SEQUENCING ADJACENT TO BssHII SITES ON HUMAN CHROMOSOME 8: A METHOD OF GENE IDENTIFICATION

by

LEAH RAE DEBELLA

BA, Sonoma State University, 1995

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE in

THE FACULTY OF GRADUATE STUDIES

Department of Genetics

THE UNIVERSITY OF BRITISH COLUMBIA

August 1997

© Leah Rae DeBella, 1997
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Genetics

The University of British Columbia
Vancouver, Canada

Date August 28/97
Abstract

Considerable research is focused on the identification of genes in the human genome. Recently, the Human Genome project has named gene identification as one of its goals to be accomplished in tandem with mapping and sequencing of the entire genome. As a result, a large program centered on the creation of expressed sequence tags (ESTs) is currently underway. Although this method of gene identification has vastly increased the number of expressed sequences in the database, it is unlikely that all genes encoded in the genome will be detected through the sequencing of cDNAs.

An alternative method of gene identification, using genomic DNA, is the sequencing of regions adjacent to rare cutting enzyme sites. Recognition sites of rare cutting enzymes are found infrequently in genomic DNA but cluster in regions known as CpG islands. CpG islands are associated with most ubiquitously expressed genes and 40% of tissue specific genes. Therefore, sequencing of these regions offers a complementary method to be used in the recognition of novel genes.

The objective of this project was to identify novel genes present in cloned DNA from human chromosome 8 using BssHI, a rare cutting enzyme. BssHI, is an excellent choice to be used in this manner, as 80% of its sites are located in CpG islands. DNA adjacent to fourteen BssHII sites, originating from eleven cosmids, was sequenced. From this set, two human ESTs were identified; two open reading frames with no apparent homologues; two CpG islands, one of which contains a translation initiation signal and is 5’ of the EST H88121; and two novel human genes. These results validate the hypothesis that the use of this method can complement other techniques in the identification of novel genes in the human genome.
Table of Contents

Abstract.................................................................................................................. ii
Table of Contents.................................................................................................... iii
List of Tables........................................................................................................... vi
List of Figures........................................................................................................... vii
List of Abbreviations............................................................................................... ix
Acknowledgments................................................................................................... xii

Chapter I: Introduction............................................................................................. 1
  1.1 Features of GC rich regions in the Human Genome................................. 1
  1.2 Characteristics of CpG Islands................................................................. 3
  1.3 Rare Cutting Enzymes............................................................................... 4
  1.4 Using CpG Islands to Identify Genes....................................................... 6
  1.5 Gene Identification.................................................................................... 7
  1.6 Project Proposal.......................................................................................... 8

Chapter II: Methods and Materials......................................................................... 11
  2.1 Isolation of cloned human genomic DNA- Minipreparation of
      Cosmid and Plasmid DNA.............................................................................. 11
  2.2 Restriction Digests...................................................................................... 12
  2.3 Electrophoresis............................................................................................ 14
  2.3.1 Agarose Gel............................................................................................ 14
  2.3.2 Polyacrylamide Gel............................................................................... 15
  2.4 Exclusion of Cosmids containing Alu Sequence........................................ 16
  2.4.1 Southern Blot.......................................................................................... 16
      2.4.1.1 Synthesis of Radiolabeled Alu Probe
             by Primer Extension.................................................................................. 17
      2.4.1.2 Hybridization.................................................................................... 18
      2.4.2 Screening by Hybridization............................................................... 19
          2.4.2.1 3' Termini Labeling of T3 Probe with T4 Kinase.. 20
  2.5 Construction of Deleted Cosmids............................................................... 21
  2.6 DNA Extraction from Agarose Gel............................................................ 21
2.7 Subcloning of Cosmid Fragments ............................................. 22
  2.7.1 Preparation of pBluescript vector ..................................... 23
  2.7.2 Ligation Reaction .......................................................... 23
2.8 Transformation of Competent E. coli ....................................... 24
2.9 DNA Purification for Sequencing ............................................ 25
2.10 Sequencing ............................................................................ 27
  2.10.1 Manual Sequencing .......................................................... 27
  2.10.2 Automated Sequencing ..................................................... 28
2.11 DNA Analysis ......................................................................... 29
  2.11.1 BLAST ........................................................................... 29
  2.11.2 GRAIL ........................................................................... 30
  2.11.3 PROSITE ......................................................................... 32
2.12 Sequence tagged sites ............................................................ 32
  2.12.1 GOR gene .................................................................... 32
  2.12.2 5-oxo-prolinase gene ....................................................... 33
  2.12.3 Human EST W90101 ......................................................... 33

Chapter III: Results ..................................................................... 45
3.1 Sequence Data ......................................................................... 45
  3.1.1 Cosmid 93C11ΔBssHII ......................................................... 45
  3.1.2 Subclone from Cosmid 13E3 ................................................. 46
  3.1.3 Subclones from Cosmid 40G1 .............................................. 46
  3.1.4 Subclones from Cosmid 46F4 .............................................. 47
  3.1.5 Subclone from Cosmid 46F11 .............................................. 48
  3.1.6 Subclone from Cosmid 77C1 .............................................. 49
  3.1.7 Subclone from Cosmid 141A6 .............................................. 49
  3.1.8 Subclone from Cosmid 156G7 .............................................. 49
  3.1.9 Subclone from Cosmid 166H7 .............................................. 50
  3.1.10 Subclone from Cosmid 175G8 ............................................ 50
  3.1.11 Subclone from Cosmid 176F5 ............................................ 51
3.2 Localization by STS mapping ................................................... 51
  3.2.1 Localization of the Human GOR gene ................................ 51
  3.2.2 Localization of Human EST W90101 ................................. 52
Chapter IV: Discussion

4.1 Gene Identification ................................................. 65
4.2 DNA Adjacent to BssHII sites ..................................... 67
  4.2.1 Repetitive DNA ................................................ 68
    4.2.1.1 Alu elements ............................................ 68
    4.2.1.2 LINE1 elements ......................................... 69
    4.2.1.3 VNTR and MER elements ............................... 69
  4.2.2 Identification of ESTs ....................................... 70
  4.2.3 Identification of Open Reading Frames .................... 71
  4.2.4 Identification of CpG Islands ............................... 71
  4.2.5 Identification of Novel Human Genes ...................... 73
    4.2.5.1 GOR gene ............................................. 73
    4.2.5.2 5-Oxo-L-Prolinase gene ................................ 75
  4.3 Conclusions .................................................... 77

References ......................................................................... 93
Appendix 1: Sequence obtained from cosmids .......................... 98
Appendix 2: Web site addresses ......................................... 104
Appendix 3: Article submitted for publication prior to completion of Thesis ........ 105
List of Tables

Table 3.1 GenBank Non-Redundant Database Search ........................................ 59
Table 3.2 GenBank Sequence Tag Sites (STTs) and Expressed Sequence Tag (ESTs) Database Searches ......................................................... 60
Table 3.3 GenBank Expressed Sequence Tag (ESTs) Database Search with the GOR gene ............................................................ 61
Table 3.4 GenBank Expressed Sequence Tag (ESTs) Database Search with the OPLAH gene ............................................................ 62
Table 3.5 PROSITE database Searches with the GOR gene .................................. 63
Table 3.6 Protein Coding Potential of BssHII-EcoRI subclones as Analyzed by GRAIL ................................................................. 64
Table 3.7 GC content of BssHII-EcoRI subclones ............................................. 64
Table 4.1 RNA splice junction sequences .......................................................... 91
Table 4.2 Potential TATA and CCAAT Boxes and Sp1 Binding Sites ................. 91
Table 4.3 Summary of Data from Cosmids ...................................................... 92
### List of Figures

| Figure 2.1 | Diagram of sCos1 vector | 35 |
| Figure 2.2 | Digestion of cosmids with restriction enzymes | 36 |
| Figure 2.3 | Detection of Alu repetitive sequence: Southern blot | 37 |
| Figure 2.4 | Detection of Alu repetitive sequence: Colony blot | 38 |
| Figure 2.5 | Apparatus for the capillary transfer method of Southern blotting | 39 |
| Figure 2.6 | Diagram of modified pBluescript vector | 40 |
| Figure 2.7 | Creation of a deleted cosmid | 41 |
| Figure 2.8 | Subcloning using the modified pKSIIAsc vector | 42 |
| Figure 2.9 | Detection of pKSIIAsc vector | 43 |
| Figure 2.10 | Restriction digests of subclones | 44 |
| Figure 3.1 | Sequence comparison between genomic sequence from cosmid 166H7 and sequence from cDNA clone 220630 | 53 |
| Figure 3.2 | Sequence comparison between genomic sequence from cosmid 176F5 and sequence from cDNA clone 418116 | 54 |
| Figure 3.3 | Chromosomal localization of the human EST W90101 by PCR against somatic cell hybrids | 55 |
| Figure 3.4 | GOR sequence comparison between human and chimpanzee for the region amplified by PCR primers | 56 |
| Figure 3.5 | Chromosomal localization of the human GOR gene by PCR against somatic cell hybrids | 58 |
| Figure 4.1 | Translations of ORFs predicted by GRAIL flanked by 3' and 5' splice sequences | 78 |
| Figure 4.2 | CpG islands with TATA boxes and transcription start sequence highlighted in italics | 79 |
| Figure 4.3 | GOR EcoRI fragment comparison between human and chimpanzee | 80 |
| Figure 4.4a | GOR sequence comparison between human and chimpanzee beginning at the 5' EcoRI site | 81 |
| Figure 4.4b | GOR sequence comparison between human and chimpanzee beginning at the 5' BssHII site | 83 |
| Figure 4.4c | GOR sequence comparison between human and chimpanzee beginning at the second BssHII site | 85 |
Figure 4.5  5-oxo-L-prolinase comparison between human and rat................. 86
Figure 4.6a  5-oxo-L-prolinase sequence comparison between human
and rat for the first exon identified in the human sequence.............87
Figure 4.6b  5-oxo-L-prolinase sequence comparison between human
and rat for the second exon identified in the human sequence........ 88
Figure 4.6c  5-oxo-L-prolinase sequence comparison between human
and rat for the third exon identified in the human sequence.......... 89
Figure 4.6d  5-oxo-L-prolinase sequence comparison between human
and rat for the fourth exon identified in the human sequence......... 90
List of Abbreviations

Amp\(^r\): ampicillin resistant

B: BssHI

BLAST: basic alignment search tool

bp: base pair

cDNA: complementary DNA

ColE1: plasmid origin of replication

cR: centiRay

dATP: deoxyriboadenosine 5'-triphosphate
dCTP: deoxyribocytidine 5'-triphosphate
ddATP: dideoxyriboadenosine 5'-triphosphate
ddCTP: dideoxyribocytidine 5'-triphosphate
ddGTP: dideoxyriboguanosine 5'-triphosphate
ddNTP: dideoxyribonucleotide 5'-triphosphate
ddTTP: dideoxyribothymidine 5'-triphosphate
dGTP: deoxyriboguanosine 5'-triphosphate
dH\(_2\)O: deionized water
dNTP: deoxynucleotide 5'-triphosphate
dTTP: deoxyribothymidine 5'-triphosphate

E: EcoRI

ESTs: expressed sequence tags

g: gram

IPTG: isopropylthiogalactoside

kb: kilobase

L: liter
lacZ: β-galactosidase
M: molar
Mb: megabase
μCi: microCurie
MCS: multiple cloning site
MER: medium reiteration frequency sequence
μg: microgram
mg: milligram
μl: microliter
ml: milliliter
μM: micromolar
mM: millimolar
mRNA: messenger RNA
ng: nanogram
nm: nanometer
NTSB: nick translation stop buffer
OPLAH: 5-oxo-prolinase
ORF: open reading frame
PCR: polymerase chain reaction
pmol: picomole
rpm: rotations per minute
SINE: short interspersed sequence retroposon
STSs: sequenced tag sites
U: unit
V: volts
VNTR: variable number tandem repeat
X-gal: 5-bromo-4-chloro-3-indolyl-β-D-galactoside
YAC: yeast artificial chromosome
Acknowledgments

To begin, I would like to thank Dr. Stephen Wood for this opportunity to pursue my interest in genetics. His support and guidance over the past two years has made my first step into this field an extremely rewarding one. As well, I am grateful for the time and valuable suggestions provided by Dr. Diana Juriloff and Dr. Ann Rose during the duration of my study. My gratitude to Mike Schertzer for his time, knowledge, and patience. Also, to the rest of my lab a special thank you for the encouragement and friendship that they gave. Alex Parker for his continual reassurance and the interest he showed in my project. And lastly, to my family for the unconditional love and understanding that they have always provided.
Chapter I: Introduction

1.1 Features of GC rich regions in the Human Genome

Even before the Human Genome Project incorporated the identification of genes into its goals (Collins et al., 1993), genetic research was centered around the search for coding sequences. In humans this can be a difficult task due to the enormous size of our genome, only a small portion of which codes for genes. Estimates of the number of genes in the human genome, based on information from genomic sequencing and expressed sequence tag (EST) analysis (Fields et al., 1994), range from 50,000-100,000.

Encoded sequences are not distributed evenly throughout the genome. Some chromosomes are rich in genes, whereas others contain very few transcribed sequences. As well, particular regions within chromosomes are richer in genes than others. Chromosome banding techniques have shown chromosomes to be composed of two distinct types of domains, R bands and G bands. R bands coincide with early replicating GC rich regions and contain 80% of identified genes, whereas G bands are AT rich and replicate later.

R bands are further subdivided into T bands, comprised of the most GC rich regions. T bands represent 15% of the genome, and contain approximately 65% of identified genes (Holmquist, 1992). Thus, GC rich regions possess a higher gene density than GC poor regions. According to Fields et al., (1994), GC rich areas contain an average of one gene every 23.4 kilobases. Gene poor regions have about a tenth the density of coding DNA, or one gene every several hundred kilobases.

Distinct sequences, called CpG islands, are also found in GC rich regions. As with coding sequences, these islands are not distributed evenly throughout the genome. R bands contain the majority, with the highest concentration of CpG islands being found within T bands. Not surprisingly, the distribution of CpG islands between and within chromosomes varies. This pattern was effectively demonstrated by Craig and Bickmore...
Chapter I: Introduction

(1994) using CpG island fragments as fluorescence in situ hybridization probes against a complete set of metaphase chromosomes.

Associations between CpG islands and coding DNA had been hypothesized previously, due to their common characteristics and distribution patterns within mammalian genomes. In 1984, several genes associated with CpG islands were identified in chickens and mice (Tykocinski and Max, 1984). An extensive examination of vertebrate coding sequences conducted three years later, analyzed the abundance and location of CpG islands in relation to genes (Gardiner-Garden and Frommer, 1987). It was concluded, from the genes available at this time, that all ubiquitously expressed genes and many of the tissue-specific genes had CpG islands. Islands located 5' of a gene usually began upstream of the transcription start site, with the distance varying from less than 100 base pairs to over 2 kilobases.

Four years later, an examination of the location of CpG islands in relation to genes in the human genome was undertaken (Aissani and Bernardi, 1991). Over 80% of the CpG islands associated with genes, in this study, were located at the 5' end of the gene. Most of these islands extended into coding DNA, with a few reaching as far as the 3' flanking sequences. 8% of the islands lay entirely within a gene, 4% extended from within the coding sequence to the 3' end, and 2% were present in the 3' flanking DNA only. In 1992, Larsen et al. confirmed the findings of the previous two studies by analyzing over 400 sequenced human genes. Similar to results obtained previously, all ubiquitously expressed genes and 40% of tissue specific genes were associated with CpG islands. To date there is no established case in which a CpG island is not associated with a gene. Thus, from information collected in these studies and elsewhere, it can be inferred that CpG islands are excellent markers for genes. Moreover, being primarily located at the 5'
end of coding sequences, CpG islands could act as landmarks in the identification of transcription start sites for genes.

1.2 Characteristics of CpG Islands

HTF (HpaII tiny fragment) islands or CpG islands account for approximately 2% of the genome and are estimated to be present 45,000 times within it (Antequera and Bird, 1993). In size, they vary from 500 to 2,000 base pairs. As well, all possess a number of distinctive properties distinguishing them from genomic DNA. Throughout most of the genome, CpG dinucleotides are methylated, but remain unmethylated in CpG islands. As mentioned previously, CpG islands are located in GC rich regions (60-70%) (Bird, 1986), as compared to bulk genomic DNA where only 40% of the nucleotides are guanine and cytosine. Most notably, the dinucleotide CpG is found one fifth as often as expected throughout most of the human genome, except within CpG islands, where CpG occurs at the predicted frequency (0.04) (Cross and Bird, 1995).

In order to explain the drastic reduction of CpG dinucleotides in bulk genomic DNA, regions with high mutation rates were investigated. Coulondre et al., (1978) studied base substitution mutations occurring in the lacI gene of Escherichia coli, a highly mutable locus. High frequencies of base substitution at 5-methylcytosines were identified as the cause. Normally, deamination of cytosine to uracil occurs spontaneously. The uracil residue is then rapidly excised by the enzyme uracil-DNA glycosidase, restoring proper base pairing. However, deamination of 5-methylcytosine residues yields thymine (5-methyl-uracil). This results in a G:T base pair, which is subject to normal mismatch repair, but is not removed by the uracil-DNA glycosidase enzyme. The increased number of unexcised G:T base pairs, generated from deamination of 5-methylcytosine, is seen as a hotspot of mutation compared to other sites in the genome (Coulondre et al., 1978).

If the transition of 5-methylcytosine to thymine has been occurring throughout the evolution of the human genome, it could explain the reduction of CpG dinucleotides in
areas where cytosine is methylated. However, if a deficiency in CpG is caused by a transition to TpG and CpA, there must be an equivalent excess of these two dinucleotides in regions where CpG is depleted. To test this theory Bird (1980) compared highly methylated vertebrate genomes with poorly methylated insect genomes. Vertebrates have a greater deficiency of CpG dinucleotides than insect genomes. It was revealed that vertebrates have an abundance of TpG and CpA dinucleotides, approximately equal to the depletion in CpG. In contrast, insects have normal frequencies for both dinucleotides. These results support the hypothesis that deamination of 5-methylcytosine has lead to decreased levels of CpG dinucleotides in the human genome.

1.3 Rare Cutting Enzymes

Due to the depletion of CpG dinucleotides in bulk genomic DNA, rare cutting enzymes can be used to locate regions possessing the unmethylated form of this dinucleotide. Enzymes of this type are called rare cutting as their recognition sequences, which include at least one CpG dinucleotide, are found infrequently within the genome. Many rare cutting enzymes can be used to identify CpG islands. However, some enzymes are not as effective at locating CpG islands as others (Bird, 1989). Due to the high GC content in island regions, rare cutting enzymes whose recognition sequence consists entirely of guanine and cytosine are more likely to locate within an island. Moreover, at least 75% of the sites for enzymes possessing two CpG dinucleotides are situated within CpG islands (Lindsay and Bird, 1987). Together these data suggest that rare cutting enzymes with recognition sequences consisting of only guanine and cytosine, and including at least two CpG dinucleotides, are the best choice for locating CpG islands.

With only a small portion of the human genome coding for genes, many of which are associated with CpG islands, an easy method for locating these landmark sequences would be invaluable. In chromosomal DNA, methylation blocks C-G enzyme sites outside of CpG islands, assisting in their identification by rare cutting enzymes. However, this

Chapter I: Introduction
advantage is lost in cloned DNA where CpG is not methylated. Lindsay and Bird (1987) using SacII, a rare cutting enzyme with the recognition sequence CCGCGG, evaluated the effectiveness of this enzyme at locating islands within cloned DNA. Four cosmids, from a human chromosome X library, were chosen because they contained a SacII site. To confirm that these sites were within CpG islands, DNA from germline and somatic tissues were analyzed for clustering of other rare cutting enzyme sites, as well as their methylation status. Each cosmid fragment contained a cluster of restriction sites and was found to be unmethylated in both tissue types. These fragments were also used as probes in Northern blots, three of which identified transcripts. Thus, SacII successfully identified CpG islands within cloned DNA, results that can be extrapolated to include all rare cutting enzymes.

Through the use of rare cutting enzymes CpG islands can be localized in cloned DNA. Five years later, Lindsay and Bird’s (1987) results were verified through the analysis of human DNA sequences in the EMBL database (Larsen et al., 1992). Sequence data was examined for the presence of a number of rare cutting enzyme sites. Estimates for each enzyme were made regarding the fraction of sites located within CpG islands, as well as the percentage of islands expected to contain an enzyme’s site. Of the twenty-six rare cutting enzymes examined, BssHII stands out as one of the most effective. BssHII, with the recognition sequence GCGCGC, had 83% of its sites in CpG islands and a site in 58% of the islands examined. BssHII’s recognition sequence is estimated to be present 17,400 times in genomic DNA, which amounts to a site roughly every 170 kilobases. However, the average fragment size in this study was 500 kilobases, due to loss of small fragments. This indicates that BssHII sites are not evenly distributed over the entire genome, but instead are clustered in GC rich regions, resulting in a variety of fragment sizes when used to cut large regions of DNA.
1.4 Using CpG islands to Identify Genes

CpG islands have been shown to be associated with approximately 57% of identified genes (Larsen et al., 1992), most often located near the 5' end. Thus, by identifying a CpG island an associated gene may also be uncovered. In fact, DNA adjacent to rare cutting sites throughout the genome has been cloned and used to identify novel genes in GC rich regions. Molecular searches for genes within a 2 megabase (Mb) YAC contig spanning Xp22.3 began with the identification of several CpG islands (Lee et al., 1994). Two were sequenced, both of which contained novel human genes. Transcripts for one of these genes (GS2) are found in all human tissues, typical of most CpG island associated genes. GS2 extends over 26 kilobases, with the CpG island located within the first of seven exons.

A study of chromosome 21, sequenced regions adjacent to 16 NotI sites (Zhu et al., 1993). NotI is a rare cutting enzyme that recognizes the sequence, GCGGCGCGC, with approximately 90% of its sites found within CpG islands (Lindsay and Bird, 1987). Using a program that identifies open reading frames, GRAIL (Uberbacher et al., 1991), five of the clones were found to have high potential as coding regions with seven others having lower probabilities. However, none of the NotI sites sequenced had extensive homologies with sequences present in the database.

A difficulty that arises with the use of NotI to identify CpG islands is that although most of its sites are located within CpG islands, NotI recognition sequences are found so infrequently that only 16% of islands contain a restriction site (Larsen et al., 1992). As a result, many CpG islands remain unidentified as do their associated genes. This problem was alleviated by using EagI and SacII rare cutting enzymes in a search for genes on chromosome 4 (John et al., 1994). Over 76% of the restriction sites for these two enzymes are located within CpG islands, with more than 57% of CpG islands containing a EagI or SacII site (Larsen et al., 1992). Forty-two regions adjacent to a rare cutting enzyme site were identified and cloned from a cosmid contig spanning one million base pairs in the
region where the Huntington gene is located. Seventeen of these clones were found to encode genes. In addition, computer searches using BLAST (Altschul et al., 1990) and GRAIL (Uberbacher et al., 1991) identified nine other clones as potentially containing coding sequences, as well as confirming fifteen of the seventeen previously identified. Thus, cloning fragments adjacent to a single rare cutting enzyme site is effective as a method of identifying genes within cloned DNA.

1.5 Gene Identification

Many techniques are available to identify genes in the human genome. One method being undertaken as part of the human genome project is the sequencing of Expressed Sequence Tags (Adams et al., 1991; Hillier et al., 1996). Expressed Sequence Tags, or ESTs, are segments of sequence from a cDNA clone reverse transcribed from a mRNA. Libraries of cDNA clones have been created from many different tissue types and stages of development in an attempt to obtain representatives of each gene. Normalized libraries have also been created (Bonaldo et al., 1996), where the frequency of all clones is within a narrow range, decreasing the redundancy of cDNAs for a given transcript. Currently, over 709,530 human derived ESTs are present in the dbEST database at the US National Center for Bioinformatics (NCBI). This is largely a result of the Merck IMAGE consortium’s effort to create and make these sequences available (Hillier et al., 1997), although ESTs from other sources are also present.

Despite the tremendous amount of sequence being generated from coding DNA, transcripts with low level expression, in particular tissues or stages of development are selected against in the sequencing of ESTs. Several other approaches are also being used to locate expressed sequences. One method, cDNA selection, involves the hybridization of an amplified cDNA library to YAC or cosmid clones immobilized on nylon membranes (Parimoo et al., 1991). cDNA inserts that hybridize to the cloned genomic DNA are eluted, amplified, cloned, and sequenced. This strategy is useful as a method of enrichment for
expressed sequences from a genomic region of interest. However, it is not a practical technique for locating novel genes throughout the genome.

Another approach, exon amplification, uses the functional sequences required for RNA splicing to isolate expressed sequences from genomic DNA (Duyk et al., 1990; Buckler et al., 1991). Cloned genomic DNA fragments are ligated into a plasmid vector containing splicing sequences. Ideally, only inserted genomic DNA containing an exon will be spliced properly into mature mRNA when present in mammalian cells in culture. However, genomic fragments with cryptic splice acceptor sites can also be identified through this method.

Many of the strategies currently being used to identify genes in the human genome involve the sequencing of expressed sequences. As a result, information from intronic and intergenic regions, such as control and regulatory sequences, are not present. Identification of genes in genomic DNA, through the use of rare cutting enzymes, has an advantage of locating genes and their associated sequences independently of time and location of expression.

1.6 Project Proposal

The majority of genes in the human genome are associated with unique sequences known as CpG islands. These regions of GC rich, unmethylated DNA are distributed unevenly throughout the genome. Fluorescence in situ hybridization studies have localized a large proportion of these landmark sequences to the R bands of chromosomes (Craig and Bickmore, 1994). Chromosome 8 has GC rich R bands distributed from one end to the other, interspersed with G bands (Holmquist, 1992). Additionally, the telomeric region of the long arm of chromosome 8 consists of T bands, which correlate with gene rich regions of the genome. Therefore, one of the characteristics of coding DNA, high GC content, indicates human chromosome 8 contains many gene rich regions.
Chapter I: Introduction

Another feature of CpG islands is their ability to be identified by rare cutting enzymes. Initially, a clustering of rare cutting sites was seen to signify a CpG island. However, it is now recognized that a large percentage of islands are identifiable through the use of a single rare cutting enzyme (Larsen et al., 1992). Rare cutting enzymes vary in their effectiveness at locating CpG islands. Recognition sequences which consist entirely of guanine and cytosine are more often found within islands than those containing adenine and thymine. In addition, enzymes with two or more CpG dinucleotides have over 80% of their recognition sites within CpG islands (Larsen et al., 1992). BssHII is a rare cutting enzyme, with a six base pair recognition sequence, possessing both of these traits. BssHII has an advantage over other restriction enzymes in this category, such as NotI and AscI, as it cuts more frequently and so is present in a larger percentage of CpG islands.

Rare cutting enzymes, such as NotI, SacII and EagI, have successfully been used in cloned DNA to identify coding sequences (Zhu et al., 1993; Lee et al., 1994; John et al., 1994). The purpose of this project is to test whether the rare cutting enzyme BssHII, can be used to identify novel genes in cosmids from a chromosome 8 library (Wood et al., 1992).

For this project, cosmids possessing a BssHII site within their human insert sequence were chosen for analysis. The regions flanking the restriction site were subcloned into a plasmid vector. This was accomplished by digesting each cosmid with EcoRI and BssHII. EcoRI restriction sites flank the human sequence embedded in the cosmid vector DNA. By digesting with BssHII and EcoRI, fragments adjacent to BssHII restriction sites become BssHII-EcoRI fragments or BssHII-BssHII fragments. These fragments of DNA were subcloned and sequenced. The sequence was analyzed using BLAST (Altschul et al., 1990), a program that searches for sequence identity matches against sequence present in the GenBank database. Additional analysis was conducted using GRAIL (Uberbacher et al., 1991), a program that recognize features unique to coding sequences.
Chapter I: Introduction

The subset of cosmids used for this project were selected from the LA08NC01 chromosome 8 library (Wood et al., 1992). This library was constructed from a human x hamster cell line, retaining human chromosomes 4, 8, and 21. Chromosome 8 was isolated by fluorescence-activated flow sorting. Fragments of human DNA, averaging 36.5 kilobases, were cloned into the sCos-1 vector and used to transfected E. coli cells. To minimize the possible coligation of non-contiguous sequences in the genome, flow-sorted insert DNA was dephosphorylated. 85% of the clones in the LA08NC01 library are human, 9% hamster, with the remaining 6% originating from neither species. At this time no chimeric cosmids have been detected. Cosmids used in this project were selected with a probe for the AscI (GGCGCGCC) recognition sequence, another rare cutting enzyme used to identify CpG islands. Due to the fact that the cosmids were not randomly selected predictions can not be made as to the percentage expected to contain a BssHII site. However, with 80% of BssHII restriction sites locating within CpG islands, approximately 80% of the cosmids identified as containing a site should be within a CpG islands. Therefore, it is predicted that this method will be efficient for identifying regions of DNA associated with genes.
Chapter II: Methods and Materials

Prior to the start of this research project, cosmids were selected from the LA08NC01 chromosome 8 library (Wood et al., 1992) for an Ascl rare cutting enzyme site. This was accomplished by probing high density filters containing each cosmid with a radiolabeled probe, NNNNGGCGCGCC, comprised of the Ascl recognition sequence. The cosmids examined in this project were a subset of the selected cosmids that did not contain an Ascl site.

2.1 Isolation of cloned human genomic DNA- Minipreparation of Cosmid and Plasmid DNA

Cosmids from the LA08NC01 chromosome 8 library (Wood et al., 1992) were used to inoculate 5 ml of LB medium containing the antibiotic ampicillin (0.25 mg/ml). Each culture was incubated overnight at 37°C while being shaken vigorously. DNA was prepared using a modification of the protocols described by Birnboim and Doly (1979) and Ish-Horowicz and Burke (1981). To begin, 1.5 ml of the culture was transferred into a eppendorf tube and centrifuged (13,000 rpm) for 5 minutes at 4°C. The liquid medium was removed by aspiration, leaving a pellet of bacteria containing genomic DNA. The pellet was completely resuspended in 100 μl of ice cold Solution I containing lysozyme (4 ng/ml) added just before being used. This solution causes lysing of the bacterial cell wall. To this mixture 200 μl of freshly prepared Solution II was added, mixed thoroughly, and stored on ice. 150 μl of ice cold Solution III was added, mixed completely, and returned to the ice for 5 minutes. The use of these solutions facilitates the removal of bacterial cell wall debris and denaturation of the protein-polysaccharide complex, while leaving nucleic acids in solution. Each tube was then centrifuged for 5 minutes at 4°C. The supernatant was transferred to a clean tube, leaving behind the protein-polysaccharide complex. To
Chapter II: Methods and Materials

precipitate the DNA, an equal volume of isopropanol was added for cosmids or 2 volumes of 95% ethanol was used with plasmids. The nucleic acids were collected by centrifugation, the supernatant removed, and the pellet rinsed with 1 ml of ice cold 70% ethanol. After each tube was centrifuged briefly, the supernatant was again removed, and the pellet dried. Each pellet of nucleic acids was redissolved in 50 μl of TE (pH 8.0) containing RNase (20 μg/ml). This solution was then incubated at 37°C for 30 minutes to allow the RNase to digest any RNA present in the minipreparation of DNA.

LB Medium
400 ml dH₂O in 1L Ehrlemeyer flask
4g peptone (tryptone)
2g yeast extract
2g NaCl
0.4g D-glucose
Place 5 ml of medium in 15 ml tube which were autoclave 20 minutes.

Solution I
50 mM glucose
25 mM TrisCl (pH 8.0)
10 mM EDTA (pH 8.0)
Add dH₂O to bring solution volume up to 200 ml, autoclave for 15 minutes, and store at 4°C.

Solution II
0.2 N NaOH freshly diluted from a 10 N stock with dH₂O
1% SDS

Solution III
5 M potassium acetate 60 ml
glacial acetic acid 11.5 ml
dH₂O 28.5 ml
Resulting in a 3 M solution of potassium and 5 M solution of acetate.

TE
10 mM Tris HCl, pH 8.0
1 mM EDTA

2.2 Restriction Digests

Restriction digest reactions were used to determine if a cosmid contained a BssHII site within the human insert. This was accomplished by digesting each selected cosmid with BssHII. The cosmid vector contains a BssHII site within it (Figure 2.1). If the
human insert also contains a site then at least two bands of DNA will be detected on the agarose gel (described in section 2.3.1). A number of cosmids were found to lack an insert fragment, or to have no $B_{ss}H_{III}$ site within the insert sequence (as is seen with cosmid 150F9 Figure 2.2). Cosmids of this type were presumably selected due to the $B_{ss}H_{III}$ site contained within the vector DNA, and were excluded from the set of cosmids to be used.

After confirming that a $B_{ss}H_{III}$ site was present within the human insert sequence, the cosmid was digested with $B_{ss}H_{III}$ and $EcoRI$. Each of these digests were completed using the following protocol.

Approximately 500 ng of DNA, purified using the minipreparation protocol described in section 2.1, was placed in a eppendorf tube. To each DNA sample 2 μl of 10X BSA, 2 μl of the appropriate buffer, and dH₂O was added, bringing the total volume to 20 μl. Half a unit of the appropriate restriction enzyme was added, the reaction mixed well, and incubated at the appropriate temperature for 1.5 hours. After the digestion reaction was completed 5 μl of gel loading buffer was added to each eppendorf tube and fragments of DNA were analyzed on a 0.7% agarose gel (as described in section 2.3).

10X BSA
bovine serum albumin fraction V 1 mg/ml

Restriction Enzymes
$B_{ss}H_{III}$ (BioLabs; 4 U/μl); NEBuffer for $B_{ss}H_{III}$ with $B_{ss}H_{III}$ alone and in double digest of $B_{ss}H_{III}$ and $EcoRI$. Each reaction was incubated at 50°C.
$EcoRI$ (GibcoBRL; 10 U/μl); 10X REact 3 (GibcoBRL). The reaction was incubated at 37°C.

Gel loading Buffer
0.25% xylene cyanol
0.25% bromophenol blue
40% sucrose in dH₂O
6 x (60 mM EDTA)
2.3 Electrophoresis

Electrophoresis is an effective method used to separate DNA fragments based on their molecular size. In this project agarose gels were employed in several ways: as a means of separating fragments of DNA resulting from restriction enzyme digests; identification; purification; as the first step in preparing a Southern blot. Polyacrylamide gels are capable of separating fragments differing in size by as little as one base pair, and were used in this project to visualize DNA fragments resulting from manual sequencing.

2.3.1 Agarose Gel

0.7% agarose gels were prepared by adding powdered agarose (GibcoBRL) to 1X TBE buffer. At this concentration of agarose, linear DNA molecules ranging in size from 800 base pairs to 10 kilobases are ideally resolved. After the agarose had dissolved in the boiling buffer and cooled, ethidium bromide was added to a final concentration of 0.5 mg/ml. The gel was poured into a mold, and allowed to set. The electrophoresis apparatus, into which the gel was placed, was filled with 1X TBE buffer. Each sample, containing 5 µl of gel loading buffer, was loaded into the wells of the gel. 5 µl (100 ng/µl) of λ DNA marker was simultaneously electrophoresed in an adjacent lane providing a ladder of known fragment sizes. Electrical voltage was applied across the gel until the marker dyes migrated the desired distance, after which DNA fragments were visualized using ultraviolet light (wavelength 302 nm).

1X TBE
89 mM Tris-borate
2 mM EDTA (pH 8.0)

λ DNA Marker
λ DNA (500 ng/µl) 40 µl
10X BSA 16 µl
10X React2 16 µl
MilliQ H₂O 88 µl
Digestion reaction proceeded for 1.5 hours after which 40 μl of gel-loading buffer was added. The eppendorf tube was stored at 4°C.

2.3.2 Polyacrylamide Gel

5% polyacrylamide gels were cast using 30% acrylamide, dH₂O, 5X TBE, and 10% ammonium persulfate. This solution was poured between two glass plates, separated by a thin spacer. A comb was inserted into the gel and the gel left to polymerize for 45 minutes. After the acrylamide had polymerized the gel was placed in a vertical electrophoresis tank filled with 1X TBE buffer. 4 μl from each of the four sequencing reactions, prepared as described in section 2.10.1, were loaded into the wells in the order TGCA. The gel was run at 50 V/cm until the marker dyes had migrated to the bottom of the gel, approximately 1.5 hours. Whatman 3MM paper was placed on the sequencing gel, after the apparatus was dismantled, allowing the removal of the polyacrylamide gel from the glass plate. The gel was covered with a second piece of paper and dried for approximately 20 minutes under a vacuum gel dryer at 80°C. Fragments resulting from the sequencing reaction were visualized by exposing the bands of radioactive DNA to film.

30% Acrylamide
29 g acrylamide
1 g N,N'-methylenebisacrylamide
H₂O to 100 ml
Heat the solution to 37°C to dissolve the chemicals.

10% Ammonium persulfate
1 g ammonium persulfate
H₂O to 10 ml
Store solution at 4°C.
Chapter II: Methods and Materials

2.4 Exclusion of Cosmids containing Alu Sequence

93C11ΔBssHII, the first cosmid to be successfully sequenced, was found to contain an Alu element. Alu is a family of repetitive DNA, found in excess of 500,000 times in the human genome (Deininger et al., 1981). A BssHII site is present in the consensus sequence of the progenitor Alu gene, as well as within subfamilies Alu-J, Alu-Sc, and Alu-Sa (Labuda and Striker, 1989). To avoid further sequencing of these repetitive regions two methods were employed. One involved the use of a transfer technique first described by Southern (1975) (section 2.4.1), with Alu DNA being used as a probe on blots of restriction digests (Figure 2.3). The other allows the screening of a large number of bacterial colonies containing a constructed deleted cosmid (section 2.4.2) for Alu sequence (Figure 2.4). BssHII-EcoRI and BssHII-BssHII fragments found to contain an Alu element were excluded from the set of DNA fragments to be sequenced.

2.4.1 Southern Blot

In this technique, DNA from each cosmid was digested with restriction enzymes and separated according to size by electrophoresis on an agarose gel. The DNA was then denatured and transferred to a solid support, in the case of this project Nylon transfer membranes were used. DNA attached to the membrane was then hybridized to radiolabeled Alu sequence enabling the identification of any bands complementary to this sequence.

To begin, after electrophoresis the gel was transferred to a glass baking dish containing several volumes of 1.5 M NaCl, 0.5 N NaOH. The gel was soaked in this solution for 30 minutes, allowing the DNA to denature. It was then rinsed in deionized water (dH₂O), and neutralized for 30 minutes by soaking it in a solution of 1 M Tris (pH 7.4), 1.5 M NaCl. In order to transfer the DNA to the Nylon transfer membrane, it must be eluted from the agarose gel and deposited on the membrane. This was accomplished using the capillary transfer method (described by Sambrook and Maniatis, 1989), in which a flow of liquid carries the DNA fragments from the gel to the membrane using capillary
action established and maintained by a stack of dry paper towels (Figure 2.5). The apparatus required for this method was assembled, with the gel placed with the membrane (Hybond-N+ Nucleic acid transfer membrane) on top of it. The Southern blot apparatus was left over night to allow a complete transfer of DNA to the solid support. Once the transfer was completed the membrane was baked for 2 hours at 80°C in a vacuum oven, to fix the DNA to the membrane.

2.4.1.1 Synthesis of Radiolabeled Alu Probe by Primer Extension

A fragment of Alu was obtained for use as a probe by digesting plasmid p51A8EBa8 with EcoRI, using the method described in section 2.2. The digest was electrophoresed on a 0.7% agarose gel for 1.5 hours at 60 volts. The 800 base pair insert fragment containing an Alu element was excised from the gel and placed in the well of a low melting point agarose gel (Sigma Type VII: Low Gelling Temperature Agarose). After allowing the DNA to enter the gel, the band was again excised and placed in an eppendorf tube to which 200 μl of TE was added. The solution was boiled for 5 minutes and stored at 4°C until needed.

15 μl of the purified linear Alu fragment was pipetted into a eppendorf tube, boiled for 5 minutes to denature the DNA, and placed on ice for 5 minutes. To this 2.5 μl of 10X BSA, 5.0 μl OLB-A, 2.5 μl α-dATP^{32} (10 μCi/μl), and 1 μl (1 U/μl) of Klenow fragment (Pharmacia Biotech) was added (Sambrook et al., 1989). The eppendorf tube was centrifuged briefly, and left to incubate overnight at room temperature. To stop the reaction an equal volume of Nick Translation Stop Buffer (NTSB) was added. The probe was then boiled for 5 minutes, and placed on ice for 5 minutes.
Chapter II: Methods and Materials

5X OLB-A
solution O: 1.25 M Tris Cl pH 8.0 and 0.125 M MgCl₂
solution A: 1 ml solution O, 18 μl B-mercaptoethanol, 5 μl of each dNTP (G/C/T)
each dNTP is 0.1 M in 3 mM Tris pH 7.0, 0.2 mM EDTA
solution B: 2 M Hepes titrated to pH 6.6 with 4 M NaOH
solution C: Hexadeoxyribonucleotides (50 units Pharmacia 27-2166-01) suspended in 550
μl TE to 90 OD units/ml
solutions A:B:C are mixed in the ratio of 100:250:150

NTSB
50 mM EDTA
20 mM NaCl
0.1% SDS
500 μg/ml Salmon Sperm DNA
d₂O

2.4.1.2 Hybridization

After fixing the nucleic acids to the membrane, as described in section 2.4.1, each
membrane was placed in a heat-sealable bag into which approximately 4 ml of warm (65°C)
prehybridization solution was added. The end of the bag was sealed, and incubated for 1
hour submerged in 65°C water. After an hour of prehybridization, during which blocking
agents anneal to the membrane suppressing background hybridization signals, the
denatured probe (prepared as described in section 2.4.1.1) was added to the
prehybridization solution. The bag was then resealed twice, to avoid radioactive
contamination of the water bath, and resubmerged in 65°C water. After the hybridization
was complete each membrane was removed from the bag and immersed in 1X SSC, 0.1%
SDS solution for 15 minutes, the membrane was then washed in 0.2X SSC, 0.1% SDS
solution for 45 minutes. After being washed, each membrane was wrapped in plastic film
wrap, and exposed to X-ray film. The film was developed, and bands of DNA hybridizing
to the probe were identified.

Prehybridization solution
6X SSC
5X Denhardt’s reagent
0.5% SDS
100 μg/ml salmon sperm DNA
2.4.2 Screening by Hybridization

A method for lysis of bacterial colonies on transfer membranes, and attachment of DNA to these membranes, was first described by Grunstein and Hogness (1975). Once fixed these nucleic acids can be hybridized to a radiolabeled probe, first described by Gillespie and Spiegelman (1965). This method allows the screening of a large number of bacterial colonies carrying cosmid or plasmid DNA containing a sequence of interest. In the case of this project, this method was used to identify cosmids containing Alu sequence and bacterial colonies containing fragments subcloned into a pKSIIAsc vector (discussed in section 2.7) eliminating false positives.

A transfer membrane (NEF-978 Colony/Plaque Screen Hybridization Transfer Membrane) was placed on an agar plate containing ampicillin. Sterile toothpicks were used to transfer one bacterial colony onto a filter and a master agar plate containing ampicillin. These colonies were numbered and placed in identical positions on each plate, allowing them to be easily identified after hybridization. Both plates were inverted and stored overnight at 37°C. After growing overnight the master plate was sealed with Parafilm and stored at 4°C.

Two pieces of Whatman 3MM paper were cut, placed in the bottom of trays, and saturated with one of the following solutions. The first paper was wet with denaturing solution (1.5 M NaCl, 0.5 N NaOH), the second paper with neutralizing solution (1.5 M NaCl, 0.5 M Tris Cl [pH 7.4]). Each transfer membrane was placed on the denaturing solution paper for 8 minutes and the neutralizing solution paper for 8 minutes, causing lysing of the bacterial colonies growing on the membrane. Lastly, each filter was placed in dH2O and wiped gently with wet tissue paper to remove excess cell debris. DNA was fixed onto the membranes by baking for 2 hours at 80°C in a vacuum oven. Hybridization of radiolabeled probes to these membranes will be discussed in section 2.4.2.1, Alu probes were prepared as described in section 2.4.1.1, and pKSIIAsc probes were prepared as described in section 2.4.2.1. After colonies were found to contain Alu or pKSIIAsc
sequence, appropriate colonies were picked from the master plate and used to inoculate LB medium.

**Agar plates containing ampicillin**
400 ml dH2O in 1L Ehrlemyer flask
4 g BactoPeptone (DIFCO laboratories 0118-17-0)
2 g Yeast Extract (DIFCO laboratories 0127-17-9)
2 g NaCl (BDH Inc.)
0.4 g Dextrose (FisherChemical D16-500)
4.8 g agar (BDH Inc.)
Autoclave for 20 minutes. Allow solution to cool, add 20 mg ampicillin, pour plates, and store at 4°C.

### 2.4.2.1 3' Termini Labeling of T3 Probe with T4 Kinase

Fragments adjacent to BssHII restriction sites were subcloned into the pKSIIAsc vector (Figure 2.6), which contains the T3 promoter allowing annealing of the T3 primer (5'-ATTACCCTCACTAAAG-3'). To identify bacterial colonies containing a pKSIIAsc vector a radiolabeled T3 probe was hybridized to transfer membranes prepared as described in section 2.4.2.

The probe was prepared by pipetting 4.5 μl of T3 primer (10 μM), 2.0 μl 5X kinase buffer (GibcoBRL), 1.0 μl T4 kinase (10 U/ml; GibcoBRL), and γATP-P\(^{32}\) (10 μCi/μl) into an eppendorf tube. The reaction was incubated for 1 hour at 37°C, followed by heating to 95°C for 5 minutes to stop the reaction. Approximately 4 ml of hybridization solution was added to a heat sealable bag containing the transfer membranes, to which the labeled T3 probe was added. The hybridization was allowed to proceed overnight at 42°C, after which each filter was washed for 2 minutes in 0.2X SSC at the same temperature. Each filter was dried briefly, covered in plastic wrap, and exposed to X-ray film for approximately 30 minutes.
Chapter II: Methods and Materials

2.5 Construction of Deleted Cosmids

Deleted cosmids were created to be used as templates for the sequencing of DNA adjacent to BssHII sites. This was accomplished by digesting selected cosmids with BssHII, as described in section 2.2, followed by religation of the resulting fragments, shown in Figure 2.7. In an eppendorf tube 2.5 μl of the digested cosmid DNA, 4 μl of 5X ligase buffer (GibcoBRL), 1 μl of T4 DNA ligase (GibcoBRL; 1 U/μl), and dH₂O were added to a final volume of 20 μl. Each reaction was left overnight at 14°C and used to transform competent bacteria, using the protocol described in section 2.8.

Religation of these BssHII fragments potentially creates three types of ligated fragments which are taken up by the competent bacteria. However, plating the bacteria out on a media containing ampicillin selects for those possessing the ampicillin resistance gene. Deletion cosmids obtained in this manner were sequenced using an ABI automated sequencer, as described in section 2.10.2.

2.6 DNA Extraction from Agarose Gel

In order to subclone regions adjacent to BssHII recognition sequences, cosmids containing a BssHII site in the human insert DNA were digested with BssHII and EcoRI, using the method described in section 2.2. These resulting fragments were separated on an agarose gel by electrophoresis, as described in section 2.3.1. Followed by excision of BssHII-EcoRI and BssHII-BssHII bands. DNA from these bands was extracted from the agarose gel using a QIAEX II Agarose Gel Extraction kit (QIAGEN).

After excision of a DNA band from the agarose gel it was placed in an eppendorf tube, to which Buffer QX1 and 10 μl of QIAEX II was added. The solution was incubated at 50°C for 10 minutes. The addition of Buffer QX1 solublized the agarose gel, as well as creating a high salt environment. QIAEX II are silica-gel particles that adsorb nucleic acids.
in the presence of high salt. Each tube was centrifuged for 30 seconds, and the supernatant removed. The pellet was washed once with 500 μl of Buffer QX1 to remove excess agarose, and twice with Buffer PE to remove salt contamination. The pellet was then air-dried, resuspended in 20 μl of dH₂O, and incubated for 5 minutes. Elution of the DNA was accomplished during this step by lowering the salt concentration allowing the nucleic acids to be released from QIAEX II silica gel particles. Each eppendorf tube was centrifuged, the supernatant containing the DNA fragment removed, and placed in a clean eppendorf tube.

2.7 Subcloning of Cosmid Fragments

A modified pBluescript (Stratagene) vector, shown in Figure 2.6, was used in this project to subclone fragments flanking BssHII sites. Modifications to pBluescriptIIKS+ replaced the SpeI and XbaI restriction sites with an AscI restriction site, as described in DeBella et al., (1997). pBluescript contains two selectable markers, an ampicillin resistance gene and the regulatory sequences and coding information for the amino-terminal region of the β-galactosidase gene (lacZ). Within the coding region of the lacZ gene is a polycloning site which does not disrupt the reading frame unless a fragment of DNA is cloned into it. When this vector is present in a host bacterial cell which encodes the carboxy-terminal portion of the β-galactosidase gene, the two fragments are able to associate forming an enzymatically active protein (Ullmann et al., 1967). Bacteria possessing the pBluescript plasmid without an insert are identified as blue colonies when grown in the presence of 5-bromo-4-chloro-3-indolyl-β-D-galactoside (X-gal). However, insertion of a fragment into the cloning site disrupts the amino-terminal segment resulting in white colonies.
2.7.1 Preparation of pBluescript vector

The pBluescript vector (pKSIIAsc) was cut with restriction enzyme(s), resulting in protruding terminal ends complementary to those of DNA fragments to be subcloned. This was accomplished by digesting 10 μl of pKSIIAsc (800 ng/μl), 5 μl 10X BSA, 5 μl of the appropriate restriction enzyme buffer, 28 μl dH₂O, with 2 μl of the appropriate restriction enzyme. The reaction was allowed to proceed for 1.5 hours. Once complete 200 μl of 95% ethanol was added, the eppendorf tube centrifuged for 5 minutes, and the supernatant removed. The resulting pellet was washed with 400 μl of ice cold 70% ethanol, centrifuged briefly, the supernatant removed, and the pellet allowed the air dry. Once dry the purified vector DNA was resuspended in 80 μl TE. 10 μl of the digested vector was run out on an agarose gel to ensure that the digestion was complete. It was then stored at -20°C.

Restriction Enzymes
Ascl (10 U/μl BioLabs); used to cut pKSIIAsc when cloning BssHII fragments; 10X NEBuffer 4
EcoRI (10 U/μl; GibcoBRL); used to cut pKSIIAsc when cloning EcoRI fragments; 10X React 3 buffer (GibcoBRL)
Ascl and EcoRI used to cut pKSIIAsc when cloning BssHII-EcoRI fragments; 10X NEBuffer 4

2.7.2 Ligation Reaction

Ligation reactions were set up in eppendorf tubes each containing 10 μl of DNA (approximately 10 ng/μl) purified using the method described in section 2.6, 5.6 μl dH₂O, 2 μl T4 DNA ligase buffer (GibcoBRL), 2 μl pKSIIAsc (10 ng/μl) cut with the appropriate enzyme(s), and 0.4 μl T4 ligase (1 U/μl; GibcoBRL). A positive control was also set up with each ligation reaction containing a pBluescript vector capable of religation and no
foreign DNA. The reaction mixtures were incubated at 14°C overnight. Figure 2.8 diagrams the steps of this process.

2.8 Transformation of Competent E. coli

Competent DH5α Escherichia coli cells (Inoue et al., 1990), stored at -70°C, were thawed on ice after which 50 μl was pipetted into a pre-chilled 15 ml tube. 10 μl of a ligation reaction (prepared as described in sections 2.5 and 2.7) was added to each tube, swirled to mix, and stored for 30 minutes on ice. Each tube was then heat shocked for 45 seconds at 42°C and returned quickly to ice for 2 minutes. 400 μl of LB medium was added to each tube, which were then incubated for 45 minutes at 37°C. 120 μl of transformed competent cells were transferred onto plates, spread over the surface of the plate, and incubated overnight at 37°C. Ampicillin plates were used for deleted cosmids and XIA plates for subcloned fragments.

Bacteria containing deleted cosmids were the only type able to grow on the selective media. These colonies were used to inoculate LB medium, and were grown overnight at 37°C. Bacteria transformed with a pKSIIAsc vector containing an insert fragment of foreign DNA appeared as white colonies on XIA plates due to an inability to α-complement (Ullmann et al., 1967). These colonies were used to inoculate LB medium, and were grown at 37°C overnight. DNA from these cultures was prepared using the minipreparation protocol (section 2.1), digested with the appropriate enzyme (section 2.2) to ensure that DNA of the correct size was present, and run out on a 0.7% agarose gel (section 2.3.1).
Chapter II: Methods and Materials

XIA Plate
400 ml dH₂O in 1L Ehrlemyer flask
4 g BactoPeptone (DIFCO laboratories 0118-17-0)
2 g Yeast Extract (DIFCO laboratories 0127-17-9)
2 g NaCl (BDH Inc.)
0.4 g Dextrose (FisherChemical D16-500)
4.8 g agar (BDH Inc.)
Autoclave for 20 minutes. Allow solution to cool, add 20 mg ampicillin, 25 mg X-gal, 60 mg IPTG, pour plates, and store in the dark at 4°C.

2.9 DNA Purification for Sequencing

Fragments adjacent to BssHII sites from ten cosmids were subcloned into a pKSIIAsc vector and used to transform competent bacteria. Bacterial colonies containing a subcloned fragment were identified either visually as white colonies on an XIA plate, or by hybridization of transfer membranes with a T3 probe present in the pKSIIAsc vector (Figure 2.9). Prior to purification of template DNA for sequencing, restriction digests were used to confirm that human insert DNA of the expected size was present (Figure 2.10).

Once a fragment was confirmed to have been cloned into the pKSIIAsc vector, plasmid DNA was purified for sequencing using a Plasmid Mini Kit (QIAGEN). This was accomplished using a modified alkaline lysis protocol, followed by binding of the nucleic acids to an anion-exchange resin present in a column (QIAGEN-tip). DNA is bound to the resin under low salt conditions, after which impurities such as RNA and proteins are removed by a medium level salt wash. DNA is then eluted in a high salt buffer at pH 7.0, concentrated, and desalted by isopropanol precipitation.

1.5 ml of an overnight culture, inoculated from a single bacterial colony confirmed to contain a subcloned fragment, was centrifuged and the pellet resuspended in 300 µl of chilled Buffer P1. 300 µl of Buffer P2 was added, mixed, and incubated for 5 minutes at room temperature enabling bacterial cell lysis. 300 µl of ice cold Buffer P3 was added, mixed, and incubated on ice for 5 minutes. Each eppendorf tube was then centrifuged for
Chapter II: Methods and Materials

10 minutes, the supernatant removed, and applied to a QIAGEN-tip. Before the supernatant was added, each column was equilibrated with 1 ml of Buffer QBT. After the supernatant had moved through the column, 4 ml of Buffer QC was added to remove any remaining contaminants. 800 μl of Buffer QF was applied to each column, raising the salt concentration, enabling nucleic acids bound to the resin to be eluted into a clean eppendorf tube. DNA was precipitated by the addition of 560 μl of isopropanol, and centrifuged for 30 minutes. Each pellet was then washed with 1 ml of 70% ethanol, centrifuged briefly, and air dried. The pellet of purified DNA was resuspended in 15 μl of dH₂O. 1 μl of each plasmid was digested for 1.5 hours with EcoRI, and run out on a gel. This was done to ensure that the plasmid was purified, as well as determining the concentration of DNA.

Buffer P1 (Resuspension Buffer)
50 mM Tris HCl, pH 8.0
10 mM EDTA
100 μg/ml RNase A

Buffer P2 (Lysis Buffer)
200 mM NaOH, 1% SDS

Buffer P3 (Neutralization Buffer)
3.0 M potassium acetate, pH 5.5

Buffer QBT (Equilibration Buffer)
750 mM NaCl
50 mM MOPS, pH 7.0
15% isopropanol
0.15% Triton X-100

Buffer QC (Wash Buffer)
1.0 M NaCl
50 mM MOPS, pH 7.0
15% isopropanol

Buffer QF (Elution Buffer)
1.25 M NaCl
50 mM Tris HCl, pH 8.5
15% isopropanol

26
2.10 Sequencing

After the template was purified, as described in section 2.9, it was sequenced manually or using an automated sequencer. Sequencing was completed using either a M13 (5'-GTAAAACGACGGCCAGT-3') or T3 (5'-ATTAACCCTCACTAAAG-3') primer, which flank the cloning sites of the SuperCos1 and pBluescript vectors.

2.10.1 Manual Sequencing

Manual sequencing was performed on a subcloned fragment by the Sanger dideoxy-mediated chain-termination method (Sanger et al., 1977) using Sequenase 2.0. Sequenase 2.0 lacks the 3'→5' exonuclease activity, resulting in an extremely stable enzyme with a higher specific activity than other versions of this enzyme. Sequencing reactions with Sequenase version 2.0 occur in two steps. The first is a polymerization reaction with limiting concentrations of dNTP's, including ATP-S\(^{35}\), extending the primer approximately 25 nucleotides. Followed by a second set of four reactions where chains of DNA are rapidly extended and terminated by the incorporation of ddNTP. After the reactions were complete, the four samples were loaded into a polyacrylamide sequencing gel, as described in section 2.3.2.

For the subclone being sequenced 2 µl of 2 M NaOH, 2 mM EDTA and 18 µl of DNA template (200-400 ng/µl) was pipetted into a 0.5 ml eppendorf tube. After 5 minutes 2 µl of 2 M ammonium acetate pH 5.4 was added to terminate the reaction. DNA was precipitated by adding 50 µl of 95% ethanol and centrifuging the sample for 15 minutes. The DNA pellet was washed by adding 200 µl of 70% ethanol to the eppendorf tube and centrifuging for 5 minutes. The pellet was dried and resuspended in 6.5 µl of water. To
each DNA sample 0.5 μl of 10 μM primer (T3) and 1 μl of DMSO was added. The eppendorf tube was boiled for 3 minutes, followed by cooling at -196°C for 5 minutes in liquid nitrogen. The eppendorf tube was quickly thawed, 2 μl of Sequenase buffer mix was added and the tube was stored at room temperature for 5 minutes. To four reaction tubes 2.5 μl of one of four dideoxynucleotides (ddATP, ddTTP, ddCTP, ddGTP) was added. In another eppendorf tube 1.0 μl of DTT, 2.0 μl of diluted GTP mix, 0.5 μl DMSO, 1.8 μl Sequenase dilution buffer, 1.0 μl S\(^{35}\) label, and 0.25 μl Sequenase enzyme were mixed. 6.3 μl of this was added to the eppendorf tube containing the DNA template. This reaction was left 4 minutes at room temperature to pre-elongate. 3.5 μl of the DNA template mixture was added to each of the 4 eppendorf tubes containing one dideoxynucleotide, and heated for 4 minutes at 37°C. 4 μl of gel-loading buffer was added to each tube, the tubes mixed, and 4 μl of each reaction loaded into a pre-heated sequencing gel.

2.10.2 Automated Sequencing

All but one of the subcloned fragments were sequenced using an automated sequencer (Applied Biosystems Model 373 Stretch). Sequence was obtained from 500 ng of QIAGEN purified double stranded DNA template using ABI's AmpliTaq FS DyeDeoxy™ Terminator Cycle Sequencing. Unlike the manual method of sequencing described in section 2.10.1, all four base reactions occur in one eppendorf tube on a thermocycler. Sequence procured through this method can be 98.0% accurate to more than 650 base pairs, whereas manual sequencing generally provides 300 to 400 base pairs of reliable sequence.
In a 0.6 ml eppendorf tube 8.0 μl of terminator premix, 500 ng of template, 3.2 pmol of primer (M13 or T3), and dH₂O are added to a final volume of 20 μl. A drop of mineral oil was added to overlay the reaction. Each tube was placed in a thermocycler preheated to 96°C where 25 cycles of the following sequence was completed. 96°C for 30 seconds; 50°C for 15 seconds; 60°C for 4 minutes; to 4°C until reaction was removed for purification. Each reaction was transferred to a 1.5 ml eppendorf tube containing 2 μl of 3M Sodium acetate, pH 4.6 and 50 μl 95% ethanol. Precipitation of the sequencing reaction with ethanol removed excess dye terminators. The tube was centrifuged for 20 minutes, the ethanol removed, pellet dried, and stored at -20°C until it was run on the sequencing gel.

2.11 DNA Analysis

After sequence was obtain from a subcloned fragment, analysis was performed using BLAST, GRAIL, and PROSITE. BLAST is available at the NCBI web site (www.ncbi.nlm.nih.gov). GRAIL and PROSITE can be found at the Baylor College of Medicine search launcher (kiwi.imgen.bcm.tmc.edu:8088).

2.11.1 BLAST

BLAST, basic local alignment search tool, was used as a method of comparison between sequence obtained from the subcloned fragments and sequence present in GenBank databases. BLAST contains a rapid database searching algorithm that optimizes local similarities between sequences and then extends these alignments based on defined match and mismatch criteria, but does not allow for gaps to improve the alignment (Altschul et al., 1990). Similarity searching begins by looking for similar segments between the query sequence and a database sequence, the statistical significance of any
match found is evaluated, and those matches found to be significance are reported back. Matches reported in this thesis had a P value of 2.8e-18 or less, signifying that the likelihood of this match representing a random alignment is 2.8 times 10 to the minus 18th power.

GenBank is the National Institute of Health database, based at the National Center for Biotechnology, which contains all known nucleotide and protein sequences. The sequence located at GenBank was obtained from numerous organisms such as homo sapiens, *C. elegans*, *S. cerevisiae*, and *mus musculus*. Nucleotide sequence data is organized into nonredundant, expressed sequence tags (ESTs), and sequence tagged sites (STSs) databases. Due to the constant addition of new sequence GenBank releases updated versions of the database. GenBank release 100 was searched in this project using BLASTN, a program that compares the nucleotide query sequence with a database of nucleotide sequences. The non-redundant database containing 310,264 sequences; EST database release 051697 containing 1,024,937 entries; STS database release 051597 containing 45,808 entries; as well as an Alu database, REPBASE (Claverie and Makalowski, 1994), containing Alu repetitive sequence, were searched. BLASTP was used to compare amino acid sequences to proteins in the SwissProt database, which contains 59,576 non-redundant protein sequences.

2.11.2 GRAIL

GRAIL 1.3 was used to locate protein coding regions through the application of a set of seven sensor algorithms designed to elicit the coding potential of a region of sequence (Uberbacher and Mural, 1991). The first of these algorithms, a frame bias matrix, enables the identification of potential coding regions and the favored reading frame. If a region codes for a protein then one of the reading frames should have a significantly better correlation to the matrix than the other two frames. The algorithm calculates the correlation coefficient between the matrix and each reading frame. The difference between
Chapter II: Methods and Materials

the best and worst coefficient is used as an indicator of coding potential. Fickett, the
second sensor, examines properties of coding sequences. The occurrence of each of the
four bases in the sequences is compared to the recurrence of the bases in coding DNA.
After which the overall base composition of the test DNA is compared to the known
compositions of coding and noncoding DNA. The third sensor, dinucleotide fractal
dimension, examines the occurrences of dinucleotides. Certain dinucleotides, such as AA
and TC, are commonly found. Their occurrence is compared to the occurrence of
dinucleotide CG. The occurrence of these dinucleotides differs between coding and
noncoding DNA enabling this sensor to compare the sequence of interest to intronic and
coding DNA. Coding 6-tuple word preferences is a sensor that examines the frequency of
nucleotide ‘words’ in a length of sequence. DNA sequence from different regions of the
genome, intronic and coding, have distinct distributions of word occurrences. Frequencies
of these words occurring in coding compared to noncoding human DNA provides an
indicator for a gene. Another GRAIL sensor compares in-frame 6-tuples from the test DNA
with in-frame 6-tuples from coding DNA. Word commonality, the sixth sensor, focuses
on the overall frequency of a given 6-tuple in bulk DNA. Intronic regions use extremely
common words whereas exons use relatively rare words. This algorithm takes this feature
into account, assisting in the detection of coding DNA. The final sensor, repetitive 6-tuple
word preferences, examines the test sequence for several classes of repetitive DNA using
6-tuple statistics. This algorithm utilizes the fact that highly repetitive DNA rarely encodes
a protein. The outputs from these sensor algorithms are integrated using a neural net that
predicts the location of a coding region within the segment of sequence.

An updated version of this program, GRAIL 1a, was used in this project. After
locating a potential open reading frame, the program evaluates a number candidate
sequences using information from the 60 base pairs flanking either side of the open reading
frame. Tests by the programmers showed that this version of GRAIL was better at finding
true exons, eliminating false positives, and finding the boundaries of coding regions.
Chapter II: Methods and Materials

GRAIL 1a was able to recognize 82% of exons, in a test set, with a false positive rate of 11%. Of the exons over 100 base pairs in length, 95% were recognized, with the correct frame assignment given to 98% of those identified.

2.11.3 PROSITE

PROSITE is a protein motif database that was used as a method for determining the potential function of a protein (Bairoch, 1993). This database was searched for particular motifs in distantly related proteins that may not have been identified by BLAST sequence alignment searches. Examination of specific regions with conserved structure and sequence can reveal biological function. The PROSITE version used in this project contained 1143 protein patterns.

2.12 Sequence tagged sites

A novel STS was developed for the following segments of DNA by designing PCR primers to be used against a cell hybrid panel specific to chromosome 8 (Wagner et al., 1991; Wood et al., 1986) and the G3 radiation hybrid cell panel (Stewart et al., 1997). Results from the G3 panel were submitted to the Stanford RH server (www-shgc.stanford.edu/rhserver2/rhserver_form.html) which provides a map location for the STS.

2.12.1 GOR gene

Polymerase chain reaction (PCR) was used to localize the novel GOR gene on chromosome 8. PCR primers, oligo320 (5'-AGGTTGCCCCAAGTCCAAGC-3') and oligo321 (5'-GCTGTCTGACCTTCCACATC-3'), were designed flanking a 286 base pair region within this gene.
Chapter II: Methods and Materials

A set of PCR reactions were set up using oligo320 and oligo321 as primers against a set of cell hybrids spanning chromosome 8 (Wagner et al., 1991; Wood et al., 1986) and the G3 radiation hybrid cell panel (Stewart et al., 1997). 25 ng of DNA was included in 25 µl amplification reaction containing 50 mM Tris Cl, pH 8.3, 0.05% Tween-20, 0.05% Nonidet-40, 2.5 mM of MgCl₂, 0.2 mM of each dinucleotide (dTTP, dGTP, dCTP, dATP), 0.5 µl of Taq polymerase (4 U/µl; BRL), and 1.0 µl of each primer (10 ng/µl).

Each reaction tube was placed in a thermocycler to undergo the following amplification reaction: 40 cycles of a 2 minute 94°C denaturing step, followed by 30 seconds at 58°C to allowing annealing of the primers, and a 1 minute extension at 72°C. In the final cycle the 72°C extension step was lengthened to 7 minutes, concluded by rapid cooling to 4°C.

2.12.2 5-oxo-L-prolinase gene

Polymerase chain reaction (PCR) will be used against the G3 radiation hybrid cell panel (Stewart et al., 1997) to localize the novel OPLAH gene on chromosome 8. PCR primers, oligo322 (5'-TTCCAAAGGCACGCAACATG-3') and oligo319 (5'-AGGGCCATCCTGCAGGTG-3'), were designed to flank a 96 base pair region within this gene.

2.12.3 Human EST W90101

Polymerase chain reaction (PCR) was used to localize the human EST W90101 on chromosome 8. PCR primers, oligo115 (5'-TTCTCCTCTCCGCTTGCTG-3') and oligo131 (5'-GAGGGACAAGTATCCAGTCC-3'), were designed flanking a 300 base pair region within this EST.
Chapter II: Methods and Materials

A set of PCR reactions were set up using oligo320 and oligo321 as primers against a set of cell hybrids spanning chromosome 8 (Wagner et al., 1991; Wood et al., 1986). 25 ng of DNA was included in 25 µl amplification reaction containing 50 mM Tris Cl, pH 8.3, 0.05% Tween-20, 0.05% Nonidet-40, 2.5 mM of MgCl<sub>2</sub>, 0.2 mM of each dinucleotide (dTTP, dGTP, dCTP, dATP), 0.5 µl of Taq polymerase (4 U/µl; BRL), and 1.0 µl of each primer (10 ng/µl). Each reaction tube was placed in a thermocycler to undergo the following amplification reaction: 40 cycles of a 2 minute 94°C denaturing step, followed by 30 seconds at 58°C to allowing annealing of the primers, and a 1 minute extension at 72°C. In the final cycle the 72°C extension step was lengthened to 7 minutes, concluded by rapid cooling to 4°C.

PCR products resulting from all reactions were examined by adding 5 µl of gel loading buffer to each reaction tube, loading the products into a 2.0% agarose gel, and electrophoresing at 8 V/cm for approximately 2 hours. 5 µl (100 ng/µl) of ϕX174 marker (HaeIII digested ϕX174 phage DNA) was simultaneously electrophoresed in an adjacent lane in order to determine the size of the PCR products by comparing band migration patterns.
Figure 2.1

Diagram of sCos1 vector.

The SuperCos1 vector has a cloning capacity of 30-42 kb. It contains an ampicillin-resistance gene to allow for selection. The cloning region is flanked by T3 and T7 primers, that can be used to sequence human insert DNA. (Wahl et al., 1987)
Figure 2.2
Digestion of cosmids with restriction enzymes.

Cosmid 150F9 has one BssHII site in the vector DNA, but no site in the inserted fragment. 147D8 has a BssHII site in the vector and insert DNA, and was included in the set of cosmids investigated in this thesis.
Figure 2.3
Detection of Alu repetitive sequence: Southern blot.

A Southern blot of cosmid restriction digests (shown on left) probed with Alu DNA. Bands containing Alu sequence appear dark in the autoradiograph (shown on right).
E: EcoRI; E/B: EcoRI/BssHII double digest
Figure 2.4

Detection of Alu repetitive sequence: Colony blot.

Autoradiograph of bacterial colonies containing constructed deleted cosmids probed with Alu DNA.

Colonies 1-4 are cosmid 46F4ΔBssHII; 5-12 are cosmid 95C11ΔBssHII; 13-20 are cosmid 124H12ΔBssHII; 21-28 are cosmid 141F3ΔBssHII; 29-36 are cosmid 156G7ΔBssHII; 37-42 are cosmid 166H7ΔBssHII. Colonies containing Alu appear as dark dots.

Cosmid 95C11ΔBssHII was excluded from the set of cosmids to be examined because it was positive with the Alu probe.
Figure 2.5

Apparatus for the capillary transfer method of Southern blotting.

Capillary transfer of DNA from agarose gel. The buffer is drawn from below through the gel into the stack of paper towels. DNA is eluted from the gel and deposited on the Nylon transfer membrane. The weight at the top ensures a tight connection between each layer in the transfer system.
Chapter II: Methods and Materials

Figure 2.6

Diagram of modified pBluescript vector.

The ampicillin resistance gene allow for antibiotic selection of the vector. ColE1 origin is a plasmid origin of replication allowing DNA replication of the plasmid. Only a portion of the lacZ gene is present enabling α-complementation for the blue/white color selection of recombinant plasmids. The presence of an inducible lac promoter upstream from the lacZ gene increases the α-peptide expression. MCS is a multiple cloning site flanked by T3 and M13 promoters.
Chapter II: Methods and Materials

Figure 2.7

Creation of a deleted cosmid.

Three types of fragments could be created during the restriction digest. Each of these religates when ligase is added, creating three forms of DNA able to transform competent bacteria. Only bacteria possessing fragments with the origin of replication and Ampicillin resistance gene are able to grow on ampicillin plates. E: EcoRI; B: BssHII
Chapter II: Methods and Materials

Subcloning using the modified pKSIIAsc vector.

Creation of a subcloned fragment by ligation of a human DNA insert and a modified pBluescript vector. In this figure a $B_{ssHII}/EcoRI$ fragment is being subcloned into a vector with complementary $Ascl/EcoRI$ ends. In this thesis a $B_{ssHII}/B_{ssHII}$ fragment was also subcloned into a vector digested with $Ascl$. 

Figure 2.8

Subcloning using the modified pKSIIAsc vector.
Figure 2.9
Detection of pKSIIAsc vector.

Autoradiograph of bacterial colonies probed with the T3 primer present in the pKSIIAsc vector. This was used as a method to isolate colonies containing a subcloned fragment. Colonies 1, 5, and 6 contained subclone 13E3 0.6; colonies 7-11 did not contain subclone 46G4 4.0; colonies 12-15 did not contain subclone 141E8 8.0; colony 16 contained subclone 176F5 1.8; colonies 23 and 28 contained subclone 175G8 3.3.
Figure 2.10

Restriction digests of subclones.

Bacterial colonies containing subcloned fragments digested with BssHII and EcoRI to ensure that the correct size fragment is present. Each lane contains a 3 kb band representing the pKSIIAsc vector and another band representing the subcloned fragment.
Chapter III: Results

3.1 Sequence Data

Initially, deleted cosmids were created as templates for sequencing. However, due to their large size (> 9 kilobases) it was difficult to obtain data and only one was successfully sequenced. The remaining sequence data presented in this thesis was obtained from subcloned fragments with a BssHII site at one or both ends. These subcloned fragments were smaller in size, alleviating the aforementioned problem.

Single-run sequence data was obtained from thirteen segments of human DNA using an automated sequencer (ABI Model 373), with a final segment being sequenced manually. During this project, sequence was acquired from the BssHII site using the M13 primer, and the EcoRI site using the T3 primer. Sequence obtained during this project is located in Appendix 1.

3.1.1 Cosmid 93C11ΔBssHII

Sequence through the BssHII site within cosmid 93C11ΔBssHII was identified as Alu-Sc repetitive DNA. Alu-Sc elements make up a subfamily of Alu repeats distributed throughout primate genomes (Labuda and Striker, 1989). BLAST analysis revealed the first 383 nucleotides of sequence shared high sequence identity with the Alu-Sc consensus sequence (Table 3.1). The BssHII site for which this cosmid was selected is located 148 base pairs from the 5’ end of an Alu-Sc element (GenBank U14571). No matches were recognized in the database for the remaining 267 base pairs of non-repetitive sequence. However, GRAIL scored the probability of an open reading frame extending from base pair 492 to 596 as excellent (Table 3.6). A search against the SwissProt database using the hypothetical protein encoded within this open reading frame identified no matches.
Chapter III: Results

Presently, no known sequence has homology with the open reading frame identified in cosmid 93C11.

3.1.2 Subclone from Cosmid 13E3

A 600 base pair *BssHII-EcoRI* fragment from cosmid 13E3 was subcloned into a pKSIIAsc vector and sequenced through the *BssHII* site (Appendix 1). BLAST inquiries revealed no significant matches with data present in GenBank’s non-redundant, EST, or STS databases. GRAIL indicated that there is a good probability of an open reading frame beginning at nucleotide 228 and ending at 333 (Table 3.6). An additional search against the SwissProt protein database did not uncover any significant homology between the hypothetical protein encoded within this open reading frame and those present in the database. At present the open reading frame identified in cosmid 13E3 does not match any known sequence.

3.1.3 Subclones from Cosmid 40G1

Two *BssHII-EcoRI* fragments from cosmid 40G1, contained within a 3.3 kilobase *EcoRI* fragment, were subcloned into a pKSIIAsc vector and sequenced. The two subcloned fragments, 1.4 and 1.6 kilobases in size, were sequenced through the *BssHII* and *EcoRI* sites (Appendix 1). Inquiries against GenBank’s non-redundant database identified one gene, the chimpanzee GOR gene (GenBank D10017), sharing high sequence identity with each sequence read (Table 3.1).

The PROSITE database of protein patterns was searched with sequence from the GOR gene in order to identify consensus protein sequences denoting possible gene function. This gene was found to contain putative phosphorylation, amidation, N-myristoylation, and N-glycosylation sites (Table 3.5). The numerous phosphorylation sites implies the encoded protein has the potential to be phosphorylated. Amidation sites are features indicative of hormone peptides. N-myristoylation sites are involved in
Chapter III: Results

associations between proteins and membranes. And N-glycosylation sites suggest that the
protein is involved in interactions with other cells or molecules (Creighton, 1993).
Although this analysis does not prove the function of the protein, it is useful information
for future research in this area.

Queries of the sequence tags site (STS) database identified an STS on chromosome
8 (GenBank L187713) with regions that matched sequence obtained from subclone 40G1
1.4 using the M13 primer (Table 3.2). This STS is reported as containing LINE1 sequence
(Gerken et al., 1993), a family of repetitive DNA that will be discussed further in Chapter
IV. Sequence obtained from the 1.4 kb BssHII-EcoRI fragment also contains LINE1
repetitive DNA, accounting for this match.

Sequence from the BssHII site of subclone 40G1 1.6 matched with four mouse
expressed sequence tags (W81932; AA034629; AA097346; AA120740) present in
GenBank’s EST database (Table 3.2). Inquiries against the EST database, using the
partially sequenced chimpanzee GOR gene, identified two additional mouse ESTs
(AA030569; AA110525) (Table 3.3). Thus, sequencing adjacent to BssHII sites in cosmid
40G1 identified a novel human gene previously identified in chimpanzees, with homology
to a number of expressed mouse sequences.

3.1.4 Subclones from Cosmid 46F4

Two BssHII-EcoRI fragments, 1.6 and 1.8 kilobases in size, were subcloned from
cosmid 46F4. Sequence obtained using the M13 primer identified a significant match with
one gene, 5-oxo-L-prolinase (GenBank U70825), previously sequenced from Rattus
norvegicus mRNA (Table 3.1). Inquiries against the EST database with the rat 5-oxo-L-
prolinase (OPLAH) gene, identified four mouse ESTs (AA27144; W29895; AA271446;
AA097778) with significant regions of homology (Table 3.4).
Chapter III: Results

GRAIL analysis recognized open reading frames in both fragments (Table 3.6). Subclone 46F4 1.6 identified two potential open reading frames, one extending from base pair 422 to 569, and the other from base pair 84 to 278. Subclone 46F4 1.8 was identified as containing one open reading frame extending from base pair 314 to 616. GRAIL failed to identify the 3' exon sequenced in the 1.8 kilobase fragment. However, analysis of this sequence, changing base 460 from C to T, predicted an exon extending from base pair 308 to 465. Thus, sequencing adjacent to BssHII sites in cosmid 46F4 successfully identified a novel human gene previously detected in rats.

3.1.5 Subclone from Cosmid 46F11

BLAST analysis of sequence from a 2.5 kilobase BssHII-EcoRI fragment revealed this subclone to contain a variable number tandem repeat (VNTR) associated with a short interspersed sequence retroposon, as well as a medium reiteration frequency sequence (MER) (Table 3.1). The first 107 nucleotides from the BssHII site, shares high sequence identity with 3' region of the short interspersed sequence retroposon (SINE-R.C2) with which it matched. The remaining 543 base pairs did not match significantly with any sequence present in the database. Sequence identity between the MER11A consensus sequence (Kaplan et al., 1991) and the sequence obtained through the EcoRI site of this subclone was quite high (Table 3.1). The GC content for the 650 base pairs from the EcoRI site was found to be 37%, typical of A+T rich MER elements. Thus, sequence obtained from this cosmid identified a number of repetitive elements, with no indications of expressed sequences being present in the subcloned fragment.
Chapter III: Results

3.1.6 Subclone from Cosmid 77C1

Sequence from subclone 77C1 2.4 (Appendix 1), obtained through the BssHII site, was used to search GenBank’s non-redundant database. Analysis revealed that 600 base pairs extending from the BssHII site of the 2.4 kilobase BssHII-EcoRI fragment is LINE1 repetitive sequence (Table 3.1). Within the L1.2 element, with which this sequence matched, a BssHII site is located 153 base pairs from the 5' end. Thus, sequence obtained from cosmid 77C1 was identified entirely as a member of the L1 repetitive element family.

3.1.7 Subclone from Cosmid 141A6

A 1.7 kilobase BssHII-EcoRI fragment from cosmid 141A6 was subcloned and sequenced through the BssHII site (Appendix 1). The first 49 base pairs of sequence is Alu-Sc repetitive DNA (Table 3.1), not detected by probing with Alu. Searches against the databases with the non-repetitive region identified no significant homologies. As well, no open reading frame was identified by GRAIL. Currently no sequence present in the GenBank database has significant homology with the 600 base pairs of non-redundant sequence from subclone 141A6 1.7.

3.1.8 Subclone from Cosmid 156G7

A 2.3 kilobase BssHII fragment was subcloned from cosmid 156G7. This region could not be sequenced using the automated sequencer and Taq polymerase. However, 135 base pairs of sequence was obtained by manual sequencing with Sequenase. Sequenase is better able to sequence difficult areas, such as GC rich regions, that have a tendency to form secondary structures. Once sequenced, subclone 156G7 2.3 was found to have a higher GC content than any of the other fragments sequenced for this project (Table 3.7).
Chapter III: Results

BLAST similarity searches with GenBank databases were unable to identify any significant regions of homology with previously sequenced DNA. GRAIL did not detect an open reading frame. It would have been unlikely for one to be detected within 135 nucleotides of sequence using this program, since GRAIL identifies coding exons 100 base pairs or greater in length. Therefore, sequence obtained from subclone 156G7 2.3 has no homology with sequence in the database and no detectable coding features.

3.1.9 Subclones from Cosmid 166H7

Two BssHII-EcoRI fragments from cosmid 166H7 were subcloned and sequenced through their BssHII sites (Appendix 1). The first fragment, 166H7 5.0, identified no significant homologies with sequence in GenBank's databases. The second fragment, 166H7 9.2, was found to match human EST H88121 (Table 3.2). Of the 221 nucleotides within the region of homology (Figure 3.1) there were 26 sequence differences (88% sequence identity). Human EST H88121 originated from a retinal cDNA library, and was sequenced at Washington University School of Medicine as part of the EST project (Hillier et al., 1995). This EST had been placed on chromosome 8 at position 731.5 cR on the WICGR radiation hybrid map (746.2 cR), between markers WI-15870 (729.9 cR) and WI-12784 (731.5 cR). Thus, sequence from cosmid 166H7 contains a human EST previously placed on chromosome 8.

3.1.10 Subclone from Cosmid 175G8

A 3.3 kilobase BssHII-EcoRI fragment from cosmid 175G8 was subcloned and sequenced through the BssHII site (Appendix 1). Inquiries against GenBank databases identified no matches with sequence present there. As well, no open reading frames were recognized by GRAIL. Therefore, at this time sequence obtained from this cosmid has no significant homologies with sequence present in the GenBank database.
Chapter III: Results

3.1.11 Subclone from Cosmid 176F5

Sequence through the BssHII site of a 1.8 kilobase BssHII-EcoRI fragment from cosmid 176F5 (Appendix 1) was compared with data at GenBank. Inquiries against the non-redundant database identified the first 109 nucleotides as Alu-J sequence (Table 3.1), a subfamily of Alu repetitive elements (Jurka and Milosavljevic, 1991).

A BLAST similarity search against GenBank’s EST database, with the non-repetitive region, identified human EST W90101 (Table 3.2). The terminal 382 base pairs of sequence obtained from this subclone shares 92% sequence identity with the 5' region of this EST (Figure 3.2). Human EST W90101 was sequenced from a fetal liver spleen cDNA library at Washington University School of Medicine as part of the EST project (Hillier et al., 1995). Thus, cosmid 176F5 contains human EST W90101 enabling this EST to be localized to chromosome 8.

3.2 Localization by STS mapping

3.2.1 Localization of the Human GOR gene

STS mapping was used as a method of positioning the human GOR gene, identified in cosmid 40G1, by designing primers to amplify a 286 bp product at the BssHII end of the 1.6 kb fragment. The forward primer, 5'-AGGTTGCCCCAAGTCCAAGC-3', is shown underlined at the top of Figure 3.4 while the reverse primer, 5'-GCTGTCTGACCTTCCACATC-3', complements the 20 nucleotides shown underlined at the bottom of Figure 3.4 that are adjacent to the GCGCGC BssHII site. These primers were used against a cell hybrid panel for chromosome 8 (Wagner et al., 1991; Wood et al., 1986). Only XVII-23Ha produced an amplified product of 286 base pairs, localizing the human GOR gene on chromosome 8q13-q22.2 (Figure 3.5). This STS was also typed in the G3 radiation hybrid panel (Stewart et al., 1997) and the results submitted to the
Chapter III: Results

Stanford RH server. The highest two point LOD score of 3.87 was reported at a distance of 52.9 cR\textsubscript{10000} from SHGC-37027 which lies within the chromosome 8 bin 67. The Généthon marker D8S1757, also located within bin 67, maps to 8q13-q22.1 on the cytogenetic map (Leach \textit{et al.}, 1996) confirming the previous localization.

3.2.2 Localization of the Human EST W90101

STS mapping was the method used to position human EST W90101, identified in cosmid 176F5, by designing primers to amplify a 300 bp product near the \textit{BssHII} site of the 1.8 kb fragment. The forward primer, 5'-TTCTCCTCTCCGCTGGGTG-3', is shown underlined at the top of Figure 3.2 while the reverse primer, 5'-GAGGGACAAGTATCCAGTCC-3', complements the 20 nucleotides shown underlined at the bottom of Figure 3.2 that are adjacent to the GCGCGC \textit{BssHII} site. These primers were used against a cell hybrid panel for chromosome 8 (Wagner \textit{et al.}, 1991; Wood \textit{et al.}, 1986). Both MGV 270 and MGV 271 produced amplified products of 300 base pairs, localizing EST W90101 on chromosome 8q24.1-qter (Figure 3.3).
Chapter III: Results

Figure 3.1

Sequence comparison between genomic sequence from cosmid 166H7 and sequence from cDNA clone 220630.

The genomic human sequence is numbered from the BssHII site in the 9.2 kb BssHII-EcoRl fragment. The cDNA sequence is numbered as reported by Hillier et al., (1995).
Figure 3.2

Sequence comparison between genomic sequence from cosmid 176F5 and sequence from cDNA clone 418116.

The genomic human sequence is numbered from the BssHII site in the 1.8 kb BssHII-EcoRI fragment. The cDNA sequence is numbered as reported by Hillier et al., (1995).
Chapter III: Results

Figure 3.3

Chromosomal localization of the human EST W90101 by PCR against somatic cell hybrids.

Lane 1 contains a marker digest, φX174 phage digested with HaeIII. The cell hybrid panel consists of: XVIII-23Ha, 8pter-8q22, lane 2; VTGHL 19, 8pter-8q13, lane 3; 1SHL 3, 8pter-q11, lane 4; 20xPO435-2, 8p23-q11, lane 5; MGV 270, 8q24.1-qter, lane 6; MGV 271, 8q22.1-qter, lane 7. Human DNA, lane 8, is the positive control, and lane 9 is a negative control.

55
Chapter III: Results

---

**Human**

- **R** L P Q V Q A I L L S F
- **human**
  - AGG TGG CCC CAA GTC CAA GCC ATC CTG CTG AGC TTT 255
  - TTA AGC GCC CAA ACC ATC NTC ATC GGG CAC AGC GTG 219
  - GAG AGC GAC CTG ATG GCC CTG AAG CTC ATC CAC AGC 183
  - GCG GCA GAC TAC CTG GGA GAC ATC ATC CAG GAC AGC 75
  - CAG GAC GGC CAC AAC TCC AGC GAG GAC GCA AAC GCC 39
  - TGC CTG CAG ATG GTG ATG TGG AAG GTC AGA CAG CGC GC 1

**Chimp**

- **Q** V Q A I L L S F
- **chimp**
  - ACC TGG CCC CAA GTG CAA GCC ATC CTG CTG AGC TTT 942
  - TTC AGC GCC CAA ACC ATC CTC ATC GGG CAC AGC CTG 978
  - GAG AGC GAT CTG CTG GCC CTG AAG CTC ATC CAC AGC 1014
  - GCG GCC GAC TAC CTG GGA GAC ATC ATC CAG GAC AGC 1122
  - CAG GAC GGC CAC AAC TCC AGC GAG GAC GCA AAC GCC 1158

**Human**

- **L** S A Q T I X I G H S V
- **human**
  - GTG CAA GCC ATC GTG CAA GCC ATC GTG CAA GCC ATC 147
  - TV V D T A V L F P Q Y
  - M G F P Y K R S L R N L
  - A A D Y L A Q I I Q D S
  - Q D G H N S S E D A N A

**Chimp**

- **chimp**
  - GTG CAA GCC ATC GTG CAA GCC ATC GTG CAA GCC ATC 1050
  - TV L D T A V L F P H Y
  - LG F P Y K R S L R N L
  - A A D Y L G Q I I Q D S
  - Q D G H N S S E D A N A
Figure 3.4

GOR sequence comparison between human and chimpanzee for the region amplified by PCR primers.

The human sequence is numbered from the \textit{BssHII} site in the 1.6 kb \textit{BssHII-EcoRI} genomic fragment. The chimpanzee sequence is numbered from the \textit{EcoRI} site of the partial cDNA as reported by Mishiro \textit{et al.}, (1990).
Figure 3.5

Chromosomal localization of the human GOR gene by PCR against somatic cell hybrids.

Lane 1 contains a marker digest, φX174 phage digested with *Hae*III. The cell hybrid panel consists of: XVIII-23Ha, 8pter-8q22, lane 2; VTGHL 19, 8pter-8q13, lane 3; ISHL 3, 8pter-q11, lane 4; 20xPO435-2, 8p23-q11, lane 5; MGV 270, 8q24.1-qter, lane 6; MGV 271, 8q22.1-qter, lane 7. Human DNA, lane 8, is the positive control, and lane 9 is a negative control.
### Table 3.1: GenBank Non-Redundant Database Search

<table>
<thead>
<tr>
<th>Subclone Name</th>
<th>Description</th>
<th>Subclone Region of Match</th>
<th>Length (bp)</th>
<th>% Identity Match</th>
<th>GenBank accession number</th>
</tr>
</thead>
<tbody>
<tr>
<td>40G1 1.4</td>
<td>Chimpanzee GOR gene, many matches with L1 repetitive element</td>
<td>14-87</td>
<td>72</td>
<td>98</td>
<td>Many including</td>
</tr>
<tr>
<td></td>
<td></td>
<td>86-332</td>
<td>247</td>
<td>92</td>
<td>D10017</td>
</tr>
<tr>
<td></td>
<td></td>
<td>456-586</td>
<td>131</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>40G1 1.4 *</td>
<td>Chimpanzee GOR gene, many matches with L1 repetitive element</td>
<td>1-48</td>
<td>48</td>
<td>100</td>
<td>Many including</td>
</tr>
<tr>
<td></td>
<td></td>
<td>48-411</td>
<td>364</td>
<td>89</td>
<td>D10017</td>
</tr>
<tr>
<td></td>
<td></td>
<td>411-433</td>
<td>23</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>424-524</td>
<td>101</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>555-578</td>
<td>24</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>40G1 1.6</td>
<td>Chimpanzee GOR gene</td>
<td>1-334</td>
<td>134</td>
<td>93</td>
<td>D10017</td>
</tr>
<tr>
<td></td>
<td></td>
<td>328-394</td>
<td>67</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>395-431</td>
<td>37</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>40G1 1.6 *</td>
<td>Chimpanzee GOR gene</td>
<td>421-465</td>
<td>45</td>
<td>84</td>
<td>D10017</td>
</tr>
<tr>
<td></td>
<td></td>
<td>487-525</td>
<td>39</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>517-633</td>
<td>117</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>46F4 1.6</td>
<td>Rattus norvegicus 5'-oxo-L-prolinase mRNA</td>
<td>81-112</td>
<td>32</td>
<td>84</td>
<td>U70825</td>
</tr>
<tr>
<td></td>
<td></td>
<td>113-232</td>
<td>120</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>306-361</td>
<td>56</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>362-429</td>
<td>68</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td>46F4 1.8</td>
<td>Rattus norvegicus 5'-oxo-L-prolinase mRNA</td>
<td>16-165</td>
<td>150</td>
<td>78</td>
<td>U70825</td>
</tr>
<tr>
<td></td>
<td></td>
<td>320-459</td>
<td>140</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>46F11 2.5</td>
<td>VNTR</td>
<td>1-54</td>
<td>54</td>
<td>98</td>
<td>Z11740</td>
</tr>
<tr>
<td></td>
<td></td>
<td>53-107</td>
<td>55</td>
<td>92</td>
<td>S36048</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>S36051</td>
</tr>
<tr>
<td>46F11 2.5 *</td>
<td>MER11A repetitive element</td>
<td>1-27</td>
<td>27</td>
<td>92</td>
<td>Z72522</td>
</tr>
<tr>
<td></td>
<td></td>
<td>42-198</td>
<td>157</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>234-348</td>
<td>115</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>342-368</td>
<td>27</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>367-386</td>
<td>20</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>386-445</td>
<td>60</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>432-506</td>
<td>75</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>77C1 2.4</td>
<td>L1.2 repetitive element</td>
<td>1-78</td>
<td>78</td>
<td>87</td>
<td>M80343</td>
</tr>
<tr>
<td></td>
<td></td>
<td>67-103</td>
<td>37</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>103-138</td>
<td>36</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>130-328</td>
<td>199</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>400-499</td>
<td>100</td>
<td>91</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>497-542</td>
<td>46</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>542-601</td>
<td>60</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>93C11</td>
<td>Alu-Sc repetitive element</td>
<td>1-82</td>
<td>82</td>
<td>91</td>
<td>U14571</td>
</tr>
<tr>
<td>ΔBssHII</td>
<td></td>
<td>82-152</td>
<td>71</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>293-393</td>
<td>101</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>141A6 1.7</td>
<td>Alu-Sc element</td>
<td>1-49</td>
<td>49</td>
<td></td>
<td>U14571</td>
</tr>
<tr>
<td>176F5 1.8</td>
<td>Alu-J element</td>
<td>1-17</td>
<td>17</td>
<td></td>
<td>U14567</td>
</tr>
<tr>
<td></td>
<td></td>
<td>46-109</td>
<td>64</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Only significant similarities with GenBank database are listed
* sequenced from the EcoRI site
## Chapter III: Results

### Table 3.2

**GenBank Sequence Tag Sites (STS) and Expressed Sequence Tag (EST) Database Searches**

<table>
<thead>
<tr>
<th>Subclone Name</th>
<th>Description</th>
<th>Region of Match (subclone sequence)</th>
<th>Length (bp)</th>
<th>Identity Match (%)</th>
<th>GenBank accession number</th>
</tr>
</thead>
<tbody>
<tr>
<td>40G1 1.4 †</td>
<td>Human chromosome 8 STS UT1959</td>
<td>427-451, 456-535</td>
<td>25, 80</td>
<td>88, 83</td>
<td>L18713</td>
</tr>
<tr>
<td>40G1 1.6 ‡</td>
<td>me93c12.r1 Soares mouse embryo NbME13.5 14.5 Mus musculus cDNA clone</td>
<td>7-334, 330-390</td>
<td>328, 61</td>
<td>74, 83</td>
<td>W81932</td>
</tr>
<tr>
<td>40G1 1.6 ‡</td>
<td>mi53h07.r1 Soares mouse embryo NbME13.5 14.5 Mus musculus cDNA clone</td>
<td>62-334, 330-390</td>
<td>273, 69</td>
<td>74, 83</td>
<td>AA034629</td>
</tr>
<tr>
<td>40G1 1.6 ‡</td>
<td>mm36e01.r1 Stratagene mouse skin (#937313) Mus musculus cDNA clone</td>
<td>314-390</td>
<td>77</td>
<td>76</td>
<td>AA097346</td>
</tr>
<tr>
<td>40G1 1.6 ‡</td>
<td>ml56g02.r1 Stratagene mouse testis (#937308) Mus musculus cDNA clone</td>
<td>330-390</td>
<td>61</td>
<td>83</td>
<td>AA120740</td>
</tr>
<tr>
<td>166H7 9.2 ‡</td>
<td>ys75c12.s1 Homo sapiens cDNA clone 220630</td>
<td>1-110, 103-168, 189-221</td>
<td>110, 66, 33</td>
<td>70, 84, 75</td>
<td>H88121</td>
</tr>
<tr>
<td>176F5 1.8 ‡</td>
<td>zh77e07.s1 Soares fetal liver spleen INFLS S1 Homo sapiens cDNA clone 418116</td>
<td>269-507, 508-649</td>
<td>239, 142</td>
<td>96, 78</td>
<td>W90101</td>
</tr>
</tbody>
</table>

Only significant similarities with GenBank database are listed
† Matched with STS
‡ Matched with EST
### Table 3.3

GenBank Expressed Sequence Tag (EST) Database Search with the GOR gene

<table>
<thead>
<tr>
<th>Gene Searched</th>
<th>Description</th>
<th>Region of Match</th>
<th>Length (bp)</th>
<th>Identity Match</th>
<th>GenBank accession number</th>
</tr>
</thead>
<tbody>
<tr>
<td>chimpanzee</td>
<td>ml56g02.r1 Stratagene mouse testis (#937308) Mus musculus cDNA clone 516050</td>
<td>352-605</td>
<td>254</td>
<td>77</td>
<td>AA120740</td>
</tr>
<tr>
<td>chimpanzee</td>
<td>me93c12.r1 Soares mouse embryo NbME13.5 14.5 Mus musculus cDNA clone 403126</td>
<td>734-1190</td>
<td>457</td>
<td>79</td>
<td>W81932</td>
</tr>
<tr>
<td>chimpanzee</td>
<td>mi53h07.r1 Soares mouse embryo NbME13.5 14.5 Mus musculus cDNA clone 467293</td>
<td>734-1135</td>
<td>402</td>
<td>79</td>
<td>AA034629</td>
</tr>
<tr>
<td>chimpanzee</td>
<td>mm36e01.r1 Stratagene mouse skin (#937313) Mus musculus cDNA clone 523608</td>
<td>744-867</td>
<td>121</td>
<td>81</td>
<td>AA097346</td>
</tr>
<tr>
<td>chimpanzee</td>
<td>mi26gl0.r1 Soares mouse embryo NbME13.5 14.5 Mus musculus cDNA clone 464706</td>
<td>8-158</td>
<td>151</td>
<td>72</td>
<td>AA030569</td>
</tr>
<tr>
<td>chimpanzee</td>
<td>mo54f10.r1 Life Tech mouse embryo 10 5dpc 10665016 Mus musculus cDNA clone 557419</td>
<td>450-551</td>
<td>102</td>
<td>82</td>
<td>AA110524</td>
</tr>
</tbody>
</table>

Only significant similarities with GenBank database are listed.
### Table 3.4

**GenBank Expressed Sequence Tag (EST)**

**Database Search with the OPLAH gene**

<table>
<thead>
<tr>
<th>Gene Searched</th>
<th>Description</th>
<th>Region of Match</th>
<th>Length (bp)</th>
<th>Identity Match (%)</th>
<th>GenBank accession number</th>
</tr>
</thead>
<tbody>
<tr>
<td>rat OPLAH</td>
<td>va86h07.r1 Soares mouse NML Mus musculus cDNA clone 738301</td>
<td>3406-3441</td>
<td>36</td>
<td>100</td>
<td>AA271445</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3454-3708</td>
<td>255</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3700-3876</td>
<td>177</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3851-3896</td>
<td>46</td>
<td>65</td>
<td></td>
</tr>
<tr>
<td>rat OPLAH</td>
<td>mc06d08.r1 Soares mouse p3NMF19.5 Mus musculus cDNA clone 338127</td>
<td>3340-3449</td>
<td>110</td>
<td>96</td>
<td>W29895</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3451-3730</td>
<td>280</td>
<td>92</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3724-3754</td>
<td>31</td>
<td>93</td>
<td></td>
</tr>
<tr>
<td>rat OPLAH</td>
<td>va86h08.r1 Soares mouse NML Mus musculus cDNA clone 738303</td>
<td>3409-3440</td>
<td>32</td>
<td>96</td>
<td>AA271446</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3449-3825</td>
<td>377</td>
<td>88</td>
<td></td>
</tr>
<tr>
<td>rat OPLAH</td>
<td>mk17c05.r1 Soares mouse p3NMF19.5 Mus musculus cDNA clone 493160</td>
<td>3340-3441</td>
<td>102</td>
<td>97</td>
<td>AA097778</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3441-3489</td>
<td>49</td>
<td>85</td>
<td></td>
</tr>
</tbody>
</table>

Only significant similarities with GenBank database are listed.
Table 3.5

PROSITE database Searches with the GOR gene

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Annotated Domains</th>
<th>Location and Sequence</th>
<th>Total matches</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOR</td>
<td>N-glycosylation site</td>
<td>379: NSSE</td>
<td>1</td>
</tr>
<tr>
<td>GOR</td>
<td>cAMP-dependent protein kinase phosphorylation site</td>
<td>13: RRPS; 219: RKDS</td>
<td>2</td>
</tr>
<tr>
<td>GOR</td>
<td>Protein kinase C phosphorylation site</td>
<td>7: SSK; 8: SKR; 19: SLK; 87: TLK; 151: SCR; 167: SGR; 229: TFK; 358: SLR</td>
<td>8</td>
</tr>
<tr>
<td>GOR</td>
<td>Casein kinase II phosphorylation site</td>
<td>55: SKQE; 262: TVVD; 293: TEAD; 339: TVLD; 380: SSED</td>
<td>5</td>
</tr>
<tr>
<td>GOR</td>
<td>Tyrosine kinase phosphorylation site</td>
<td>169: RCVRDQLCY; 277: KPDNEIVDY</td>
<td>2</td>
</tr>
<tr>
<td>GOR</td>
<td>N-myristoylation site</td>
<td>93: GLTPSS; 101: GLSRAA; 190: GGRVSQ; 204: GSVGCQ; 207: GCQVAK 255: GELTR; 291: GVTEAD; 443: GAVDGR 473: GLSPSL</td>
<td>9</td>
</tr>
<tr>
<td>GOR</td>
<td>Amidation site</td>
<td>217: DGRK; 446: DGRR</td>
<td>2</td>
</tr>
</tbody>
</table>
Chapter III: Results

Table 3.6
Protein Coding Potential of BssHII-EcoRI subclones as Analyzed by GRAIL

<table>
<thead>
<tr>
<th>Subclone Name</th>
<th>frame</th>
<th>ORF start</th>
<th>ORF end</th>
<th>Score</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>13E3 0.6</td>
<td>2</td>
<td>228</td>
<td>333</td>
<td>55.00</td>
<td>good</td>
</tr>
<tr>
<td>46F4 1.6</td>
<td>-1</td>
<td>422</td>
<td>569</td>
<td>90.00</td>
<td>excellent</td>
</tr>
<tr>
<td>46F4 1.6</td>
<td>-2</td>
<td>84</td>
<td>278</td>
<td>80.00</td>
<td>excellent</td>
</tr>
<tr>
<td>46F4 1.8</td>
<td>1</td>
<td>314</td>
<td>616</td>
<td>65.00</td>
<td>good</td>
</tr>
<tr>
<td>93C11ΔBssHII</td>
<td>-2</td>
<td>492</td>
<td>596</td>
<td>85.00</td>
<td>excellent</td>
</tr>
</tbody>
</table>

Table 3.7
GC content of BssHII-EcoRI subclones

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>13E3 0.6</td>
<td>650</td>
<td>56</td>
<td>49</td>
<td>28</td>
<td>51</td>
<td>46</td>
<td>0.57</td>
</tr>
<tr>
<td>40G1 1.4</td>
<td>650</td>
<td>47</td>
<td>33</td>
<td>18</td>
<td>30</td>
<td>27</td>
<td>0.54</td>
</tr>
<tr>
<td>40G1 1.6</td>
<td>650</td>
<td>57</td>
<td>48</td>
<td>31</td>
<td>47</td>
<td>55</td>
<td>0.64</td>
</tr>
<tr>
<td>40G1 1.6*</td>
<td>650</td>
<td>58</td>
<td>46</td>
<td>29</td>
<td>44</td>
<td>31</td>
<td>0.63</td>
</tr>
<tr>
<td>46F4 1.6</td>
<td>650</td>
<td>64</td>
<td>64</td>
<td>30</td>
<td>60</td>
<td>49</td>
<td>0.47</td>
</tr>
<tr>
<td>46F4 1.8</td>
<td>650</td>
<td>66</td>
<td>68</td>
<td>43</td>
<td>63</td>
<td>46</td>
<td>0.63</td>
</tr>
<tr>
<td>46F11 2.5*</td>
<td>649</td>
<td>37</td>
<td>32</td>
<td>5</td>
<td>21</td>
<td>45</td>
<td>0.15</td>
</tr>
<tr>
<td>176F5 1.8</td>
<td>650</td>
<td>51</td>
<td>42</td>
<td>16</td>
<td>37</td>
<td>52</td>
<td>0.38</td>
</tr>
</tbody>
</table>

* Sequenced from the EcoRI site
§ Values over 0.60 are considered indicators of CpG islands. Bulk genomic DNA has frequencies of 0.20.
Chapter IV: Discussion

4.1 Gene Identification

An important focus of research in human genetics is the identification of genes. Eventually this will be accomplished by sequencing the entire genome, predicted to be completed in 2005, followed by identification and localization of genes on genetic maps. For a number of organisms this process is well underway. *Saccharomyces cerevisae* is the first eukaryote to have its entire genome sequenced and all of the genes identified (Goffeau et al., 1996). Current research on genome organization and gene function is pointing the way for future work to be done in the human genome. Sequencing of *Caenorhabditis elegans*’ genome is nearing completion, with roughly 30% remaining. Similarity searches against the database as well as programs that identify coding features, are being used to uncover the genes within this sequence. Functional analysis of genes identified in these genomes will assist in analysis of homologous human genes. Even now over half of cloned genes known to cause human heritable diseases have similarities to a gene in either the yeast or *C. elegans* genome (Goffeau et al., 1996; Riddle et al., 1997). Sequence comparison between organisms enables important features such as genes to be detected, providing an indispensable tool to assist in analysis of the human genome.

As large scale sequencing of the human genome begins researchers are looking ahead to the next steps in the process of characterizing the genome (Lander, 1996). Re-sequencing of selected regions has been suggested to identify allelic variations in populations, as well as variations in coding regions underlying predisposition and susceptibility to disease. As genes are identified and placed on genetic maps, more information about genome organization will become available providing new insights into the human genome.
Chapter IV: Discussion

While sequencing of the human genome awaits completion, efforts are being taken to sequence the estimated 50,000 to 100,000 genes encoded within it. Numerous techniques, such as cDNA selection and exon trapping, are currently being used to isolate genes within a small fragment (1 Mb) of DNA (review Monaco 1994). Large scale sequencing of expressed sequence tags (ESTs), as part of the human genome project (Adams et al., 1991; Hillier et al., 1997), has been responsible for identifying 709,530 ESTs. However, due to the difficulty in establishing cDNA libraries with genes transcribed at low frequencies, particular stages of growth, or only in certain tissues, genes of this type are likely to be missed. As well, other genes have been found to have a number of ESTs representing them in the database. To reduce this redundancy comparisons were made among ESTs and 3' ends of known genes (Schuler et al., 1996).

From this approximately 50,000 unique sequences were identified (www.ncbi.nlm.nih.gov/UniGene/) more than 16,000 of which have been positioned on genetic maps spanning the human genome. As more genes are identified these will also be placed on genetic maps providing a framework onto which more information can be added.

Searches for genes in the human genome revealed that regions which posses a high GC content also have a clustering of coding sequences (Fields et al., 1994). Larsen et al., (1992) found all ubiquitously expressed genes and 40% of tissue specific genes, present in the EMBL DataBank (release 28), were associated with regions known as CpG islands.

CpG islands are easily identifiable through the use of a rare cutting enzyme, due to the high frequency of CpG dinucleotides within island DNA compared to genomic DNA (Cross and Bird, 1995). BssHII is a rare cutting enzyme with 80% of its sites in CpG islands (Larsen et al., 1992), making it particularly useful for locating these sequences. As CpG islands are able to be discerned within cloned genomic DNA (Lindsay and Bird, 1987), their identification offers an alternative method of gene detection that could be used to identify genes. The intent of this project was to identify novel human genes by sequencing adjacent to BssHII sites in chromosome 8. Through this method two ESTs, two open reading
Chapter IV: Discussion

frames, two CpG islands and two novel human genes were identified, demonstrating the effectiveness of this approach.

Although sequencing of expressed sequences has identified a large number of genes, approximately half of those predicted to be in the human genome remain to be found. Sequencing of the entire genome will provide us with the sequence for all of the genes, but the task of recognizing them will still remain. Programs, such as GRAIL, assist in the detection of genes by searching for features of coding DNA in sequence. Identification of CpG islands as a method of gene identification offers the advantage of being a useful tool before and after sequencing of the human genome is complete. Sequencing DNA adjacent to a rare cutting enzyme site has proven a successful method of gene detection in this project as well as others (Zhu et al., 1993; Lee et al., 1994; John et al., 1994). Moreover, after sequencing of the genome is complete, identification of CpG islands can be accomplished by searching for rare cutting enzyme recognition sequences. Thus, identification of genes through the use of a rare cutting enzyme is a method which can be used throughout the sequencing of the human genome.

4.2 DNA Adjacent to BssHIII sites

Sequence obtained from regions adjacent to BssHIII sites, present in cosmids selected from the LA08NC01 chromosome 8 library, was examined and coding DNA features described. By comparing sequence obtained to those present in GenBank's databases along with identification of open reading frames, two EST, two open reading frames, and two novel human genes were identified. As well, two CpG islands were detected, one of which is believed to be the 5' island for human EST H8121 partially sequenced during this project. Comparisons were made among protein data present in the SwissProt database and the hypothetical proteins coded for by the open reading frames. This was done to identify similarities with better understood genes in other organisms, providing insight into possible functions for these genes. Unfortunately, no significant
homologies were found. Using the PROSITE database of protein motifs, a number of protein consensus sites were identified within the amino acid sequence of the GOR gene suggesting possible post-translational modifications in the protein. Thus, sequence obtained during this project identified a number of genes and gene associated regions endorsing the use of the rare cutting enzyme BssHII as a method of gene detection.

4.2.1 Repetitive DNA

4.2.1.1 Alu elements

Some of the cosmids selected to be used in this project were found to contain repetitive DNA. The Alu family of repetitive DNA, found in excess of 500,000 times in the human genome (Deininger et al., 1981) is known to contain a BssHII site. Cosmid 93C11ΔBssHII was found to include Alu sequence, after which Alu repetitive DNA was used as a probe against cosmid DNA. This was done to eliminate DNA containing an Alu element from the set of fragments to be sequenced. Two subcloned cosmid fragments, 141A6 1.7 and 176F5 1.8, were identified as containing a small region of Alu sequence. It is probable that due to the modest amount of Alu sequence, 49 and 109 base pairs respectively, initial screening did not produce an identifiable signal.

An individual Alu element is approximately 300 base pairs in length, composed of two tandemly arranged halves separated by an A-rich region (Batzer et al., 1990). Each element possesses a 3' oligo-dA tail, the length of which varies, and is flanked by short direct repeats. It is believed that formation and evolution of Alu elements has involved reverse transcription of RNA, which is then reintegrated into a new site in the genome (Deininger and Daniels, 1986). Ancestrally, Alu sequences are believed to have originated from the 7SL RNA gene, due to sequence and secondary structure domains present in both. Divergence from the Alu progenitor gene has resulted in two major subfamilies Alu-J and Alu-S (Jurka and Milosavljevic, 1991), with Alu-S being further subdivided into at least
five distinct subfamilies. Folding patterns predicted for RNA derived from each Alu subfamily are the same (Labuda and Striker, 1989) despite deviations in sequence, suggesting an important function for these elements not yet discovered.

4.2.1.2 LINE1 elements

Another type of repetitive DNA, LINE1, was detected throughout subclone 77C12.4. Long interspersed repetitive sequences (LINE1) comprise a family of redundant DNA accounting for approximately 5% of the human genome (Dombroski et al., 1991). As was proposed for Alu elements, LINE1s appear to be formed through the process of retroposition (Deininger and Daniels, 1986). A complete and transpositionally active L1 element is 6 to 7 kilobases long, has a 5' untranslated region, two open reading frames, one of which contains a region similar to reverse transcriptase, and a 3' untranslated region followed by an A-rich tail (Smit et al., 1995; Jurka, 1989). In the human genome these elements are present approximately 100,000 times, however only about 3500 are complete (Dombroski et al., 1991). Most members of the L1 family are truncated in the 5' region, such that only part of the 3' untranslated region remains. Comparisons between L1 sequences allow subfamilies to be identified based on variation in nucleotides, as is done with Alu elements. L1.2, determined to match sequence from subclone 77C11 2.4, was identified as a full-length L1 element (Dombroski et al., 1991), containing both of the open reading frames.

4.2.1.3 VNTR and MER elements

Representatives from two families of repetitive elements were found within subclone 46F11 2.5. The first of these, variable number tandem repeats, are segments of sequence repeated in tandem a differing number of times resulting in various sized alleles
Chapter IV: Discussion

for these markers. The VNTR present in this BssHII-EcoRI fragment is associated with a short interspersed sequence retroposon (SINE-R.C2), predicted to be present 4,000 to 5,000 times in the human genome (Zhu et al., 1992).

Sequence obtained from the EcoRI site was identified as a MER11A element. MER repeats are found in the human genome hundreds to thousands of times, representing one of the more frequent type of repeat in the genome (Jurka et al., 1993). The origin of these sequences is unknown, however it is presumed that they represent pseudogenes which arose from an active gene by a process that has not yet been determined. Sequences representative of this family of repeats vary in length from 150 to 650 base pairs (Jurka, 1989) and are predominantly A+T rich.

Within the set of cosmids sequenced, two were found to include a significant proportion of repetitive DNA. The remaining nine cosmids provided non-repetitive sequence for analysis and were used towards the goal of this project.

4.2.2 Identification of ESTs

Examination of the expressed sequence tag (EST) database identified two human ESTs with significant homology to sequence obtained during this project. The first of these, H88121, was identified in a 9.2 kilobase BssHII-EcoRI fragment from cosmid 166H7. This EST (Hillier et al., 1995) had previously been placed near the telomere on the long arm of chromosome 8 at position 731.5 cR on the WICGR radiation hybrid map. A second human EST, W90101, was identified in a 1.8 kilobase BssHII-EcoRI fragment from cosmid 176F5. Localization of this EST was accomplished by STS mapping using primers designed from DNA adjacent to the BssHII site (Figure 3.2). Both MGV 270 and MGV 271, from a cell hybrid panel spanning chromosome 8 (Wager et al., 1991; Wood et al., 1986), produced a product localizing the EST to 8q24.1-qter (Figure 3.3). Although the goal of this project was to identify novel genes, by sequencing DNA adjacent to BssHII sites two formerly identified human ESTs were also recognized.
Chapter IV: Discussion

4.2.3 Identification of Open Reading Frames

DNA adjacent to BssHII sites within subclone 13E3 0.6 and the non-repetitive region of 93C11ΔBssHII, are predicted by GRAIL to contain open reading frames. 3' (TGGTGGTGTGGAG/G) and 5' (AG/GTGCTCCCTGCT) consensus splice sequences (Shapiro and Senapathy, 1987) were found to flank the predicted open reading frame within 13E3 0.6 (Table 4.1). The open reading frame in 93C11ΔBssHII is also flanked by consensus splice sequences, AANANANTGCNCAG/G at the 3' junction and AG/GAGATNGAGAC at the 5' location (Table 4.1). Translations resulted in putative proteins, encoded by 13E3 0.6 and 93C11ΔBssHII, 32 and 34 amino acids respectively. Examination of the SwissProt database uncovered no significant matches, suggesting homologous proteins have not yet been identified. Thus, sequencing uncovered open reading frames within two coding regions with no homologous sequence identifiable in GenBank’s databases.

4.2.4 Identification of CpG islands

Previous work with CpG islands (Aissani and Bernardi, 1991; Gardiner-Garden and Frommer, 1987; Larsen et al., 1992) concluded that 80% were located at the 5' end of a gene. Most of these islands began upstream of the associated gene's transcription start site, with distances varying from under 100 base pairs to over 2 kilobases. CpG islands are known to range in size from 500 to 2,000 base pairs (Antequera and Bird, 1993), with the BssHII site located anywhere within this sequence. Thus, it is not surprising that sequence obtained from four subcloned fragments (141A6 1.7, 156G7 2.3, 166H7 5.0, 175G8 3.3) did not match with genes in the databases or contain an open reading frame.
Chapter IV: Discussion

Presumably, these represent CpG islands located upstream of the first exon, such that less than one hundred base pairs of the coding DNA is within 650 base pairs of the BssHII site.

Assuming that a CpG island was identified through these BssHII sites, the sequence should posses certain features. CpG islands have a higher GC content than genomic DNA (Bird, 1986). This was true for three of the four subclones (156G7 2.3, 166H7 5.0, and 175G8 3.3) shown in Table 3.7. As well, the dinucleotide CpG is found at a higher frequency in island DNA than in the rest of the genome (Cross and Bird, 1995). Sequence from two of the subclones, 156G7 2.3 and 166H7 5.0, had a ratio of observed CpG dinucleotides over expected greater than 0.60. This value has been used as the lower limit indicative of a CpG island (Gardiner-Garden and Frommer, 1987; Larsen et al., 1992). Genomic DNA has a value of approximately 0.20. Thus, sequence analysis supports the assertion that sequencing adjacent to BssHII sites identifies CpG islands, each of which is likely to be associated with a gene.

As was mentioned previously, most CpG island are located upstream of transcription initiation. Preceding this start site in genes transcribed by RNA polymerase II are two consensus sequences known as TATA and CCAAT elements. Sp1 transcription factor binding sites (GGGCGG) may also be located a short distance upstream of the transcription start site. These consensus sequences were searched for within sequence from the four subclones. TATA consensus sequences are present in three (Table 4.2), however no CCAAT elements or Sp1 binding site consensus sequences were found. Translation initiation typically begins a short distance downstream of a TATA element. Translation begins at an ATG start codon that is flanked by sequence favorable to initiation (Kozak, 1996). The optimal context for initiation in vertebrates is GCCACCCatgG, however variations such as YNNatgG are also found (Y represents pyrimidine). A reasonable initiation sequence (TGAatgG) was identified 8 base pairs downstream of the TATA element in subclone 166H7 5.0 (Figure 4.2). Thus, it appears that a gene and its
Chapter IV: Discussion

associated CpG island were identified using this method. Primers designed during this project are currently being used to localized this CpG island on chromosome 8.

4.2.5 Identification of Novel Human Genes

4.2.5.1 GOR Gene

Investigation of the non-redundant database identified a novel human gene within cosmid 40G1. 2859 base pairs of this cDNA had previously been sequenced in chimpanzees as an EcoRI fragment (Mishiro et al., 1990). The corresponding region within the human homologue, identified in this thesis, is located within a 3.3 kilobase EcoRI fragment. Contained within this fragment are two BssHII sites resulting in two BssHIII-EcoRI fragments, of 1.4 and 1.6 kilobases, and a 300 base pair BssHIII fragment (Figure 4.3).

Few variations in sequence were found between the human and chimpanzee genes (Figures 4.4 a-c). Sequence obtained from the EcoRI site of the 1.6 kilobase fragment (Figure 4.4 a) shows more variation in sequence between human and chimpanzee (61% identity) due in part to lower quality sequence obtained for the last 200 base pairs. Among the 76 amino acids compared there are 9 amino acid differences and 19 triplets not able to be translated. Variations in amino acids were due to substitutions at codon position 1 in thirteen cases, position 2 in seven cases, and position 3 in eight cases. Three additional silent substitutions were found at position 3. Among the 168 amino acids compared in Figure 4.4b there are 30 differences (82% identity). Substitutions at codon position 1 accounts for sixteen of the differences, position 2 for six, and position 3 for eight. An additional ten silent substitutions were observed, nine at position 3 and one at position 1. Among the 20 amino acids compared in Figure 4.4c 2 amino acid differences were identified (90% identity). These differences were due to a substitution at position 1 and position 2.
Localization of this novel human gene was accomplished by STS mapping using primers, shown underlined in Figure 4.4b, designed from DNA adjacent to the *BssHII* site in the 1.6 kilobase fragment. Only XVIII-23Ha, from a cell hybrid panel spanning chromosome 8 (Wager *et al.*, 1991; Wood *et al.*, 1986), produced a product localizing GOR to 8q13-q22.2 (Figure 3.5). STS typing in the G3 radiation hybrid panel confirmed the location of this gene by linking it to Généthon marker D8S1757 assigned to the same region of chromosome 8.

The 5’ chimpanzee *EcoRI* recognition site is not present in the human DNA sequenced. As a result, the human *EcoRI* fragment extends 5’ of the chimpanzees’ by 417 bases, to the preceding *EcoRI* site. Within the human sequence a potential 3’ splice junction was identified, AAGACGCGCAACAG/G, beginning 398 base pairs into the 1.6 kilobase fragment. Subsequent to human base pair 421, where the chimpanzee sequence begins, no intronic sequence is present. Therefore, it appears the 3’ exon of a novel human gene has been identified, extending approximately 3 kilobases. In both humans and chimpanzees this 3’ exon contains a truncated L1 element starting at base pair 1807 of the chimpanzee sequence.

Isolation of the chimpanzee cDNA clone, GOR-47-1 containing the GOR epitope (GRRGQKAKSNPNRPL), was first reported by Mishiro *et al.*, (1990). The protein encoded by GOR was identified due to its high levels in the blood plasma of chimpanzees infected with hepatitis C virus. The GOR epitope within this protein was found to cross-reacts with antibodies from patients infected with hepatitis C, and was suggested as a method to be used in the detection of low level virus infection in blood (Mishiro *et al.*, 1990).
Chapter IV: Discussion

Inquiries against the EST database, with the human and chimpanzee sequences, identified six mouse ESTs (W81932; AA034629; AA097346; AA120740; AA030569; AA110524). These ESTs originated from various mouse cDNA libraries, indicating that homologous mouse genes are ubiquitiously expressed. Although the function of the GOR gene product is unknown, its high level of conservation between mammalian genomes suggests it has an important function.

4.2.5.2 5-Oxo-L-Prolinase Gene

Inquiries against the non-redundant database with two subcloned fragments from cosmid 46F4 recognized a novel human gene previously identified in rats. The gene (OPLAH) encodes an enzyme, 5-oxo-L-prolinase, responsible for catalyzing the cleavage of 5-oxo-L-proline to form L-glutamate. The 5-oxo-L-prolinase cDNA, isolated from a rat kidney library, was sequenced in an effort to characterize this enzyme in rats (Guo-jie et al., 1996).

Sequence obtained from the two subcloned fragments matched the OPLAH cDNA between base pair 2857 and 3395 in the rat. The region subcloned in the human homologue consists of two BssHII-EcoRI fragments contained within a 4.2 kilobase EcoRI fragment (Figure 4.5) Within the sequence obtained four exons were detected, flanked on either side by introns. The sequences identified as 3' and 5' splice junctions (Shapiro and Senapathy, 1987) are listed in Table 4.1.

Comparisons between the rat enzyme protein and a hypothetical yeast protein YKL215C, encoded by a homologous gene (Guo-jie et al., 1996), found them to be almost 50% identical. Comparisons between regions of the rat protein and the four homologous human exons offered the following results. Among the 40 amino acids from exon a there were 7 differences (83% identity) (Figure 4.6a). Variations were due to substitutions in the codon at position 2 in five cases, and position 3 in two cases. Five silent substitutions were identified, one at position 1 and the other four at position 3. Three differences were
identified in the 49 amino acids of exon b (94% identity) (Figure 4.6b). Two of these were due to a substitution at position 1, and the third a substitution at position 2. As well, there were nineteen silent substitutions, sixteen occurring at position 3, and three at position 1. Exon c has 5 differences among the 45 amino acids (89% identity) (Figure 4.6c) The differences were due to substitutions at position 1 in four cases, and position 2 in one case. An additional seventeen silent substitutions were identified, fifteen at position 3 and two at position 1. Of the 46 amino acids in exon d (Figure 4.6d) 10 differences were identified (78% identity) all of which were due to substitutions at codon position 1. An additional eighteen silent substitution were recognized, seventeen at codon position 3 and one at position 1. The high level of conservation for this gene throughout eukaryotic evolution demonstrates the importance of this enzyme’s function.

Subsequent investigation of the EST database with sequence from the rat OPLAH cDNA identified four mouse ESTs (AA271445; W29895; AA271446; AA097778). Thus, sequencing adjacent to BssHII sites in cosmid 46F4 detected a novel human gene with homologous genes identified in yeast, mouse and rat. Primers designed for this project and are currently being used to localize the OPLAH gene on chromosome 8.
4.3 Conclusions

The goal of this project was to identify human genes not previously detected by sequencing DNA adjacent to a rare cutting enzyme site. DNA adjacent to \textit{BssHII} sites in eleven cosmids from the LA08NC01 chromosome 8 library were sequenced. Through this method two ESTs were identified; two open reading frames with no apparent homologues; two CpG islands, one of which contains a translation initiation signal and is 5' of human EST H88121; and two novel human genes (Table 4.3). The identification of novel human genes and open reading frames with no homologues is notable considering the number of expressed sequences present in the database (709,530 human; 182,549 mouse; 30,196 nematode; 9,206 fruit fly; 6,650 rat; 3,042 baker's yeast). As predicted, approximately 80% of regions adjacent to \textit{BssHII} sites were found to be associated with genes. In conclusion, this thesis demonstrates that sequencing of DNA adjacent to \textit{BssHII} sites is an effective approach for the identification of novel genes.
Chapter IV: Discussion

Figure 4.1

Translations of ORFs predicted by GRAIL flanked by 3’ and 5’ splice sequences (Shapiro and Senapathy, 1987).
Figure 4.2

CpG islands with TATA boxes and potential transcription start sequence (Kozak, 1996) highlighted in italics.
Figure 4.3

GOR EcoRI fragment comparison between human and chimpanzee.

The human and chimpanzee EcoRI fragments are shown to scale. The human genomic fragment extends 421 base pairs 5' of the chimpanzee fragment. The location of the PCR product is shown on the 1.6 kb fragment. E: EcoRI; B: BssHII.
Chapter IV: Discussion

human GAATTC CCCACTGCCCTCTCNCANCTGCCTGCCGCTGCTGGGCTCT 46
human GGGTCCCCAGATGCACTTACNCATCTCTATTTGTGACATCCA 92
human GNCGGANCANACANCGAGCGGCTCATCCATATTTGTGACATCCA 138
human AATATGGTGAGGAGGTGCTACTTCCTGCGCACANGGGATCGG 184
human CCTGCCGCTCGCAACCACCAGCCAGGGCTGACAGAGCCCCTCGGG 230
human TAGTCCGCATCTCAACCAGTGAAGAAGGAAGACGATCGCCAC 276
human TCTTCCACCCTTGCTGTCNTGTCATGTTACACAGATCCANCCN 322
human ACCCGGGTGCAGCGACAGCGCAGCGGCTCCGCTGCTCCANGGCTCN 368
human GCANACAGCCANGGGAAGACGCGCAACAGGTCAAGGGGATGGC 410

S G D G
human XQEXXTAXTXXSXSKX 445
human ATNCAA GAA NC ACC GCC ACC ANN NG TCT AAG CNA
chimp GAA TTC CCC ACC ACC ATC AGC TCT AAG CCA

EFPSPTTISKS
human IXTRRPSLPSLXNLN 479
human ATCNCTCGTCGTCCTCACTCGACGGCCCGTGGCTCCANGGCTCN
chimp ATCGTCGTCGTCCTCACTCGACGGCCCGTGGCTCCANGGCTCN

IVRRPSLSLSLKL
human KPIILLRSXSQGC 516
human TTGAATA AAAAAAACTATTATCCCTCGACGANC TCTGGGTGC
chimp AAAAAAATTTATTATCCCTCGACGANC TCTGGGTGC

QIPTVLRRSYGAL
human CAA ATCCACCGCGTCTGCTCGCAGGCTACGACCGTCGCTTCCANGGCTCN 553
chimp CAA GTCACCACCGCGTCTGCTCGCAGGCTACGACCGTCGCTTCCANGGCTCN
**Chapter IV: Discussion**

<table>
<thead>
<tr>
<th>L</th>
<th>F</th>
<th>T</th>
<th>X</th>
<th>K</th>
<th>F</th>
<th>L</th>
<th>X</th>
<th>F</th>
<th>C</th>
<th>A</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>CTG</td>
<td>TTC</td>
<td>ACC</td>
<td>NAA</td>
<td>AAA</td>
<td>TTT</td>
<td>CTC</td>
<td>AAN</td>
<td>TTT</td>
<td>TGC</td>
<td>GCC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chimp</td>
<td>CTG</td>
<td>TTC</td>
<td>ACC</td>
<td>GAA</td>
<td>GAG</td>
<td>TGT</td>
<td>CTC</td>
<td>AAG</td>
<td>TTC</td>
<td>TGC</td>
<td>GCC</td>
</tr>
<tr>
<td>L</td>
<td>F</td>
<td>T</td>
<td>E</td>
<td>E</td>
<td>C</td>
<td>L</td>
<td>K</td>
<td>F</td>
<td>C</td>
<td>A</td>
<td>S</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
<td>E</td>
<td>A</td>
<td>X</td>
<td>X</td>
<td>K</td>
<td>G</td>
<td>N</td>
<td>E</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>human</td>
<td>AAN</td>
<td>CAN</td>
<td>GAAG</td>
<td>GCC</td>
<td>NAG</td>
<td>GAN</td>
<td>AAG</td>
<td>GCC</td>
<td>TC</td>
<td>AAC</td>
<td>GAA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chimp</td>
<td>AAG</td>
<td>CAG</td>
<td>GA</td>
<td>G</td>
<td>GCC</td>
<td>GTG</td>
<td>GAG</td>
<td>AAG</td>
<td>GCG</td>
<td>CTG</td>
<td>AAC</td>
</tr>
<tr>
<td>K</td>
<td>Q</td>
<td>E</td>
<td>A</td>
<td>V</td>
<td>E</td>
<td>K</td>
<td>A</td>
<td>L</td>
<td>N</td>
<td>E</td>
<td>E</td>
</tr>
<tr>
<td>X</td>
<td>V</td>
<td>E</td>
<td>C</td>
<td>P</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>human</td>
<td>AAN</td>
<td>GTG</td>
<td>CC</td>
<td>TA</td>
<td>GAA</td>
<td>TGC</td>
<td>NC</td>
<td>CCC</td>
<td>ACN</td>
<td>A</td>
<td>650</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chimp</td>
<td>AAG</td>
<td>GTG</td>
<td>GCC</td>
<td>TAC</td>
<td>GAC</td>
<td>TGC</td>
<td>AGC</td>
<td>CCC</td>
<td>AAC</td>
<td>A</td>
<td>229</td>
</tr>
<tr>
<td>K</td>
<td>V</td>
<td>A</td>
<td>Y</td>
<td>D</td>
<td>C</td>
<td>S</td>
<td>P</td>
<td>N</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 4.4a*

GOR sequence comparison between human and chimpanzee beginning at the 5' EcoRI site.

The human sequence is numbered from the EcoRI site in the 1.6 kb EcoRI-BssHII genomic fragment. The chimpanzee sequence is numbered from the EcoRI site of the partial cDNA as reported by Mishiro et al., (1990).
### Chapter IV: Discussion

<table>
<thead>
<tr>
<th></th>
<th>K</th>
<th>K</th>
<th>X</th>
<th>L</th>
<th>S</th>
<th>R</th>
<th>E</th>
<th>F</th>
<th>I</th>
<th>Q</th>
<th>G</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>AAG</td>
<td>AAA</td>
<td>GAN</td>
<td>TTG</td>
<td>TCC</td>
<td>AGA</td>
<td>GAG</td>
<td>TTT</td>
<td>ATT</td>
<td>CAG</td>
<td>GGA</td>
<td>ATT</td>
</tr>
<tr>
<td>chimp</td>
<td>AAG</td>
<td>AAA</td>
<td>GAN</td>
<td>TTG</td>
<td>TCC</td>
<td>AGA</td>
<td>GAC</td>
<td>GCT</td>
<td>TAT</td>
<td>CCA</td>
<td>GGA</td>
<td>ATC</td>
</tr>
<tr>
<td></td>
<td>473</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Y</td>
<td>A</td>
<td>L</td>
<td>X</td>
<td>C</td>
<td>E</td>
<td>X</td>
<td>W</td>
<td>Y</td>
<td>H</td>
<td>T</td>
<td>H</td>
</tr>
<tr>
<td>human</td>
<td>TAAC</td>
<td>GCC</td>
<td>TTG</td>
<td>GAN</td>
<td>TGT</td>
<td>GAG</td>
<td>ANG</td>
<td>TGG</td>
<td>TAA</td>
<td>CAC</td>
<td>ACG</td>
<td>CAT</td>
</tr>
<tr>
<td>chimp</td>
<td>TA</td>
<td>C</td>
<td>GCC</td>
<td>TTG</td>
<td>GAC</td>
<td>TGT</td>
<td>GAG</td>
<td>ATG</td>
<td>TGC</td>
<td>TAC</td>
<td>ACC</td>
<td>ACG</td>
</tr>
<tr>
<td></td>
<td>437</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>X</td>
<td>E</td>
<td>X</td>
<td>A</td>
<td>G</td>
<td>F</td>
<td>T</td>
<td>V</td>
<td>V</td>
<td>D</td>
<td>X</td>
</tr>
<tr>
<td>human</td>
<td>GGC</td>
<td>NTA</td>
<td>GAG</td>
<td>NTG</td>
<td>GCC</td>
<td>GGG</td>
<td>TTC</td>
<td>ACC</td>
<td>GTG</td>
<td>GTG</td>
<td>GAC</td>
<td>GCN</td>
</tr>
<tr>
<td>chimp</td>
<td>GGC</td>
<td>CTA</td>
<td>GAG</td>
<td>CTG</td>
<td>ACC</td>
<td>CGC</td>
<td>GTC</td>
<td>ACC</td>
<td>GTG</td>
<td>GTG</td>
<td>GAC</td>
<td>GCC</td>
</tr>
<tr>
<td></td>
<td>400</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>M</td>
<td>G</td>
<td>V</td>
<td>V</td>
<td>Y</td>
<td>D</td>
<td>T</td>
<td>X</td>
<td>V</td>
<td>K</td>
<td>P</td>
</tr>
<tr>
<td>human</td>
<td>GAC</td>
<td>ATG</td>
<td>GGA</td>
<td>GTG</td>
<td>GTG</td>
<td>TAC</td>
<td>GAC</td>
<td>TAC</td>
<td>GAC</td>
<td>ACT</td>
<td>TTN</td>
<td>GTG</td>
</tr>
<tr>
<td>chimp</td>
<td>GAC</td>
<td>ATG</td>
<td>CGA</td>
<td>GTG</td>
<td>GTG</td>
<td>TAC</td>
<td>GAC</td>
<td>ACC</td>
<td>TTC</td>
<td>GTG</td>
<td>AAG</td>
<td>GCC</td>
</tr>
<tr>
<td></td>
<td>364</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>X</td>
<td>V</td>
<td>T</td>
<td>E</td>
<td>A</td>
<td>D</td>
<td>V</td>
<td>A</td>
<td>K</td>
<td>X</td>
<td>S</td>
<td>X</td>
</tr>
<tr>
<td>human</td>
<td>NGA</td>
<td>GTC</td>
<td>ACC</td>
<td>GAG</td>
<td>GCC</td>
<td>GAT</td>
<td>GTC</td>
<td>GCC</td>
<td>AAG</td>
<td>ANG</td>
<td>AGC</td>
<td>ATN</td>
</tr>
<tr>
<td>chimp</td>
<td>GGA</td>
<td>GTC</td>
<td>ACC</td>
<td>GAG</td>
<td>GCC</td>
<td>GAC</td>
<td>GTC</td>
<td>GCC</td>
<td>AAG</td>
<td>ACG</td>
<td>AGC</td>
<td>ATC</td>
</tr>
<tr>
<td></td>
<td>291</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>L</td>
<td>P</td>
<td>Q</td>
<td>V</td>
<td>Q</td>
<td>A</td>
<td>I</td>
<td>L</td>
<td>L</td>
<td>S</td>
<td>F</td>
</tr>
<tr>
<td>human</td>
<td>AGG</td>
<td>TTG</td>
<td>CCC</td>
<td>CAA</td>
<td>GTC</td>
<td>CAA</td>
<td>GCC</td>
<td>ATC</td>
<td>CTG</td>
<td>CTG</td>
<td>AGC</td>
<td>TTT</td>
</tr>
<tr>
<td>chimp</td>
<td>ACC</td>
<td>TTG</td>
<td>CCC</td>
<td>CAA</td>
<td>GTG</td>
<td>CAA</td>
<td>GCC</td>
<td>ATC</td>
<td>CTG</td>
<td>CTG</td>
<td>AGC</td>
<td>TTT</td>
</tr>
<tr>
<td></td>
<td>255</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>S</td>
<td>A</td>
<td>Q</td>
<td>T</td>
<td>I</td>
<td>X</td>
<td>I</td>
<td>G</td>
<td>H</td>
<td>S</td>
<td>V</td>
</tr>
<tr>
<td>human</td>
<td>TTA</td>
<td>AGC</td>
<td>GCC</td>
<td>CAA</td>
<td>ACC</td>
<td>ATC</td>
<td>NTC</td>
<td>ATC</td>
<td>GGG</td>
<td>CAC</td>
<td>AGC</td>
<td>GTG</td>
</tr>
<tr>
<td>chimp</td>
<td>TTC</td>
<td>AGC</td>
<td>GCC</td>
<td>CAA</td>
<td>ACC</td>
<td>ATC</td>
<td>CTG</td>
<td>ATC</td>
<td>GGG</td>
<td>CAC</td>
<td>AGC</td>
<td>CTG</td>
</tr>
<tr>
<td></td>
<td>219</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>S</td>
<td>A</td>
<td>Q</td>
<td>T</td>
<td>I</td>
<td>L</td>
<td>I</td>
<td>G</td>
<td>H</td>
<td>S</td>
<td>L</td>
</tr>
<tr>
<td>chimp</td>
<td>TTC</td>
<td>AGC</td>
<td>GCC</td>
<td>CAA</td>
<td>ACC</td>
<td>ATC</td>
<td>CTG</td>
<td>ATC</td>
<td>GGG</td>
<td>CAC</td>
<td>AGC</td>
<td>CTG</td>
</tr>
<tr>
<td></td>
<td>978</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

83
**Chapter IV: Discussion**

<table>
<thead>
<tr>
<th></th>
<th>E</th>
<th>S</th>
<th>D</th>
<th>L</th>
<th>M</th>
<th>A</th>
<th>L</th>
<th>K</th>
<th>L</th>
<th>I</th>
<th>H</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>GAG</td>
<td>AGC</td>
<td>GAC</td>
<td>CTG</td>
<td>ATG</td>
<td>GCC</td>
<td>CTG</td>
<td>AAG</td>
<td>CTG</td>
<td>ATC</td>
<td>CAC</td>
<td>AGC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chimp</td>
<td>GAG</td>
<td>AGC</td>
<td>GAT</td>
<td>CTG</td>
<td>CTG</td>
<td>GCC</td>
<td>CTG</td>
<td>AAG</td>
<td>CTG</td>
<td>ATC</td>
<td>CAC</td>
<td>AGC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>V</td>
<td>V</td>
<td>D</td>
<td>T</td>
<td>A</td>
<td>V</td>
<td>L</td>
<td>F</td>
<td>P</td>
<td>Q</td>
<td>Y</td>
</tr>
<tr>
<td>human</td>
<td>ACC</td>
<td>GTG</td>
<td>GTG</td>
<td>GAC</td>
<td>ACG</td>
<td>GCC</td>
<td>GTG</td>
<td>CTC</td>
<td>TTC</td>
<td>CCG</td>
<td>CAA</td>
<td>TAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chimp</td>
<td>ACC</td>
<td>GTG</td>
<td>CTG</td>
<td>GAC</td>
<td>ACG</td>
<td>GCC</td>
<td>GTG</td>
<td>CTC</td>
<td>TTC</td>
<td>CCG</td>
<td>CAC</td>
<td>TAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>V</td>
<td>L</td>
<td>D</td>
<td>T</td>
<td>A</td>
<td>V</td>
<td>L</td>
<td>F</td>
<td>P</td>
<td>E</td>
<td>Y</td>
</tr>
<tr>
<td>human</td>
<td>ATG</td>
<td>GGT</td>
<td>TTC</td>
<td>CCC</td>
<td>TAC</td>
<td>AAG</td>
<td>CGC</td>
<td>TCC</td>
<td>CTC</td>
<td>AGG</td>
<td>AAT</td>
<td>CTC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chimp</td>
<td>CTG</td>
<td>GGT</td>
<td>TTC</td>
<td>CCC</td>
<td>TAC</td>
<td>AAG</td>
<td>CGT</td>
<td>TCC</td>
<td>CTC</td>
<td>AGG</td>
<td>AAT</td>
<td>CTC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>human</td>
<td>ACC</td>
<td>GCA</td>
<td>GAC</td>
<td>TAC</td>
<td>CTG</td>
<td>GCA</td>
<td>CAG</td>
<td>ATC</td>
<td>ATC</td>
<td>CAG</td>
<td>GAC</td>
<td>AGC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chimp</td>
<td>GCG</td>
<td>GCC</td>
<td>GAC</td>
<td>TAC</td>
<td>CTG</td>
<td>GCA</td>
<td>CAG</td>
<td>ATC</td>
<td>ATC</td>
<td>CAG</td>
<td>GAC</td>
<td>AGC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>human</td>
<td>AGA</td>
<td>ATY</td>
<td>LAQ</td>
<td>I</td>
<td>I</td>
<td>Q</td>
<td>I</td>
<td>Q</td>
<td>D</td>
<td>S</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chimp</td>
<td>AAG</td>
<td>D</td>
<td>Y</td>
<td>L</td>
<td>A</td>
<td>Q</td>
<td>I</td>
<td>I</td>
<td>Q</td>
<td>D</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>human</td>
<td>CAG</td>
<td>GAC</td>
<td>GCC</td>
<td>CAC</td>
<td>AAC</td>
<td>TCC</td>
<td>AGC</td>
<td>GAC</td>
<td>GAC</td>
<td>GCA</td>
<td>AAC</td>
<td>GCC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chimp</td>
<td>CAG</td>
<td>GAC</td>
<td>GCC</td>
<td>CAC</td>
<td>AAC</td>
<td>TCC</td>
<td>AGC</td>
<td>GAC</td>
<td>GAC</td>
<td>GCA</td>
<td>AAC</td>
<td>GCC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>human</td>
<td>CAG</td>
<td>GAC</td>
<td>GCC</td>
<td>CAC</td>
<td>AAC</td>
<td>TCC</td>
<td>AGC</td>
<td>GAC</td>
<td>GAC</td>
<td>GCA</td>
<td>AAC</td>
<td>GCC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chimp</td>
<td>CAG</td>
<td>GAC</td>
<td>GCC</td>
<td>CAC</td>
<td>AAC</td>
<td>TCC</td>
<td>AGC</td>
<td>GAC</td>
<td>GAC</td>
<td>GCA</td>
<td>AAC</td>
<td>GCC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>human</td>
<td>TGC</td>
<td>CTG</td>
<td>CAG</td>
<td>ATG</td>
<td>CTG</td>
<td>ATG</td>
<td>TGG</td>
<td>AAG</td>
<td>GTC</td>
<td>AGA</td>
<td>CAG</td>
<td>CGC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chimp</td>
<td>TGC</td>
<td>CTG</td>
<td>CAG</td>
<td>ATG</td>
<td>CTG</td>
<td>ATG</td>
<td>TGG</td>
<td>AAG</td>
<td>GTC</td>
<td>CGA</td>
<td>CAG</td>
<td>CGC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4.4b**

GOR sequence comparison between human and chimpanzee beginning at the 5' BssHII site.

The human sequence is numbered from the BssHII site in the 1.6 kb EcoRI-BssHII genomic fragment. The chimpanzee sequence is numbered from the EcoRI site of the partial cDNA as reported by Mishiro et al., (1990).
Chapter IV: Discussion

GOR sequence comparison between human and chimpanzee beginning at the second BssHII site.

The human sequence is numbered from the BssHII site in the 1.4 kb BssHII-EcoRI genomic fragment. The chimpanzee sequence is numbered from the EcoRI site of the partial cDNA as reported by Mishiro et al., (1990).

Figure 4.4c
Figure 4.5

5-oxo-L-prolinase comparison between human and rat.

The 650 base pairs of sequence obtained from the 1.6 and 1.8 kb BssHII-EcoRI subclones are shown as a line where the sequence is intronic and as a dark boxes, labeled a-d, where it is an exon. The location of these exons in the rat cDNA are shown as white boxes.

B: BssHII
### Figure 4.6a

5-oxo-L-prolinase sequence comparison between human and rat for the first exon identified in the human sequence.

The human sequence is numbered from the *BssHII* site in the 1.6 kb *EcoRI-BssHII* genomic fragment. The rat sequence is numbered from the complete cDNA as reported by Guo-jie *et al.*, (1996).
Chapter IV: Discussion

Figure 4.6b

5-oxo-L-prolinase sequence comparison between human and rat for the second exon identified in the human sequence.

The human sequence is numbered from the BssHII site in the 1.6 kb EcoRI-BssHII genomic fragment. The rat sequence is numbered from the complete cDNA as reported by Guo-jie et al., (1996).
## Figure 4.6c

5-oxo-L-prolinase sequence comparison between human and rat for the third exon identified in the human sequence.

The human sequence is numbered from the BssHII site in the 1.8 kb BssHII-EcoRI genomic fragment. The rat sequence is numbered from the complete cDNA as reported by Guo-jie et al., (1996).

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>S</th>
<th>A</th>
<th>V</th>
<th>F</th>
<th>D</th>
<th>F</th>
<th>S</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>GCCTAGGGCGCGCAG</td>
<td>GGC</td>
<td>AGC</td>
<td>GCT</td>
<td>GTG</td>
<td>TTT</td>
<td>GAC</td>
<td>TTC</td>
<td>AGC</td>
</tr>
<tr>
<td>rat</td>
<td>TCAACCTGAGTCAG</td>
<td>GGC</td>
<td>AGT</td>
<td>GCG</td>
<td>GTC</td>
<td>TTT</td>
<td>GAC</td>
<td>TTA</td>
<td>ACT</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>T</th>
<th>G</th>
<th>P</th>
<th>E</th>
<th>V</th>
<th>F</th>
<th>F</th>
<th>G</th>
<th>N</th>
<th>L</th>
<th>N</th>
<th>A</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>GGC</td>
<td>ACT</td>
<td>GGG</td>
<td>CCG</td>
<td>GAG</td>
<td>GTG</td>
<td>TTT</td>
<td>GGT</td>
<td>AAT</td>
<td>CTC</td>
<td>AAT</td>
<td>AAC</td>
<td>GCC</td>
<td></td>
</tr>
<tr>
<td>rat</td>
<td>GGT</td>
<td>TCC</td>
<td>GGG</td>
<td>TCT</td>
<td>GAG</td>
<td>GTG</td>
<td>TTT</td>
<td>GCC</td>
<td>AAT</td>
<td>CTC</td>
<td>AAT</td>
<td>GCC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>A</th>
<th>V</th>
<th>T</th>
<th>L</th>
<th>S</th>
<th>A</th>
<th>L</th>
<th>I</th>
<th>Y</th>
<th>C</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>CCG</td>
<td>CGG</td>
<td>GCC</td>
<td>GTA</td>
<td>ACC</td>
<td>CTG</td>
<td>TCC</td>
<td>GCC</td>
<td>CTC</td>
<td>ATC</td>
<td>TAC</td>
<td>TGC</td>
<td></td>
</tr>
<tr>
<td>rat</td>
<td>CCG</td>
<td>AGA</td>
<td>GCC</td>
<td>ATA</td>
<td>ACA</td>
<td>CTG</td>
<td>TCT</td>
<td>GCT</td>
<td>CTC</td>
<td>ATC</td>
<td>TAT</td>
<td>TGC</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>L</th>
<th>R</th>
<th>C</th>
<th>L</th>
<th>V</th>
<th>G</th>
<th>R</th>
<th>X</th>
<th>I</th>
<th>P</th>
<th>L</th>
<th>N</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>CTG</td>
<td>CGC</td>
<td>TGT</td>
<td>CTG</td>
<td>GTG</td>
<td>GGC</td>
<td>CGC</td>
<td>NAC</td>
<td>ATC</td>
<td>CCA</td>
<td>CTC</td>
<td>AAC</td>
<td></td>
</tr>
<tr>
<td>rat</td>
<td>TTA</td>
<td>CGC</td>
<td>TGT</td>
<td>CTA</td>
<td>GTG</td>
<td>GGC</td>
<td>CGT</td>
<td>GAC</td>
<td>ATC</td>
<td>CCA</td>
<td>CTT</td>
<td>AAC</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Q</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>CAG</td>
<td>GTTCGCAGGGG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rat</td>
<td>CAG</td>
<td>GTTTGCTGGGC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

89
Figure 4.6d

5-oxo-L-prolinase sequence comparison between human and rat for the fourth exon identified in the human sequence.

The human sequence is numbered from the BssHII site in the 1.8 kb BssHII-EcoRI genomic fragment. The rat sequence is numbered from the complete cDNA as reported by Guo-jie et al., (1996).
Chapter IV: Discussion

Table 4.1

RNA Splice Junction Sequences

<table>
<thead>
<tr>
<th>Subclone Name</th>
<th>3' splice junction</th>
<th>5' splice junction</th>
</tr>
</thead>
<tbody>
<tr>
<td>13E3 0.6</td>
<td>TGGTGGTGCTGGAG/G</td>
<td>AG/GTGCTCCCTGCT</td>
</tr>
<tr>
<td>93C11ΔBssHII</td>
<td>AANANANTGCNCAG/G</td>
<td>AG/GAGATNGAGAC</td>
</tr>
<tr>
<td>46F4: exon a</td>
<td>CCAGAAACCTGCAG/G</td>
<td>AG/GTGGGCTGGG</td>
</tr>
<tr>
<td>46F4: exon b</td>
<td>CTGNTCCCTGCCAG/G</td>
<td>AG/GTGGGCCCCTG</td>
</tr>
<tr>
<td>46F4: exon c</td>
<td>GCCTAGGCGCGCAG/G</td>
<td>AG/GTTCGCAGGGG</td>
</tr>
<tr>
<td>46F4: exon d</td>
<td>CCGCCCCCGCAGCAG/G</td>
<td></td>
</tr>
</tbody>
</table>

Splice site consensus sequences from each intron/exon boundary identified in the human 5-oxo-L-prolinase gene, listed as exons a-d, and two open reading frames.

Table 4.2

Potential TATA and CCAAT Boxes and Sp1 Binding Sites

<table>
<thead>
<tr>
<th>Subclone Name</th>
<th>TATA box</th>
<th>CCAAT box</th>
<th>Sp1 binding site</th>
</tr>
</thead>
<tbody>
<tr>
<td>141A6 1.7</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>156G7 3.3</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>166H7 5.0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>175G8 3.3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 4.3

Summary of Data from Cosmids

<table>
<thead>
<tr>
<th>COSMID</th>
<th>Fragment Isolated (kb)</th>
<th>Match in Data Base</th>
<th>Information from Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>13E3</td>
<td>0.6</td>
<td>-</td>
<td>ORF</td>
</tr>
<tr>
<td>40G1</td>
<td>1.4</td>
<td>+</td>
<td>GOR gene</td>
</tr>
<tr>
<td>40G1</td>
<td>1.6</td>
<td>+</td>
<td>GOR gene</td>
</tr>
<tr>
<td>46F4</td>
<td>1.6</td>
<td>+</td>
<td>OPLAH gene</td>
</tr>
<tr>
<td>46F4</td>
<td>1.8</td>
<td>+</td>
<td>OPLAH gene</td>
</tr>
<tr>
<td>46F11</td>
<td>2.5</td>
<td>+</td>
<td>Repetitive DNA</td>
</tr>
<tr>
<td>77C1</td>
<td>2.4</td>
<td>+</td>
<td>L1 sequence</td>
</tr>
<tr>
<td>93C11</td>
<td>-</td>
<td></td>
<td>ORF</td>
</tr>
<tr>
<td>141A6</td>
<td>1.7</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>156G7</td>
<td>2.3</td>
<td>-</td>
<td>CpG island</td>
</tr>
<tr>
<td>166H7</td>
<td>5.0</td>
<td>-</td>
<td>CpG island</td>
</tr>
<tr>
<td>166H7</td>
<td>9.2</td>
<td>+</td>
<td>human EST</td>
</tr>
<tr>
<td>175G8</td>
<td>3.3</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>176F5</td>
<td>1.8</td>
<td>+</td>
<td>human EST</td>
</tr>
</tbody>
</table>
References


Appendix

13E3 0.6 (M13)
```
GCGCGCATGCTAGGCTGGTGGAGAGCTCATCTGGAAAGACGCAAAGCCTCTCTCTCCT
CCATCTCAGAGGGTCTGGTGCAGAGAAGCTTCGGAAGACGGCAAACGTCT
5 5
CCACCAACGAGACAGCACGATCGTGAGGCACGTCCTGGATGAGGTGACAACGATT
110
CCATTTTCACAGCCATGGGCACCAACGCCACTCGCAGCATGTTCGAGGACGCCAG
165
ACTGGCAGCTGACCTGCTTCTGCTGAGGCCGACGCTTGGAGGCGGCCTGGACGGGCC
220
GAAGCCTTCTGCCGGGAATCCCTG
275
GTGCAGTCCAGCTGTCCTAGCCTACCCCTGCTGCTGGAGCCCAGATGTGCGTCCC
330
AGGTGCTCCTGCTCTTACTANAAACGATCTCCCAGGTGAAGGCGACNGCCCTANNA
385
ACCATTGTAGGCTTCGGCAAACGTGCTCTCTCGCATTCAAATAGGTTAAGTNGAC
440
TCTTACCTGCTTTCTCCTACAAAAGCTCCATTGACACCGCAAGGAGACTCTT
495
AAGCCGACATGTCACACGCTTACCCGCTGAGGCAGAGGCCTGCTTCTGCAC
550
TTGAGAGATCTCTGAGGGAGGTTGCCACGCTTTTTCTGTAGAGGCACCAC
605
AACAGCAGCTGGTCTAGGGGAAGGTGGGCAATTCAAAAAACAAA
650
```

40G1 1.4 (M13)
```
GCGGCGCTCGCCGCGCTGCTCCATCCCTCTGCCCATCCGAGACCTCTGTCCTTACA
5 5
CCACTAGCCACCCACGTTGGGACTTCCATGGCTTCTGAGTACAAGGCCAGCCCC
110
CCGGCCCAACCAGCTTTGCGAATGCCCCTGTTACCTCTTTTCTGTAGAGGCCACCAC
165
AGGGAGGTGGTGAAGCTTCGGCTCTGGAGTTACAGATCTGGGTTCAAGGCCA
220
AATTCCACCACTTACTAGGTGTAAATATGGAGATAAGCTCTTTGCGCTTCT
275
ACCTTTTGGTCTTTTAATGTTGATCAAAAAAGACTTAGACACCCACATAATAA
330
ATAATAAATATGGCAGAACCTTAAACCCAACACTGCTCAATATTACAGACACCAAA
385
GACAGAGAAAGTTAAAAAGATATACCCGGGGAATTTGAGCTTACCTCTGCACCA
440
AGCGGGACCTAGGANCATACCTACAGAACGTCCACCCCAATACCAAGATATAC
495
ATCTTTCCAGCACCACACTACACTATTATTTCCCCTACATTGAGGGAAT
550
TAAACACCTCTCTGTTAAAGTCTAATACAGAAAATTATTACAAACCGTCTCTCC
605
AACCACATGGAACACCAACTAATAAACCTCCTGAGTNGAAAAACACT
650
```

40G1 1.4 (T3)
```
GAAATCCGCTGTGAATCCGCTGCTTCTGATATTCATTGCTCTATTCTTAAAT
5 5
TTATTGCCCTCAATTTCCAGAAGCTGTATTGCTTTAGCATTCCAACATTCTTCTC
110
CTTGTATTACCTCGGGAGTGGTGGCATGTGTCCTCAAGAATTTATCTTCTGTA
165
TTTTCCGAGTTTTTGTGCTGCTTAGAGGTGTTGACAGTATCTTCATCTAGTGAT
220
CTTCTGTCGGGATCGTATGTTATGCTCTTTTATATCTTGATTGAGTTCCTGTG
275
ATTCTCTCTTCCTTTCAATTTAAATAGCTGTCTGCTAGTGGTCTACATTCTTTT
330
ATGTTTCCAAAAACGCCCTCCCTGGAATTGATTTTATTTTAGAAAAGGGTTTTT
385
TGCTCTATCCTTCCAGTCTGCTCGGNAATTGATTATTTCTTGGCCTCTGCTCTA
440
GCTTGTGATGTTGCTCTTCTTCTCTTCTTATTAAATGGTATGTTANGG
495
TAGGCTTTTTGATCTTTTCTGCTTTCCCCTTTGGGCATTTAAAGGCAATTATA
550
TTTCCCTTCCATACACTGCTGNTTTAAATGTGGTTCCTCCANANANTCCGGTAA
605
GTGCTCTGCCATGCTTCCANAAAAATATCTTTATTTCTGCNT
650
```
Appendix

40G1 1.6 (M13)
GCGCGCTGTCTGACCTTCCACATCACCATCTGCAGGCAGGCGTTTGCNTCCTCGC 5
TGGANTTGTGGCCGTCCTGGCTGTCCTGNATGATCTGTGCCAGNTAGTCTGCCGC 110
GANATNCCTGAGGGGTGCTTGGAGGGAACCCATGTATTGGCCGGAAGGACACG 165
GCCGCTGCACCCACCGGCTGCTGTGNTAGTGTCTGCAACATCGTCAGTCTCCTCA 220
CGCCTGACCNTAGANATGGTGGTCTGGCAGGCTTAAAAANCTCAGCAANGATGCTTG 275
GACTTTGGGCAACCTNTAGTCTGNTGCTGCGATCTGCTGTGACNTCGA 330
AAACCTGTGGTTGAGTCACACCTCACAATCTCGTGTTGGGCTGACNAAAAGGTGCTG 385
ACCACCTCNCATGTCNCCGTCACACCGTGAAACCCGGCAACNTCNATNGCCTAG 440
CGTNGTGTTAACCACANTCTCACAANTCTCCAGGCTTTAATCTCCTGAATAACTCTC 495
TGGACAAANTCTTCTCTCGAAGTCTCNCCCAGAAATCCTNTGANGCTCTCCCTTTGCG 550
GCCCTCCCCACGTTCCTGCTTGGCACTTGCANCCCAACATCCCAGTAACC 605
NNCTGCCCAACAGGTTTTTTTGTGTTAAACCCCGCTCNCNACCNCNC 650

40G1 1.6 (T3)
GAATTCCCACTGCCCTCNTCANCCCTGCACCCANAGCTGGCTCTGGTGGGCC 5
AGATGCACNANCAGGCCTCATCATTTGTGGACATCCAGCAGGANCCANAACNG 110
GGGTCCGGCGGTGCCCCCAACGTGGCCCAATATGGTGACGGAGTGCTGCTACTTC 165
CTTGCGACANGGGATCGGCCCTGCCGCTTGCCAACCACCCCAAGCTGACAGACA 220
GGCCCTCGGGTAGTCCGCATCTCAACCCCCAGTGAAGAAGGAAGACGATCGCCCA 275
CTCTTCCACCCCTTGGTNGTCTGGTCATGTTACACAGATNCCNNGNANAACCGGG 330
GCCCGAGNACAGCCAAAGCTCCGCTCCGTCCNANGTNCGCNCANACCGCGGAAA 385
GAGCGGCAACAGGGGTGCTGCTGACTNCAAGANCAACCACCCACANNNGCTCTA 440
AGCNAACTTCNCTGGCGTCTGCACATTTGCAGTTTGAATTTGAAATACCTATT 495
ATCTCCGAANCTCTGTTGGTCGCAATAACCCNACCCTGCTCTGCCAGCTATCTC 550
CACAAGTCCACCNCANAAAAATCTTCRANTCTGGCTCTCAANANCAGAGGCGNAG 605
GANAAGGCGTCACAGAAGAAGNTGCGCTGAATGCNCCCCACNA 650

46F4 1.6 (M13)
GCGCGCCCGAGGGNAAGGGAGAGGCTGTCAGCGGCCGCACCGTGCCCCTGCCGC 5
CTAAAGGCACCTCTCCACCCAGGCGCACCCTGACTCATGACTGATCGACACGGAGGC 110
GGGATGGGGGAACCGTCGTCCATGTGGTCTTCCGAGGACACCTCCAGGGGAGCAG 165
CCCGGGCCTGCGCGGAGGTTTCCAAAGGCCAGCACAATGTCTCGACGGGCGGAGCTC 220
AGCGTGTGCTGGACAGGGANACAGGATCAGTGTGGGCGAGTCACCTGACACGGATGG 275
CCCTGCGCCACACCTCTGCTGCGCCAGGCCCAACCTGAATAGTGCGCGCTAGAGCC 330
TGGACACCCAGTCGCTACGTGGCCAAATAGAAGCTGCCACACAGCGGTGATGCC 385
TTCTGTTGGCTGACCTGCGGAGGTGCTGCGAGTTCGCTGCTCAGGTCTTT 440
CTGTCTCCGCTGACTTTGGGGAACCTTGCTGGCGCGCAGGCTGCTCGTCCGC 495
TGGATGACGAGTGCTCCTCAACCCCTTTGCGCAACCTCGTGCACCCCGC 550
ACCAAGTTTTCACCANATTCAAAAACAGGTCTCTTGGCGCCCTTAAAACACCCCA 605
GCGCGGCTGGGAATTTGCTCAACACCTCAGGAAACCCCTCCTCTCCT 650
Appendix

46F4 1.8 (M13)
GCGCGCCCTAGGCGCGCAGGGCAGCGCTGTGTTTGACTTCAGCGGCACTGGGCCGG 55
AGGTTTGGTAAATCTCAAGCAGCAGCGGCGTGTAACCCTTGCTCCCTCTATCTTA 110
CTGCTCGCTGCTGTCTGGTGCTGCCGTACATCCTACATCCTGTCAGCAGCTGGG 165
GTGGTGCGGAGGGGGCACCACNNCCGAGTTCTACCTGAGTGCGGCTCGTGAGCCC 220
CAGCCAGTCTGAGACGAAGGTTGCTGCTGCTGCTGCTGTCTGGTGCAGGG 275
CGAAGGCCACACCCCAANNCCTCAAGAAGCCGCCCAGCCACAGGGCTGCTCCT 330
GCCGCCCTGCGCCTCCTCCTCCCTGGCCCCCTCCCTAAATCTGGGCCGCGGGGTGGG 385
AATCTGGGGGGGGAATAGGAGGGAGAGGGAGAGGGAGAGGGAGAGGGAGAGGG 440
TAGGGGACCTTTCGCGCTCCGCCACACCCCAACCCCCAAGCAAATATCCCTGCAAC 500
GCCAACAAACTTTACCTTGGGACACCCCAACCCCCCAGCGGACAGCATACACTGGG 550
GCCAACAAACTTTACCTTGGGACACCCCAACCCCCCAGCGGACAGCATACACTGGG 605
GCCAACAAACTTTACCTTGGGACACCCCAACCCCCCAGCGGACAGCATACACTGGG 660

46F11 2.5 (M13)
GCGCGCCCTAGGCGCGCAGGGCAGCGCTGTGTTTGACTTCAGCGGCACTGGGCCGG 55
AGGTTTGGTAAATCTCAAGCAGCAGCGGCGTGTAACCCTTGCTCCCTCTATCTTA 110
CTGCTCGCTGCTGTCTGGTGCTGCCGTACATCCTACATCCTGTCAGCAGCTGGG 165
GTGGTGCGGAGGGGGCACCACNNCCGAGTTCTACCTGAGTGCGGCTCGTGAGCCC 220
CAGCCAGTCTGAGACGAAGGTTGCTGCTGCTGCTGCTGTCTGGTGCAGGG 275
CGAAGGCCACACCCCAANNCCTCAAGAAGCCGCCCAGCCACAGGGCTGCTCCT 330
GCCGCCCTGCGCCTCCTCCTCCCTGGCCCCCTCCCTAAATCTGGGCCGCGGGGTGGG 385
AATCTGGGGGGGGAATAGGAGGGAGAGGGAGAGGGAGAGGGAGAGGGAGAGGG 440
TAGGGGACCTTTCGCGCTCCGCCACACCCCAACCCCCAAGCAAATATCCCTGCAAC 500
GCCAACAAACTTTACCTTGGGACACCCCAACCCCCCAGCGGACAGCATACACTGGG 550
GCCAACAAACTTTACCTTGGGACACCCCAACCCCCCAGCGGACAGCATACACTGGG 605
GCCAACAAACTTTACCTTGGGACACCCCAACCCCCCAGCGGACAGCATACACTGGG 660
GCCAACAAACTTTACCTTGGGACACCCCAACCCCCCAGCGGACAGCATACACTGGG 720

46F11 2.5 (T3)
GAATTCCAGCATATTGTCCCTACATTGGCACTTCCATATTATGCTGCTCCTGCAAGAC 55
AAAAAAATATGCGTCTATTTCTGCGCAACCACAGGCACTACCTCTATTGCTGTTAT 110
CTTCCCCATTGCTGCCCTTTATGATGTCTTGTCTCTTCTGCAGGCTGTTACTGA 165
TTTCTGATTGCTGCAACACATATTTTCTAACTATGGGTATTTCATGTGCTAAACA 220
CATGTTTCTACAACTGCTGTTGCAATTGTCTGCTTGGATGGGCTCACTAGTCTCAGC 275
TTAGAAAGTAAAGGCAAGATGGATTTTAAGGAAGATGAAAGTTGAAAGGAAGATAATA 330
AGAGATTGATTTGCAAGTAAATAATAGTCAATGTGATTGTAACAGGCTAACTGATA 385
TTTCAAGATGAAAGGCAAGATGGATTTTAAGGAAGATGAAAGTTGAAAGGAAGATAATA 440
GAATGGGATATATCATTGCAAAACTTTATATTATATGCTGCTCCTCAGCTC 495
CAGCGCCGCTCCCTCTTATTGCGGTCTCCTGCAATTCTGGCAACACAGGAAATTTT 550
GTTATTATCCCTCTTGCAACACCTATTTCTGCTCAATTGGTCNAACCTCATTCTC 605
CTGTCGAATTATTTATGCTCCCTTGGGCTGGGAAATTGTTGCGGAA 660

100
Appendix

77C1 2.4 (M13)
GCGCGCACCAGGCGCACAGGCCACAGGGCGAGGCATATCTGCCATCTGGGAGAGG
CAAGGGCTAACTCAGGAAAATGGCGAGCTTGCTGACGCTCGCTGATGCCAGG
CTGAGAACTTCCAGACCTCCCTAAGGGCGGCAGGACGAGCTCTGGGAGAGG
AAAAAGCTGAGAGACGAGCTGAGAGACGAGCTGAGAGACGAGCTGAGAGAC
GAAGCTGAGAGACGAGCTGAGAGACGAGCTGAGAGACGAGCTGAGAGAC

93C11 (M13)
GCGCGCACCAGGCGCACAGGCCACAGGGCGAGGCATATCTGCCATCTGGGAGAGG
CAAGGGCTAACTCAGGAAAATGGCGAGCTTGCTGACGCTCGCTGATGCCAGG
CTGAGAACTTCCAGACCTCCCTAAGGGCGGCAGGACGAGCTCTGGGAGAGG
AAAAAGCTGAGAGACGAGCTGAGAGACGAGCTGAGAGACGAGCTGAGAGAC
GAAGCTGAGAGACGAGCTGAGAGACGAGCTGAGAGACGAGCTGAGAGAC

141A6 1.7 (M13)
GCGCGCACCAGGCGCACAGGCCACAGGGCGAGGCATATCTGCCATCTGGGAGAGG
CAAGGGCTAACTCAGGAAAATGGCGAGCTTGCTGACGCTCGCTGATGCCAGG
CTGAGAACTTCCAGACCTCCCTAAGGGCGGCAGGACGAGCTCTGGGAGAGG
AAAAAGCTGAGAGACGAGCTGAGAGACGAGCTGAGAGACGAGCTGAGAGAC
GAAGCTGAGAGACGAGCTGAGAGACGAGCTGAGAGACGAGCTGAGAGAC

156G7 2.3 (T3)
GCGCGCACCAGGCGCACAGGCCACAGGGCGAGGCATATCTGCCATCTGGGAGAGG
CAAGGGCTAACTCAGGAAAATGGCGAGCTTGCTGACGCTCGCTGATGCCAGG
CTGAGAACTTCCAGACCTCCCTAAGGGCGGCAGGACGAGCTCTGGGAGAGG
AAAAAGCTGAGAGACGAGCTGAGAGACGAGCTGAGAGACGAGCTGAGAGAC
GAAGCTGAGAGACGAGCTGAGAGACGAGCTGAGAGACGAGCTGAGAGAC

101
Appendix

166H7 5.0 (M13)
GCGCGCCGNNAGTTATGGGTCTCNGCGTCTGGCCGGAANCTGCGGAGGTTGGAA 55
ACTGANCGCTGTGATGTCCCGAGCCATGGGGACCGATGCCCGGCGCGATTTGGCC 110
GCTGTCGCGCCCTACCAATGCTTACATGCTTACGCTGACTTGGGCC 165
ACTGCTTCCACNTGGCGCTGGCCAACCTCAANCTNGTCAAGCGCCGCCGCTTC 220
TCATGCGGCCCTAGGGTCTGANGCCGGCCCAGCCAGCTTCTCTACAGCTNGC 275
GTCACTATAAGGCGCTGATTGGGGAGGCAGCGGCCCCAGGCTGCGGAGGTTGGAA 330
CTCTCTCTTTTNNATCGGATATGCTGCTTGGGATCGGTGATGCTGCTTGGCC 385
GAAACNCCCTCAGCCCAATTGGGATCAGAGCGGTGCTGAGCAGGGACCGCCGCGG 440
TCCTGCGGATCTCTTNACTGTGTAATCTGNTNNGGGCATTNTGNAAGAATCTTCCAAT 495
NCTAATTNACAAAACCACACTCAANCGAATCTTGCAAGAAAAATAATTNTTAA 550
ATGGAAGATTATTNGAGGAAACCACCTTGTCTGATGTCTTGAATCTCACCACCCTCC 605
TTNCAACCTTTTTTTNTNGNCCCCCNCAATTCTTTCTCCCCCTCTCCGC 650

166H7 9.2 (M13)
GCGCGCCGCCCACCACTACTCCTAGC GGCGCTGCTCTCCAAACCTCAATCAACGCCCTGC 55
GGCGGCCCTCCCTCTCTCTCAGGCACCCGGGCGNTGTGAACCCCAGGGCGCCCGCGGG 110
CGGGTCTGGGACGACGAAAGTTAGGGGGGACCGCCCTAANTGACATCTCCCAACATG 165
TGCCCTTGGTNNNCCCTCGCTGCCCTTACCANTTACCCCACTCTCTCTCTGACT 220
CCACGCGCTCNGGGTGGTCCGGACCTGGCCCTCTTACCCANTTACCCCCACTCCTGTACT 275
CCACGCGCTCNGGGAATTTGCTGCTTGGGATCGGTGATGCTGCTTGGCC 330
CAGGCTCAGGTCCACCTCCACCCTCCTCCTTNTTNTNCTNNGTCTCCCTNCTTNTNCTTNT 385
CACCNGGGGCTTGGCTCCACCTGGGAACCTTCTTTTCTCNATNNTNCTNGTCTCCCTNT 440
TCCCTCCNTCTCNATCCCCCTCNATGGGACANTCCCGCAANTGGGTCCCTCCNTNCNNNG 495
NANCCTNNANCCTGCGNANNACTCCTCCCTCCCCCTTTTTGGCTNCTCTCCCTCC 550
CTCCTNCGCNTCNTGGTACNGAATTCTTNGCTCCTGCGNANCNGAATCTNANACNGC 590
GNTNNCTTAAANTNNNNNTANCTNNTCAGANACCTCTTNTAATTCTTNGCTCTNTA 605
NCANGGAANGGGAATNNCTANTTNTNAATTNNAACTAATAATNN 650

175G8 3.3 (M13)
GCGGCCAACTCCTCTTTTGGTGACCCCGCATGGGATGGTGCTTTCTCAC 55
CATTCTGTACCGCAAGCAGCATGAACAAATATTCAACAAGCAAACTTGGGACGCA 110
TCCCTGAATATCTTAGCTGTCCCTTTGTGTGGTCTCTCCAGACATAGGTAGCCCTGT 165
TGGAACTCGAGCCACAAACAGAGCGCGCATGGAGACAGAACAACACATGGCCCG 220
TGATTACCGCAAGTTTGGCGACAGAGATCCAGACAGACACACTTCTGGCCCGGAAA 275
GCACCGAAGCCACTCCACATTTGTACGGCAGCTTGGGGGCTATGGGGATCTTGGG 330
TGAGACTTCTTGGTCTCTGTCTAGCAGACACAGACAGACTTTGCTCTCTGCGCAGAAAAAAA 385
TCNNATTACAGCCACAGAGATGGGCTTACCTTAGGTTCTCGTAGGGAGAGAGAAGC 440
CAGATTCTGGGCTCTACAGCAGACACAGAGAACTCCATTGCGGAAGAACCCACACAGAAAGGGATACATGTCTATGGG 495
CACCCTTCTCTCTTCTCTCTCTCAGACATGGGATGGTGCTTTCTCAC 550
GCGCTGCGCCCTCCCTCTCTCTCAGGCACCCGGGCGNTGTGAACCCCAGGGCGCCCGCGGG 605
AAAGGAAGGAAGGAGAGGAGGTTTGGAGAAATTAAGAAGCCAGAAGG 650
### Appendix 1

Sequence obtained from cosmids beginning at a rare cutting enzyme site. BssHII (GCGCGC) with the M13 primer and EcoRI (GAATTC) with the T3 primer.
Appendix

Baylor College of Medicine search launcher: http://kiwi.imgen.bcm.tmc.edu:8088

Appendix 2

Web site addresses.
Identification of a Novel Human Gene (GOR) Localized to 8q13-8q22

Leah R. DeBella, Michael Schertzer and Stephen Wood

Department of Medical Genetics, University of British Columbia, Vancouver, British Columbia, Canada, V6T 1Z3

Subject category: Gene discovery and disease genes

Short Communication

Appendix 3

Article submitted for publication prior to completion of Thesis.
Appendix

Transcribed sequences are not dispersed evenly between and within chromosomes, but appear to be concentrated in particular regions. Areas that are rich in genes have a higher GC content than gene poor regions and may contain landmark regions called HTF-islands (Lindsay and Bird, 1987) or CpG regions (Aissani et al., 1991). These islands are associated with 40% of tissue specific genes and with most ubiquitously expressed sequences (Larsen et al., 1992). Each CpG island that has been investigated has been found to be located near or within a gene.

Rare cutting enzymes are a class of enzyme whose recognition sequences are found frequently in areas with high GC and gene content. Although the recognition sequences for this class of enzymes vary, each contains at least one CpG dinucleotide. Lindsay and Bird (1987) demonstrated that a single rare cutting enzyme could be used to identify CpG islands in cloned DNA, suggesting a method for locating gene rich regions. One of these rare cutting enzymes, BssHII, is an excellent choice to be used in this manner, since 80% of its sites are located in GC rich regions (Larsen et al., 1992).

Cosmid 40G1, selected from the LA08NC01 flow-sorted chromosome 8 library (Wood et al., 1992), contains two BssHII sites within a 3.3 kb EcoRI fragment. Both EcoRI/BssHII fragments, 1.6 and 1.4 kb in size, were subcloned into a modified pBluescript (Stratagene) vector. The multiple cloning site in the pBluescript vectors is present as a BssHII fragment making these vectors unsuitable for cloning doubly digested fragments with a BssHII end. The pBluescript II KS(+) vector was modified by cleavage with XbaI and BamHI. The vector was then religated in the presence of two oligonucleotides; 5'-GATCCACTAGTT-3' and 5'-CTAGGCGCGCCG-3'. The modified vector, pKSIIAsc, lacks the XbaI and SpeI sites but contains an AscI site (Figure 1). The AscI recognition sequence is a superset of BssHII sites and BssHII ends ligate to the AscI cut ends of the modified vector. Each fragment was then sequenced from both ends using an automated sequencer (Applied Biosystems Model 373). The sequence obtained was compared to genes present in the GenBank database using BLAST (Altschul et al., 1990).
Appendix

Only one gene, GOR (GenBank D10017), identified in chimpanzees, was found to share high sequence identity with the 40G1 cosmid sequences.

Few variations in sequence were found between the human and chimpanzee genes. Among the 95 amino acids compared in Figure 2 there are 9 amino acids differences (91% identity). The 9 amino acid differences were due to substitutions at codon position 1 in 6 cases, position 2 in one case and position 3 in two cases. An additional 5 silent substitutions were observed 1 at position 1 and 4 at position 3.

A novel STS was developed by designing PCR primers to amplify a 286 bp product at the BssHII end of the 1.6 kb fragment. The forward primer, 5'-AGGTTGCCCCAAGTCCAAGC-3', is shown underlined at the top of Figure 2 while the reverse primer, 5'-GCTGTCTGACCTTCCACATC-3', complements the 20 nucleotides shown underlined at the bottom of Figure 2 that are adjacent to the GCGCGC BssHII site. These primers were used to localize the human GOR gene on chromosome 8 using a somatic cell hybrid panel (Wagner et al., 1991; Wood et al., 1986). Only XVIII-23Ha produced a product localizing GOR to 8q13-q22 (Figure 3). This STS was also typed, in duplicate, in the G3 radiation hybrid panel (Stewart et al., 1997) and the results submitted to the Stanford RH server (http://www-shgc.stanford.edu/rhserver2/rhserver_form.html). The highest two point LOD score of 3.87 was reported at a distance of 52.9 cR from SHGC-37027 which lies within the chromosome 8 bin 67. We then used multi-point mapping, RHMAP 3.0 (Boehnke et al., 1991) available from http://www.sph.umich.edu/group/statgen/software/, locally and found that GOR could not be ordered within bin 67. The Généthon marker D8S1757, also located within bin 67, maps to position 94.5 cM on the 165.8 cM sex-average Généthon linkage map (Dib et al., 1996) compatible with the cytogenetic localization for GOR. Thus GOR seems to map close to autotaxin (ATX) and most likely proximal to the carbonic anhydrase cluster (Leach et al., 1996).
Appendix

A 2859 bp EcoRI partial cDNA clone containing the GOR epitope (GRRGQKAKSNPNRPL), was isolated by Mishiro et al., (1990) and sequenced. The corresponding human genomic EcoRI fragment is 3.3 kb in size. The human homologue has lost the EcoRI recognition site corresponding to the 5' site of the chimpanzee fragment. Consequently the human EcoRI fragment extends 5' of the chimpanzee fragment by 417 bases to the preceding EcoRI site. We have identified a potential 3' splice junction, tgttacacagA at bp 311 of the human EcoRI fragment. No intronic sequences were found in the region corresponding to the chimpanzee cDNA. Thus we have identified the 3' exon of the human GOR gene. In both humans and chimpanzees this 3' exon contains a truncated L1 element starting at bp 1807 of the chimpanzee sequence which is downstream from the UAG stop codon at 1576/8 bp.

The protein encoded by GOR was identified due to its high levels in the blood plasma of chimpanzees infected with hepatitis C virus. The GOR epitope within this protein was found to cross-react with antibodies from patients infected with hepatitis C, and was suggested as a complementary method to be used in the detection of low level virus infection in blood (Mishiro et al., 1990).

Additional BLAST searches of the EST database identified six mouse ESTs (W81932, AA034629, AA097346, AA120740, AA030569, AA110524), sharing high sequence identity with the chimpanzee and/or human GOR genes. Interestingly no human ESTs for GOR were present in the database, which validates the value of the rare cutter CpG island approach as a complementary method for gene identification. Although the function of the GOR gene product is unknown, its high level of conservation between mammalian genomes suggests it has an important function.
ACKNOWLEDGEMENTS

This work was supported by grant GO-12753 from the Canadian Genome Analysis and Technology Programme.

REFERENCES


Appendix

Appendix

pBluescript II KS+

PstI  SmaI  BamHI  SpeI  XbaI  NotI
5'...CTGCAGCCCCGGGGAATCCACTAGTTCTAGAGCGGCCGC...3'
3'...GACGTCGGGCCCCCTAGGTGATCAAGATCTCGCCGGCGGCG...5'

pKSIIAsc

PstI  SmaI  BamHI  Ascl  NotI
5'...CTGCAGCCCCGGGGAATCCGGCGCGCTAGAGCGGCCGC...3'
3'...GACGTCGGGCCCCCTAGGTGATCAAGATCTCGCCGGCGGCG...5'

Figure 1

Construction of the modified Bluescript plasmid pKSIIAsc.

The relevant region of the multiple cloning site is shown. The region in bold in pBluescript II KS+ has been excised by restriction digestion and replaced in pKSIIAsc by the region shown in bold.
Figure 2

GOR sequence comparison between human and chimpanzee for the region amplified by the human GOR STS.

The human sequence is numbered from the BssHII site in the 1.6 kb EcoRI/BssHII genomic fragment. The chimpanzee sequence is numbered from the EcoRI site of the partial cDNA as reported by Mishiro et al (1990).
Figure 3

Chromosomal localization of the human GOR gene by STS typing of somatic cell hybrids.

Lane 1 contains a marker digest, φX174 phage digested with HaeIII. The cell hybrid panel consists of: XVIII-23Ha, 8pter-8q22, lane 2; VTGHIL 19, 8pter-8q13, lane 3; 1SHL 3, 8pter-q11, lane 4; 20xPO435-2, 8p23-q11, lane 5; MGV 270, 8q24.1-qter, lane 6; MGV 271, 8q22.1-qter, lane 7. Human DNA, lane 8, is the positive control, and lane 9 is a negative control.