttctacttta

~~~~~~~~~~~~~~~intron 13, sequence 67005-67524 not shown~~~~~~~~~~~~~~~

contig2.2  > 67525     cacatctgttggcctgcctggcagtgagagcttgcccaca

contig2.2  > 67565     atttgagggtgacagctgtgtttgctcccatgacag|gatg
                                                           M

contig2.2  > 67605     aaggacatgcgcgtgcagactttcagcattcattttgggt
                        K  D  M  R  V  Q  T  F  S  I  H  F  G

contig2.2  > 67645     tcaagcacaagtttctggccagcgacgtggtctttgccac
                       F  K  H  K  F  L  A  S  D  V  V  F  A  T

contig2.2  > 67685     catgtctttgatggagagccccgagaaggatggctcaggg
                         M  S  L  M  E  S  P  E  K  D  G  S  G

contig2.2  > 67725     acagatcacttcatccaggctctggacagcctctccag|gt
                        T  D  H  F  I  Q  A  L  D  S  L  S  R

contig2.2  > 67765     agcaggagggctgtgggtttgcctcatggggccaccactg

~~~~~~~~~~~~~~~intron 14, sequence 67805-73764 not shown~~~~~~~~~~~~~~~

contig2.2  > 73765     tttaccaagagttcctgacaagcaccaccaaagccatgtg

contig2.2  > 73805     tctgccctgttacag|gagtaacctggacaagctgtaccat
                                       S  N  L  D  K  L  Y  H

contig2.2  > 73845     ggcctggaactcgccaagaagcagctgcgagccacccagc
                        G  L  E  L  A  K  K  Q  L  R  A  T  Q

contig2.2  > 73885     agaccattgccagctgcctttgcaccaacctcgtcatctc
                       Q  T  I  A  S  C  L  C  T  N  L  V  I  S

contig2.2  > 73925     ccaggggcctttcctgtactgctctctcatggag|gtcagg
                         Q  G  P  F  L  Y  C  S  L  M  E

contig2.2  > 73965     cttcccacagagcacaggcgcctggtcacagcaccagtgg

~~~~~~~~~~~~~~~intron 15, ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

contig2.2  > 74005     cttcgtctgaccatctgtctcgcctccgcag|ggcactcca
                                                    G  T  P
```

```
contig2.2  > 74045     gatgtcatgctgttctctaggccggcatccctaagcctgc
                        D  V  M  L  F  S  R  P  A  S  L  S  L

contig2.2  > 74085     tcagcaaacacctgctcaagtcctttgtgtgttcg|gtgag
                       L  S  K  H  L  L  K  S  F  V  C  S

contig2.2  > 74125     gggccaggcgggtgctgtgggtacaggagggcaggtgctt

----------------intron 16, sequence 74165-75524 not shown----------------

contig2.2  > 75525     gttgggctctgggctctgagtgttgagctggggcccacca

contig2.2  > 75565     taccctgacggagggtgctctccgactccatag|acaaaga
                                                          T  K

contig2.2  > 75605     accggcgctgcaaactgctgcccctggtgatggctgcccc
                       N  R  R  C  K  L  L  P  L  V  M  A  A  P

contig2.2  > 75645     cctgagcatggagcatggcacagtgaccgtggtgggcatc
                         L  S  M  E  H  G  T  V  T  V  V  G  I

contig2.2  > 75685     cccccagagaccgacagctcggacaggaagaa|gtgagcag
                        P  P  E  T  D  S  S  D  R  K  N

contig2.2  > 75725     cttccactcgtcctgggactggagggtcgggggtgttggg

----------------intron 17, sequence 75765-75844 not shown----------------

contig2.2  > 75845     attgcaggccgcctggcccccgacatgtcttgtgtgtcag

contig2.2  > 75885     cag|ctttttgggagggcgtttgagaaggcagcggaaagc
                              F  F  G  R  A  F  E  K  A  A  E  S

contig2.2  > 75925     accagctcccggatgctgcacaaccattttgacctctcag|
                        T  S  S  R  M  L  H  N  H  F  D  L  S

contig2.2  > 75965     gtgagagtctcctgccactctgccacactttcccacctga

----------------intron 18, sequence 76005-77844 not shown----------------

contig2.2  > 77845     gggagttctgtgccctgtcttgtgttcccactcctcccttt

contig2.2  > 77885     ctcacggctgtttttctttcattacttcag|taattgagct
                                                    V  I  E  L

contig2.2  > 77925     gaaagctgaggatcggagcaagtttctggacgcacttatt
                         K  A  E  D  R  S  K  F  L  D  A  L  I

contig2.2  > 77965     tccctcctgtcctagg|gtgagttacaggggttctgcaggg
                        S  L  L  S ^^^

contig2.2  > 78005     gtggctgcagcagcccccctcagagcccgaccctgatgccc

----------------intron 19, sequence 78045-79484 not shown----------

contig2.2  > 79485     cagtttcttcaattcggtggcacatcaacatcgtttgaaa
```

```
contig2.2  > 79525     cttgtttttttcttgttttgttttctag|aatttgattcttc

contig2.2  > 79565     cagaatgaccttcttatttatgtaactggctttcatttag

contig2.2  > 79605     attgtaagttatggacatgatttgagatgtagaagccatt

contig2.2  > 79645     ttttatta|aataaa|atgcttattttaggctccgtccccat
                              Poly(A) signal

contig2.2  > 79685     tgtggctctggtctcgaggccagtgcttgagtgtgctctg

contig2.2  > 79725     agccttctgtgggtcttgggccccaggagccctcagtgcc

contig2.2  > 79765     tggcaggaaatcagcatggggattggcccaaaggacaaga

contig2.2  > 79805     tctccttactctgtcctccaggccctccgggtggcagtgg

contig2.2  > 79845     acagaggtcaggacccctcggcatccccgcctcctcatag

contig2.2  > 79885     cacctgcaccggccactgcctggtggtgccagccagcacc

contig2.2  > 79925     tggctctgtccacctaccacttcctctggtccacctggcc

contig2.2  > 79965     tcagccacatagcaagtccacccctctcagccatccctcca

contig2.2  > 80005     cctgtgggtctcaccatccagcactcttttccaggtcttt
```

Note: Not all sequences are shown. "." indicates the identity between genomic sequence and T34235. "|" indicates the splice site. "^^^" indicates stop codon. Poly(A) signal is bold. The potential regulatory sites of the 5' promoter region are outlined.

The 5' region of this putative gene, the first exon that contains the 5' UTR, the second exon, the first intron and part of second intron all are within a CpG island (position 38939-39379 according to Xgrail prediction). The G/C content of this CpG island is 67.6% and the overall G/C content of contig2.2 is 48.9%. The potential promoter region from 38725 to 39018 was analyzed by SIGNALSCAN and the TRANSFAC data base. The putative regulatory elements and transcription factor binding sites identified are three CCAAT boxes, one CREB (cyclic AMP regulatory element), three GR (glucocorticoid receptor), one LF-A1(liver specific trans-acting factor) and three Sp1 (see Figure 28). There was no TATA box present but three Sp1 transcription factor binding sites were present in different locations as required for interaction of transcription factor IID (TFIID) with TATA-less promoters.

The deduced amino acid sequence of this putative gene was used to search for similarities in the non-redundant GenBank protein database with the Blastp program and four high score hits were found. These were to the yeast protein similar to *C. elegans* hypothetical protein F34D10.2 (U53876), yeast ORF YLR103c (Z73275), *C. elegans* hypothetical protein F34D10.2 (Z34799), and *Ustilago maydis* protein Tsd2 (U50276). U53876 and Z73275 are actually the same protein. The alignment of putative gene amino acid sequence and yeast protein similar to *C. elegens* hypothetical protein F34D10.2 was done with the ClustalW program and the result is provided in Figure 29.

Figure 29: Alignment of yeast F34D10.2-like protein and amino acid sequence of putative gene in contig 2.2 by CLUSTAL W

```
CLUSTAL W(1.60) multiple sequence alignment

putgene_62      --MFVSDFRKEFYEVVQS-------QRVLLFVASDVDALCAYKILQALFQCDHVQYTLVP
yeastf34d1      MYYGISQFSEAYNKILRNSSSHSSCQLVIFVSCLNIDALCATKMLSLLFKKQLVQSQIVP
                    .* *   .  ...        * *.   .  .***** *.*  **.   **  .**

putgene_62      VSGWQELETAFLEHKEQFHYFILINCGANVDLLDILQPDEDTIFFVCDTHRPVNVVNIKL
yeastf34d1      IFGYSELRRHYSQLDDNINSLLLVGFGGVIDLEAFLEIDP--QEYVIDTDEKSGEQSFRR
                . *  **   . .  .. .  .*.. *  .**   *. *     .* **    .    .

putgene_62      LIKQDDDLEVPAYEDIFRDEEEDEEHSGNDSDGS-EPSEKRTRLEELRKDGSLGSHWAFA
yeastf34d1      DIYVLDAHRPWNLDNIFGSQIIQCFDDGTVDDTLGEQKEAYYKLLELDEES--GDDELSG
                 *   *       . **  .       *.  *   *  *   .* **  .   *

putgene_62      GVDMDVLGDSGANHAEEAAARVGGPEHFSKIFGFLYCYFLLKTCILCHYRRDILFDYEQY
yeastf34d1      DENDNNGGDDEATDADEVTDED--EEDEDETISNKRGNSSIGPNDLSKRKQ--RKK-QIH
                       **  *. *.* .      *                  .    *.. ..     .

putgene_62      EYHGTSSAMVMFELAWMLSKDLNDMLWYVAPAAAVWEYLLPWWAIVGLTDQWVQDKILQM
yeastf34d1      EYEG--VLEEYYSQGTTVVNSISAQIYSLLSAIGETNLSNLWLNILGTTS---LDIAYAQ
                ** *       .    .   .   .  *          *  *.* *     *

putgene_62      KYVTDVGVLQRHVSRHNHRNEDEENTLSVDCTRISFEPSLRLVLYQHWSLHDSLCNTSYT
yeastf34d1      VYNRLYPLLQDEVKRLTPSSRN--SVKTPDTLTLNIQPDYYLFLLRHSSLYDSFYYSNYV
                 *     .**  * * .          . *.  .  .*   * * .* ** **   . *


putgene_62      AARFKLWSVHGQKRLQEFLADMGLPLKQVKQKFQAMDISLKENLREMIEESANKFGMKDM
yeastf34d1      NAKLSLWNENGKKRLHKMFARMGIPLSTAQETWLYMDHSIKRELGIIFDKNLDRYGLQDI
                 *.  **  .*.***.   * **.**   ..    ** *.*  *  . .    ..*..*.

putgene_62      RVQTFSIHFGFKHKFLASDVVFATMSLMES---PEKD-----------GSG---------
yeastf34d1      IRDGFVRTLGYRGSISASEFVEALTALLEVGNSTDKDSVKINNDNNDDTDGEEEEDNSAQ
                    *   *..    **. * *  .*.*     .**             *

putgene_62      ---------TDHFIQALDSLSRSNLDKLYHGLELAKKQLRATQQTIASCLCTNLVISQGP
yeastf34d1      KLTNLRKRWVSNFWLSWDALDDRKVELLNRGIQLAQDLQRAIFNTGVAILEKKLIKHLRI
                           .*  . *.*    .. * .*..**.   **  .*  . *   *.

putgene_62      FLYCSLMEGTPDVMLFSRPASLSLLSKHLLKSFVCSTKNRRCKLLPLVMAAPLSMEHGTV
yeastf34d1      YRLCVLQDG-PDLDLYRNPLTLLRLGNWLIECCAESEDKQ---LLPMVLASIDENKD-TY
                .  * * .* **. *.  * .*  *   *. .   *   .   ***.*.*.       *

putgene_62      TVVGIPPETDS-----SDRK---NFFGRAFEKAAESTSSRMLHNHFDLSVIELKAEDRSK
yeastf34d1      LVAGLTPRYPRGLDTIHTKKPILNNFSMAFQQITAETDAKVRIDNFESSIIEIRREDLSP
                 * *. *           .*   * *  **.. .  * ...   .*. *.**.. ** *

putgene_62      FLDALISLLS-
yeastf34d1      FLEKLTLSGLL
                **. *
```

Note: "*" indicates identity, "." indicates similarity, " " (space) indicates dissimilarity, and "-" indicates deletion/insertion.

The similarity between human putative protein and yeast protein is only 40.2%. There is a large gap (29 amino acids) in the middle of the human putative protein compared with yeast F34D10.2-like protein. This gap is within exon 14 of the putative gene, and thus within a predicted open reading frame. Because of the low similarity of these two proteins, it is possible that these two proteins may carry out slightly different functions but belong to the same gene family.

The deduced amino acid sequence of this putative gene then was analyzed by the Online Protein Secondary Structure Prediction Program which is available at the web site of the Human Genome Center, Baylor College of Medicine. Because the 5' end of this putative gene (protein) is identical to EST35508 (T31599), which is similar to heparan sulfate proteoglycan, a putative transmembrane protein, searching for possible transmembrane helices on this putative protein was reasonable and the result of this search was as expected. Four inside to outside helices and three outside to inside helices were found by this program. As shown in Table 9 and the graphic output of this prediction (Figure 30), four significant helices easily can be seen.

Table 9: Possible transmembrane helices in putative gene

```
The sequence positions in brackets denominate the core region.
Only scores above 500 are considered significant.

      Inside to outside helices :   5 found
           from         to     score center
       19 (  19)  45 (  37)     520     29
      201 ( 201) 217 ( 217)     723    209
      252 ( 257) 278 ( 274)     964    265
      479 ( 487) 505 ( 503)    1096    495
      540 ( 545) 561 ( 561)     147    553

      Outside to inside helices :   4 found
           from         to     score center
      201 ( 201) 217 ( 217)    1142    209
      255 ( 258) 276 ( 274)    1085    266
      479 ( 484) 501 ( 501)     698    493
      491 ( 494) 515 ( 512)      77    504
```

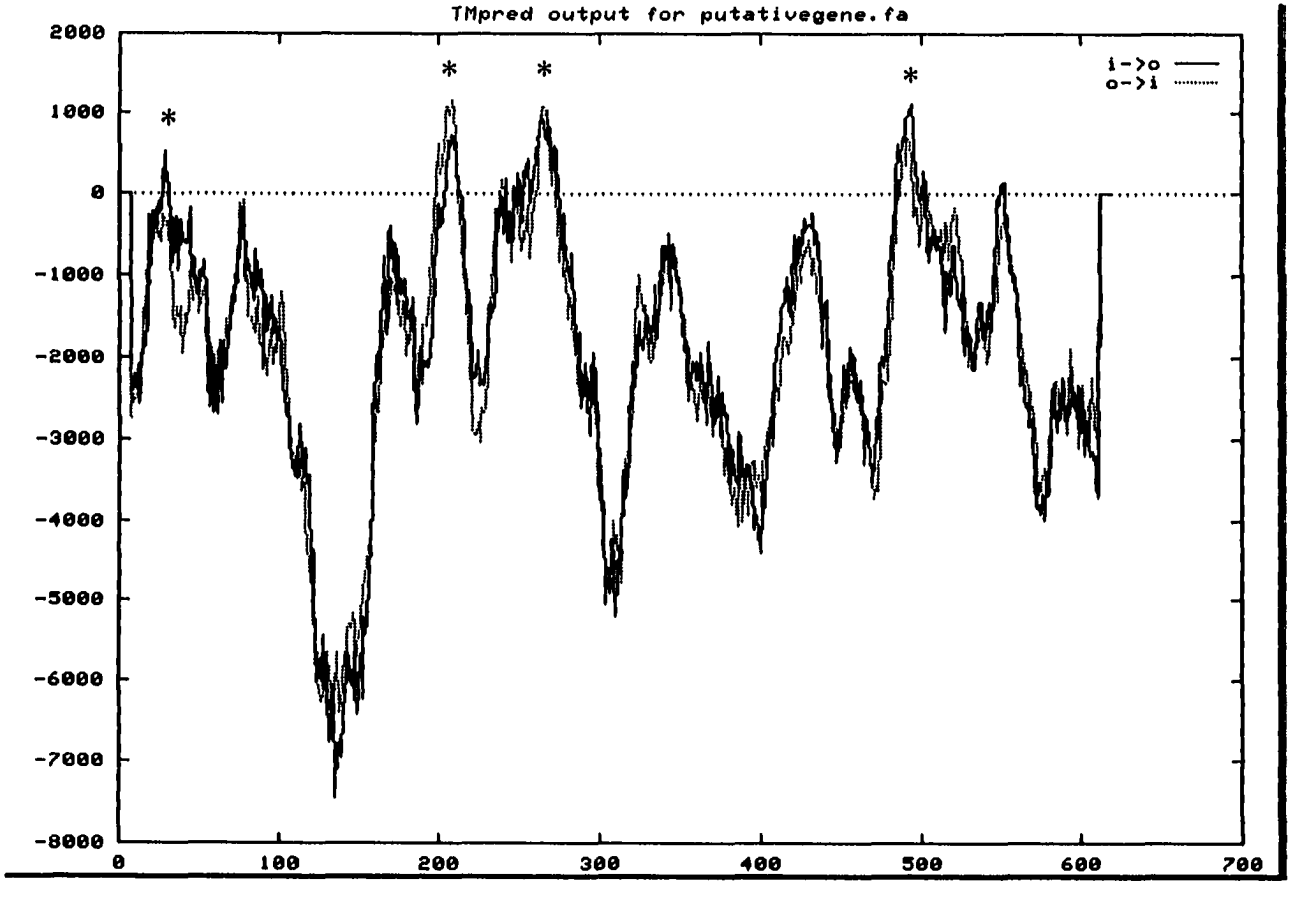


Figure 30: Graphic output of transmembrane domain prediction of putative gene. X-axis: amino acid number, Y-axis: score. "*" represents the peaks of helices.

It also can be seen from Table 9 that several regions were predicted as both inside-to-outside and outside-to-inside helices. The scores for each helix were analyzed and the preference of each helix orientation was calculated and is given in Table 10:

Table 10: Preferences of the orientation of helices in putative gene

```
Helices shown in brackets are considered insignificant.
A "+"-symbol indicates a preference of this orientation.
A "++"-symbol indicates a strong preference of this orientation.


          inside->outside | outside->inside
 19-  45 (27)  520 ++    |
201- 217 (17)  723        |   201- 217 (17) 1142 ++
                          |(  234- 250 (17)  169 ++ )
252- 278 (27)  964        |   255- 276 (22) 1085  +
479- 505 (27) 1096 ++     |   479- 501 (23)  698
                          |(  491- 515 (25)   77 ++ )
(540- 561 (22)  147 ++ ) |
```

According to the above table and considering only significant transmembrane segments (with scores higher than 500), a possible model for transmembrane topology is suggested as in Table 11. This suggestion is purely speculative, because it is based on the assumption that all transmembrane helices have been found.

Table 11: Suggested model for transmembrane topology of putative gene

```
STRONGLY preferred model: N-terminus inside
4 strong transmembrane helices, total score :       3324
# from   to length score orientation
1   19   45 (27)     520 i-o
2  201  217 (17)    1142 o-i
3  252  278 (27)     964 i-o
4  479  501 (23)     698 o-i
```

The 5' end of this putative gene is similar to that of heparan sulfate proteoglycan. It also probably is expressed in the embryogenesis stage. [271] The function of several membrane-inserted heparan sulfate proteoglycans is to mediate the adhesion of cells in the embryogenesis stage, [272-276] and recalling that DGS is an early developmental field disorder involving improper neural crest migration and

attachment to the third and fourth pouch in embryogenesis, it suggests that this putative heparan sulfate proteoglycan-like gene play a role in the etiology of DGS.

3.4.2.3 Repetitive elements in contig 2.2

Contig2.2 contains the interspersed repetitive elements Alu, L1, Mer, Mir, Mlt, Mst and Thelb as found by the Powerblast program. This contig contains 60 Alu elements which include full, half and quarter-length Alu fragments with a similarity higher than 70% and 12 L1 elements, all of which were less than 550 bp. Powerblast also identified 3 Mers, 2 Mirs, 2 Mlts, 2 Msts and 2 Thelbs. The total repetitive element content was 22.6% (21057/93386) and compared with contig1 and contig2.1, contig2.2 contains less total repetitive element sequence since contig1 had 31.4% total repeat in which 19.5% were Alu and 11.9% were non-Alu and contig2.1 had totally 32.0% in which 17.9% were Alu and 14.1 were non-Alu. The content of Alu elements in all three contigs is comparative while the non-Alu contents differ significantly (11.9%, 14.1% and 6.2% respectively). A summary of all repetitive elements in contig2.2 is listed in Table 12.

Table 12: Interspersed repetitive elements in contig 2.2

Family	number	bases	% region
Alu	**60**	**15276**	**16.4**
L1	12	3288	3.5
Mer/Mir/Mlt/Mst	9	1982	2.1
Thelb	2	511	0.6
Total non-Alu	**23**	**5781**	**6.2**

3.5 Analysis of contig 3

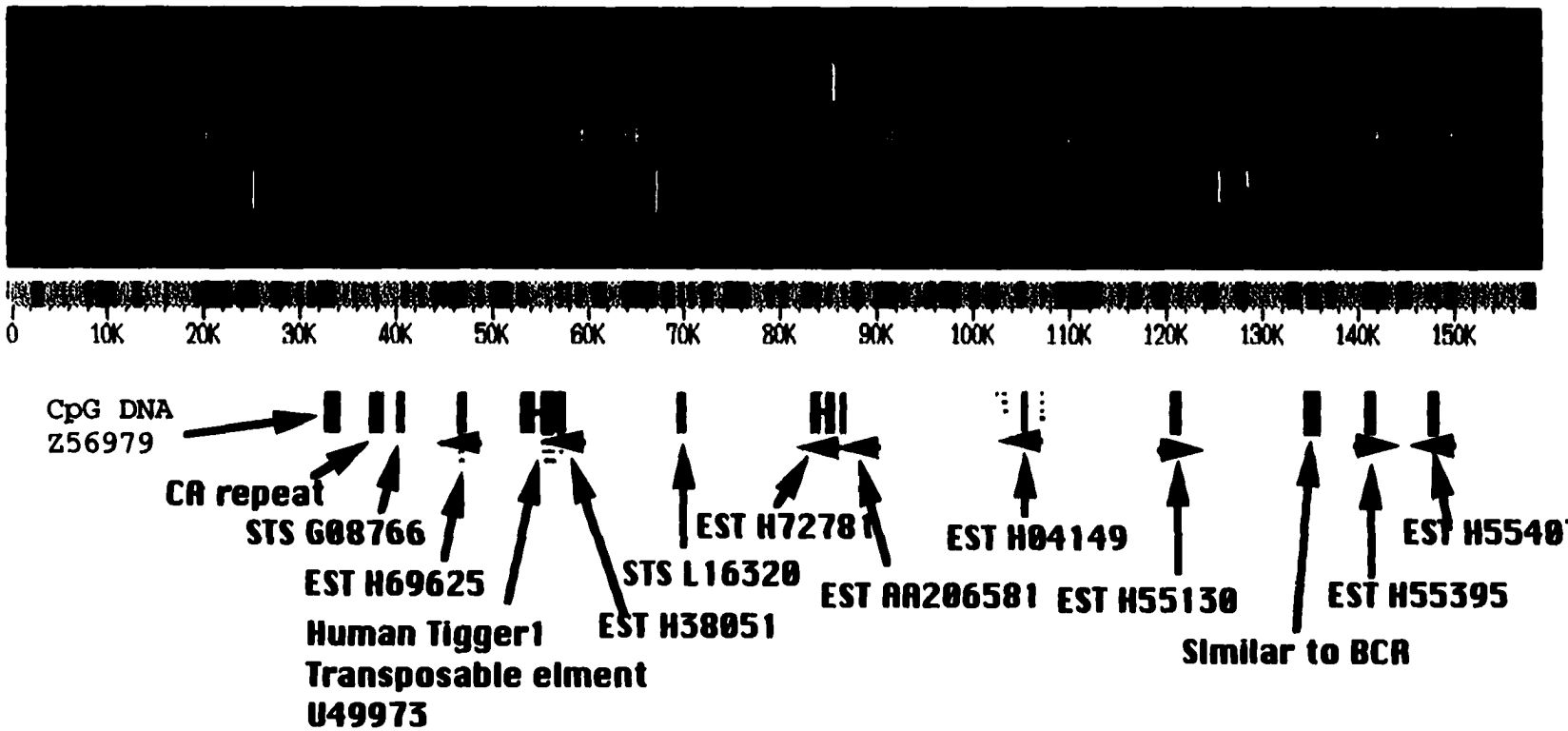
Contig3 contains 157,885 bp and the Xgrail and Powerblast results for this contig are given in Figure 31. No known genes were revealed in this region by a Powerblast search vs. the non-redundant GenBank database, but three regions with EST homology were found. These three ESTs correspond to no open reading frame predicted by Xgrail but they correspond to a region similar to human Tigger1 transposon element (U49973) at positions 53280- 57520 and at positions 134577-135352 to the human breakpoint cluster region(BCR) gene.

3.5.1 Major similarities found in database by Powerblast in contig3

EST H55130 begins at position 121251. The entire EST (127 bp) matches perfectly with contig3. This EST was prepared by an expression-independent method and should correspond to an open reading frame. [277, 278] However, Xgrail predicted an excellent exon in this position but in the opposite orientation and thus this EST most likely would be expressed in the opposite orientation as well, but this was the only Xgrail predicted exon within 35 kb of either side of this region.

A second region of EST homology was observed to begin at position 141140. The similarity between this EST, EST H55395, and contig3 is 100% (with a one base deletion in EST). The origin of this EST is the same as EST H55130 but Xgrail did not predict any exon in this region.

Figure 31: Result of Powerblast and Xgrail analysis of contig 3



Note: Xgrail result is on the top. The gray line is DNA sequence. The peaks are predicted exons. The small color bars above or below the peaks indicate the quality of exons, Green for excellent, blue for good and red for marginal. Peak above the gray line means the exon is on forward strand while peak under the line, reverse. The intensity of the gray line means the GC contents. The lighter the region is, the higher the GC contents. The pink boxes on the gray line are CpG islands.
Powerblast result is on the bottom. The black-gray line is DNA sequence. The gray regions are repeat sequence. The small colored bars are matches found in GenBank. The blue bars are matches with very high identity. The pink bars are matches with relative low similarity.

Another region of EST homology to EST H55407 begins at position 147588 that matches perfectly with contig3. The direction of this EST is opposite to two previously discussed ESTs and the origin of this EST is the same as those just mentioned. Even though this EST corresponds to no Xgrail predicted exon, there is a cluster of Xgrail predicted exons nearby. However, about half of those exons contain internal stop codons. Furthermore, several of the exons without stop codons are located in repetitive regions. It still is possible that there is a putative gene around this region but analysis of the genomic sequence alone fails to provide sufficient evidence to predict it.

Xgrail also predicts many open reading frames throughout this contig in addition to those discussed above but most of them are located either in repetitive regions or contain internal stop codons and no significant EST homologies were observed.

However, beginning at position 33476, there is a region with a perfect match to GenBank accession number Z56979. Z56979 was purified from genomic DNA by using a methylated DNA binding column. [279] Although this method does not screen the DNA fragments directly by the density of CpG dinucleotides,the restriction enzyme MseI (recognition site TTAA) was used to cleave out the assumed non-CpG region. Thus, this DNA fragment represents a putative CpG island containing a short stretch of DNA with a high density of non-methylated CpG dinucleotides, predominantly associated with the 5' region of a human gene. However, Xgrail failed to consider this region a CpG island. Because of the shortcoming of the method of purification of Z56979 as a CpG island, this CpG island could be false.

Starting at position 53280, there is a region similar to human Tigger1 transposon element (U49973) which also closely resembles *Drosophila melanogaster* DNA transposon pogo. [280] A DNA transposon element is a fragment of DNA which may represent a mobile genomic element that can be excised out of and reintegrated into the genome. It is different from other interspersed repetitive elements because its reiteration does not involve retrotransposition, which is a transposition mediated by RNA. A DNA transposon usually has terminal inverted repeats whose sizes vary and may contain internal inverted repeats. The DNA fragment carried by a transposon usually has coding capacity and may code for proteins for a transposase. However, the inverted repeats and the coding region may be altered during evolution and no longer be conserved and expressed. [81]

The human Tigger1-like transposon element found in this region is 4240 bp long and the terminal inverted repeats of this transposon are imperfect and short. A dotplot comparison result between human Tigger1 and this region is provided in Figure 32. The comparison showed that the Tigger1-like transposon contains all the sequence elements the human Tigger1 has, but instead of being contiguous, the sequence similar to human Tigger1 transposon in this region is broken into five fragments. The sequences between fragments are unique to Tigger1-like transposon, and this transposon has multiple internal inverted repeats. Almost all of these internal inverted repeats are within regions unique to this transposon, i.e. regions not found in human Tigger1. The positions of major internal inverted repeats are shown in Figure 34. Human Tigger1 transposon has two open reading frames and there is no gap between them. The first one is a putative transposase similar to the pogo element and the function of the second one is unknown. [281, 282] However, in this Tigger1-like transposon, the coding regions have been altered and contain multiple stop codons in all reading frames. The translation of the coding region is provided in Figure 33.

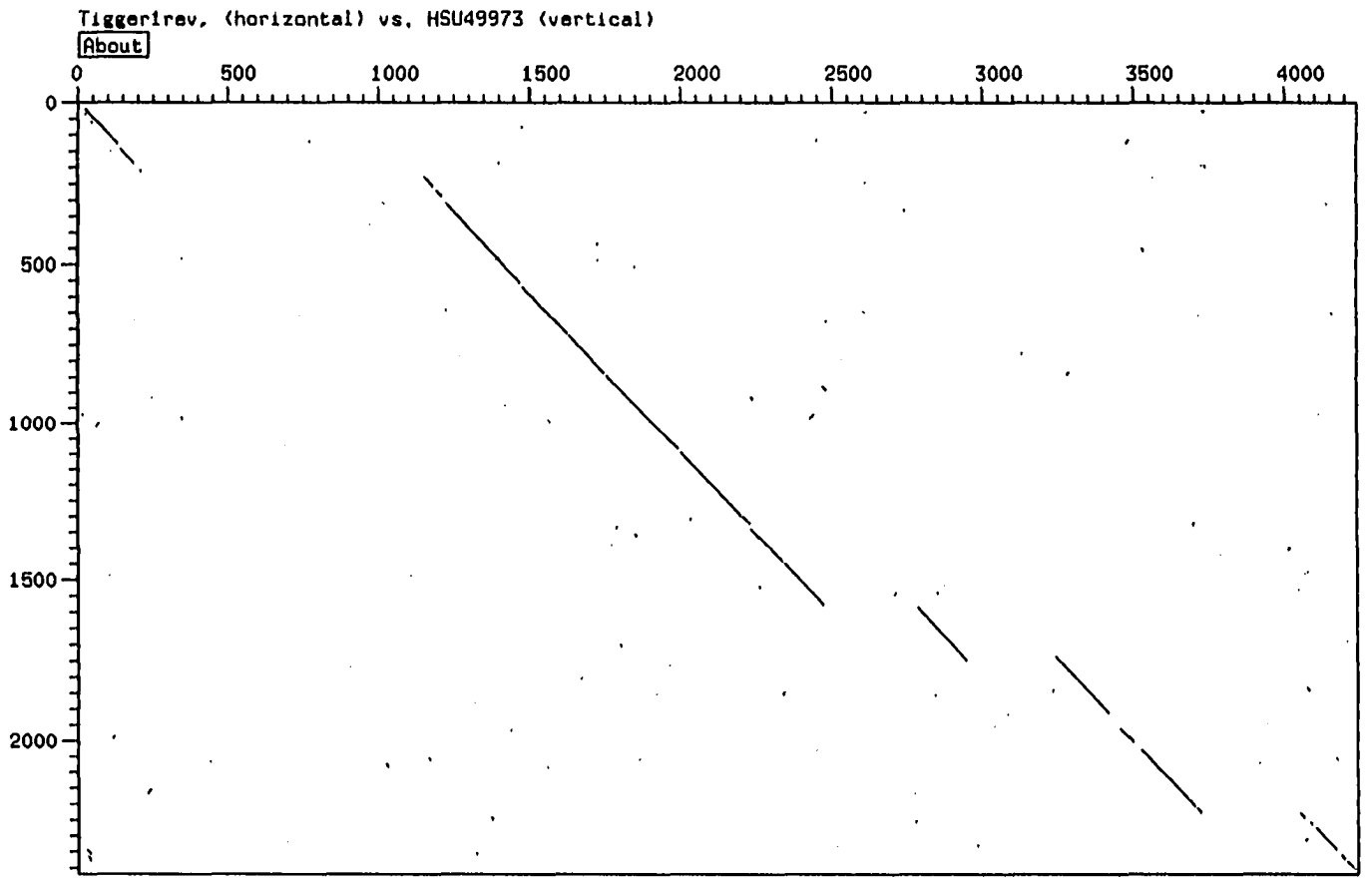


Figure 32: Dot plot of human Tiggerl vs. the genomic region sequences. Tigger 1-like sequence in chromosome 22 is horizontal and Tigger 1 sequence is vertical.

Figure 33: Translation of human Tigger 1-like transposon coding region

```
(Linear) MAP of: Tiggerlrev.Gcg  check: 2950  from: 1  to: 4240

   1 caggcacacctgtattttattgtagtttgctttattgcattttatagatattgtgttttc

  61 tacaaattgaaggcttgtggcaagcctgtgtctagcaagtctcttggtgccatttttcca

 121 acagtatgtgacctctggctctctgtgtcacattggtaattctcataatatgtcatactt

 181 tttccttattaggtatccatcatggtgatcggtgactgtcgccggtagagggtcgtgact

 241 gcaagttgtccaggttcttggtgttttgaacaaagaattggacaaaacacccagcaaagc

 301 aaagaaagaatgaagcaacaaaagaatgaaagcaggtgtttattgaaaatgaaagaacac

 361 gccacagtgtgggagcgggcccaagcagtggctcagggcccagatacagaatcttcttgg

 421 gttcaaataccccaagaagtttcccattggccacttcatgctcatctcatgtaaatgaa

 481 gtgatagcccacaatcattctgattggttgcagaaagcagccaaccagaggctgaagtaa

 541 tgttacaaaggtcacactcctgtgcaaacatctgattggttgcaaaaagcaatcagaggc

 601 taggatgaagttacagagttatacttctatgcaaacaaagactctgcccacaatcagtct

 661 gattgattgtggacagcaaccattccaaggccagagtgaagttaccaagttgcaaacaaa

 721 gactccaccggcaatcagtcagatttgttgaagacagccaatttcccatctgccagtgca

 781 gaaaaggtgaggatttaggttccctgcttccagaccctattctcctgcctcatgagtaat

 841 gatttttTTTTTTTTTTTGAGACAGGGTCTCACTCTGTCACCCAGGCTGGAGTGCAGTA

 901 CTGCAGTCATGGTTCACTGCAGCCTCAATCTCCCAGGCTCAGGCGATCCTCCCACCTCAG

 961 CCTCCCGAGTAGCTGGGACCACAGGCACACGTGTGCCACCATGCCTGGATAACTTTTGTA

1021 TTTTTTGTAGAAACTGGGTTTCACCATGTTTCCCAGGCTGGTCTCAAACTCCTGGGCTCA

1081 AGCAATCTACGTGCTTCAGCCTCCCAAAGTGTTGAGATTACAGGTGTGAGCCACTGCACC

1141 cagcctgtgatctttgatgttactattgtcattgttttgggcccggaaccatgcccgtat

1201 aagacagtgaacttctgataatatgtgtgttctgactgctccactgactgtccatgcccc

1261 atctctctctctctccttaggcctccctattccctgagacacaacaatattgaaattagg

1321 ccaattaataaccctacaatggcctctaagtgttcaagtgaaaggaagggttgcacattt

a                    M  A  S  K  C  S  S  E  R  K  G  C  T  F  -
b                     W  P  L  S  V  Q  V  K  G  R  V  A  H  F  -
c                      G  L  *  V  F  K  *  K  E  G  L  H  I  S -

1381 ctcactttaagtcaaaagttagaaatgattaagcttagtgaggaaggctgtcaaaagttg

a       L  T  L  S  Q  K  L  E  M  I  K  L  S  E  E  G  C  Q  K  L  -
b        S  L  *  V  K  S  *  K  *  L  S  L  V  R  K  A  V  K  S  *  -
c         H  F  K  S  K  V  R  N  D  *  A  *  *  G  R  L  S  K  V  E -
```

```
    1441 agacaggctgaaagctagacctcttgcaccaaataatcaaattgtgaatgcaaaggtaaa

a       R  Q  A  E  S  *  T  S  C  T  K  *  S  N  C  E  C  K  G  K  -
b        D  R  L  K  A  R  P  L  A  P  N  N  Q  I  V  N  A  K  V  K -
c         T  G  *  K  L  D  L  L  H  Q  I  I  K  L  *  M  Q  R  *  S -

    1501 gttcctgaaggagattaaaagtgctcctccagtgaagacacaaatgataagaaagtgaaa

a       V  P  E  G  D  *  K  C  S  S  S  E  D  T  N  D  K  K  V  K  -
b        F  L  K  E  I  K  S  A  P  P  V  K  T  Q  M  I  R  K  *  N -
c         S  *  R  R  L  K  V  L  L  Q  *  R  H  K  *  *  E  S  E  T -

    1561 cagccttattgctgatatggagaatgtggtagtggtctggatagaagatcaaaccagcca

a       Q  P  Y  C  *  Y  G  E  C  G  S  G  L  D  R  R  S  N  Q  P  -
b        S  L  I  A  D  M  E  N  V  V  V  V  W  I  E  D  Q  T  S  H -
c         A  L  L  L  I  W  R  M  W  *  W  S  G  *  K  I  K  P  A  M -

    1621 tgcatttccttaagtcaaagcctaattcagagcaagatcctacttctcttcaattctatg

a       C  I  S  L  S  Q  S  L  I  Q  S  K  I  L  L  L  F  N  S  M  -
b        A  F  P  *  V  K  A  *  F  R  A  R  S  Y  F  S  S  I  L  *  -
c         H  F  L  K  S  K  P  N  S  E  Q  D  P  T  S  L  Q  F  Y  E -

    1681 aaggctgagagaggtgaggaagctgcagaagaaaacttagaagctagtagagatttgtgc

a       K  A  E  R  G  E  E  A  A  E  E  N  L  E  A  S  R  D  L  C  -
b        R  L  R  E  V  R  K  L  Q  K  K  T  *  K  L  V  E  I  C  A -
c         G  *  E  R  *  G  S  C  R  R  K  L  R  S  *  *  R  F  V  H -

    1741 atgagggtaaggaaggaagccatctacataacataaaagtgcaaggtgaaggagcaagtg

a       M  R  V  R  K  E  A  I  Y  I  T  *  K  C  K  V  K  E  Q  V  -
b        *  G  *  G  R  K  P  S  T  *  H  K  S  A  R  *  R  S  K  C -
c         E  G  K  E  G  S  H  L  H  N  I  K  V  Q  G  E  G  A  S  A -

    1801 ctgatgtagaagctgccgcaagttatccagaagatctggctaagatcactgatgaaggtg

a       L  M  *  K  L  P  Q  V  I  Q  K  I  W  L  R  S  L  M  K  V  -
b        *  C  R  S  C  R  K  L  S  R  R  S  G  *  D  H  *  *  R  W -
c         D  V  E  A  A  A  S  Y  P  E  D  L  A  K  I  T  D  E  G  G -

    1861 gccacactaaacaacagatcgtcaatgtaggtgaaacagccttctatcagaagaggatgc

a       A  T  L  N  N  R  S  S  M  *  V  K  Q  P  S  I  R  R  G  C  -
b        P  H  *  T  T  D  R  Q  C  R  *  N  S  L  L  S  E  E  D  A -
c         H  T  K  Q  Q  I  V  N  V  G  E  T  A  F  Y  Q  K  R  M  P -

    1921 catctaggactttcctagctagagaggagaagtcaatgactggcttcaaagtttcaaagg

a       H  L  G  L  S  *  L  E  R  R  S  Q  *  L  A  S  K  F  Q  R  -
b        I  *  D  F  P  S  *  R  G  E  V  N  D  W  L  Q  S  F  K  G -
c         S  R  T  F  L  A  R  E  E  K  S  M  T  G  F  K  V  S  K  E -

    1981 aaaggctggctctgttaggggctaatgcagctggtgactttatgttgaagccaatgttta

a       K  G  W  L  C  *  G  L  M  Q  L  V  T  L  C  *  S  Q  C  L  -
b        K  A  G  S  V  R  G  *  C  S  W  *  L  Y  V  E  A  N  V  Y -
c         R  L  A  L  L  G  A  N  A  A  G  D  F  M  L  K  P  M  F  I -
```

```
    2041 ttcgccattccagaaatcctagggccttaagaatcatgccaaattttctctgcctgcgct

a        F  A  I  P  E  I  L  G  P  *  E  S  C  Q  I  F  S  A  C  A  -
b         S  P  F  Q  K  S  *  G  L  K  N  H  A  K  F  S  L  P  A  L  -
c          R  H  S  R  N  P  R  A  L  R  I  M  P  N  F  L  C  L  R  S -

    2101 ctagaaatggaacaagaaaacctagatgacagcacatttgtttacagcagggcttactga

a        L  E  M  E  Q  E  N  L  D  D  S  T  F  V  Y  S  R  A  Y  *  -
b         *  K  W  N  K  K  T  *  M  T  A  H  L  F  T  A  G  L  T  E  -
c          R  N  G  T  R  K  P  R  *  Q  H  I  C  L  Q  Q  G  L  L  N -

    2161 atattttaagcccactgttgagaaatgctgctcagaaaaaaaaaaaagaatcctttcaaaa

a        I  F  *  A  H  C  *  E  M  L  L  R  K  K  K  R  I  L  S  K  -
b         Y  F  K  P  T  V  E  K  C  C  S  E  K  K  K  E  S  F  Q  N  -
c          I  L  S  P  L  L  R  N  A  A  Q  K  K  K  K  N  P  F  K  I -

    2221 tattactgcacctagtcatccaggagctgtgatggagatggacaaggagatgaaagctgt

a        Y  Y  C  T  *  S  S  R  S  C  D  G  D  G  Q  G  D  E  S  C  -
b         I  T  A  P  S  H  P  G  A  V  M  E  M  D  K  E  M  K  A  V  -
c          L  L  H  L  V  I  Q  E  L  *  W  R  W  T  R  R  *  K  L  C -

    2281 gtgcatgcctgctaacacaacatccattccgcagcccatggactgaagagtaatttcaat

a        V  H  A  C  *  H  N  I  H  S  A  A  H  G  L  K  S  N  F  N  -
b         C  M  P  A  N  T  T  S  I  P  Q  P  M  D  *  R  V  I  S  I  -
c          A  C  L  L  T  Q  H  P  F  R  S  P  W  T  E  E  *  F  Q  F -

    2341 ttcaagtcttattattcaagaaatacatttcataaggctatagttgccatagatggtgat

a        F  K  S  Y  Y  S  R  N  T  F  H  K  A  I  V  A  I  D  G  D  -
b         S  S  L  I  I  Q  E  I  H  F  I  R  L  *  L  P  *  M  V  I  -
c          Q  V  L  L  F  K  K  Y  I  S  *  G  Y  S  C  H  R  W  *  F -

    2401 tcctctgatggattcgggcgaagtacattgaaaatcttttggaaaagattcaccagtcta

a        S  S  D  G  F  G  R  S  T  L  K  I  F  W  K  R  F  T  S  L  -
b         P  L  M  D  S  G  E  V  H  *  K  S  F  G  K  D  S  P  V  *  -
c          L  *  W  I  R  A  K  Y  I  E  N  L  L  E  K  I  H  Q  S  R -

    2461 gatgccatgaaaagcagcaagttgtggtgacttacacctgttaattccaacaatttggga

a        D  A  M  K  S  S  K  L  W  *                                 -
b         M  P  *  K  A  A  S  C  G  D                                 -
c          C  H  E  K  Q  Q  V  V  V  T                                 -

    2521 ggctgaggcagaagaatcacttgagtccaggagtttgagaccagccagggcaacatatgg

    2581 agactctgtttataaaaaaaaaaaacaaaacaaacaaaaaaaaaaaactagccaggtttgg

    2641 tggtgcatgcctatagttccagctaattgggaggctgaggcagggctcacctgagcccag

    2701 aaggttgttgatgtcctcgcgatcacatcactgcactccagactgggcaacgagtgagac

    2761 cctgtctcaaaaaaaagagaacatttttgcttcatgggaagaggtcaaaatatcaacatt

a                          H  F  C  F  M  G  R  G  Q  N  I  N  I  -
b                           I  F  A  S  W  E  E  V  K  I  S  T  L  -
c                            F  L  L  H  G  K  R  S  K  Y  Q  H  * -
```

```
    2821 aataggagtttggaagaagtagattccaaccctcatggatgactttgagggttcaggact

a        N  R  S  L  E  E  V  D  S  N  P  H  G  *  L  *  G  F  R  T  -
b         I  G  V  W  K  K  *  I  P  T  L  M  D  D  F  E  G  S  G  L  -
c          *  E  F  G  R  S  R  F  Q  P  S  W  M  T  L  R  V  Q  D  F-

    2881 tcagtggagaaagtcactgcagatgtgggggaaacagcaggagaaccagatttagaagtg

a        S  V  E  K  V  T  A  D  V  G  E  T  A  G  E  P  D  L  E  V  -
b         Q  W  R  K  S  L  Q  M  W  G  K  Q  Q  E  N  Q  I  *  K  W  -
c          S  G  E  S  H  C  R  C  G  G  N  S  R  R  T  R  F  R  S  G-

    2941 gagcctagaccaGGTGCAGTGGCTCACACTGGTAATCCCAGCACTTTGGGAGTCCGAGGT

    3001 GGGTGGATCACGAGGTCAGGAGTTCGAGACCATCCTGGCTAACACGGTGAAACCCCATCT

    3061 CTACTAAAAATACAAAAAATTAGCTGGGCGTGATGGCGGGCGCCTGTAGTCCCAGCTACT

    3121 CAGGAGGCTGAGGGAGGAGGATGGCGTGAACCCGGGAGGCGGAGCTTGCAGTGAGCCAAG

    3181 ATTGCACCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCGTCTCAAAAAAAAAAAA

    3241 gaagtggagcccgaagaaggggctgaattgctgcaatctcaggatccaacttgaacacat

a        E  V  E  P  E  E  G  A  E  L  L  Q  S  Q  D  P  T  *  T  H  -
b         K  W  S  P  K  K  G  L  N  C  C  N  L  R  I  Q  L  E  H  M  -
c          S  G  A  R  R  R  G  *  I  A  A  I  S  G  S  N  L  N  T  *-

    3301 gaggagttgcttcttagggatgaacaaagaaagtggttccttgagacggaacgtcctcct

a        E  E  L  L  L  R  D  E  Q  R  K  W  F  L  E  T  E  R  P  P  -
b         R  S  C  F  L  G  M  N  K  E  S  G  S  L  R  R  N  V  L  L  -
c          G  V  A  S  *  G  *  T  K  K  V  V  P  *  D  G  T  S  S  W-

    3361 ggggcagatgctgtgaacatggttgcaatgacaacaaagcgttcagaatatttcataacc

a        G  A  D  A  V  N  M  V  A  M  T  T  K  R  S  E  Y  F  I  T  -
b         G  Q  M  L  *  T  W  L  Q  *  Q  Q  S  V  Q  N  I  S  *  P  -
c          G  R  C  C  E  H  G  C  N  D  N  K  A  F  R  I  F  H  N  R-

    3421 gaaaaaacacagcaggcttgagaggatttctaattttgaaagaccttctactgtgggtaa

a        E  K  T  Q  Q  A  *  E  D  F  *  F  *  K  T  F  Y  C  G  *  -
b         K  K  H  S  R  L  E  R  I  S  N  F  E  R  P  S  T  V  G  K  -
c          K  N  T  A  G  L  R  G  F  L  I  L  K  D  L  L  L  W  V  K-

    3481 aatggtatcaagcagcctccagtgccgggggccatctcctgtgaaaggcagagtcaatca

a        N  G  I  K  Q  P  P  V  P  G  A  I  S  C  E  R  Q  S  Q  S  -
b         M  V  S  S  S  L  Q  C  R  G  P  S  P  V  K  G  R  V  N  Q  -
c          W  Y  Q  A  A  S  S  A  G  G  H  L  L  *  K  A  E  S  I  N-

    3541 atgtggtaaacttcactgttgtctgatttgaatacattgtcacagcccctctacccttca

a        M  W  *  T  S  L  L  S  D  L  N  T  L  S  Q  P  L  Y  P  S  -
b         C  G  K  L  H  C  C  L  I  *  I  H  C  H  S  P  S  T  L  Q  -
c          V  V  N  F  T  V  V  *  F  E  Y  I  V  T  A  P  L  P  F  S-

    3601 gcaaccaccaccctgatcagtcagcagccatcagcaccgaggcaaggccctccaccagca

a        A  T  T  T  L  I  S  Q  Q  P  S  A  P  R  Q  G  P  P  P  A  -
b         Q  P  P  P  *  S  V  S  S  H  Q  H  R  G  K  A  L  H  Q  Q  -
c          N  H  H  P  D  Q  S  A  A  I  S  T  E  A  R  P  S  T  S  K-
```

```
    3661 aaaagattctgactcactgaagacttggatgatcattagtatttttagcagtaaagtttt

a         K  R  F  *  L  T  E  D  L  D  D  H  *                      -
b          K  D  S  D  S  L  K  T  W  M  I  I  S                     -
c           K  I  L  T  H  *  R  L  G  *  S  L  V                    -

    3721 tttttctttttctttctttttttttttttttttagacggagtctcgctctgtcatccagtc

    3781 tggagtgcaatggcgtgatctcggctcactaccacctccgtctcctgggttcaagagatt

    3841 cttctgccttagcctcctgagtagctgggattacaggcgcgcgacaccatgcccagctaa

    3901 ttttttttttttgtatttttagtagagacggggtttcaccatattgaccaggctggtctt

    3961 gaattcctgaggtcatgatcctcccgcctcaacctcccaaaatgctgggattacaggcat

    4021 gagccaccacgctcggccaagtatttttaaatgaaggcatataaaaattgtttttttag

    4081 atgtgatgctatcctatatttaatagactacagtttagtgtaaacataacttttatatgt

    4141 gctgggaaaccaaaaaattcatatgacccacttcattatgatattcactttactgcagtg

    4201 tggtctggaactgaactcacactatctccaaagtatgcct
```

Note: Three translation frames are used. Stop codons are shown as "*". Uppercase regions are the largest internal inverted repeats.

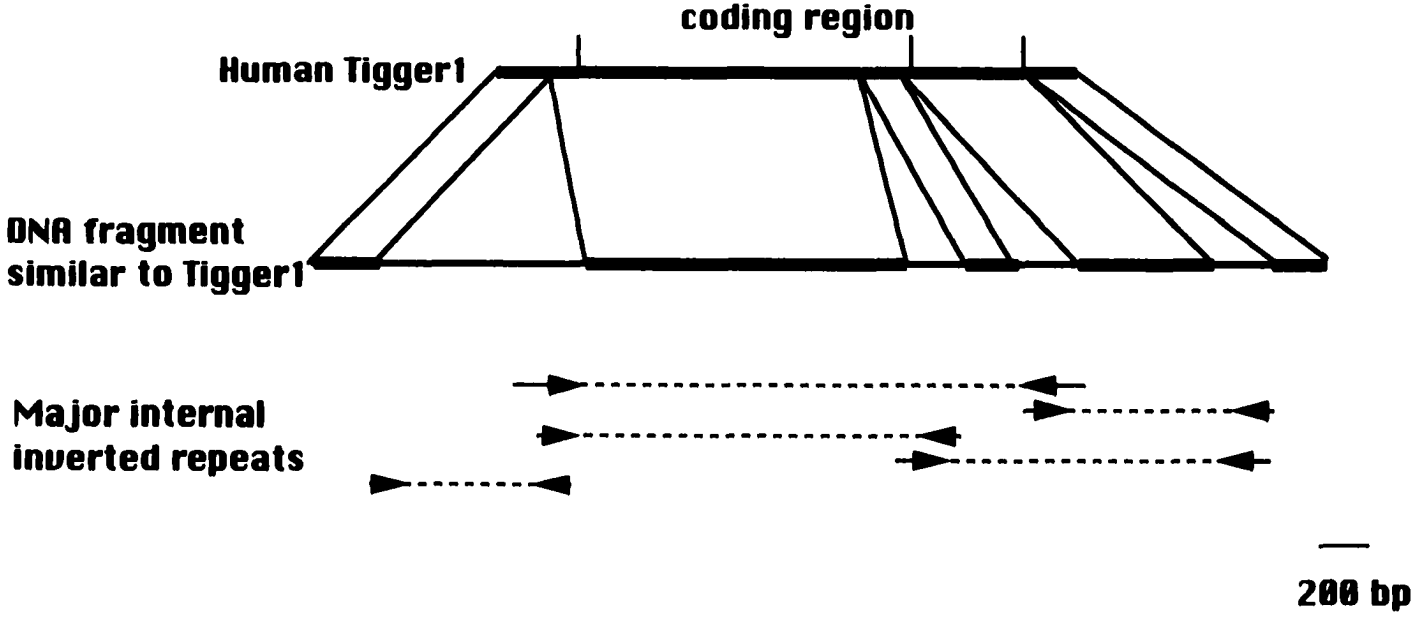


Figure 34: Position of major internal inverted repeats in Tigger 1-like transposon

At positions 134577-135352, there was an observed Powerblast homology to the human breakpoint cluster region (BCR) gene (U07000). The similarity is approximately 63% with 189 single base mutation and 100 indels. The similar region in the BCR gene is within the large intron between alternative exon b and exon 2. There is no breakpoint in this region and the region contains no known repetitive elements and thus may represent a previously undescribed low copy number repeat element.

3.5.2 Repetitive elements in contig 3

The interspersed repetitive elements found in contig3 include Alu, L1, Mer, Mir, Mlt, Mst and The1b. A summary of these repetitive elements is listed in Table 13.

Table 13: Interspersed repetitive elements in contig 3

Family	number	bases	% region
Alu	**126**	**35601**	**22.6**
L1	50	17239	10.9
Mer/Mir/Mlt/Mst	25	7875	5.0
The1b	2	443	0.3
Total non-Alu	**77**	**25557**	**16.2**

This region has a high density of non-Alu repeats than contig1 and contig2. Contig1 has 11.9% total non-Alu repeats and contig2 has 10.6%. The Alu repeats in this contig are slightly higher than those in contig1 and contig2 as contig1 has 19.5%

and contig2 has 17.2% total Alu, respectively. The total interspersed repetitive elements in contig3 is 38.8%.

3.6 Analysis of contig 4

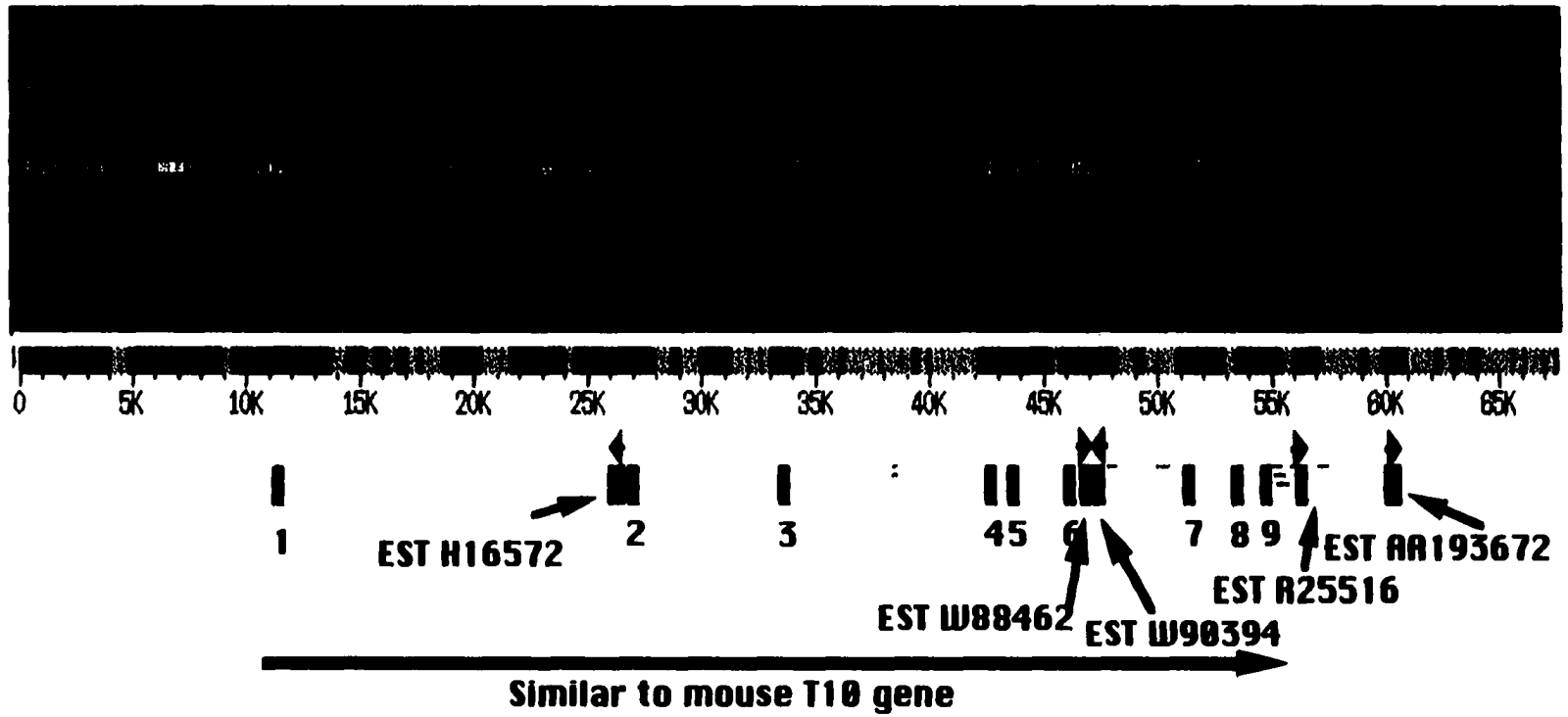
Contig4 is 67,378 bp and is located near the distal end of the DGCR. It contains two overlapping cosmid, 68f and 24b. The consensus of this contig is shown in the direction from centromere to telomere. The results of Xgrail and Powerblast are shown in Figure 35. This contig contains the genomic organization information of a gene highly similar to the mouse T10 gene. The transcription direction of this putative human T10 gene is from centromere to telomere.

3.6.1 T10 gene

Positions 26936 to 54800 in contig4 contain eight coding regions with high similarity to the coding regions of the mouse T10 gene. Positions 54801 to 56053 contain the region similar to the 3' UTR of the mouse T10 gene but the similarity is very low and there is a poly(A) signal sequence at the end of this region. A region homologous to the 5' UTR of the mouse T10 gene was not found in contig4.

A further study of the Powerblast results revealed that the first coding region of the mouse T10 gene also is highly similar to the 3' end half of EST T53435 in the database. The 5' end half of this particular EST is highly similar to contig4 in positions 11280-11390. This EST contains the 5' end of a human cDNA from a placental cDNA library. In the largest part of this region and the region right proximal of this region, from 10993 to 11361, there is a CpG island which contains a CAAT box and several Sp1 binding sites, which are the major features of a promoter. In contrast, the close

Figure 35: Result of Powerblast and Xgrail analysis of contig 4



Note: Xgrail result is on the top. The gray line is DNA sequence. The peaks are predicted exons. The small color bars above or below the peaks indicate the quality of exons, Green for excellent, blue for good and red for marginal. Peak above the gray line means the exon is on forward strand while peak under the line, reverse. The intensity of the gray line means the GC contents. The lighter the region is, the higher the GC contents. The pink boxes on the gray line are CpG islands.
Powerblast result is on the bottom. The black-gray line is DNA sequence. The gray regions are repeat sequence. The small colored bars are matches found in GenBank. The blue bars are matches with very high identity. The pink bars are matches with relative low similarity. The green bars are exons of T10 gene and the number under them are exon numbers.

upstream region of the first coding region of the T10 gene does not show any specific feature of a promoter. Other expression studies showed that the transcript of the T10 gene is larger than the cDNA sequence reported and suggested that T10 has a longer 5' untranslated region than reported. [65] The region from positions 11280 to 11390 therefore may be a potential first exon for the human T10 gene. The alignment of contig4, EST T53435 and mouse T10 cDNA is provided in Figure 36.

Figure 36: Alignment of contig 4, EST T53435, and mouse T10 gene (X74504)

```
contig4    > 11240    gggcgcggcacccggaagtcggcggcggtggcggaggcgg

contig4    > 11280    tgagtgcgcggctccggggctggccgactccgctagtggc
T53435     > 17       ......................n...................


contig4    > 11320    ccggccggcctgggctcgggggctccgggctctgggctct
T53435     > 57       ............t............................


contig4    > 11360    gggtgcgcggaccgggccaggctgcttgaag|gtgggtggg
T53435     > 97       ............................


contig4    > 11400    ctggccgcacttcctcaccctgcctcagtttccccctttct

~~~~~~~~~~~~~~~~intron1, sequence 11441-26839 not shown~~~~~~~~~~~~~~~~

contig4    > 26840    gagggcagccctgagtgtctgccatccgctcaactcagtg


contig4    > 26880    ttttccttttcccgcag|acctcgcgacctgtgtcagcaga

                                start codon
contig4    > 26920    gccgccctgcaccacc**atg**tgcatcatcttctttaagttt
X74504     > 48                    ..........................


contig4    > 26960    gatcctcgccctgtttccaaaaacgcgtacag|gtaacccc
X74504     > 74       ........t..............


contig4    > 27000    ctcgctctgcatctgctgcgccctgcagggtcctgggtgc

~~~~~~~~~~~~~~~~intron2, sequence 27041-33439 not shown~~~~~~~~~~~~~~~~

contig4    > 33440    acttggctcgctcgtttccatctgaagccatcagtgatgc


contig4    > 33480    tttcctcttgcag|gctcatcttggcagccaacagggatga
X74504     > 103                 ........................a.....
```

```
contig4    > 33520     attctacagccgaccctccaagttagctgacttctggggg
X74504     > 133       ......t.a.a.g..g........g..a.............

contig4    > 33560     aacaacaacgagatcctcagtg|gtgagtcttcctgtgtgc
X74504     > 173       .......gt...............

contig4    > 33600     tcagcggtggctgcgccttgttgcttccctagcccctgcg

----------------intron 3, sequence 33641-42519 not shown----------------

contig4    > 42520     tgcgtctcccagggctgggaaccaggtgggggactggaat

contig4    > 42560     gggtggcatcaatggtggtaacactgtcatctgccacag|g

contig4    > 42600     gctggacatggaggaaggcaaggaaggaggcacatggctg
X74504     > 200           ...................c......a.........

contig4    > 42640     ggcatcagcacacgtggcaagctggcagcactcaccaact
X74504     > 236       ..................g.......gg.............

contig4    > 42680     acctgcagccgcagctggactggcaggcccgagggcgag|g
X74504     > 276       ....

contig4    > 42720     taaggcgagtggggtggggccaaggtgagacagggtgggg

----------------intron 4, sequence 42761-43519 not shown----------------

contig4    > 43520     agaagagcacatgcctgcagctgtggacacagcatctgtc

contig4    > 43560     ccctgcctacag|gtgaacttgtcacccactttctgaccac
X74504     > 313                 ............gt.t............g

contig4    > 43600     tgacgtggacagcttgtcctacctgaagaaggtctctatg
X74504     > 343       ....a........c.......................ca

contig4    > 43640     gagggccatctgtacaatggcttcaacctcatagcagccg
X74504     > 383       ........c.....t..c.....t...a..........t.

contig4    > 43680     acctgag|gtgggtctgccagaggggatggcggctctgcc
X74504     > 423       .......

contig4    > 43720     cagcactgcctcgggggcaggcctcaggctcatcaggaag

----------------intron 5, sequence 73761-45999 not shown----------------

contig4    > 46000     gggctgtgagatcaccgggcagtgtcggggaggacgccct

contig4    > 46040     atggctgtagtctgacattctctcccccttggcctgcag|ca
```

```
contig4     > 46080     cagcaaagggagacgtcatttgctactatgggaaccgagg
                        T  A  K  G  D  V  I  C  Y  Y  G  N  R  G

contig4     > 46120     ggagcctgatcctatcgttttgacgccag|gtgagcctgcc
                          E  P  D  P  I  V  L  T  P

contig4     > 46160     ctggcagcctgatggggtgggggactgtttctatgcagag

~~~~~~~~~~~~~~~intron 6, sequence 46201-51599 not shown~~~~~~~~~~~~~~~

contig4     > 51600     caggtggcaatttgccctgacatggcacagcagggcctct

contig4     > 51640     gcatggcccgctgattgctcctcacag|gcacctacgggct
                                                    G  T  Y  G  L

contig4     > 51680     gagcaacgcgctgctggagactccctggaggaagctgtgc
                          S  N  A  L  L  E  T  P  W  R  K  L  C

contig4     > 51720     tttgggaagcagctcttcctggaggctgtggaacggagcc
                         F  G  K  Q  L  F  L  E  A  V  E  R  S

contig4     > 51760     aggcgctgcccaaggatgtgctcatcgccagcctcctgga
                        Q  A  L  P  K  D  V  L  I  A  S  L  L  D

contig4     > 51800     tgtgctcaacaatgaagaggc|gtgagtgggcgggtcctgc
                          V  L  N  N  E  E  A

contig4     > 51840     tggggtgagccccagtgtcccgccaccagggcagagggaa

~~~~~~~~~~~~~~~intron 7, sequence 51881-53399 not shown~~~~~~~~~~~~~~~

contig4     > 53400     ggtgggctggggctagacccggctgcgggcgggccacaga

contig4     > 53440     gggactaatggccaccacttctctccatcctgcag|gcagc
                                                          Q

contig4     > 53480     tgccagacccggccatcgaggaccagggtggggagtacgt
                        L  P  D  P  A  I  E  D  Q  G  G  E  Y  V

contig4     > 53520     gcagcccatgctgagcaagtacgcggctgtgtgcgtgcgc
                          Q  P  M  L  S  K  Y  A  A  V  C  V  R

contig4     > 53560     tgccctggctacggcaccag|gtattgcagcaccgtgggtg
                         C  P  G  Y  G  T  R

contig4     > 53600     cgccacctcctatcccatgtcccacctcccacgctagagg

~~~~~~~~~~~~~~~intron 8, sequence 53641-54599 not shown~~~~~~~~~~~~~~~

contig4     > 54600     tcttgccctatttgtggctgtgacaggcagggcagggctg

contig4     > 54640     agggacaccaggtgaacgagggcccctgctctctttcag|a

contig4     > 54680     accaacactatcatcctggtagatgcggacggccacgtga
X74504      > 761       ........c........t..g.....aa.t..........
```

```
contig4    > 54720    ccttcactgagcgtagcatgatggacaaggacctctccca
X74504     > 801      .g........a..g......c.......a...ac...t.g


                                                        stop codon
contig4    > 54760    ctgggagaccagaacctatgagttcacactgcagagctaa
X74504     > 841      ..........ac..................

contig4    > 54800    ccccacctctgggcctggccagtgggctcctgggggccc

contig4    > 54840    tgccttgaggggcactgtggacaggaaaccttcctttgcc

contig4    > 54880    atactgcattgcactgcccgtggcttggccagcatccccc

contig4    > 54920    ggatcagggccctgtggtttgcgtgttacccatctgtgtc

contig4    > 54960    cccatgcccagttcagggtctgcctttatgccagtgagga

contig4    > 55000    gcagcagagtctgatactaggtctaggaccggccgaggta

contig4    > 55040    taccatgaacatgtggatacacctgagcccactcttgcac

contig4    > 55080    atgtacacaggcactcacatggcacacacatacactcctg

contig4    > 55120    cgtgtgcacaagcacacacatgcaaggcatatacatggac

contig4    > 55160    accgacacaggcacatgtacatgcacaggtgtgctacaca

contig4    > 55200    tgtgcacacatgcacagttgcacagacacacacacacagg

contig4    > 55240    tgcacacacacgatgccgaacaaggcagaagggcgactct

contig4    > 55280    cacctctcatgtgcttctggccagtaggtctttgttctgg

contig4    > 55320    tccaacgacaggagtaggcttgtatttaaaagcggccccc

contig4    > 55360    cctctcctgtggccacagaacacaggcgtgcttggatttt

contig4    > 55400    tgacaagcagacctgctcctgcagaggagacagccacatt

contig4    > 55440    tggaattgggcaccgagaagacctgagaaaaacccactct

contig4    > 55480    ctctttttttttttttgagacggagtcttgctctgtcac

contig4    > 55520    ccaggctggagtgcagtggcacgatctcggctcactgcaa

contig4    > 55560    cctccgcctcccaagttcaagcaattctcctgccccagcc

contig4    > 55600    tcctgattagctgggattacaggcgtgtgccaccatgccc

contig4    > 55640    agctaatttttgtgtttttagtagagacggggtttcacca

contig4    > 55680    tgttggtcaggctggtcttgaactcctgacgtggtgatcc

contig4    > 55720    acctgcctcggcctcccaaagtgctgggattacaggcgtg

contig4    > 55760    agccaccacgcctggccgaaaaacccactctcataacaga

contig4    > 55800    agtgcagactcattgctagattcagtgcccttgagtgtgc

contig4    > 55840    cagaggtcctctgtgtttgagacaatcctgtgtgtgccag
```

```
contig4    > 55880    gaggctccgtgtgcaccaggggctctcagaatcccgctta
contig4    > 55920    cccagctggagaccatgcctctggcagccccatctcagcc
contig4    > 55960    agccctgctctctccctcttccctccaggtgaggcaaact
contig4    > 56000    tcataggaatctgtacctgaatgtgagctcctgatAATAA
                                                        poly(A)
contig4    > 56040    Aactctgaggctttggtgagcgcatttcgaggcctttccc
```

Note: Not all sequences are shown. "." indicates the identity. "|" indicates the splice site. Powerblast program failed to find exons 5, 6 and 7. Start codon and stop codon are bold. Poly(A) signal is capitalized.

In summary, contig4 contains the human T10 gene. The transcription direction of this 9 exon containing gene is from centromere to telomere, which disagreed with that previously reported by hybridization. [65] The first putative exon is similar to a human EST and the coding region starts in exon 2. Comparative analysis of the human genomic sequence and the mouse cDNA sequence revealed the position and the size of all 9 exons (Table 14) except for the first and last exons because the boundaries of the 5' and 3' UTR can not be determined by the information provided only by the genomic sequence. The boundaries of the exon/intron were determined with the available genomic data and the nearly invariant consensus 5' GT and 3' AG dinucleotides of intron boundaries. The comparison of all splice sites and the consensus is provided in Table 15.

Table 14: Position and size of exons and size of introns of human T10 gene in contig 4

Exons				Introns	
Exon	Start	End	Size	Intron	Size
1	unknown	11390	-	1	15496
2	26897	26991	95	2	6501
3	33493	33581	89	3	9017
4	42599	42718	120	4	853
5	43572	43686	115	5	2391
6	46078	46148	71	6	5518
7	51667	51820	154	7	1654
8	53475	53579	105	8	2127
9	54689	unknown	-		

Note: The size of exon 1 and exon 9 are unknown because the genomic sequence data can not define the boundaries of the 5' and 3' untranslated regions. Start codon ATG begins at position 40 of the second exon. Stop codon TAA begins at position 89 of the last exon.

Table 15: Exon/intron boundary of human T10 gene

```
          exon                  intron                  exon
     1 ...CTGCTTGAAG|gtgggtgggc... 1 ...tttcccgcag|ACCTCGCGAC... 2
     2 ...ACGCGTACAG|gtaacccccc... 2 ...cctcttgcag|GCTCATCTTG... 3
     3 ...ATCCTCAGTG|gtgagtcttc... 3 ...tctgccacag|GGCTGGACAT... 4
     4 ...CGAGGGCGAG|gtaaggcgag... 4 ...ctgcctacag|GTGAACTTGT... 5
     5 ...CCGACCTGAG|gtgggtctgc... 5 ...tggcctgcag|CACAGCAAAG... 6
     6 ...TTGACGCCAG|gtgagcctgc... 6 ...ctcctcacag|GCACCTACGG... 7
     7 ...ATGAAGAGGC|gtgagtgggc... 7 ...catcctgcag|GCAGCTGCCA... 8
     8 ...ACGGCACCAG|gtattgcagc... 8 ...tctctttcag|AACCAACACT... 9
consensus   (C/A)AG|gt(a/g)agt.....(t/c)nn(c/t)ag|G
```

Note: Upper case letters are exons and lower case letters, introns.

The similarity of the coding region between mouse T10 gene and human T10 gene is 83.8% identical at the DNA level and at the amino acid level, they are 87% identical and 93.5% similar. Using the deduced amino acid sequence of the human T10 gene to search the non-redundant GenBank protein database with Blastp revealed no similarities for this protein to any protein except for the mouse T10 gene and a weakly similar *C. elegens* hypothetical gene (1483279) products. Alignment of amino acid sequences of these three proteins with ClustalW program is provided in Figure 37. The similarity between human T10 gene and *C. elegens* putative gene is only 29.7% identical and 47.8% similar.

Figure 37: CLUSTAL W(1.60) multiple sequence alignment among human, mouse and yeast T10 gene products.

```
CLUSTAL W(1.60) multiple sequence alignment


h vs m.          ******************************** ************ ************ ***
ht10pep__2       MCIIFFKFDPRPVSKNAYRLILAANRDEFYSRPSKLADFWGNNNEILSGLDMEEGKEGGT
mmt10pep         MCIIFFKFDPRPVSKNAYRLILAANRDEFYNRPSKLADFWGNNSEILSGLDMEEGKAGGT
yt10like_        MCITFIKTAKS--SLDKYKLILLNNRDEQLLRPTKEMH-WHDG--ILSGVDKQEN-ARGT
h/m/y            *** * *      *   *.***  ****   **.*    *  .  ****.* .*.   **


h. vs m.         ********** ********. . -- ********.****.*.********** *******
ht10pep__2       WLGISTRGKLAALTNYLQPQLDW--QARGRGELVTHFLTTDVDSLSYLKKVSMEGHLYNG
mmt10pep         WLGISTRGKLGALTNYLQPRQEP--DARGRGELVSHFLTSDMDSLSYLKKVSTEGHLYNG
yt10like_        WLGLNENGRIGMMLSITQTQESKNLHAPSRGGIVNEFLNAN-DTSKMIESLKKCASKYNG
h/m/y            ***.   *.. .   * .      *  ** .*  **..  *.   .  .      ***


h. vs m.         **.*******.****.******--***.********************.*********.*
ht10pep__2       FNLIAADLSTAKGDVICYYGNR--GEPDPIVLTPGTYGLSNALLETPWRKLCFGKQLFLE
mmt10pep         FNIIAADLSTSKGDVVCYYGNR--GEPEPIVLTPGTYGLSNALLETPWKKLCFGKQLFME
yt10like_        FQLVAVEKNSTGLYEVRTLANQQVDEIEVCQLKDEYHVVSNSPPTKPYQKAVQGKKLLRE
h/m/y            *...* .  ..    .    *.  * .  *      .**.    * .*   **.*  *


h. vs m.         ***.*.*******.. ********************** ---*****.* *****  **
ht10pep__2       AVERSQALPKDVLIASLLDVLNNEEAQLPDPAIEDQGG---EYVQPMLSKYAAVCVRCPG
mmt10pep         AVEQSEALPKDVLVTQLLDVLNNEEAQLPDPAIEDQGQ---EYVQPILNKYAAVWCRCAS
yt10like_        HLENSDQFSVDQIFEKLLSIAKNTTQWYPDAQLQYQTQNVEEYNRPLSAIFIKYPEGTRM
h/m/y            .* *     * .   ** .  *     **  .. *     ** .*.   .      .
```

```
h. vs m.        ************ **********.*** *.*** ********
ht10pep__2      YGTRTNTIILVDADGHVTFTERSMMDKDLSHWETRTYEFTLQS-
mmt10pep        YGTRTNTIILVDANGHVTFTERSMLDKDTSRWETNTYEFTLQS-
yt10like_       YGTRCHTLITVDQKDKINILERRLLP-EQSTWHDARFEFVLNGS
h/m/y           ****..*.* **   ...  ** ..  . * *    .** *.
```

Note: "*" indicates identity, "." indicates similarity, " " (space) indicates dissimilarity, and "-" indicates deletion/insertion. Similarity parameters above amino acid sequences indicate the comparison between human and mouse and those below indicate similarity among all three.

The potential 5' promoter region of the human T10 gene was analyzed by using the SIGNALSCAN against the TRANSFAC database. A potential promoter region was found within a Xgrail predicted CpG island, and the size of this promoter region is about 300 bp. The position of the CpG island is from 10993 to 11361 in contig4 and from -288 to 81 if the beginning of the match of EST 53435 is defined as +1. The C/G content of this CpG island is 78.6% and the total C/G content of contig4 is 54.2%.

No TATA box or the potential binding sites for transcription factor IID (TFIID) were found in this region, but at least five putative binding sites of transcription factor Sp1 are present in this region. Other putative regulatory elements and transcription factor binding sites found by SIGNALSCAN are: NF-1 (nuclear factor 1), cac-binding protein, and SRF (serum response factor). The position of these putative features are provided in Figure 38.

Figure 38: Genomic sequence of 5' region of the human T10 gene

```
cccatggcgtccctgcaggcagggagccgcggagtccccgaaggccgctggatcaggaag      -240

cgctcagACGCCCgcgcagcccgaagctttggcgcagccccGGGGGCGGGGCGCCAggaG      -180
        Sp1                                    Sp1     NF-1

GTGGcagcctccgagcgacagcgcgcccAGCCAAcgggacgccagcatgcggcgcgCCGC      -120
cac-binding pro              NF-1

CCCGCCCCtcgcTGCGCggcCCAATggtgaacgcgccggcttcgggctGGGCGGtactgg      -60
  Sp1        Sp1     SRF                            Sp1

gctggctgcggtggcgcgcgggcgcggcacccggaagtcggcggcggtggcggaggcgg      potential -1
```

Note: The starting point of match of EST 53435 is defined as +1.

The function of T10 gene is unknown and no putative function has been proposed. However, its expression analysis has been studied and reported. [65] The expression level of T10 gene is highest in fetal liver, but is also detected in lung, heart and kidney. *In situ* hybridization on mouse embryos from 8.5 days to 15.5 days showed that T10 gene is expressed during early embryogenesis at a time slightly later than when the structures commonly affected in DGS are forming. Therefore it is not likely that the T10 gene product plays a role in the DGS phenotype.

3.6.2 Repetitive elements in contig 4

Contig4 has an overall G/C content of 54.2%. Compared with the other three contigs discussed previously, this G/C content is slightly higher, since none of these three earlier contigs has a G/C content higher than 50% (G/C% of contig1=45.2, G/C% of contig2=47.2 and G/C% of contig3=48.2). As shown in Table 16, the Alu repeat

element density found in contig4 is slightly more than the other three contigs but the non-Alu repeat density is much less.

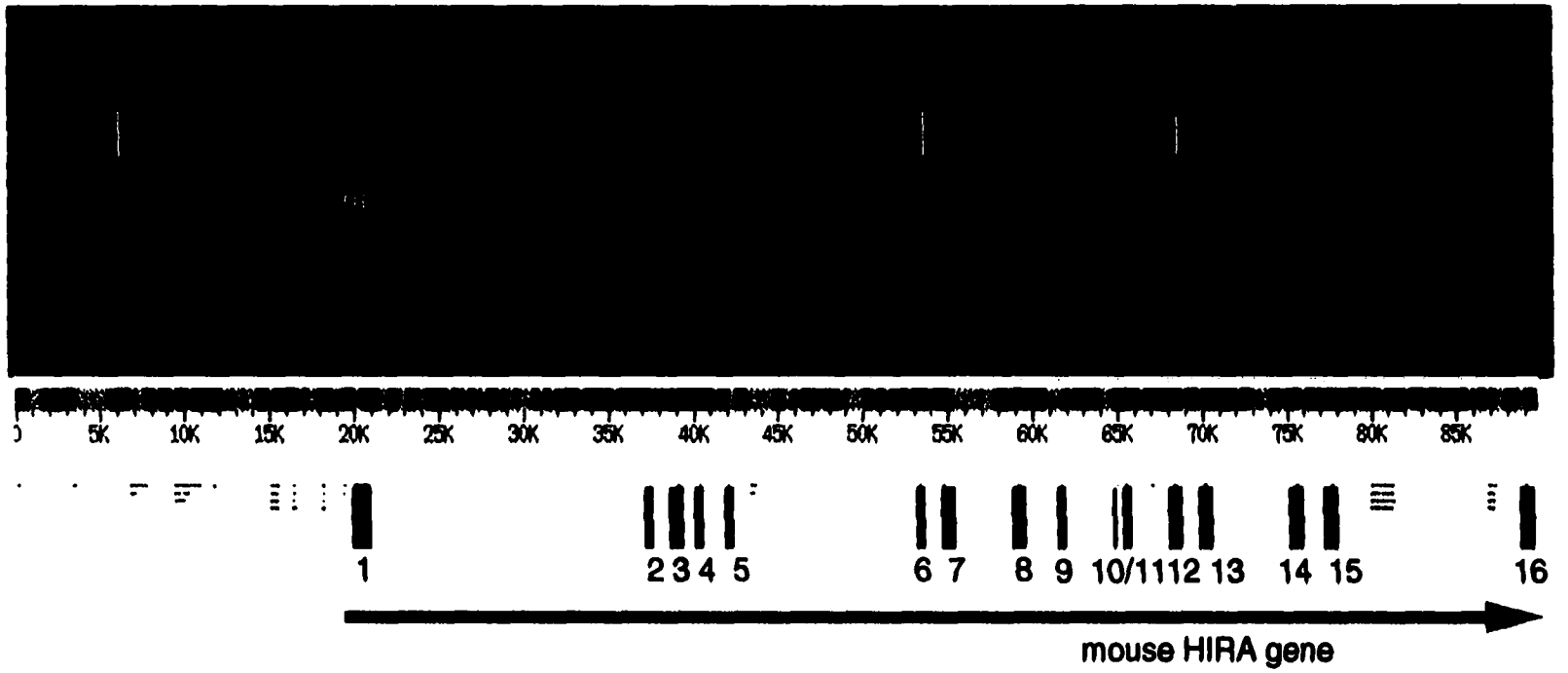
Table 16: Interspersed repetitive elements in contig 4

Family	number	bases	% region
Alu	**53**	**16524**	**24.5**
L1	3	878	1.3
Mer/Mlt/Mst	9	2066	3.1
Thelb	3	743	1.1
Total non-Alu	**15**	**3687**	**5.5**

3.7 Analysis of a mouse BAC containing the mouse HIRA gene

The 89508 bp mouse BAC b264n1 was sequenced and analyzed. This BAC was shown to contain exons 1 through 16 of the mouse HIRA gene. The results of Xgrail and Powerblast are given in Figure 39. Comparing the results of Xgrail and Powerblast, only one of the known exons of the mouse HIRA gene, namely exon 10 could not be found by Powerblast. This exon is relatively small (71 bp in length) and very close to exon 11 (intron 10 is 466 bp in size). In contrast, Xgrail failed to predict exons 1, 4, 9 and 15 but instead predicted several additional exons which are not present in the mouse HIRA cDNA sequence. The majority of these additional predicted exons lie in the opposite orientation and within three large introns on the mouse HIRA gene, namely exon 1, 17 kb, intron 5, 11 kb, and intron 15, 11 kb.

Figure 39: Result of Powerblast and Xgrail analysis of mouse BAC b264n1



Note: Xgrail result is on the top. The gray line is DNA sequence. The peaks are predicted exons. The small color bars above or below the peaks indicate the quality of exons, Green for excellent, blue for good and red for marginal. Peak above the gray line means the exon is on forward strand while peak under the line, reverse. The intensity of the gray line means the GC contents. The lighter the region is, the higher the GC contents. The pink box on the gray line is CpG islands.
Powerblast result is on the bottom. The black-gray line is DNA sequence. The gray regions are repeat sequence. The small colored bars are matches found in GenBank. The numbers under green bars are exon numbers of mouse HIRA gene.

3.7.1 Mouse HIRA gene

If it is assumed that the region of mouse chromosome 16 in BAC b264n1 and the corresponding region of human chromosome 22 are in the same orientation, then the transcription direction of this gene is from centromere to telomere. Comparative analysis of the genomic sequence and the cDNA sequence also reveals the positions and the sizes of exons 1 to 16 and introns 1 to 15, as listed in Table 17. The boundaries of exon/intron were determined assuming that the nearly invariant consensus 5' GT and 3' AG dinucleotides of intron boundaries must present at the cleavage point. A list of the deduced splice sites and the consensus splice sequences is provided in Table 18.

The alignment of the mouse genomic DNA sequence determined here and the mouse HIRA cDNA sequence (Genbank accession number X92590) is shown in Figure 40. These two sequences match perfectly except for one mismatch at position 89111, where a "c" occurs in the genomic sequence instead of the "t" found in cDNA sequence. This discrepancy was checked carefully and there was six-fold redundancy in genomic DNA sequence with gel readings in both orientations. Thus, at least in the sequence determined here, a "c" occurs at position 89111 and would result in a Leucine rather than a Phenylalanine in X92590 cDNA sequence. Immediately after exon 3, at positions 39174 to 39350, there is an alternatively spliced exon, exon 3a [263] and at positions 65451 to 65556, which is in intron 11, there is another alternatively spliced exon, 11a [263]. Both of these alternatively spliced exons introduce stop codons as indicated in Figure 40.

Table 17: Positions and sizes of exons and size of introns of the mouse HIRA gene

Exons				Introns	
Exon	Start	End	Size	Intron	Size
1	20093	20349	257	1	17033
2	37383	37445	63	2	1617
3	39063	39173	111	3	1133
(3a	39174	39350	177	3a	0)
4	40307	40397	91	4	1574
5	41972	42066	95	5	11198
6	53265	53360	96	6	1302
7	54663	54823	161	7	4213
8	59037	59204	168	8	2368
9	61573	61686	114	9	3227
10	64914	64984	71	10	466
11	65451	65556	106	11	2688
(11a	66933	67039	107	11a	1376)
12	68245	68460	216	12	1574
13	70035	70120	86	13	5344
14	75465	75659	195	14	1956
15	77615	77776	162	15	11159
16	88936	89137	202	16	unknown

Note: Exon 1 contains the 5' untranslated region. The start codon ATG begins at position 231 of the first exon. The size of intron 16 is unknown and the alternatively spliced exons 3a and 11a are in parentheses. Intron 3 includes exon 3a and intron 3a. Intron 11 includes exon 11a and intron 11a.

Table 18: Exon/intron boundaries of the mouse HIRA gene

	exon	intron	exon	
1	...AACCACAA**T**G|gtgagtgcac...	1	...ttgtttccag|GGAAGCCAAT...	2
2	...GGAGGACAAG|gtatgtttgt...	2	...ttcattccag|GGCAGGATTC...	3
3	...AATCACTTAG|gtattacaga...	3	...ttcttcccag|CATGTGTGAA...	4
4	...AGCGGGCTA**C**|gtaagcatca...	4	...tttcacacag|GTACATTGGG...	5
5	...CACTCAGGC**G**|gtaagtaggc...	5	...tttcttccag|ATGTGATGGA...	6
6	...AAGTTCCCAG|gtctggggtt...	6	...tccttcacag|AAATTCTTGC...	7
7	...TTTTGATGAG|gtaacggaaa...	7	...tttatttcag|TGTGGAGGAA...	8
8	...GACTGTTGT**G**|gtgagtgtcc...	8	...ttgacttcag|AAATTCAACC...	9
9	...CTCTGTCTG**G**|gtaagcctga...	9	...ccccttgtag|CTCACATGTT...	10
10	...ATATTTCCT**G**|gtaagatgga...	10	...tctctttcag|GACTCTGAAT...	11
11	...GGAGGAAAAG|gtaggcttgc...	11	...cttcctacag|AGCCGAATTC...	12
12	...TATCAGGAAG|gtgagggcat...	12	...actttgccag|AATCTTTTGA...	13
13	...TGGACACT**G**G|gtactcactc...	13	...ttaatgacag|GGACTTCTCC...	14
14	...GTAAGGACAG|gttagtaaac...	14	...tttcccccag|TATGAATGCT...	15
15	...CTGTTGAAAG|gtaggtggta...	15	...ttgatttcag|GTTGAAAGAG...	16
16	...GTCCGTCCAG|gtgaggccta...	16		

consensus: (C/A)AG|gt(a/g)agt.....$(t/c)_n$n(c/t)ag|G

Note: Upper case letters represent the exons portion and lower case letters, introns. The bold letters are those nucleotides which do not match with the consensus.

Figure 40: Alignment of mouse genomic DNA and mouse HIRA cDNA

```
mbconsphas > 20085     aggaggagcggcgcaggggatgcggctgtggcggcggcgg
X92590     > 1                 ..............................

mbconsphas > 20125     cggcggccgagcgcaggaggcggctgtggcggcggctggg
X92590     > 33        ........................................

mbconsphas > 20165     ggcacgggccggcgatggcgcggcggcgctgagggcaagg
X92590     > 73        ........................................

mbconsphas > 20205     gtgggcggcggccggagggcgggcggcgcgggaggaagct
X92590     > 113       ........................................

mbconsphas > 20245     gcggcagctccatggcccaggcgtgctgagggacacggct
X92590     > 153       ........................................

                                              start codon
mbconsphas > 20285     ctggcttcagcccggcagcggccgaacaatgaagctcttg
X92590     > 193       ........................................
HIRA prote > 1                                   M  K  L  L

mbconsphas > 20325     aagccaacctgggtcaaccacaatg|gtgagtgcacgcagg
X92590     > 233       ......................... .
HIRA prote > 5          K  P  T  W  V  N  H  N   G

mbconsphas > 20365     agtcgtgggcggccgaacctgcatcggggtggggtctggg

-------------------intron 1, sequence 20405-37324 not shown-------------

mbconsphas > 37325     agtttattgttttctagtgaataaaagccaagcttgatat

mbconsphas > 37365     gtttttctttgtttccag|ggaagccaatttttttcagttga
X92590     > 257                         . ......................
HIRA prote > 13                           G  K  P  I  F  S  V  D

mbconsphas > 37405     tattcaccctgatgggaccaagtttgcaactggaggacaa
X92590     > 280       ........................................
HIRA prote > 21           I  H  P  D  G  T  K  F  A  T  G  G  Q

mbconsphas > 37445     g|gtatgtttgtagagtgaataccacttccctgttgccgta
X92590     > 320       . .
HIRA prote > 34           G

mbconsphas > 37485     ctctgcttcttttggatccatggttagctgtaatgagaca

-------------------intron 2, sequence 37525-39004 not shown-------------

mbconsphas > 39005     gttatcattgtcgtttccacttcttgcatattgagctttt
```

```
mbconsphas > 39045     cttgctccttcattccag|ggcaggattctgggaaggttgt
X92590     > 319                       .. ......................
HIRA prote > 34                          G  Q  D  S  G  K  V  V


mbconsphas > 39085     gatctggaatatgtctccagtcctccaggaggatgacgag
X92590     > 343       ........................................
HIRA prote > 42          I  W  N  M  S  P  V  L  Q  E  D  D  E


mbconsphas > 39125     aaggatgaaaatattcccaagatgctttgccagatggaca
X92590     > 383       ........................................
HIRA prote > 55         K  D  E  N  I  P  K  M  L  C  Q  M  D


mbconsphas > 39165     atcacttag|gtattacagagtggccctttgcaggctttgc
X92590     > 423       .........
HIRA prote > 68        N  H  L
X99713     > 1                  ..............................          |
                                                                        |
                                                                        |
mbconsphas > 39205      gtgttggcagagtgcccaaggttagcacatgtgtttgtaa         |
X99713     > 32         ........................................        |
                                          stop          stop          |
                                                                        |
mbconsphas > 39245      ttagaaagataaggccccgtgcggcgttaactcccctcgg         exon 3a
X99713     > 72         ........................................        |
                               stop                                     |
                                                                        |
mbconsphas > 39285      cctgctgtctgaaggcacagatgttggctgcacacctggg         |
X99713     > 112        ........................................        |
                                                                        |
                                                                        |
mbconsphas > 39325      tgagttattcggcagaagactcggcc|cgcccacagaacca        |
X99713     > 152        ........................                        |
                       stop

mbconsphas > 39365     tctttggacagtctgcaggtacctctggtgatttcctgct

~~~~~~~~~~~~~~~~~~intron 3, sequence 39405-40244 not shown~~~~~~~~~~~~~

mbconsphas > 40245     ctacataatgcaaatgggcatgcctattggattttcatca


mbconsphas > 40285     ccttatttttttttcttcccag|catgtgtgaactgtgtgc
X92590     > 430                          .. ...................
HIRA prote > 71                             A  C  V  N  C  V


mbconsphas > 40325     ggtggtcaaacagtgggatgtatttagcttctgggggaga
X92590     > 450       ........................................
HIRA prote > 77        R  W  S  N  S  G  M  Y  L  A  S  G  G  D


mbconsphas > 40365     tgacaaactgattatggtgtggaagcgggctac|gtaagca
X92590     > 490       ................................. ...
HIRA prote > 91          D  K  L  I  M  V  W  K  R  A  T   Y


mbconsphas > 40405     tcattttccttttttcacatttaatttttagaagcttcttt

~~~~~~~~~~~~~~~~~~intron 4, sequence 40445-41924 not shown~~~~~~~~~~~~~

mbconsphas > 41925     tatctagaacactccttgagtgcatatctcatgcctcttt
```

```
mbconsphas > 41965     cacacag|gtacattgggcccagcactgtgtttggttccag
X92590     > 523               ..............................
HIRA prote > 102                 Y  I  G  P  S  T  V  F  G  S  S

mbconsphas > 42005     tggtaagcttgccaatgtggagcaatggcggtgtgtctcc
X92590     > 556               .....................................
HIRA prote > 113                G  K  L  A  N  V  E  Q  W  R  C  V  S

mbconsphas > 42045     atcctccggagtcactcaggcg|gtaagtaggctcctttgt
X92590     > 596               ..................
HIRA prote > 126                I  L  R  S  H  S  G

mbconsphas > 42085     ttgggtgagtggatccctttttttttttttttttcgag

~~~~~~~~~~~~~~~~~~intron 5, sequence 42125-53204 not shown~~~~~~~~~~~~~

mbconsphas > 53205     accatcacccaagtgcataccaggtaaaaggtcttaccct

mbconsphas > 53245     cacaccctgttttcttccag|atgtgatggatgtagcatgg
X92590     > 617                        . ..................
HIRA prote > 133                          D  V  M  D  V  A  W

mbconsphas > 53285     tctccccacgacgcctggctggcctcatgcagcgtggata
X92590     > 638               .....................................
HIRA prote > 140                S  P  H  D  A  W  L  A  S  C  S  V  D

mbconsphas > 53325     acactgttgtcatttggaatgccgtgaagttcccag|gtct
X92590     > 678               ................................
HIRA prote > 153             N  T  V  V  I  W  N  A  V  K  F  P

mbconsphas > 53365     ggggttcccccttactgtgtagcaactgaagaggcctgctg

~~~~~~~~~~~~~~~~~~intron 6, sequence 53405-54604 not shown~~~~~~~~~~~~~

mbconsphas > 54605     tctttttcttatagataacacctgttcatattttatttta

mbconsphas > 54645     ttttattttccttcacag|aaattcttgcaactctgagagg
X92590     > 700                        ... ..................
HIRA prote > 161                        P  E  I  L  A  T  L  R  G

mbconsphas > 54685     tcattctggcttagtaaaaggtttgacttgggatcccgtt
X92590     > 736               .....................................
HIRA prote > 173                 H  S  G  L  V  K  G  L  T  W  D  P  V

mbconsphas > 54725     ggtaaatatattgcctctcaagctgatgatcgaagtttga
X92590     > 776               .....................................
HIRA prote > 186                G  K  Y  I  A  S  Q  A  D  D  R  S  L

mbconsphas > 54765     aggtatggaggacgctggactggcagctagagactagcat
X92590     > 816               .....................................
HIRA prote > 199             K  V  W  R  T  L  D  W  Q  L  E  T  S  I
```

```
mbconsphas > 54805     caccaagccttttgatgag|gtaacggaaacagtctagagc
X92590     > 856       ..................
HIRA prote > 213         T  K  P  F  D  E

mbconsphas > 54845     ttggcctctacccctgctatacttcttgaacagtgaggga

~~~~~~~~~~~~~~~~~~intron 7, sequence 54885-58964 not shown~~~~~~~~~~~~~~

mbconsphas > 58965     aaatcacagtttgatttatctttgaggtgagaattttctt

mbconsphas > 59005     gaacttctttcaagcttatattttatttacag|tgtggagg
X92590     > 873                                    .. ........
HIRA prote > 218                                    E  C  G  G

mbconsphas > 59045     aacgacgcatgttctccggcttagttggtcacctgatggc
X92590     > 883       ....................................
HIRA prote > 222         T  T  H  V  L  R  L  S  W  S  P  D  G

mbconsphas > 59085     cattacctggtatctgcccatgccatgaataattctggcc
X92590     > 923       ....................................
HIRA prote > 235        H  Y  L  V  S  A  H  A  M  N  N  S  G

mbconsphas > 59125     ccactgctcagatcatcgaaagagagggctggaagaccaa
X92590     > 963       ....................................
HIRA prote > 248       P  T  A  Q  I  I  E  R  E  G  W  K  T  N

mbconsphas > 59165     catggactttgtgggtcaccggaaagctgtgactgttgtg|
X92590     > 1003      ....................................
HIRA prote > 262         M  D  F  V  G  H  R  K  A  V  T  V  V

mbconsphas > 59205     gtgagtgtccctcacctcagaaatcctagctatcagtacc

~~~~~~~~~~~~~~~~~~intron 8, sequence 59245-61524 not shown~~~~~~~~~~~~~~

mbconsphas > 61525     catcttggtatttgattacttattttaactgtggttattt

mbconsphas > 61565     gacttcag|aaattcaacccaaaaatcttcaagaagaagca
X92590     > 1042            . ................................
HIRA prote > 275                K  F  N  P  K  I  F  K  K  K  Q

mbconsphas > 61605     gaagaatgggagctctacaaagcccagctgcccatactgc
X92590     > 1075      ....................................
HIRA prote > 286         K  N  G  S  S  T  K  P  S  C  P  Y  C

mbconsphas > 61645     tgctgtgctgttggcagcaaggaccgctcactctctgtct
X92590     > 1115      ....................................
HIRA prote > 299        C  C  A  V  G  S  K  D  R  S  L  S  V

mbconsphas > 61685     gg|gtaagtgcctgagctttggctgtacggtaggttaagaa
X92590     > 1155      ..
HIRA prote > 312       W
```

```
mbconsphas > 61725     ctgggaagacagagactctggaagacttcaaggaaagaga

~~~~~~~~~~~~~~~~~~intron 9, sequence 61765-64844 not shown~~~~~~~~~~~~~

mbconsphas > 64845     gatggttaaatttagtgcacagactggttctcagtgattt


mbconsphas > 64885     cattaatttcaacttttttccccttgtag|CTCACATGTTT
AA437523   > 479                                    . ..........
                                                     L  T  C  L


mbconsphas > 64925     GAAACGGCCTCTGGTTGTCATCCATGAACTGTTTGACAAG
AA437523   > 491       .....................................
                        K  R  P  L  V  V  I  H  E  L  F  D  K


mbconsphas > 64965     TCCATCATGGATATTTCCTG|gtaagatggattcttactac

                        S  I  M  D  I  S  W


mbconsphas > 65005     actgcattggctactccaacctgagacctcttggctactt

~~~~~~~~~~~~~~~~~~intron 10, sequence 65045-65404 not shown~~~~~~~~~~~~~

mbconsphas > 65405     ttcaggtacattgtccacctaggctcctgacagctgtctc


mbconsphas > 65445     tttcag|gactctgaatgggttgggtatcctggtgtgctcc
X92590     > 1227          . ................................
HIRA prote > 336           W   T  L  N  G  L  G  I  L  V  C  S


mbconsphas > 65485     atggacggctctgtggcgttccttgatttctctcaggatg
X92590     > 1262      .....................................
HIRA prote > 348        M  D  G  S  V  A  F  L  D  F  S  Q  D


mbconsphas > 65525     aactcggagacccccctgagtgaggaggaaaag|gtaggctt
X92590     > 1302      ............................
HIRA prote > 361       E  L  G  D  P  L  S  E  E  E  K


mbconsphas > 65565     gtggcctatgtgctatgatactggctctgagaacctattc

~~~~~~~~~~~~~~~~~~intron 11a, sequence 65605-66884 not shown~~~~~~~~~~~~

mbconsphas > 66885     gctctgctttacctgctgtagacacagaaactgttctttc


mbconsphas > 66925     taccgacag|acttttgccaaccgttaatggggtattatat        |
X99722     > 1               . ..............................        |
                                          stop                       |
                                                                     |
mbconsphas > 66965     ccatttgtctgaatggagctctgtggaagcccagagcagg        exon 11a
X99722     > 33        .....................................          |
                                                                     |
                                                                     |
mbconsphas > 67005     tcagcagctatccacttagtgatgctgttgctgggg|gtaa       |
X99722     > 73        ..................................            |
                                  stop stop
```

```
mbconsphas > 67045     gctacctcttaatctactgtgcgacaccaccaccctcaac

~~~~~~~~~~~~~~~~~~intron 11, sequence 67085-68164 not shown~~~~~~~~~~~~

mbconsphas > 68165     tatattgcacctgctggttagcaacttcagccatgatgat


mbconsphas > 68205     actcaaacccatgtatcaaacctctccctgcttcctacag
X92590      > 1332                                             ..
HIRA prote > 371                                              K


mbconsphas > 68245     |agccgaattcaccagtctacctatggcaagagcctggcaa
X92590      > 1334      ....................................
HIRA prote > 372         S  R  I  H  Q  S  T  Y  G  K  S  L  A


mbconsphas > 68285     taatgactgaggcccagctttccacagctgttattgagaa
X92590      > 1374     ....................................
HIRA prote > 385       I  M  T  E  A  Q  L  S  T  A  V  I  E  N


mbconsphas > 68325     ccctgagatgctcaagtaccagcggagacagcaacagcag
X92590      > 1414     ....................................
HIRA prote > 399         P  E  M  L  K  Y  Q  R  R  Q  Q  Q  Q


mbconsphas > 68365     cagctggatcagaagaatgccactactagggagacaagct
X92590      > 1454     ....................................
HIRA prote > 412        Q  L  D  Q  K  N  A  T  T  R  E  T  S


mbconsphas > 68405     cagcatcctcagtcacgggtgtggtcaatggggaaagtct
X92590      > 1494     ....................................
HIRA prote > 425       S  A  S  S  V  T  G  V  V  N  G  E  S  L


mbconsphas > 68445     agaagatatcagaaag|gtgagggcatggtccagtggcact
X92590      > 1534     ..............
HIRA prote > 439         E  D  I  R  K


mbconsphas > 68485     aaaagacttgagcacatgcccagagacagggcaaaggggg

~~~~~~~~~~~~~~~~~~intron 12, sequence 68525-69964 not shown~~~~~~~~~~~~

mbconsphas > 69965     ttgagtttttgatatttggatgtggttagctgaaacttttt


mbconsphas > 70005     gtattcacttatttgactctactttgccag|aatcttttga
X92590      > 1548                                 .. .........
HIRA prote > 443                                  K  N  L  L


mbconsphas > 70045     agaaacaagtggaaactcggacagcagatggtcggaggag
X92590      > 1560     ....................................c..
HIRA prote > 447       K  K  Q  V  E  T  R  T  A  D  G  R  S  R


mbconsphas > 70085     aatcacgcctctttgcatagcacagctggacactgg|gtac
X92590      > 1600     ..................................  .
HIRA prote > 461         I  T  P  L  C  I  A  Q  L  D  T  G
```

```
mbconsphas > 70125     tcactccttcaccaagctttctgtggccttgtctggggag

------------------intron 13, sequence 70165-75404 not shown------------

mbconsphas > 75405     gttataaaataactacatgtgacatttttgtgctgattta


mbconsphas > 75445     cttttccttgttaatgacag|ggacttctccacggcattct
X92590      > 1635                          . ..................
HIRA prote > 472                             G  D  F  S  T  A  F


mbconsphas > 75485     tcaacagcatcccactctccagctccctagcaggcaccat
X92590      > 1656     ........................................
HIRA prote > 479       F  N  S  I  P  L  S  S  S  L  A  G  T  M


mbconsphas > 75525     gctctcctctcctagtggtcagcagctactaccactggac
X92590      > 1696     ........................................
HIRA prote > 493         L  S  S  P  S  G  Q  Q  L  L  P  L  D


mbconsphas > 75565     tccagtacccccctccttcggcgcctcaaagccttgcacag
X92590      > 1736     ........................................
HIRA prote > 506        S  S  T  P  S  F  G  A  S  K  P  C  T


mbconsphas > 75605     aaccagtggcagccaccagtgccaggcctacaggcgaatc
X92590      > 1776     ........................................
HIRA prote > 519       E  P  V  A  A  T  S  A  R  P  T  G  E  S


mbconsphas > 75645     tgtcagtaaggacag|gttagtaaacaccaaggaccagcca
X92590      > 1816     ................
HIRA prote > 533         V  S  K  D  S


mbconsphas > 75685     tgcccttggggaaagtttaggttccagggataagagtttc

------------------intron 14, sequence 75725-77564 not shown------------

mbconsphas > 77565     atcaggtgctttgttgcttgtgtttgacgaatctcccatc


mbconsphas > 77605     tttcccccag|tatgaatgctacctctactcctgctgcatc
X92590      > 1828               ... ............................
HIRA prote > 537                   S  M  N  A  T  S  T  P  A  A  S


mbconsphas > 77645     gtcaccctctgtgttaacaaccccatccaagattgaaccc
X92590      > 1861     ........................................
HIRA prote > 548         S  P  S  V  L  T  T  P  S  K  I  E  P


mbconsphas > 77685     atgaaagcatttgattcccggttcacagaacggtccaaag
X92590      > 1901     ........................................
HIRA prote > 561        M  K  A  F  D  S  R  F  T  E  R  S  K


mbconsphas > 77725     ccacaccaggtgctccttccttgaccagtgtgattccaac
X92590      > 1941     ........................................
HIRA prote > 574       A  T  P  G  A  P  S  L  T  S  V  I  P  T
```

```
mbconsphas > 77765     agctgttgaaag|gtaggtggtagctactgtatctcctcgg
X92590     > 1981      ............ ..
HIRA prote > 588         A  V  E  R


mbconsphas > 77805     gacttcaccatgcatcctgcctcctttaggagagcaaggc

~~~~~~~~~~~~~~~~~~intron 15, sequence 77845-88884 not shown~~~~~~~~~~~~

mbconsphas > 88885     tcatagaatgaatggctttgtagacagagttacagatgtc


mbconsphas > 88925     tttgatttcag|gttgaaagagcagaacctcgtcaaggagc
X92590     > 1991                .. ..........................
HIRA prote > 591                   R   L  K  E  Q  N  L  V  K  E


mbconsphas > 88965     tgaggtcccgggaactggagagcagcagtgacagcgatga
X92590     > 2022      ........................................
HIRA prote > 601       L  R  S  R  E  L  E  S  S  S  D  S  D  E


mbconsphas > 89005     gaaggtccacctagccaagccctcttcactgtccaagcgc
X92590     > 2062      ........................................
HIRA prote > 615         K  V  H  L  A  K  P  S  S  L  S  K  R


mbconsphas > 89045     aaacttgagcttgaggtagagacggtggaaaagaagaaga
X92590     > 2102      ........................................
HIRA prote > 628        K  L  E  L  E  V  E  T  V  E  K  K  K

                                                 L (a. a. according to
                                                    genomic  sequence)
mbconsphas > 89085     aaggtcgccctaggaaggattcacgtcttttacccatgtc
X92590     > 2142      ........................t.............
HIRA prote > 641       K  G  R  P  R  K  D  S  R  F  L  P  M  S
                                                 (a. a. in mouse cDNA
                                                    X92590)

mbconsphas > 89125     tctgtccgtccag|gtgaggcctatggactgtcttgcctgt
X92590     > 2182      ............
HIRA prote > 655         L  S  V  Q

mbconsphas > 89165     cccaccctgcctccttgggaggcccttgtgtgtggagggt

~~~~~~~~~~~~~~~~~~intron 16, size not available~~~~~~~~~~~~~~~~~~~~~~~~
```

Note: Not all sequences are shown. "." indicates the identity. "|" indicates the splice site. Capitalized sequence is exon 10 which Powerblast program failed to find as mouse HIRA gene. Sequence from 39174 to 39350 and sequence from 66933 to 67039 are alternatively spliced exon 3a and 11a.

3.7.2 Comparison of the human and mouse HIRA gene

As discussed above, the human and mouse cDNA have an identity of 87.53% as determined by Cross_match and identity in the coding region of 89.80%. The identity at the amino acid level is 97.60%. Alignment of the amino acid sequences of the human (X89887) and mouse (X92590) HIRA gene products is provided in Figure 41. According to the mouse genomic sequence, amino acid 650 in the mouse HIRA protein is a Leucine which agrees with that in the human HIRA protein.

Figure 41: CLUSTAL W(1.60) sequence alignment between human (X89887) and mouse (X92590) HIRA gene products.

```
CLUSTAL W(1.60) multiple sequence alignment


hshira(X89887)      MKLLKPIWVNHNGKPIFSVDIHPDGTKFATGGQGQDSGKVVIWNMSPVLQEDDEKDENIP
mmhira(X92590)      MKLLKPIWVNHNGKPIFSVDIHPDGTKFATGGQGQDSGKVVIWNMSPVLQEDDEKDENIP
                    ************************************************************

hshira(X89887)      KMLCQMDNHLACVNCVRWSNSGMYLASGGDDKLIMVWKRATYIGPSTVFGSSGKLANVEQ
mmhira(X92590)      KMLCQMDNHLACVNCVRWSNSGMYLASGGDDKLIMVWKRATYIGPSTVFGSSGKLANVEQ
                    ************************************************************

hshira(X89887)      WRCVSILRNHSGDVMDVAWSPHDAWLASCSVDNTVVIWNAVKFPEILATLRGHSGLVKGL
mmhira(X92590)      WRCVSILRSHSGDVMDVAWSPHDAWLASCSVDNTVVIWNAVKFPEILATLRGHSGLVKGL
                    ******** ************************************************

hshira(X89887)      TWDPVGKYIASQADDRSLKVWRTLDWQLETSITKPFDECGGTTHVLRLSWSPDGHYLVSA
mmhira(X92590)      TWDPVGKYIASQADDRSLKVWRTLDWQLETSITKPCDECGGTTHVLRLSWSPDGHYLVSA
                    ********************************** ************************

hshira(X89887)      HAMNNSGPTAQIIEREGWKTNMDFVGHRKAVTVVKFNPKIFKKKQKNGSSAKPSCPYCCC
mmhira(X92590)      HAMNNSGPTAQIIEREGWKTNMDFVGHRKAVTVVKFNPKIFKKKQKNGSSTKPSCPYCCC
                    ************************************************.*********

hshira(X89887)      AVGSKDRSLSVWLTCLKRPLVVIHELFDKSIMDISWTLNGLGILVCSMDGSVAFLDFSQD
mmhira(X92590)      AVGSKDRSLSVWLTCLKRPLVVIHELFDKSIMDISWTLNGLGILVCSMDGSVAFLDFSQD
                    ************************************************************

hshira(X89887)      ELGDPLSEEEKSRIHQSTYGKSLAIMTEAQLSTAVIENPEMLKYQRRQQQQQLDQKSAAT
mmhira(X92590)      ELGDPLSEEEKSRIHQSTYGKSLAIMTEAQLSTAVIENPEMLKYQRRQQQQQLDQKNATT
                    ************************************************************ *.*

hshira(X89887)      REMGSATSVAGVVNGESLEDIRKNLLKKQVETRTADGRRRITPLCIAQLDTGDFSTAFFN
mmhira(X92590)      RETSSASSVTGVVNGESLEDIRKNLLKKQVETRTADGRSRITPLCIAQLDTGDFSTAFFN
                    **  **.**.****************************** ********************
```

```
hshira(X89887)      SIPLSGSLAGTMLSSHSSPQLLPLDSSTPNSFGASKPCTEPVVAASARPAGDSVNKDSMN
mmhira(X92590)      SIPLSSSLAGTMLSSPSGQQLLPLDSSTP-SFGASKPCTEPVAATSARPTGESVSKDSMN
                    ***** ********* *  ********** ************ *.****.*.** *****

hshira(X89887)      ATSTPAALSPSVLTTPSKIEPMKAFDSRFTERSKATPGAPALTSMTPTAVERLKEQNLVK
mmhira(X92590)      ATSTPAASSPSVLTTPSKIEPMKAFDSRFTERSKATPGAPSLTSVIPTAVERLKEQNLVK
                    ******* ********************************.***. **************

hshira(X89887)      ELRPRDLLESSSDSDEKVPLAKASSLSKRKLELEVETVEKKKKGRPRKDSRLMPVSLSVQ
mmhira(X92590)      ELRSR-ELESSSDSDEKVHLAKPSSLSKRKLELEVETVEKKKKGRPRKDSRFLPMSLSVQ
                    *** *  *********** *** ****************************** .*.*****
```

The "f" in mouse (X92590) protein disagrees with the human (X89887) protein, and according to the mouse genomic sequence, it is an "L" which agrees with the human protein.

```
hshira(X89887)      SPAALTAEKEAMCLSAPALALKLPIPSPQRAFTLQVSSDPSMYIEVENEVTVVGGVKLSR
mmhira(X92590)      SPAALSTEKEAMCLSAPALALKLPIPGPQRAFTLQVSSDPSMYIEVENEVTTVGGIRLSR
                    *****..****************** ********************** ***..***

hshira(X89887)      LKCNREGKEWETVLTSRILTAAGSCDVVCVACEKRMLSVFSTCGRRLLSPILLPSPISTL
mmhira(X92590)      LKCNREGKEWETVLSSRVLTAAGSCDVVCVACEKRMLSVFSTCGRRLLPPILLPSPISTL
                    **************.**.****************************** ***********

hshira(X89887)      HCTGSYVMALTAAATLSVWDVHRQVVVVKEESLHSILAGSDMTVSQILLTQHGIPVMNLS
mmhira(X92590)      HCTGPYVMALTAAATLSVWDVHRQVVVVKEESLHSILSGSDMTVSQILLTQHGIPVMNLS
                    **** ********************************.**********************

hshira(X89887)      DGKAYCFNPSLSTWNLVSDKQDSLAQCADFRSSLPSQDAMLCSGPLAINQGRTSNSGRQA
mmhira(X92590)      DGKAYCFNPSLSTWNLVSDKQDSLAQCADFRNSLPSQDAMLCSGPLAIIQGRTSNSGRQA
                    ******************************** *************** ***********

hshira(X89887)      ARLFSVPHVVQQETTLAYLENQVAAALTLQSSHEYRHWLLVYARYLVNEGFEYRLREICK
mmhira(X92590)      ARLFSVPHVVQQETTLAYLENQVAAALTLQSSHEYRHWLLLYARYLVNEGFEYRLREICK
                    ****************************************.******************

hshira(X89887)      DLLGPVHYSTGSQWESTVVGLRKRELLKELLPVIGQNLRFQRLFTECQEQLDILRDK
mmhira(X92590)      DLLGPVHCSTGSQWESTVVGLRKRELLKELLPVIGQNLRFQRLFTECQGQLDILWGK
                    ******* ****************************************** *****  *
```

Note: "*" indicates identity, "." indicates similarity, " " (space) indicates dissimilarity, and "-" indicates deletion/insertion. Mouse amino acid sequence does not include alternative spliced exons.

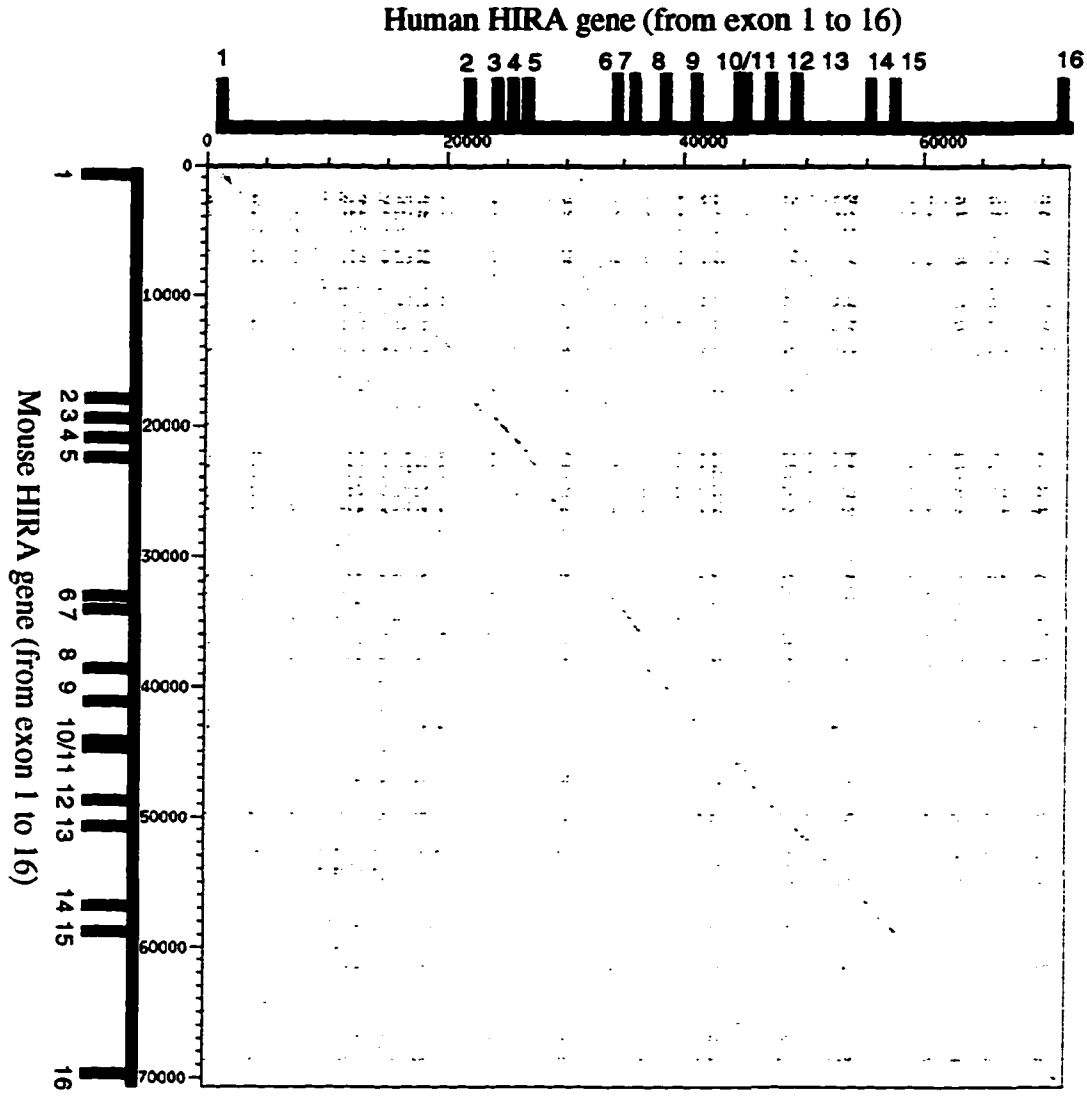


Figure 42: Dotplot comparison of the human and mouse HIRA genomic DNA. The human sequence is on X-axis and the mouse sequence is on Y-axis.

Comparison of the genomic sequences of the human and mouse HIRA genes determined as part of this dissertation research were studied in detail. The nucleotide sequences were compared by the GCG Bestfit program and the Cross_match program. The similarity between the genomic sequences in the introns of these two genes is so low that both programs cannot find any similarity except in the coding regions. The Dotplot comparison program of GCG package gives the comparison shown in Figure 42. From the Dotplot results, it can be seen that the similarity in genomic sequence level between the human and mouse HIRA gene is definitely clear but very low. The sizes of introns of the human and mouse HIRA gene are listed and compared in Table 19. Although the sizes of all introns are different, the relative distances from each exon to neighboring exons are clearly proportional in the human and mouse HIRA gene, i. e., larger introns in the human HIRA gene also are larger in the mouse gene and vice versa. This result indicates that the genomic organizations of the human and the mouse HIRA gene are similar and interestingly conserved during at least human and mouse evolution. A graphical version of the comparison of the size and position of introns of the human and mouse HIRA gene is given in Figure 43.

Table 19: Size of introns of the human and mouse HIRA genes.

intron number	size	
	human	mouse
1	20662	17033
2	2122	1617
3	1208	1133
4	1303	1574
5	7699	11198
6	1044	1302
7	2277	4213
8	2127	2368
9	3545	3227
10	666	466
11	1974	2688
12	1814	1574
13	5552	5344
14	2075	1956
15	13697	11159

Note: Only known introns are compared.

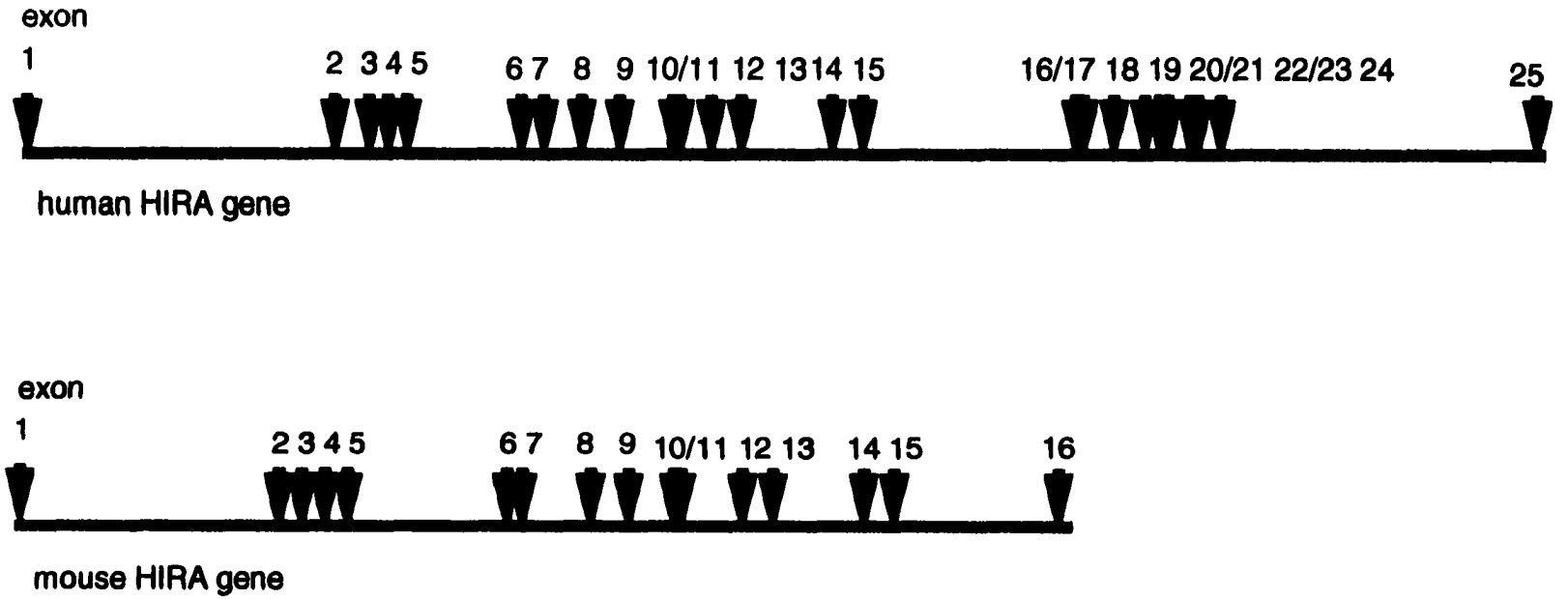


Figure 43: Comparison of size and position of introns of the human and mouse HIRA genes. The green lines are the genomic sequences of the human and mouse HIRA genes. The filled vertical arrows are exons. The numbers above the arrows are the numbers of the exons. The sequence of the rest of the mouse HIRA gene is not available.

The promoter regions of the human and mouse HIRA genes were compared by Dotplot program (see Figure 44). From this initial comparison, it can be seen that several regions 5' to the first exon are highly similar. A more detailed comparison of these two regions was performed by the GCG Bestfit program and the highly similar regions are displayed in Figure 45. The 5' region of the mouse HIRA gene was analyzed by the SIGNALSCAN program that uses the TRANSFAC database to identify the potential transcription factor binding sites. As found above with the human HIRA gene promoter region, the mouse HIRA gene promoter region lacks a TATA box but does contain several GC boxes (GGGCGG). The GC boxes bind the Sp1 transcription factor and have been shown to be required for interaction of transcription factor IID (TFIID) when TATA promoters [264, 266] are absent. There also are two putative histone H4 gene-specific transcription factors present in the mouse HIRA gene promoter region. Thus, in the highly similar promoter regions of both the human and mouse HIRA gene, potential transcription factor binding sites can be found for the: CAC-binding protein, H4TF-2 (histone H4 gene-specific transcription factor), LF-A1 (liver specific trans-acting factor), NF-1 (nuclear factor 1), NF-1/L (nuclear factor 1-like), NF-E (erythroid specific nuclear factor), Pit-1 (pituitary transcription factor-1), Sp1, and SRF (serum response factor) as shown in Figure 45.

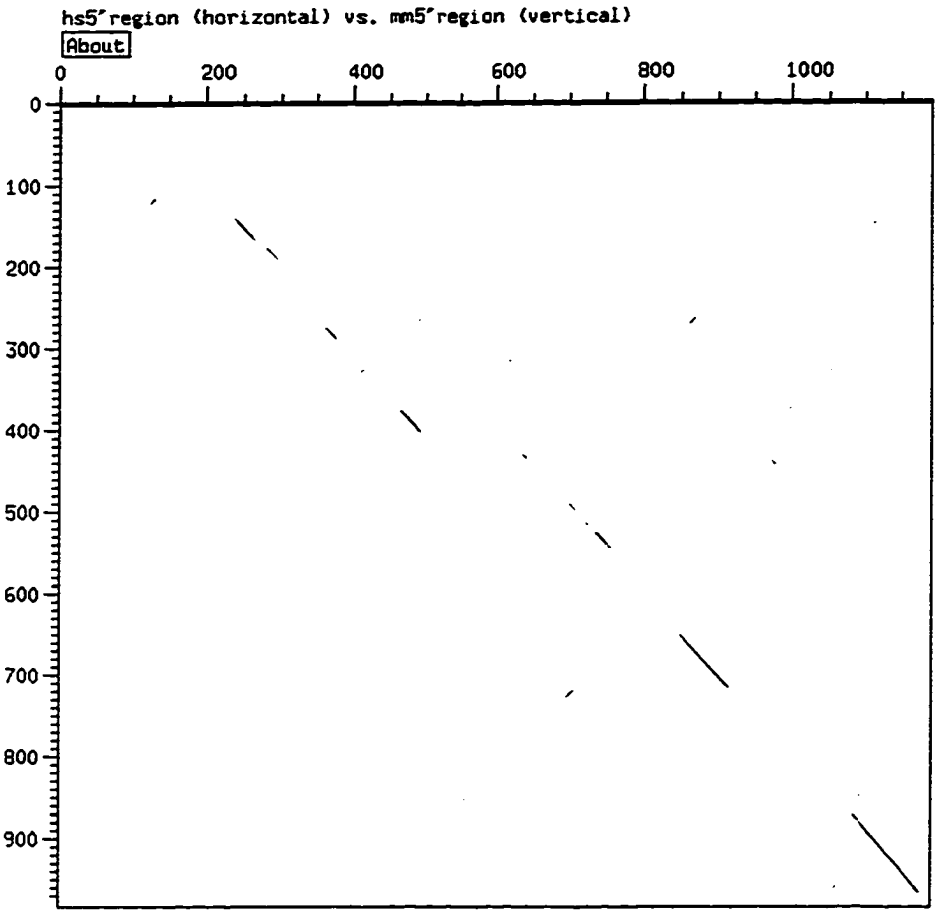
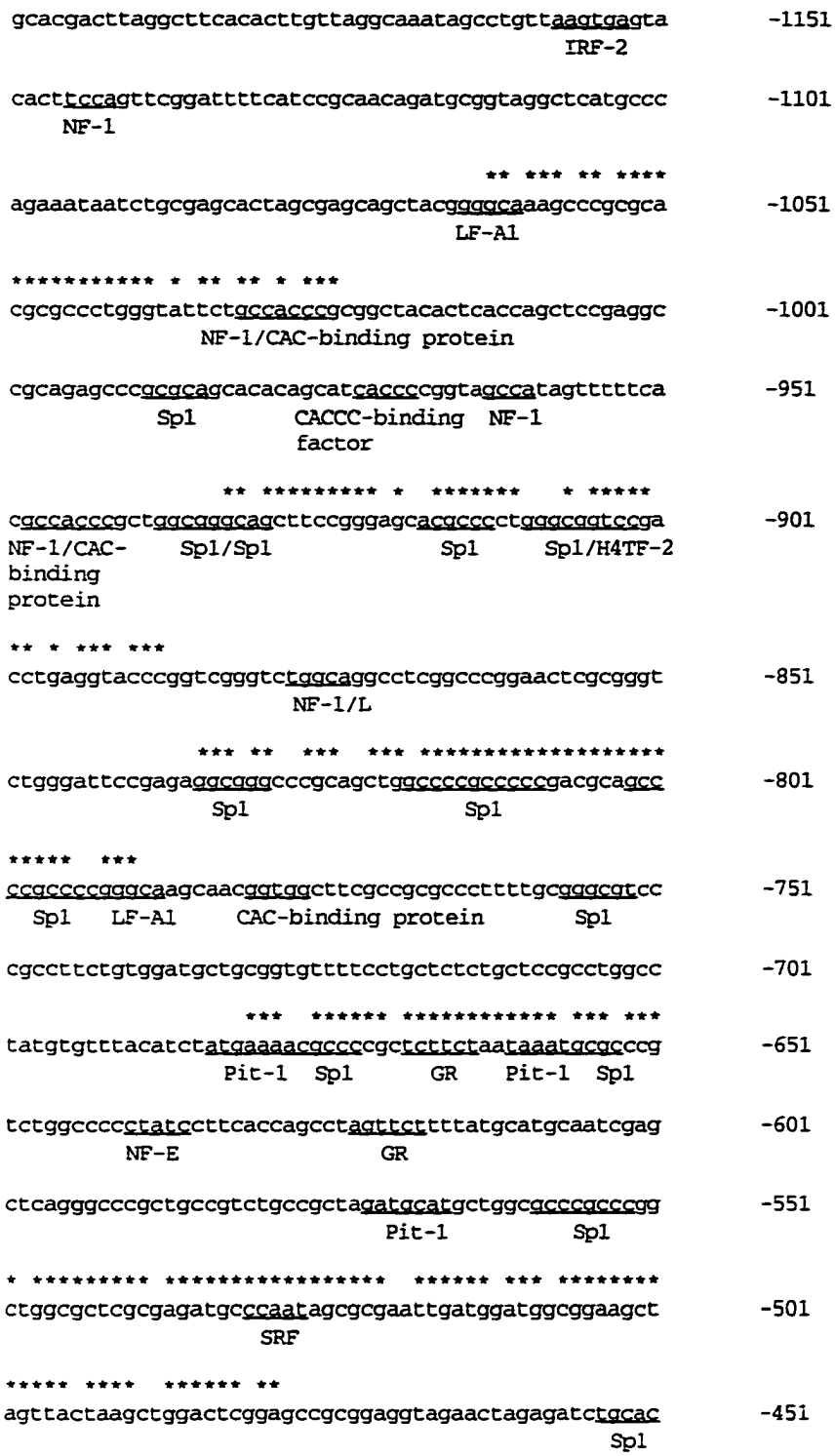


Figure 44: Dotplot comparison of promoter regions of human and mouse HIRA gene. Human sequence is horizontal and mouse sequence is vertical.

Figure 45: Promoter region of the mouse HIRA gene



```
accaccctggttctgcatcgccgggcgggatctcaaggctttgccagcct        -401
CAC-binding protein      Sp1                  NF-1

gctcacaccgaccctcatcggcgaccacatcagtggcagcggagtatccc        -351
                                  NF-1/L

          ****** *  * ** ** *** ******** ********
cgtcctcccccccctgtctaccgaagaaccgaggcggcccggtccgcgta        -301
             NF-E                 Sp1    H4TF-2

**************************************************
gggagcggtttcccgccggggcgatttggcaggtgcgcgccgtgacttcc        -251
          Sp1             NF-1/L   Sp1

******************** ********* start of exon 1
ggcgttgcccgggagccgccggaggaggag|cggcgcaggggatgcggctg       -201

tggcggcggcggcggcggccgagcgcaggaggcggctgtggcggcggctg        -151

ggggcacgggccggcgatggcgcggcggcgctgagggcaagggtgggcgg        -101

cggccggagggcgggcggcgcgggaggaagctgcggcagctccatggccc        -51

aggcgtgctgagggacacggctctggcttcagcccggcagcggccgaaca        -1

start of coding region
|ATGAAGCTCTTGAAGCCAACCTGGGTCAACCACAATG
```

Note: Position 1 is defined as starting position of coding region. Potential transcription factor binding sites are underlined. "*" above the sequence indicates identity to human sequence.

Chapter IV

Conclusion

The results presented in this dissertation provide support for the notion that a pUC plasmid-based double stranded random shotgun sequencing strategy employing genetically engineered thermally stable Taq DNA polymerase and dye-labeled terminator chemistry is feasible for large-scale genomic DNA sequencing. Through continuous improvements in the overall random shotgun sequencing strategy, this approach yields a highly accurate final sequence that could be generated in a reasonable amount of time and with a cost within the range of 50 to 60 cents per base as targeted by the NIH Human Genome Project. Specifically, as part of this dissertation research, efforts to improve the methods of DNA shearing, DNA template isolation, subcloning and sequencing were carried out. Also, by employing various different computer assembly programs, such as XGAP, FAK2, CAPII, and Phred/Phrap, and collecting new data until all gave identical assemblies, one gains confidence that the final sequence assembly is accurate. Once the sequence of a region was accurately defined, approaches that include database searches (Powerblast) and computer based exon and feature prediction (Xgrail) programs were implemented to determine the genomic organization, by locating the presence of known genes by their homology with the same or similar genes already in the database, and to predict possible new genes.

When this dissertation research began, it was questionable if the random shotgun sequencing approach was feasible to determine the sequence of large regions of the human genome. It also was not known how many shotgun sequencing reactions should be generated during the random phase and when one should switch to a more directed sequencing approach to finish a large scale DNA sequencing project more

efficiently. It was however accepted that the accuracy of a determined DNA sequence depends largely on the level of sequence redundancy, i. e., the number of times an individual base in the final DNA sequence has been determined, and confidence that a sequence is accurate also is strengthened by determining the sequence of each base from both strands, i. e., both orientations or directions. Thus, for regions where shotgun sequencing yields only reach in one orientation, directed sequencing methods could be used to sequence the opposite strand with a custom synthesized primer. However, because the approach of random shotgun sequencing is sophisticated, it is relatively easy and quick to generate a high number of random sequences once one is trained in the techniques. Considering the expenses of large scale DNA sequencing, it also is less expensive to generate random sequences rather than synthesize custom primers until the redundancy reaches a level of 5 to 6 fold and the number of gaps is small. This level has been practically determined during the course of this research to be sufficient to maintain a highly accurate sequence because most of the gel readings, which were collected after redundancy reaches a high level and most of the project is defined, will fall into the already sequenced regions and will not help to cover the gaps between the known regions and the regions with less accuracy based on probability rules. Furthermore, some of the gaps and less accurate regions exist because they are regions that, for either known or unknown reasons, are difficult to sequence. In this case, directed sequencing using different approaches is the only viable choice to obtain these sequences. Because the bottleneck of the large scale DNA sequencing is the stage of directed sequencing, much of the work in this dissertation was concentrated on reducing the expense and time needed to select and synthesize the oligonucleotides, that could be used as the primers for PCR and subsequent sequencing by primer walking. In addition, because other difficulties are encountered in the gap closure stage, this stage is considered as the most challenging and time consuming step in large scale DNA sequencing. Therefore a sophisticated, systematic strategy for gap closure is crucial for

economical and efficient large scale DNA sequencing. In the course of this dissertation, a feasible and effective gap closure strategy has been developed. This strategy entails resequencing the ends of each contig using long gels followed by primer walking at the ends of contigs with custom primers, PCR amplification to estimate the size of the gap, sequencing directly off the PCR product, and if necessary nebulizing, subcloning and sequencing the shotgun cloned PCR product.

In summary, with the recently available technology and expensive and time-consuming methods of oligonucleotide synthesis, large scale DNA sequencing can be completed after only 5 to 6-fold redundancy, followed by custom primer synthesis and directed sequencing to obtain the final 15% to 20%. This timely and cost effective approach results in a final accuracy of at least 99.99%.

During this dissertation work, the above strategy was employed to sequence four segments of human DiGeorge syndrome critical region genomic DNA covering 43,995, 211,386, 157,885, and 67,378 base pairs in contigs 1, 2, 3, and 4 respectively, and 89,500 base pairs of a mouse genomic BAC containing DNA that is syntenic to the human HIRA gene region. These studies revealed the presence of a clathrin heavy chain-like gene (CLTCL), a repressor of histone gene transcription (HIRA), a ubiquitin fusion-degradation protein 1-like gene (UFD1L), a putative gene with unknown function, and a gene homologous to the mouse T10 gene. A ribosomal L34 pseudogene and a human Tigger 1-like transposon also were found in these regions. These human genomic DNA segments have a gene density of slightly more than one gene per 100 kb and approximately 32% of repeated sequences. In addition, the mouse BAC was shown to contain 16 of the 25 exons of the mouse HIRA gene. Functionally, all of the genes in these regions are involved in various biochemical pathways, with the CLTCL most likely involved in endocytosis, the HIRA being a

negative regulator of core histone gene transcription, and the UFD1L probably involved in protein degradation. A predicted putative gene distal to the UFD1L gene encodes four putative transmembrane domains but the function of this gene is unknown, as is that of the T10 gene. The presence of different families of interspersed repetitive elements, such as Alu, L1, Mer, Mir, Mlt, and The1b also was observed. There also are countless simple sequence repeats (including di- or tri-oligonucleotides repeats or other simple combinations of tandem repeats) within these regions. [283-285]

During the course of this research, a team effort was employed to sequence the entire 1.25 mb DiGeorge syndrome critical region. Initially, the team consisted of Dr. Zhili Wang, Dr. Guozhong Zhang, the author, and more recently Axin Hua and Lingzhi Chu joined the team. Drs. Nientung Ma and Steve Toth also contributed during the early stage of this project. To date, approximately 1.25 mb of sequence covering the entire DiGeorge syndrome critical region has been completed. There are a total of 27 genes in this region, although two are ribosomal protein pseudogenes that most likely are non-functional. The minimum DiGeorge syndrome region, which is about 300 kb in size and located at the proximal end of DiGeorge syndrome region, contains 12 genes. The gene density of this minimum region is relatively high but the genes in the minimum region contain various numbers of exons. Among these 12 genes, CLTCL (CHC2) has 34 exons, the LAN, DGS-I, and CTP genes have 10 exons, 9 exons and 9 exons, respectively, the DGCR-5, Gsc1, and DGS-D genes have only 5, 3, and 2 exons, respectively, and the remaining 5 genes only have one exon. In the entire DiGeorge syndrome region, there is only one single exon gene located outside of the minimum region, and it is a putative transmembrane protein that is deleted in Velocardiofacial syndrome, named TMVCF. Interestingly, TMVCF is distal to another predicted putative transmembrane protein gene that is transcribed on the opposite strand. In total of the 27 genes, 15 are transcribed in the direction from distal

to proximal and 12 are transcribed from proximal to distal. A list of all the genes in this 1.3 mb region showing their transcription direction, number of exons and relevant references, as well as a diagram showing their relative positions in this region of human chromosome 22, are provided at the end of this chapter.

Approximately 36 kb of genomic sequences from two individuals also were sequenced from overlapping clones. When compared, these sequences differed by approximately 0.2%. Surprisingly, these regions of microheterogeneity between the two individuals are not distributed evenly but are clustered into two places. One of these two regions is within an Alu element and the other is within an L1 repeat. In the Alu region, a total of 8 variations occur within a region of 67 bp and all 8 variations are transitions (6 "t" <-> "c" and 2 "a" <-> "g"). In the L1 region, 8 variations occur within a 120 bp region, but six of these 8 variations are indel (6 "a" and 1 "g"), one is transition ("a" <-> "g") and one is transversion ("c" <-> "a").

The genomic sequences of the human and mouse HIRA genes also were determined and compared. Although the cDNAs of these two genes are highly similar, with an identity of 87.53%, and the similarity of the translated amino acid sequences is even higher, with an identity of 97.60%, the similarity of the genomic sequences is much lower. Even so, the genomic organizations of these two genes are extremely similar and conserved since the sizes of their respective introns are similar. Thus, even though these introns contain very little if any identical sequences, the large introns in the mouse HIRA gene also are large in human and vice versa. This raises the possibility of conservation of intron length and distribution in other syntenic regions of the human and mouse genome. Thus, additional comparative sequencing studies for other human and mouse genes would give support for the hypothesis that the size and organization of these introns are conserved throughout evolution. This possibility

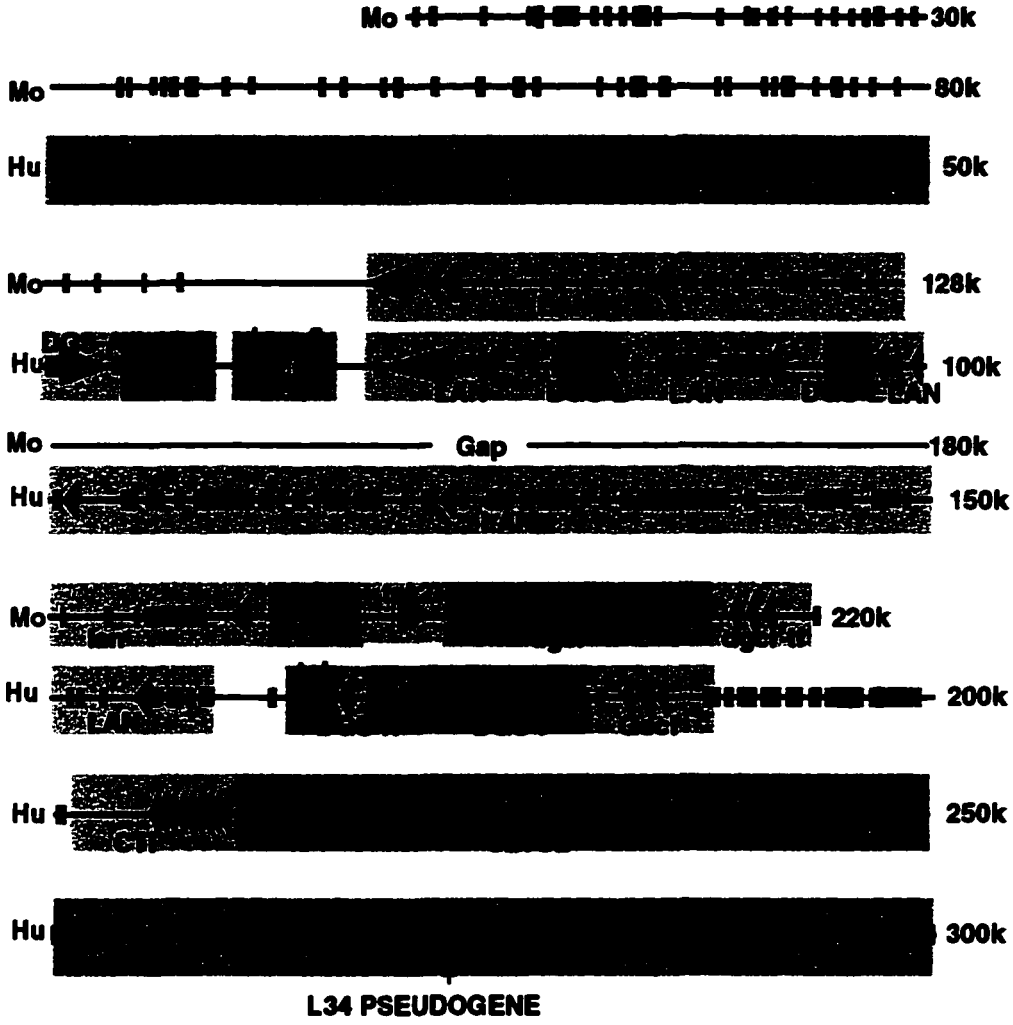
raises the question, if they are conserved, how did this conservation occur while the sequences in these syntenic introns differ? It is hypothesized here that the intronic regions evolved by a mechanism of random insertion over the millennium, after the exons first were split from each other by some unknown events. as the human and mouse genomes further evolved, the intron length of each species increased as a function of the length of the primordial initial intron by random insertions. Thus, based on probability, given a sufficient number of insertion events, each intron would be expanded in proportion to its original size. The fact that noncoding region conservation at the nucleotide level is rare, any small regions conserved in the noncoding segments most likely were maintained by functional selection. In addition, it also is likely that the proportional spacing of these regions is selected, since this spacing most likely is required for proper gene expression. The similarity of promoter regions of these two genes also is strikingly similar as both genes lack the TATA promoter element but instead contain the typical GC rich promoters of the housekeeping gene family members. The nucleotide sequence conservation observed in the regions adjacent to the transcription sites indicates that proper expression of these genes most likely requires that these sequences be conserved and that the spacing between these region may be needed to maintain their function. In conclusion, at present, large scale sequencing of genomic DNA is both robust and efficient, and through such studies, a greater understanding of the function of coding and noncoding sequences of a large complex gene region can be obtained and additional insight into previously unknown events that occurred during evolution also can be gained.

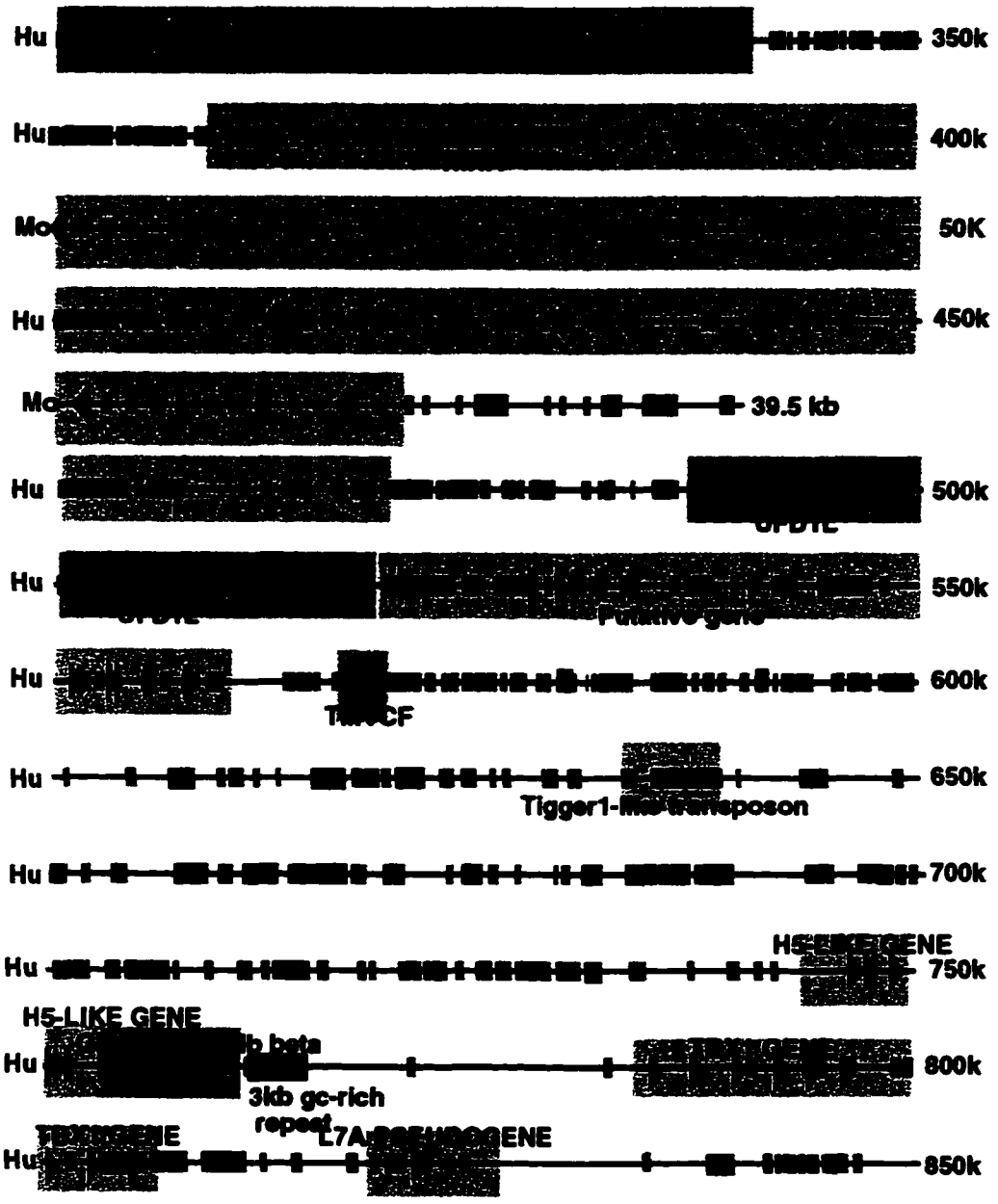
List of Genes in the DiGeorge Syndrome Critical Region

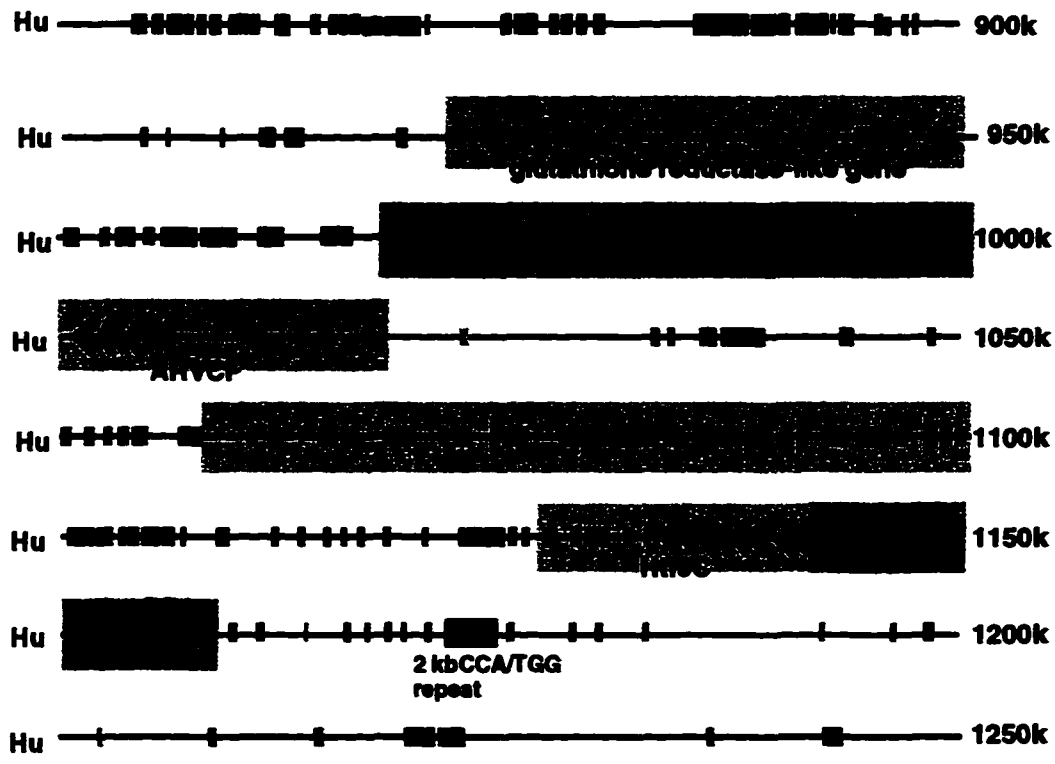
# of gene	Abb. Name	Full Name	Orientation	# of exons	reference
1.	DGCR-5		C -> T	5	57
2.	DGS-A	DiGeorge Syndrome Gene A	C -> T	1	57
3.	DGS-B	DiGeorge Syndrome Gene B	C -> T	1	56
4.	LAN	DiGeorge Syndrome LAN Gene	T -> C	10	57, 58, 59
5.	DGS-D	DiGeorge Syndrome Gene D	T -> C	2	57
6.	DGS-E	DiGeorge Syndrome Gene F	T -> C	1	57
7.	TSK-1	Threonine/Serine Kinase Gene	C -> T	1	57
8.	DGS-H	DiGeorge Syndrome Gene H	T -> C	1	57
9.	DGS-I	DiGeorge Syndrome Gene I	T -> C	9	57
10.	Gscl	Goosecoid-Like Gene	T -> C	3	283
11.	CTP	Citrate Transport Protein Gene	T -> C	9	57
12.	CHC2	Clathrin Heavy Chain 2 Gene	T -> C	34	61
13.		Ribosomal Protein L34 Pseudogene	T -> C	1	251
14.	HIRA/TUPLE	Histone Gene Transcription Repressor Gene	T -> C	25	62, 63
15.	UFD1L	Ubiquitin Fusion-Degradation Protein 1-like Gene	T -> C	12	268
16.		Putative Transmembrane protein Gene	C -> T	20	predicted
17.	TMVCF	Transmembrane Protein Deleted in VCFS	T -> C	1	284
18.	H5-like	H5 Like Gene	C -> T	6	285
19.		Platelet Glycoprotein Ib beta Gene	C -> T	5	285
20.	TBX1	T-box 1 Gene	C -> T	9	286
21.		Ribosomal Protein L7A Pseudogene	C -> T	1	287
22.		Glutathione Reductase-like Gene	T -> C	12	predicted
23.	COMT	Catechol O-methyl Transferase Gene	C -> T	6/4	69, 70, 71
24.	ARVCF	Armadillo Repeat Protein Gene	T -> C	16	288
25.	T10	Mouse T10-like Gene	C -> T	9	65
26.	Htf9C		T -> C	10	289
27.	RanBP1	Ran-Binding Protein 1 Gene	C -> T	5	290

Note: Predicted - Gene model built based on the EST matches from Genbank, Xgrail predicted exons and potential promoter.

Figure 46: Human and the Syntenic Mouse Sequences in the 1.25 Mb DiGeorge Critical Region




Hu
350k
Hu
400k
Mo
50K
Hu
450k
Mo
39.5 kb
Hu
500k
Hu
550k
Hu
600k
Hu
650k
Tigger1-
Hu
700k
Hu
750k
H5-LIKE GENE
Hu
800k
3kb gc-rich
repeat
Hu
850k


Hu
900k
Hu
950k
Hu
1000k
Hu
1050k
Hu
1100k
Hu
1150k
Hu
1200k
2 kbCCA/TGG
repeat
Hu
1250k
Human genomic
sequence
Mouse genomic
sequence
Repeat sequences
Exons
EST hits
STS hits
special repeat element

Chapter V

Literature Cited

1. Cooper, M. D.; Peterson, R. D. A.; Good, R. A.: A new concept of the cellular basis of immunology. J. Pediat. 67: 907-908, 1965.

2. DiGeorge, A. M.: Congenital absence of the thymus and its immunologic consequences: concurrence with congenital hypothyroidism. In: Good, R. A.; Bergsma, D.: Birth Defects Orig. Art. Ser.. New York: National Foundation-March of Dimes (pub.) 4 1968. Pp. 116-121.

3. Strong, W. B.: Familial syndrome of right-sided aortic arch, mental deficiency, and facial dysmorphism. J. Pediat. 73: 882-888, 1968.

4. Kinouchi, A.; Mori, K.; Ando, M.; Takao, A.: Facial appearance of patients with conotruncal abnormalities. Pediat. Jpn. 17: 84, 1976.

5. Takao, A.; Ando, M.; Cho, K.; et al.: Etiologic categorization of common congenital heart diseases. In: Van Praagh, R.; Takao, A. (eds.): Etiology and Morphogenesis of Congenital Heart Disease. Mount Kisco, New York: Futura Publishing (pub.) 1980. Pp. 262-269.

6. Shimizu, T.; Takao, A.; Ando, M.; Hirayama, A.: Conotruncal face syndrome: its heterogeneity and association with thymus involution. In: Nora, J. J.; Takao, A. (eds): Congenital Heart Disease: Cause and Processes. Mount Kisco, New York: Futura Publishing (pub.) 1984. Pp. 29-41.

7. Wilson, D. I.; Burn, J.; Scambler, P.; Goodship, J.: DiGeorge Syndrome, part of CATCH 22. J. Med. Genet. 30: 852-856, 1993.

8. Shprintzen, R. J.: Velocardiofacial syndrome and DiGeorge syndrome. (letter) J. Med. Genet. 31: 423-424, 1994.

9. Hall, J. G.: CATCH 22. J. Med. Genet. 30: 801-802, 1993.

10. Driscoll, D. A.; Salvin, J.; Sellinger, B.; Budarf, M. L.; McDonald-McGinn, D. M.; Zackai, E. H.; Emanuel, B. S.: Prevalence of 22q11 microdeletions in DiGeorge and Velocardiofacial syndromes: implication for genetic counselling and prenatal diagnosis. J. Med. Genet. 30: 813-817, 1993.

11. Taitz, L. S.; Sarate-Salvador, C.; Schwartz, E.: Congenital absence of the parathyroid and thymus glands in an infant (III and IV pouch syndrome). Pediat. 38: 412, 1966.

12. Kretschmer, R.; Say, B.; Brown, D.; Rosen, F. S.: Congenital aplasia of the thymus gland (DiGeorge syndrome). New Eng. J. Med. 279: 1295-1301, 1968.

13. Beckwith, J. B.; Mancer, J. F. K.; Gorin; L.; Tenckhoff, L.: Aorta arch anomalies with thymic hypoplasia or agenesis. Teratology 3: 198, 1970.

14. Harvey, J. C.; Dungan, W. T.; Elders, J. J.; Hughes, E. R.: Third and fourth pharyngeal pouch syndrome, associated vascular anomalies and hypocalcemic seizures. Clin. Pediat. 9: 496-499, 1970.

15. Freedom, R. M.; Rosen, F. S.; Nadas, A. S.: Congenital cardiofacial disease and anomalies of the third and fourth pharyngeal pouch. Circulation 46: 165-172, 1972.

16. Gatti, R. A.; Gershanik, J. J.; Levkoff, A. H.; Wertelecki, W.; Good, R. A.: DiGeorge syndrome associated with combined immunodeficiency. J. Pediat. 81: 920-926, 1972.

17. Steele, R. W; Limas, C.; Thurman, G. B.; Schuelein, M.; Bauer, H.; Bellanti, J. A.: Familial thymus aplasia: attempted reconstitution with fetal thymus in a millipore diffusion chamber. New Eng. J. Med. 287: 787-791, 1972.

18. Rose, J. S.; Levin, D. C.; Goldstein, S.; Laster, W.: Congenital absence of the pulmonary valve associated with congenital aplasia of the thymus (DiGeorge syndrome). Am. J. Roentgenol. 122: 97-102, 1974.

19. Vesterhus, P.; Eide, J.; Froland, S. S.; Haneberg, B.; Jacobsen, K. B.: Maldescent of the thymus in a hypoparathyroid infant with pharyngeal pouch syndrome. Acta. Paediat. Scand. 64: 555-558, 1975.

20. Finley, L. P.; Collins, G. F.; De Chadarevian, J. P.; Williams, R. L.: DiGeorge syndrome presenting as severe congenital heart disease in the newborn. Can. Med. Assoc. J. 116: 635-640, 1977.

21. Conley, M. E.; Beckwith, J. B.; Mancer, J. F. K.; Tenckhoff, L.: The spectrum of the DiGeorge syndrome. J. Pediat. 94: 883-890, 1979.

22. Raatikka, M.; Rapola, J.; Tuuteri, L.; Louhimo, I.; Savilahti, E.: Familial third and fourth pharyngeal pouch syndrome with truncus arteriosus: DiGeorge syndrome. Pediatrics 67: 173-175, 1981.

23. Atkin, J. F.; Hsia, Y. E.; Sommer, A.: Familial DiGeorge syndrome in 7 children. Am. J. Hum. Genet. 34: 80A, 1982.

24. Rohn, R. D.; Leffell, M. S.; Leadem, P.; Johnson, D.; Rubio, T.; Emanuel, B. S.: Familial third-fourth pharyngeal pouch syndrome with apparent autosomal dominant transmission. J. Pediat. 105: 47-51, 1984.

25. Keppen, L. D.; Fasules, J. W.; Burks, A. W.; Gollin, S. M.; Sawyer, J. R.; Miller, C.: Confirmation of autosomal dominant transmision of the DiGeorge malformation complex. J. Pediat. 113: 506-508, 1988.

26. Stevens, C. A.; Carey, J. C.; Ahigeoka, A. O.: DiGeorge anomaly and velocardiofacial syndrome. Pediatrics 85: 526-530, 1990.

27. Wilson, D. I.; Cross, I. E.; Goodship, J. A.; Coulthard, S.; Carey, A. H.; Scambler, P. J.; Bain. H. H.; Hunter, A. S.; Carter, P. E.; Burn, J.: DiGeorge syndrome with isolated aortic coarctation and isolated ventricular septal defect in three sibs with a 22q11 deletion of maternal origin. Brit. Heart J. 66: 308-312, 1991.

28. de la Chapelle, A.; Herva, R.; Koivisto, M.; Aula, P.: A deletion in chromosome 22 can cause DiGeorge syndrome. Hum. Genet. 57: 253-256, 1981.

29. Rosenthal, I. M.; Bocian, M.; Krmpotik, E.: Multiple anomalies including thymic aplasia associated with monosomy 22. Pediat. Res. 6: 358A, 1972.

30. Kelly, R. I.; Zackai, E. H.; Emanuel, B. S.; Kistenmacher, M.; Greenberg, F.; Punnett, H. H.: The association of the DiGeorge anomaly with partial monosomy of chromosome 22. J. Pediat. 101: 197-200, 1982.

31. Greenberg, F.; Crowder, W. E.; Paschall, V.; Colon-Linares, J.; Lubianski, B.; Ledbetter, D. H.: Familial DiGeorge syndrome and associated partial monosomy of chromosome 22. Hum. Genet. 65: 317-319, 1984.

32. Augusseau, S.; Jouk, S.; Jalbert, P.; Prieur, M.: DiGeorge syndrome and 22q11 rearrangements. (letter) Hum. Genet. 74: 206, 1986.

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

34. Wilson, D. I.; Cross, I. E.; Goodship, J. A.; Brown, J.; Scambler, P. J.; Bain, H. H.; Taylor, J. F. N.; Walsh, K.; Bankier, A.; Burn, J.; Wolstenholme, J.: A prospective cytogenetic study of 36 cases of DiGeorge syndrome. Am. J. Hum. Genet. 51: 957-963, 1992.

35. Carey, A. H.; Kelly, D.; Halford, S.; Wadey, R.; Wilson, D.; Goodship, J. Burn, J.; Paul, T.; Sharkey, A.; Dumanski, J.; Nordenskjold, M.; Williamson, R.; Scambler, P. J.: Molecular genetic study of the frequency of monosomy 22q11 in DiGeorge syndrome. Am. J. Hum. Genet. 51: 964-970, 1992.

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

37. Monaco, G.; Pignata, C.; Rossi, E.; Mascellaro, O.; Cocozza, S.; Ciccimarra, F.: DiGeorge anomaly associated with 10p deletion. Am. J. Med. Genet. 39: 215-216, 1991.

38. Shapira, M.; Borochowitz, Z.; Bar-El, H.; Dar, H.; Etzioni, A.; Lorber, A.: Deletion of the short arm of chromosome 10 (10p13): report of a patient and review. Am. J. Med. Genet. 52: 34-38, 1994.

39. Schuffenhauer, S; Seidel, H.; Oechsler, H.; Helohradsky, B.; Bernsau, U.; Murken, J.; Meitinger, T.: DiGeorge syndrome and partial monosomy 10p: case report and review. Ann. Genet. 38: 162-167, 1995.

40. Lindgren, V.; Rosinsky, B.; Chin, J.; Berry-kravis, E.: Two patients with overlapping *de novo* duplication of the long arm of chromosome 9, including one case with DiGeorge syndrome. Am. J. Med. Genet. 49: 67-73, 1994.

41. Greenberg, F.; Courtney, K. B.; Wessel, R. A.; Huhta, J.; Carpenter, R. J.; Rich, D. C.; Ledbetter, D. H.: Prenatal diagnosis of deletion 17p13 associated with DiGeorge anomaly. Am. J. Med. Genet. 31: 1-4, 1988.

42. van Essen, A. J.; Schoots, C. J.; van Lingen, R. A.; Mourits, M. J.; Tuerlings, J. H.; Leegte, B.: Isochromosome 18q in a girl with holoprosencephaly, DiGeorge anomaly, and streak ovaries. Am. J. Med. Genet. 47: 85-88, 1993.

43. Lai, M. M. R.; Scriven, P. N.; Ball, C.; Berry, A. C.: Simultaneous partial monosomy 10p and trisomy 5q in a case of hypoparathyroidism. J. Med. Genet. 29: 586-588, 1992.

44. Fukushima, Y.; Ohashi, H.; Wakui, K.; Nishida, T.; Nakamura, Y.; Hoshino, K.; Ogawa, K.; Oh-ishi, T.: DiGeorge syndrome with del(4)(q21.3q25): possibility of the fourth chromosome region responsible for DiGeorge syndrome. (Abstract) Am. J. Hum. Genet. 51 (suppl.): A80, 1992.

45. Faced, M. J.; Robertson, J.; Beck, J. S.; Cater, J. I.; Bose, B.; Maldlom, M. M.: Features of DiGeorge syndrome in a child with 45,XX,-3,-22,-der(3),t(3,22)(p25;q11). J. Med. Genet. 24: 225-227, 1987.

46. Dallapiccola, B.; Marino, B.; Giannotti, A.; Valorani, G.: DiGeorge anomaly associated with partial deletion of chromosome 22. Report of a case with X/22 translocation and review of the literature. Ann. Geneti. 32: 92-96, 1989.

47. Pinto, M. R.; Leite, R. P.; Areias, A.: Features of Turner's and DiGeorge's syndrome in a child with an X;22 translocation. J. Med. Genet. 26: 778-780, 1989.

48. Couillin, P.; Zucman, J.; Le Guern, E.; Reguigne, I.; Grisard, M. C.; Delattre, O.; Novelli, G.; Dallapiccola, B.; Boue, A.: Molecular studies of a translocated (X;22) DiGeorge patient using somatic cell hybridization. Ann. Genet. 35: 140-145, 1992.

49. Lupski, J. R.; Langston, C.; Friedman, R.; Ledbetter, D. H.; Greenberg, F.: DiGeorge anomaly associated with a *de novo* Y;22 translocation resulting in monosomy del(22)(q11.2). Am. J. Med. Genet. 40: 196-198, 1991.

50. Fibison, W. J.; Emanuel, B. S.: Molecular mapping in DiGeorge syndrome. (Abstract) Am. J. Hum. Genet. 41: A119, 1987.

51. Fibison, W. J.; Budarf, M.; McDermid, H.; Greenberg, F. F.; Emanuel, B. S.: Molecular studies of DiGeorge syndrome. Am. J. Hum. Genet. 46, 888-895, 1990.

52. Desmaze, C.; Prieur, M.; Amblard, F.; Aikem, M.; LeDeist, F.; Demczuk, S.; Zucman, J.; Plougastel, B.; Delattre, O.; Croquette, M.-f.; Breviere, G.-M.;Huon, C.; Le Merrer, M.; Mathieu, M.; Sidi, D.; Stephan, J.-L.; Aurias, A.: Physical mapping by FISH of the DiGeorge critical region (DGCR): involvement of the region in familial cases. Am. J. Hum. Genet. 53: 1239-1249, 1993.

53. Scambler, P.; Dumanski, J. P.; Nordenskjold, M.; Williamson, R.; CArey, A.: Molecular detection of 22q11 deletion in patients with DiGeorge syndrome and normal karyotype. (Abstract) Am. J. Hum. Genet. 47 (suppl.): A235, 1990.

54. Emanuel, B. S.; Budarf, M. L.; Sellinger, B.; Goldmuntz, E.; Driscoll, D. A.: Detection of microdeletions of 22q11.2 with fluorescence *in situ* hybridization (FISH): diagnosis of DiGeorge syndrome (DGS), velo-cardio-facial (VCF) syndrome, CHARGE association and conotruncal cardiac malformations. Am. J. Hum. Genet. 51: 3A, 1992.

55. Lindsay, E. A.; Halford, S.; Wadey, R.; Scambler, P. J.; Baldini, A.: Molecular cytogenetic characterization of the DiGeorge syndrome region using fluorescence *in situ* hybridization. Genomics 17: 403-407, 1993.

56. Budarf, M. L.; Collins, J.; Gong, W.; Roe, B.; Wang, Z.; Bailey, L. C.; Sellinger, B.; Michaud, D.; Driscoll, D. A.; Emanuel, B. S.: Cloning a balanced translocation associated with DiGeorge syndrome and identification of a disrupted candidate gene. Nature Genet. 10: 269-278, 1995.

57. Gong, W.; Emanuel, B. S.; Collins, J.; Kim, D. H.; Wang, Z.; Chen, F.; Zhang, G.; Roe, B.; Budarf, M. L.: A transcription map of the DiGeorge and velo-cardio-facial syndrome minimal critical region on 22q11. Hum. Mol. Genet. 5: 789-800, 1996.

58. Demczuk, S.; Aledo, R.; Zucman, J.; Delattre, O.; Desmaze, C.; Dauphinot, L.; Jalbert, P.; Rouleau, G. A.; Thomas, G.; Aurias, A.: Cloning of a balanced translocation breakpoint in the DiGeorge syndrome critical region and isolation of a novel potential adhesion receptor gene in its vicinity. Hum. Mol. Genet. 4: 551-558, 1995.

59. Wadey, R.; Daw, S.; Taylor, C.; Atif, U.; Kamath, S.; Halford, S.; O'Donnell, H.; Wilson, D.; Goodship, J.; Burn, J.; Scambler, P.: Isolation of a gene encoding an integral membrane protein from the vicinity of a

balanced translocation breakpoint associated with DiGeorge syndrome. Hum. Mol. Genet. 4: 1027-1033, 1995.

60. Cannizzaro, A. L.; Aroson, M. M.; Emanuel, B. S.: *In situ* hybridization and translocation breakpoint mapping II. Two unusual t(21;22) translocations. Cytogenet. Cell Genet. 39: 173-178, 1995.

61. Holmes, S. E.; Riazi, M. A.; Gong, W.; McDermid, H. E.; Sellinger, B. T.; Hua, A.; Chen, F.; Wang, Z.; Zhang, G.; Roe, B.; Gonzalez, I.; McDonald-McGinn, D. M.; Zackai, E.; Emanuel, B. S.; Budarf, M. L.: Disruption of the clathrin heavy chain-like gene (CLTCL) associated with features of DGS/VCFS: a balanced (21;22)(p12;q11) translocation. Hum. Mol. Genet. 6: 357-367, 1997.

62. Halford, S.; Wadey, R.; Roberts, C.; Daw, S. C. M.; Whiting, J. A.; O'Donnell, H.; Dunham, I.; Bentley, D.; Lindsay, E.; Baldini, A.; Francis, F.; Lehrach, H.; Williamson, R.; Wilson, D. I.; Goodship, J.; Cross, I.; Burn, J.; Scambler, P. J.: Isolation of a putative transcriptional regulator from the region of 22q11 deleted in DiGeorge syndrome, Shprintzen syndrome and familial congenital heart disease. Hum. Mol. Genet. 2: 2099-2107, 1993.

63. Lamour, V.; Lecluse, Y.; Desmaze, C.; Spector, M.; Bodescot, M.; Aurias, A.; Osley, M. A.; Lipinski, M.: A human homolog of the *S. cerevisiae* HIR1 and HIR2 transcriptional repressors cloned from the DiGeorge syndrome critical region. Hum. Mol. Genet. 4: 791-799, 1995.

64. Aubry, M.; Demczuk, S.; Desmaze, C.; Aikem, M.; Aurias, A.; Julien, J. P.; Rouleau, G. A.: Isolation of a zinc finger gene consistently deleted in DiGeorge syndrome. Hum. Mol. Genet. 2: 1583-1587, 1993.

65. Halford, S.; Wilson, D. I.; Daw, S. C. M.; Roberts, C.; Wadey, R.; Kamath, S.; Wickremasinghe, A.; Burn, J.; Goodship, J.; Mattei, M.-G.; Moormon, A. F. M.; Scambler, P. J.: Isolation of a gene expressed during early embryogenesis deleted in DiGeorge syndrome. Hum. Mol. Genet. 2: 1577-1582, 1993.

66. Kurahashi H.; Akagi, K.; Inazawa, J.; Ohta, T.; Niikawa N.; Kayatani, F.; Sano, T.; Okada, S.; Nishisho, I.: Isolation and characterization of a novel gene deleted in DiGeorge syndrome. Hum. Mol. Genet. 4: 541-549, 1995.

67. Pizzuti, A.; Novelli, G.; Mari, A.; Ratti, A.; Colosimo, A.; Amati, F.; Penso, D.; Sangiuolo, F.; Calabrese, G.; Palka, G.; Silani, V.; Gennarelli, M.; Mingarelli, R.; Scambler, P.; Dallapiccola, B.: Human homologue sequences to the *Drosophila* dishevelled segments-polarity gene are deleted in the DiGeorge syndrome. Am. J. Hum. Genet. 58: 722-729, 1996.

68. Demczuk, S.; Thomas, G.; Aurias, A.: Isolation of a novel gene from the DiGeorge syndrome critical region with homology to *Drosophila* gdl and to human LAMC1 genes. Hum. Mol. Genet. 5: 633-638, 1996.

69. Brahe, C.; Bannetta, P.; Meera, K. P.; Arwert, F.; Serra, A.: Assignment of the catechol-O-methyltransferase gene to human chromosome 22 in somatic cell hybrids. Hum. Genet. 74: 230-234, 1986.

70. Winqvist, R.; Lundstrom, K.; Salminen, M.; Laatikainen, M.; Ulmanen, I.: The human catechol-O-methyltransferase (COMT) gene maps to band q11.2 of chromosome 22 and shows a frequent RFLP with BglI. Cytogenet. Cell Genet. 59: 253-257, 1992.

71. Lachman, H. M.; Morrow, B.; Shprintzen, R.; Veit, S.; Parsia, S. S.; Faedda, G.; Goldberg, R.; Kucherlapati, R.; Papolos, D. F.: Association of codon 108/158 catechol-O-methyltransferase gene polymorphism with the psychiatric manifestation of velo-cardio-facial syndrome. Am. J. Med. Genet. 67: 468-472, 1996.

72. Demczuk, S.; Levy, A.; Aubry, M.; Croquette, M. F.; Philip, N.; Prieur, M.; Sauer, U.; Bouvagnet, P.; Rouleau, G. A.; Thomas, G.; Aurias, A.: Excess of deletions of maternal origin in the DiGeorge/velo-cardio-facial syndromes. A study of 22 new patients and review of the literature. Hum. Genet. 96: 9-13, 1995.

73. Chisaka, O.; Capecchi, M. R.: Regionally restricted developmental defects resulting from targeted disruption of the mouse homeobox gene hox-1.5. Nature 350: 473-479, 1991.

74. Ammann, A. J.; Wara, D. W.; Cowan, M. J.; Barrett, D. J.; Stiehm, E. R.: The DiGeorge syndrome and the fetal alcohol syndrome. Am. J. Dis. Child. 136: 906-908, 1982.

75. Lammer, E. J.; Chen, D. T.; Hoar, R. M.; Agnish, N. D.; Benke, P. J.; Braun, J. T.; Curry, C. J.; Fernhoff, P. M.; Grix, A. W., Jr.; Lott, I. T.; Richard, J. M.; Sun, S. C.: Retinoic acid embryopathy. New Eng. J. Med. 313: 837-841, 1995.

76. Burn, J.; Wilson, D. I.; Cross, I.; Atif, U.; Scambler, P.; Takao, A.; Goodship, J.: The clinical significance of 22q11 deletion. In: Clark, E. B.; Markwald, R. R.; Takao, A. (eds.): Developmental Mechanisms of Heart Disease. Armonk, New York: Futura Publ. (pub.) 1995. Pp. 559-567.

77. Sutherland, H. F.; Wadey, R.; McKie, J. M.; Taylor, C.; Atif, U.; Johnson, K. A.; Halford, S.; Kim, U. J.; Goodship, J.; Baldini, A.; Scambler, P. J.: Identification of a novel transcript disrupted by a balanced translocation associated with DiGeorge syndrome. Am. J. Hum. Genet. 59: 23-31, 1996.

78. Goldmuntz, E.; Wang, Z.; Roe, B. A.; Budarf, M. L.: Cloning, genomic organization, and chromosomal localization of human citrate transport protein to the DiGeorge/velocardiofacial syndrome minimal critical region. Genomics 33: 271-276, 1996.

79. Watson, J. D.: Molecular Biology of the Gene, 3rd Ed.. W. A. Benjamin, Inc. (pub.) Menlo Park, Ca. 1976.

80. Watson, J. D.; Hopkins, N. H.; Roberts, J. W.; Steitz, J. A.; Weiner, A. M.: Molecular Biology of the Gene, 4th ed.. The Benjamin/Cummings Publishing Company, Inc. Menlo Park, Ca. 1988.

81. Lewin, B.: Genes IV. Oxford University Press (pub.) Oxford. 1990.

82. Sutton, W. S.: The chromosomes in heredity. Biol. Bulletin 4: 231-251, 1903.

83. Morgan, T. H.: Sex-linked inheritance in *Drosophila*. Science 32: 120-122, 1910.

84. Morgan, T. H.; Sturtevant, A. H.; Muller, H. J.; Bridges, C. B.: The Mechanism of Mendelian Heredity. Holt, Rinehart & Winston. (pub.) New York. 1915.

85. Garrod, A. E.: Inborn Errors of Metabolism. Lancet 2: 1-7, 73-79, 142-148, 214-220, 1908.

86. Beadle, G. W.; Tatum, E. L.: Genetic control of biochemical reactions in *Neurospora*. Proc. Natl. Acad. Sci., USA 27: 499-506, 1941.

87. Griffith, F.: The significance of Pneumococcal types. J. Hyg. 27: 113-159, 1928.

88. Avery, O. T.; MacLeod, C. M.; McCarty, M.: Studies on the chemical nature of the substance inducing transformation of Pneumococcal types. J. Exp. Med. 79: 137-158, 1944.

89. Hershey, A. D.; Chase, M.: Independent functions of viral protein and nucleic acid in growth of bacteriophage. J. Gen. Physiol. 36: 39-52, 1952.

90. Chargaff, E.; Lipschitz, R.; Green, C.; Hodes, M. E.: The composition of the deoxyribonucleic acid of salmon sperm. J. Biol. Chem. 192: 223-230, 1951.

91. Watson, J. D.; Crick, F. H. C.: Molecular structure of nucleic acid. A structure of the deoxyribose nucleic acid. Nature 171: 737-738, 1953.

92. Watson, J. D.; Crick, F. H. C.: Genetical implications of the structure of deoxyribonucleic acid. Nature 171: 964-967, 1953.

93. Wilkins, M. H. F.; Strokes, A. R.; Wilson, H. R.: Molecular structure of deoxypentose nucleic acid. Nature 171: 738-740, 1953.

94. Arnott, S.; Hutchinson, F.; Spencer, M.; Wilkins, M. H. F.; Fuller, W.; Langridge, R.: X-ray diffraction studies of double helical ribonucleic acid. Nature 211: 227-232, 1966.

95. Wang, A. H.-J.; Fujii, S.; van Boom, J. H.; Rich, A.: Molecular structure of the octamer d(G-G-C-C-G-G-C-C-): modified A-DNA. Proc. Nat. Acad. Sci., USA. 79: 3968-3972, 1982.

96. Wang, A. H.-J.; Fujii, S.; van Boom, J. H.; van der Marel, G. A.; van Boeckel, S. A. A.; Rich, A.: Molecular structure of r(GCG)d(TATACGC): A DNA-RNA hybrid helix joined to double helical DNA. Nature 299: 601-604, 1982.

97. Dickerson, R. E.: Base sequence and helix structure variation in B and A DNA. J. Mol. Biol. 166: 419-441, 1983.

98. Conner, B. N.; Yoon, C.; Dickerson, J. L.; Dickerson, R. E.: Helix geometry and hydration in an A-DNA tetramer: CCGG. J. Mol. Biol. 174: 663-695, 1984.

99. Poll, F. M.; Jovin, T. M.: Salt-induced co-operative conformational change of a synthetic DNA: Equilibrium and kinetic studies with poly(dG-dC). J. Mol. Biol. 67: 375-396, 1972.

100. Wang, A. H.-J.; Quigley, G. J.; Kolpak, F. J.; Crowford, J. L.; van Boom, J. H.; van der Marel, G.; Rich, A.: Molecular structure of a left-handed DNA fragment at atomic resolution. Nature 282: 680-686, 1979.

101. Rich, A.; Nordheim, A.; Wang, A. H.-J.: The chemistry and biology of left-handed Z-DNA. Ann. Rev. Biochem. 53: 791-846, 1984.

102. Azorin, F.; Rich, A.: Isolation of Z-DNA binding protein from SV40 minichromosomes: Evidence for binding to the viral control region. Cell 41: 365-374, 1985.

103. Kmiec, E. B.; Angelides, K. J.; Holloman, W. K.: Left-handed DNA and the synaptic pairing reaction promoted by ustilago Rec1 protein. Cell 40: 139-145, 1985.

104. Behe, M.; Felsenfeld, G.: Effects of methylation on synthetic polynucleotides. The B-Z transition in poly(dG-m5dC) poly(dG-m5dC). Proc. Nat. Acad. Sci., USA 78: 1619-1623, 1981.

105. Doerfler, W.: DNA methylation and gene activity. Ann. Rev. Biochem. 52: 93-124, 1983.

106. Miller, F. D.; Rattner, J. B.; van der Sande, J. H.: Nucleosome-core assembly on B and Z forms of poly[d(G-m5C)]. Cold Spring Harber Symp. Quant. Biol. 47: 571-576, 1983.

107. Crick, F. H. C.: On protein synthesis. Symp. Soc. Exp. Biol. 12: 548-555, 1958.

108. Brachet, J.: Biochemical Embryology. Academic Press, New York. 1957.

109. Caspersson, T.: Cell Growth and Cell Function. Norton, New York. 1950.

110. Kornberg, A.: Biological synthesis of deoxyribonucleic acid. Science 131: 1503-1508, 1960.

111. McMacken, R.; Ueda, K.; Kornberg, A.: Migration of Escherichia coli dnaB protein on the template DNA strand as a mechanism in initiating DNA replication. Proc. Nat. Acad. Sci., USA 74: 4190-4194, 1977.

112. Arai, K.; Kornberg, A.: Unique primed start of phage ϕX174 DNA replication and mobility of the primosome in a direction opposite chain synthesis. Proc. Nat. Acad. Sci., USA 78: 69-73, 1981.

113. Okazaki, T.; Okazaki, R.: Mechanism of DNA chain growth. IV. Direction of synthesis of T4 short DNA chains as revealed by exonucleolytic degradation. Proc. Nat. Acad. Sci., USA 64: 1242-1248, 1969.

114. Watson, J. D.: Involvement of RNA in synthesis of protein. Science 140: 17-26, 1963.

115. Hoagland, M. B.; Stephenson, M. L.; Scott, J. F.; Hecht, L. I.; Zamecnik, P. C.: A soluble ribonucleic acid intermediate in protein synthesis. J. Biol. Chem. 231: 241-257, 1958.

116. Burgess, R.; Travers, A. A.; Dunn, J. J.; Bautz, E. K. F.: Factor stimulating transcription by RNA polymerase. Nature 222: 537-540, 1969.

117. Rodriguez, R.; Chamberlin, M.; eds.: Promoters: Structure and Function. New York: Praeger. 1982.

118. Youderian, P.; Bouvier, S.; Susskind, M.: Sequence derterminants of promoter activity. Cell 30: 843-853, 1982.

119. Hawley, D. K.; McClure, W. R.: Compilation and analysis of *Escherichia coli* promoter DNA sequences. Nucleic Acid Res. 11: 2237-2255, 1983.

120. Shenk, T.: Transcriptional control Regions: Nucleotide sequence requirements for initiation by RNA polymerase II and III. Current Topics in Microbiology and Immunology, Vol. 93. New York: Springer-Verlag, pp. 25-46, 1981.

121. Weil, P. A.; Luse, D. S.; Segall, J.; Roeder, R. G.: Selective and accurate initiation of Transcription of the AD2 major late promoter in a soluble system dependent on purified RNA polymerase II and DNA. Cell 18: 469-484, 1979.

122. Manley, J. L.; Fire, A.; Cano, A.; Sharp, P. A.; Gefter, M. L.: DNA-dependent transcription of adenovirus genes in a soluble whole-cell extract. Proc. Nat. Acad. Sci., USA 77: 3855-3859, 1980.

123. Leff, S. E.; Rosenfeld, M. G.: Complex transcriptional units: diversity in gene expression by alternative RNA processing. Annu. Rev. Biochem. 55: 1091-1117, 1981.

124. Breathnach, R..; Chambon, P.: Organization and expression of eucaryotic split genes coding for proteins. Annu. Rev. Biochem. 50: 349-383, 1981.

125. Cochran, M. D.; Weissmann, C.: Modular structure of the β-globin and the TK promoters. EMBO J. 3: 2453-2459, 1984.

126. McKnight, S. L.; Kingsbury, R. C.; Spence, A.; Smith, M.: The distal transcription signals of the Herpes virus tk Gene share a common hexanucleotide control sequence. Cell 37: 253-262, 1984.

127. Dynan, W. S.; Tijan, R.: Control of eukaryotic messenger RNA synthesis by sequence-specific DNA-binding proteins. Nature 316: 774-778, 1985.

128. Khoury, G.; Gruss, P.: Enhancer elements. Cell 33: 313-314, 1984.

129. Schaffner, W.; Serfling, E.; Jasin, M.: Enhancers and eukaryotic gene transcription. Trends in Genetics 1: 224-230, 1985.

130. Sergeant, A.; Bohmann, D.; Zentgraf, H.; Weiher, H.; Keller, W.: A transcription enhancer acts *in vitro* over distances of hundreds of base-pairs on both circular and linear templates but not on chromatin-reconstituted DNA. J. Mol. Biol. 180: 577-600, 1984.

131. Zaret, K. S.; Yamamoto, K. R.: Reversible and persistent changes in chromation structure accompany activation of a glucocorticoid-dependent enhancer element. Cell 38: 29-38, 1984.

132. Yamamoto, K. R.: Hormone-dependent transcriptional enhancement and its implications for mechanisms of multifactor gene regulation. In Molecular Developmental Biology: Expressing Foreign Genes, 43rd Symposium of the Society for Developmental Biology, ed. L. Bogorad, and G. Adelman. New York: Alan Liss, pp. 131-148, 1985.

133. Birnstiel, M. L.; Busslinger, M.; Strub, K.: Transcription termination and 3' processing: The end is in site! Cell 41: 349-359, 1985.

134. Falck-Pedersen, E.; Logan, J.; Shenk, T.; Darnell, J. E. Jr.: Transcription termination within the E1A gene of adenovirus induced by insertion of the mouse β-major globin terminator element. Cell 40: 897-905, 1985.

135. Moore, C. L.; Sharp, P. A.: Accurate cleavage and polyadenylation of exogenous RNA substrate. Cell 41: 845-855, 1985.

136. Lai, E. C.; Woo, S. L.; Dugaiczyk, A.; Catterall, J. F.; O'Malley, B. W.: The ovalbumin gene: Structural sequences in native chicken DNA are not contiguous. Proc. Nat. Acad. Sci., USA. 75: 2205-2209, 1978.

137. Tilghman, S. M.; Tiemeier, D. C.; Seidman, J. G.; Peterlin, B. M.; Sullivan, M.; Maizel, J. V.; Leder, P.: Intervening sequence of DNA identified in the structural portion of a mouse beta-globin gene. Proc. Nat. acad. Sci., USA. 75: 725-729, 1978.

138. Breathnach, R.; Chambon, P.: Organization and expression of eucaryotic split genes coding for proteins. Ann. Rev. Biochem. 50: 349-383, 1981.

139. Wozney, J.; Hanahan, D.; Tate, V.; Boedtker, H.; Doty, P.: Structure of the Pro alpha (2) (I) collagen gene. Nature 294: 129-135, 1981.

140. Cech, T. R.: RNA splicing: three themes with variations. Cell 34: 713-716, 1983.

141. Mount, S. M.: A catalogue of splice junction sequences. Nucleic Acid Res. 10: 459-472, 1982.

142. Wieringa, B.; Meyer, F.; Reiser, J.; Weissmann, C.: Unusual splice sites revealed by mutagenic inactivation of an authentic splice site of the rabbit beta-globin gene. Nature 301: 38-43, 1983.

143. Keller, W.: The RNA lariat: A new ring to the splicing of mRNA precursors. Cell 39: 423-425, 1984.

144. Brody, E.; Abelson, J.: The 'spliceosome'; Yeast premessenger RNA associates with a 40S complex in a splicing dependent reaction. Science 228: 963-967, 1985.

145. Ruskin, B.; Krainer, A. R.; Maniatis, T.; Green, M. R.: Excision of an intact intron as a novel lariat structure during pre-mRNA splicing in vitro. Cell 38: 317-331, 1984.

146. Frendewey, D.; Keller, W.: The stepwise assembly of a pre-mRNA splicing complex requires U-snRNPs and specific intron sequences. Cell 42: 355-367, 1985.

147. Black, D. L.; Chabot, B.; Steitz, J. A.: U2 as well as U1 small Nuclear Ribonucleoprotein are involved in premessenger RNA splicing. Cell 42: 737-750, 1985.

148. Rogers, J.; Wall, R.: A mechanism for RNA splicing. Proc. Nat. Acad. Sci., USA. 77: 1877-1879, 1980.

149. Kornberg, R. D.; Klug, A.: The nucleosome. Sci. Amer. 244: 52-64, 1981.

150. McGhee, J. D.; Felsenfeld, G.: Nucleosome structure. Ann. Rev. Biochem. 49: 1115-1156, 1980.

151. Igo-Kemenes, T.; Horz, W.; Zachau, H. G.: Chromatin. Ann. Rev. biochem. 51: 89-121, 1982.

152. Paulson, J. R.; Leammli, U. K.: The structure of histone-depleted metaphase chromosomes. Cell 12: 817-828, 1977.

153. Thoma, F.; Koller, T.; Klug, A.: Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. J. Cell Biol. 83: 403-427, 1979.

154. Brown, S. W.: Hetrochromatin. Science 151: 417-425, 1966.

155. Pederson, D. S.; Thoma, F.; Simpson, R.T.: Core particle fiber and transcriptionally active chromatin structure. Ann. Rev. Cell Biol. 2: 117-147, 1986.

156. Robertson, H. D.; Barrell, B. G.; Weith, H. L.; Donelson, J. E.: Isolation and sequence analysis of a ribosome-protected fragment from bacteriophage ϕX174 DNA. Nat. New Biol. 241: 38-40, 1973.

157. Ziff, E. B.; Sedat, J. W.; Galibert, F.: Determination of the nucleotide sequence of a fragment of bacteriophage ϕX 174 DNA. Nat. New Biol. 241: 34-37, 1973.

158. Maxam, A. M.; Golbert, W.: A new method for sequencing DNA. Proc. Natl. Acad. Sci., USA 74: 560-564, 1977.

159. Sanger, F.; Nicklen, S.; Coulson, A. R.: DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci., USA 74: 5463-5467, 1977.

160. Mizusawa, S.; Nishimura, S.; Seela, F.: Improvement of the dideoxy chain termination method of DNA sequencing by use of deoxy-7-deazaguanosine triphosphate in place of dGTP. Nucl. Acids Res. 14: 1319-1324, 1986.

161. Lee, L. G.; Connell, C. R.; Woo, S. L.; Cheng, R. D.; McArdle, B. F.; Fuller, C. W.; Halloran, N. D.; 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.

162. Klenow, H.; Overgaard-Hansen, K.; Patkar, S. A.: Proteolytic cleavage of native DNA polymerase into two different catalytic fragments. Influence of assay conditions on the change of exonuclease activity and polymerase activity accompanying cleavage. Eur. J. Biochem. 22: 371-381, 1971.

163. Tabor, S.; 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.

164. Stenesh, J.; 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.

165. Innis, M. A.; Myambo, K. B.; Gelfand, D. H.; Brow, M. A. D.: DNA sequencing with *Thermus aquaticus* DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA. Proc. Natl. Acad. Sci., USA 85: 9436-9440, 1988.

166. Protocol in ABI PRISM double-stranded cycle sequencing kit with AmpliTaq DNA polymerase.

167. Tabor, S.; Richardson, C. C.: DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. Proc. Natl. Acad. Sci., USA 84: 4767-4771, 1987.

168. Atkinson, M. R.; Deutscher, M. P.; Kornberg, A.; Russell, A. F.; Moffatt, J. G.: Enzymatic synthesis of deoxyribonucleic acid. XXXIV. Termination of chain growth by a 2', 3'-dideoxyribonucleotide. Biochemistry 8: 4897-4909, 1969.

169. Tabor, S.; Richardson, C. C.: A single residue in DNA polymerase of *Escherichia coli* DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. Proc. Natl. Acad. Sci., USA 92: 6339-6343, 1995.

170. Korolev, S.; Nayal, M.; Barnes, W. M.; DiCera, E.; Waksman, G.: Crystal structure of the large fragment of *Thermus aquaticus* DNA polymerase I at 2.5-A resolution: structure basis for thermostability. Proc. Natl. Acad. Sci., USA 92: 9264-9268, 1995.

171. Barnes, W. M.: U. S. Patent No. 5,436,149. Thermostable DNA polymerase with enhanced thermostability and enhanced length and efficiency of primer extension. 1995.

172. Barnes, W. M.: The fidelity of Taq polymerase catalyzing PCR is improved by an N-terminal deletion. Gene 112:29-35, 1992.

173. Protocol in ABI PRISM double-stranded cycle sequencing kit with AmpliTaq FS DNA polymerase.

174. Smith, L. M.; Sander, J. Z.; Kaiser, R. J.; Hughes, P.; Dodd, C.; Connell, C. R.; Heiner, C.; Kemt, S. B. H.; Lood, L. E.: Fluorescence detection in automated DNA sequence analysis. Nature 321: 674-679, 1986.

175. Prober, J. M.; Trainor, G. L.; Dam, R. J.; Hobbs, F. W.; Robertson, C. W.; Zagursky, R. J.; Cocuzza, A. J.; Jensen, M. A.; Baumeister, K.: A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science 238: 336-341, 1987.

176. McBride, L. J.; Koepf, S. M.; Gibbs, R. A.; Salser, W.; Mayrand, P. E.; Hunkapiller, M. W.; Kronick, M. N.: Automated DNA sequencing methods involving polymerase chain reaction. Clin. Chem. 35: 2196-2201, 1989.

177. Craxton, M.: Linear amplification sequencing, a powerful method for sequencing DNA. Methods: Companion Methods Enzymol. 3: 20-26, 1991.

178. Henikoff, S.: Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene 28: 351-359, 1984.

179. Bankier, A. T.; Barrell, B. G.: in Nucleic Acids Sequencing: A practical approach, C. J. Howe and E. S. Ward, eds. IRL Press, Oxford, 1989. Pp. 37-38.

180. Deininger, P. L.: Random subcloning of sonicated DNA: application to shotgun sequence analysis. Anal. Biochem. 129: 216-223, 1983.

181. Bankier, A. T.; Weston, K. M.; Barrell, B. G.: Random cloning and sequencing by the M13/dideoxynucleotide chain termination method. Meth. Enzymol. 155: 51-93, 1987.

182. Bodenteich, A.; Chissoe, S.; Wang, Y. F.; 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, 1994.

183. Surzycki, S. Department of Biology, Indiana University, personal communication. 1994.

184. Phadnis, S. H.; Huang, H. V.; 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.

185. Anderson, S.; Bamkier, 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.; Young, I. G.: Sequencing and organization of the human mitochondrial genome. Nature 290: 457-465, 1981.

186. Fitzgerald, M. C.; Skowron, P.; Van Etten, J. L.; Smith, L. M.; Mead, D. A.: Rapid shotgun cloning utilizing the two base recognition endonuclease CviJI. Nucl. Acids Res. 20: 3753-3762, 1992.

187. Sambrooks, J.; Fritsch, E. F.; Maniatis, T.: Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1989.

188. Scott, J. R.: Regulation of Plasmid replication. Microbiol. Rev. 48: 1-23, 1984.

189. Cesareni, G.; Banner, D. W.: Regulation of plasmid copy number by complementary RNAs. Trends Biochem. Sci. 10: 303-, 1985.

190. Covarrubias, L.; Cervantes, L.; Covarrubias, A.; Soberon, X.; Vichido, I.; Blanco, A.; Kupersztoch-Portnoy, Y. M.; Bolivar, F.: Construction and characterization of new cloning vehicles. V. Mobilization and coding properties of pBR322 and several deletion derivatives including pBR327 and pBR328. Gene 13:25-35, 1981.

191. Minton, N. P.; Chambers, S. P.; Prior, S. E.; Cole, S. T.; Garnier, T.: Copy number and mobilization properties of pUC plasmids. Bethesda Res. Lab. Focus 10(3): 56-, 1988.

192. Vieira, J.; Messing, J.: The pUC plasmids and M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. Gene 19: 259-268, 1982.

193. Davies, J.; Smith, D. I.: Plasmid-determined resistance to antimicrobial agents. Annu. Rev. Microbiol. 32: 469-518, 1978.

194. Ullmann, A.; Jacob, F.; Monod, J.: Characterization by *in vitro* complementation of a peptide corresponding to an operon-proximal segment of the β-galactosidase structural gene of *Escherichia coli*. J. Mol. Biol. 24: 339-343, 1967.

195. Herskowitz, I.: Control of gene expression in Bacteriophage lamda. Ann. Rev. Genet. 7: 289-324, 1973.

196. Hendrix, R. W.; Roberts, J. W.; Stahl, F. W.; Weisberg, R. A. eds: Lamda II. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York. 1983.

197. Messing, J.; Gronenborn, B.; Muller-Hill, B.; Hofschneider, P. H.: Filamentous coliphage M13 as a cloning vehicle: Insertion of a HindII fragment of the lac regulatory region in M13 replicative form *in vitro*. Proc. Natl. Acad. Sci., USA 74: 3642-3646, 1977.

198. Messing, J.; Vieira, J.: A new pair of M13 vectors for selecting either DNA strand of double-digest restriction fragments. Gene 19: 269-276, 1982.

199. Messing, J.: New M13 vectors for cloning. Meth. Enzymol. 101: 20-78, 1983.

200. Meyer, T. K.; Geider, K.; Kurz, C.; Schaller, H.: Cleavage site of bacteriophage fd gene II-protein in the origin of viral strand replication. Nature 278: 365-367, 1979.

201. Horiuchi, K.: Origin of DNA replication of bacteriophage f1 as the signal for termination. Proc. Natl. Acad. Sci., USA 77: 5226-5229, 1980.

202. Mazur, B. J.; Model, P.: Regulation of coliphage f1 single-stranded DNA synthesis by a DNA-binding protein. J. Mol. Biol. 78: 285-300, 1973.

203. Lerner, T. J.; Model, P.: The "steady state" of coliphage f1: DNA synthesis late in infection. Virology 115: 282-294, 1981.

204. Marvin, D. A.; Wachtel, E. J.: Structure and assembly of filamentous bacterial viruses. Nature 253: 19-23, 1975.

205. Webster, R. E.; Cashman, J. S.: Morphogenesis of the filamentous single-stranded DNA phage. In The Single-stranded DNA Phages (ed. D. T. Denhardt et al.), p. 557. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York. 1978.

206. Messing, J.: M13mp2 and derivatives: A molecular cloning system for DNA sequencing, strand specific hybridization, and *in vitro* mutagenesis. In Recombinant DNA: Proceedings of the Third Cleveland Symposium on Macromolecules (ed. A. G. Walton), p. 143. Elsevier, Amsterdam. 1981.

207. Bagdasarian, M.; Lurz, R.; Ruckert, B.; Franklin, F. C.; Bagdasarian, M. M.; Frey, J.; Timmis, K. N.: Specific-purpose plasmid cloning vectors. II. Broad host range, high copy number, RSF1010-derived vectors, and a host-vector system for gene cloning in *Pseudomonas*. Gene 16: 237-247, 1981.

208. Ish-Horowicz, D.; Burke, J. F.: Rapid and effient cosmid cloning. Nucleic Acids Res. 9: 2989-2998, 1981.

209. Bates, P. F.; Swift, R. A.: Double *cos* site vectors: simplified cosmid cloning. Gene 26: 137-146, 1983.

210. Brady, G.; Jantzen, H. M.; Bernard, H. U.; Brown, R.; Schutz, G.; Hashimoto-Gotoh, T.: New cosmid vectors developed for eukaryotic DNA cloning. Gene 27: 223-232, 1984.

211. Burke, D. T.; Carle, G. F.; Olson, M. V.: Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science 236: 806-812, 1987.

212. Shero, J. H.; McCormick, M. K.; Antonarakis, S. E.; Hieter, P.: Yeast artificial chromosome vectors for efficient clone manipulation and mapping. Genomics 10: 505-508, 1991.

213. Shizuya, H.; Birren, B.; Kim, U. J.; Mancino, V.; Slepak, T.; Tachiiri, Y.; Simon, M.: Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. Proc. Natl. Acad. Sci., USA 89: 8794-8797, 1992.

214. Ioannou, P. A.; Amemiya, C. T.; Garnes, J.; Kroisel, P. M.; Shizuya, H.; Chen, C.; Batzer, M. A.; de Jong, P. J.: A new bacteriophage P1-derived

vector for the propagation of large human DNA fragments. Nat. Genet. 6: 84-89, 1994.

215. Ansorge, W.; Sproat, B.; Stegemann, J.; Schwager, C.: A non-radioactive automated method for DNA sequence determination. J. Biochem. Biophys. Meth. 13: 315-322, 1986.

216. Ansorge, W.; Sproat, B.; Stegemann, J.; Schwager, C.; Zenke, M.: Automated DNA sequencing: ultrasensitive detection of fluorescent bands during. electrophoresis. Nucl. Acids Res. 15: 4593-4602, 1987.

217. Middendoff, L. R.; Bruce, J. C.; Eckles, R. D.; Grone, D. L.; Roemer, S. C.; Sloniker, G. D.; Steffens, D. L.; Sutter, S. L.; Brumbaugh, J. A.; et al.: Continous, on-line DNA sequencing using a versatile infrared laser scanner/electrophoresis apparatus. Electrophoresis 13(8): 487-494, 1992.

218. McIndoe, R. A.; Hood, L.; Bumgarnar, R. E.: An analysis of the dynamic range and linearity of an infrared DNA sequencer. Electrophoresis 17(4): 652-658, 1996.

219. Applied Biosystems Incorporated Users Manual ABI 373, 1994.

220. Applied Biosystems Incorporated Users Manual ABI 377, 1996.

221. Birnboim, H. C.; Doly, J.: A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucl. Acid Res. 7: 1513-1523, 1979.

222. Tanaka, T.; Kuroda, M.; Sakaguchi, K.: Isolation and characterization of four plasmids from *Bacillus subtilis*. J. Bacteriol. 129(3): 1487-1494, 1977.

223. Carter, J. M.; Milton, I. D.: An inexpensive and simple method for DNA purifications on silica particles. Nucl. Acid Res. 21: 1044-1044, 1993.

224. Cohen, S. N.; Chang, A. C. Y.; Hsu, L.: Nonchromosomal antibiotic resistance in bacteria: genetic transformation of *Escherichia coli* by R-factor DNA. Proc. Natl. Acad. Res., USA 69: 2110-2114, 1972.

225. Gleeson, T. J.; Staden, R.: An X window and UNIX implementation of our sequence analysis package. Compt. Appl. Biosci. 7: 398-401, 1991.

226. Meyers, E. W.: Incremental alignment algorithms and their application (Technical report 86-2). Department of Computer Science, University of Arizona, Tucson, AZ 85721, 1986.

227. Huang, X.: An improved sequence assembly program. Genomics 33: 21-31, 1996.

228. Dear, S.; Staden, R.: A sequence assembly and editing program for efficient management of large projects. Nucl. Acids Res. 19(14): 3907-3911, 1991.

229. Smith, T. F.; Waterman, M. S.; Fitch, W. M.: Comparative biosequence metrics. J. Mol. Evol. 18(1): 38-46, 1981.

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

231. Madden, T. L.; Tatusov, R. L.; Zhang, J.: Applications of network BLAST server. Methods Enzymol. 266: 131-141, 1996.

232. Uberbacher, E. C.; 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.

233. Wisconsin Package Version 8.0, Genetics Computer Group (GCG), Madison, Wisconsin.

234. Thompson, J. D.; Higgins, D. G.; Gibson, T. J.: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl. Acids Res. 22(22): 4673-4680, 1994.

235. Staden, R.: The current status and portability of our sequence handling software. Nucl. Acids Res. 14: 217-231, 1986.

236. Prestridge, D. S.: SIGNAL SCAN: A computer program that scans DNA sequences for eukaryotic transcriptional elements. CABIOS 7:203-206, 1991.

237. Hofmann, K.; Stoffel, W.: TMBASE: A database of membrane spanning protein segments. Biol. Chem. Hoppe-Seyler, 374: 166, 1993.

238. Mullis, K. B.; Faloona, F. A.: Specific synthesis of DNA *in vitro* via a polymerase-catalyzed chain reaction. Method Enzymol. 155: 335-350, 1987.

239. Winship, P. R.: An improved method for directly sequencing PCR amplified material using dimethyl sulfoxide. Nucl. Acids Res. 17: 1266,1989.

240. Sun, Y.; Hegamyer, G.; Colburn, N. H.: PCR-direct sequencing of a GC-rich region by inclusion of 10% DMSO: application to mouse c-jun. Biotechniques 15: 372-374, 1993.

241. Sommer S. S.: Assessing the underlying pattern of human germline mutations: lessons from the factor IX gene. FASEB J. 6(10): 2767-2774, 1992.

242. Kawasaki, K.; Minoshima, S.; Nakato, E.; Shibuya, K.; Shintani, A.; Schmeits, J. L.; Wang, J.; Shimizu, N.: One-megabase sequence analysis of the human immunoglobulin l gene locus. Genome Res. 7(3): 250-261, 1997.

243. Sirotkin, H.; Morrow, B.; DasGupta, R.; Goldberg, R.; Ratanjali, S. P.; Shi, G.; Cannizzaro, L.; Shprintzen, R.; Weissman, S. M.; Kucherlapati, R.: Isolation of a new clathrin heavy chain gene with muscle-specific expression from the region commonly deleted in velo-cardio-facial syndrome. Hum. Mol. Genet. 5: 617-624, 1996.

244. Brodsky, F. M.: Living with clathrin: its role in intracellular membrane traffic. Science 242: 1396-1402, 1988.

245. Liu, S.; Wong, M. L.; Craik, C. S.; Brodsky, F. M.: regulation of clathrin assembly and trimerization defined using recombinant triskelion hubs. Cell 83:257-267, 1995.

246. Kedra, D.; Peyrard, M.; Fransson, I.; Collins, J.; Dunham, I.; Roe, B. A.; Dumanski, J. P.: Charaterization of a second human clathrin heavy chain polypeptide gene (CLH-22) from chromosome 22q11. Hum. Mol. Geneti. 5: 625-631, 1996.

247. Monk, R. J.; Meyuhas, O.; Perry, R. P.: Mammals have multiple genes for individual ribosomal proteins. Cell 24: 301-306,1981.

248. Wiedemann, L. M.; Perry, R. P.: Characterization of the expressed gene and several processed pseudogenes for the mouse ribosomal protein L30 gene family. Mol. Cell. Biol. 4: 2518-2528, 1984.

249. Dudov, K. P.; Perry, R. P.: The gene family encoding the mouse ribosomal protein L32 contains a uniquely expressed intron-containing gene and an unmutated processed gene. Cell 37: 457-468, 1984.

250. Klein, A.; Meyuhas, O.: A multigene family of intron lacking and containing genes, encoding for mouse ribosomal protein L7. Nucl. Acids Res. 12: 3763-3776, 1984.

251. Aoyama, Y.; Chan, Y-L.; Wool, I. G.: The primary structure of rat ribosomal protein L34. FEBS Letters 249: 119-122, 1989.

252. Rommens, J. M.; Durocher, F.; McArthur, J.; LeBlanc, J. F.; Allen, T.; Samson, C.; Ferri, L.; Narod, S.; Morgan, K.; et al.: Generation of a transcription map at the HSD17B locus centromeric to BRCA1 at 17q21. Genomics 28: 530-542,1995.

253. Schmid, C. W.; Deininger, P. L.: Sequence organization of the human genome. Cell 6: 345-358, 1975.

254. Deininger, P. L.; Schmid, C. W.: An electromicroscope study of the DNA sequence organization of the human genome. J. Mol. Biol. 106: 773-790, 1976.

255. Jurka, J.; Smith, T.: A fundamental division in the Alu family of repeated sequence. Proc. Natl. Acad. Sci., USA 85: 4775-4778, 1988.

256. Batzer, M. A.; Deininger, P. L.: A human-specific subfamily of Alu sequences. Genomics 9: 481-487, 1991.

257. Lorain, S.; Demczuk, S.; Lamour, V.; Toth, S.; Aurias, A.; Roe, B. A.; Lipinski, M.: Structural Organization of the WD repeat protein-encoding gene HIRA in the DiGeorge syndrome critical region of human chromosome 22. Genome Res. 6: 43-50, 1996.

258. Sherwood, P. W.; Tsang, S. V-M.; Osley, M. A.: Charaterization of HIR1 and HIR2, two genes required for regulation of histone gene transcription in *Saccharomyces cerevisiae*. Mol. Cell. Biol. 13: 28-38, 1993.

259. Neer, E. J.; Schmidt, C. J.; Nambudripad, R.; Smith, T.: The ancient regulatory-protein family of WD-repeat proteins. Nature 371: 297-300, 1994.

260. Robbins, J.; Dilworth, S. M.; Laskey, R. A.; Dingwall, C.: Two interdependent basic domains in nucleoplasmin nuclear targeting sequence: identification of a class bipartite nuclear targeting sequence. Cell 64: 615-623, 1991.

261. Dingwall, C.; Laskey, R. A.: Nuclear targeting sequences--a consensus? Trends Biochem. Sci. 16: 478-481, 1991.

262. Wilming, L. G.; Sylvia Snoeren, C. A.; van Rijswijk, A.; Grosveld, F.; Meijers, C.: The murine homologue of HIRA, a DiGeorge syndrome

candidate gene, is expressed in embryonic structures affected in human CATCH22 patients. Hum. Mol. Genet. 6: 247-258, 1997.

263. Roberts, C.; Daw, S. C. M.; Halford, S.; Scambler, P. J.: Cloning and developmental expression analysis of chick Hira (Chira), a candidate gene for DiGeorge syndrome. Hum. Mol. Genet. 6: 237-245, 1997.

264. Mavrothalassitis, G. J.; Watson, D. K.; Papas, T. S.: Molecular and function characterization of the promoter of EST2, the human c-est-2 gene. Proc. Natl. Acad. Sci., USA 87: 1047-1051, 1990.

265. Herrick, K. R.; Gorin, F. A.; Park, E. A.; Tait, R. C.: Characterization of the 5' flanking region of the gene encoding rat liver glycogen phosphorylase. Gene 126: 203-211, 1993.

266. Azizkhan, J. C.; Jensen, D. E.; Pierce, A. J.; Wade, M.: Transcription from TATA-less promoters: dihydrofolate reductase as a model. Crit. Rev. Eukaryot. Gene Expr. 3: 229-254, 1993.

267. Smith, T. M.; Lee, M. K.; Szabo, C. I.; Jerome, N.; McEuen, M.; Taylor, M.; Hood, L.; King, M-C.: Complete genome sequence and analysis of 117 kb of human DNA containing the gene BRCA1. Genome Res. 6: 1029-1049, 1996.

268. Pizzuti, A.; Novelli, G.; Ratti, A.; Amati, F.; Mari, A.; Calabrese, G.; Nicolis, S.; Marino, B.; Scarlato, G.; Ottolenghi, S.; Dallapiccola, B.: UFD1L, a developmentally ubiquitination gene, is deleted in CATCH 22 syndrome. Hum. Mol. Genet. 6: 259-265, 1997.

269. Johnson, E. S.; Bartel, B.; Seufert, W.; Varshavsky, A.: Ubiquitin as a degradation signal. EMBO J. 11: 497-505, 1992.

270. Johnson, E. S.; Ma, P. C. M.; Ota, I. M.; Varshavsky, A.: A proteolytic pathway that recognizes ubiquitin as a degradation signal. J. Biol. Chem. 270: 17442-17456, 1995.

271. Adams, M. D.; Kerlavage, A. R.; Fleischmann, R. D.; Fuldner, R. A.; Bult, C. J.; Lee, N.; Kirkness, E. F.; et al.: EST T31599 from Initial assessment of human diversity and expression patterns based upon 52 million basepairs of cDNA sequence. Unpublished, 1995.

272. Schubert, D.; LaCorbiere, M.: Isolation of an adhesion-mediating protein from chick neural retina adherons. J. Cell. Biol. 101: 1071-1077, 1985.

273. Cole, G. J.; Burg, M.: Characterization of a heparan sulfate proteoglycan that copurifies with the neural cell adhesion molecule. Exp. Cell. Res. 182: 44-60, 1989.

274. Burg, M.; Cole, G. J.: Characterization of cell-associated proteoglycan synthesized by embryonic neural retinal cells. Arch. Biochem. Biophys. 276: 396-404, 1990.

275. Carson, D. D.; Tang, J. P.; Julian, J.: Heparan sulfate proteoglycan (perlecan) expression by mouse embryos during acquisition of attachment competence. Dev. Biol. 155: 97-106, 1993.

276. Tsen, G.; Halfter, W.; Kroger, S.; Cole, G. J.: Agrin is a heparan sulfate proteoglycan. J. Biol. Chem. 270: 3392-3399, 1995.

277. Buckler, A. J.; Chang, D. D.; Graw, S. L.; Brook, J. D.; Haber, D. A.; Sharp, P. A.; Housman, D. E.: Exon amplification: a strategy to isolate mammalian genes based on RNA splicing. Proc. Natl. Acad. Sci., USA 88: 4005-4009, 1991.

278. Trofatter, J. A.; Long, K. R.; Murrell, J. R.; Stotler, C. J.; Gusella, J. F.; Buckler, A. J.: An expression-independent catalog of genes from human chromosome 22. Genome Res. 5: 214-224, 1995.

279. Cross, S. H.; Charlton, J. A.; Nan, X.; Bird, A. P.: Purification of CpG islands using a methylated DNA binding column. Nat. Genet. 6: 236-244, 1994.

280. Tudor, M.; Lobocka, M.; Goodell, M.; Pettitt, J.; O'Hare, K.: The pogo transposable element family of *Drosophila melanogaster*. Mol. Gen. Genet. 232: 126-134, 1992.

281. Smit, A. F.; Riggs, D.: Tiggers and DNA transposon fossils in the human genome. Proc. Natl. Acad. Sci., USA 93: 1443-1448, 1996.

282. Robertson, H. M.: Members of the pogo superfamily of DNA-mediated transposons in the human genome. Mol. Gen. Genet. 252: 761-766, 1996.

283. Galili, N.; Baldwin, H. S.; Lund, J.; Reeves, R.; Gong, W.; Wang Z.; Roe, B. A.; Emanuel, B. S.; Nayak, S.; Mickanin, C.; Budarf, M. L.; Buck, C. A.: A region of mouse chromosome 16 is syntenic to the DiGeorge, Velocardiofacial syndrome minimal critical region. Genome Res. 7: 17-26, 1997.

284. Sirotkin, H.; Morrow, B.; St. Jore, B.; Puech, A.; Das Gupta, R.; Patanjali, S.; Skoultchi, A.; Weissman, S. M.; Kucherlapati, R.: Identification, characterization and precise mapping of a human gene encoding a novel membrane spanning protein from the 22q11 region deleted in Velo-Cardio-Facial syndrome. Genomics 42: 245-251, 1997.

285. Ware, J.; Zieger, B.: Alternative expression of platelet glycoprotein Ib(beta) mRNA from an adjacent gene with an imperfect polyadenylation signal sequence. J. Clin. Invest. 99: 520-525, 1997.

286. Chapman, D. L.; Garvey, N.; Hancock, S.; Alexiou, M.; Agulnik, S. I.; Gibson-Brown, J. J.; Cebra-Thomas, J.; Bollag, R. J.; Silver L. M.;

Papaioannou, V. E.: Expression of the T-box family genes, Tbx1-Tbx5, during early mouse development. Dev. Dyn. 206: 379-390, 1996.

287. Ben-Ishai, R.; Scharf, R.; Sharon, R.; Kapten, I.: A human cellular sequence implicated in trk oncogene activation is DNA damage inducible. Proc. Natl. Acad. Sci., USA 87: 6039-6043, 1990.

288. Sirotkin, H.; O'Donnell, H.; DasGupta, R.; St. Halford, S.; Jore, B.; Puech, A.; Parimoo, S.; Morrow, B.; Skoultchi, A.; Weissman, S.; Scambler, P.; Kucherlapati, R.: Identification of a new human gene family member (ARVCF) from the region deleted in Velo-Cardio-Facial syndrome. Genomics 41: 75-83, 1997.

289. Bressan, A.; Somma, M. P.; Lewis, J.; Santolamazza, C.; Copeland, N. G.; Gilbert, D.; J.; Jenkins, N. A.; Lavia, P.: Characterization of the opposite-strand genes from the mouse bidirectionally transcribed HTF9 locus. Gene 103: 201-209, 1991.

290. Di Matteo, G.; Fuschi, P.; Zerfass, K; Moretti, S.; Ricordy, R.; Cenciarelli, C.; Tripodi, M.; Jansen-Durr, P.; Lavia, P.: Transcriptional control of the Htf9-A/RanBP-1 gene during the cell cycle. Cell Growth Differ. 6: 1213-1224, 1995.