1994

Re-examining Subfamily Classifications For the Alu Family of Repeated DNA Sequences

William A. York
University of Central Florida, lago@pobox.com

RE-EXAMINING SUBFAMILY CLASSIFICATIONS FOR THE ALU FAMILY OF REPEATED DNA SEQUENCES

BY

William A. York
B.A., North Park College, 1989

THESIS

Submitted in partial fufillment of the requirements
for the degree of Master of Science in Biology
in the graduate studies program
of the College of Arts and Sciences
University of Central Florida
Orlando, Florida

Fall, 1994

ABSTRACT

The primate *Alu* family of repetitive elements has been widely characterized. This ubiquitous class of retroposons has been found to occupy some 5% of the human genome. This heterogenous group of Short Interspersed Nucleic acid Elements (SINEs) has been theorized to possess an identifiable subfamily structure between and within various taxonomic levels in primates. It has been postulated that humans possess up to 6 *Alu* subfamilies in their genome; this number, however, has varied according to the method of analysis performed on the data. Quentin (1988) analyzed 127 aligned *Alu* sequences and found evidence supporting the amplification/fixation theory in 5 subfamilies. The research presented in this thesis posits that Quentin's method of alignment used in the correspondence analysis is questionable. A re-examination using an alternative, perhaps more tenable, alignment of the *Alu* sequences may allow for a more lucid and accurate identification of *Alu* subfamily structure in the human genome.

To the Walters Clan

ACKNOWLEDGEMENTS

I would like to acknowledge the following people without whom none of this project would have been possible. Janet, thanks for your infinite flexibility in our many hours of need. Thanks to Dr. Daniels for a starting point and direction in which to move. Dr. Kuhn, thank you for the chance to make something of myself. Dr. Sweet, thanks for the diverting conversations and your jumper cables. Special thanks to Dr. Vickers for your confidence in me. Your friendship helped get me through. I also feel that the remarkably efficient administrative staff of the department of biological sciences deserves special mention, namely Barbara and Rita. Thanks.

To my newly acquired extended family for holiday gatherings and exacerbating my fear of red meat. Kay, for her fabulous cooking (and iced tea); Heather for being around to make fun of things with and for walks in Philly; and Rodney, the coolest father-in-law ever to put up with someone taking away his little girl. Also to my Grandma, who stuck by me, in spite of everything.

Calvin, Dave, and Willi, three friends who give new meaning to the word distraction. Boog, if you need any more places to lounge, you'll need your own apartment. Cyndy, for haircuts, plant-doctoring, and love. Tim, for memorable Halloweens. Most especially to Norba, for putting me up, and putting up with me, I love you. Finally to my wife, Melissa, for getting me out, getting me in, and getting me through; for not sparing my feelings in her countless rereads of this manuscript; and for enduring through the years past, and those to come.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The *Alu* family, named for its typical internal *Alu I* endonuclease restriction site, is the most ubiquitous family of repetitive DNA sequences in the human genome. Its occurrence of over 500,000 copies has led to its being one of the most widely studied of any nucleic acid sequence and has resulted in the association of this sequence with cell-specific transcription regulation (Brini, *et al.*), unequal crossing over events (Shaw, *et al.*), pseudoautosomal boundary definition (Ellis, *et al.*), and hypercholesterolemia (Lehrman, *et al.*). Although its methods of mobility and many of its insertional effects are, at least, partially understood, its origins and evolution are still a subject of controversy. One controversial aspect regarding the *Alu* family is the process by which recently discovered subfamilies of *Alu* sequences have come into existence. It is believed that if the method of subfamily formation and organization is understood, the ubiquitousness of the *Alu* family can be fathomed.

Several theories have been advanced to explain the existence of subfamily formation. Slagel, *et al.* (1987) proposed that either there were relatively few copies of the *Alu* repeat that were capable of amplification when the family arose, or variant members appeared that were more capable of amplification. Alternatively, Willard, *et al.* (1987) proposed that over the course of primate evolution, new members of the *Alu* family arose from different progenitors successively and were then amplified and fixed in their new locations in the genome. Quentin (1988) found support for this amplification/fixation theory for subfamily formation via correspondence analysis which resulted in the identification of 5 distinct subfamilies.

The iterative nature of correspondence analysis, however, makes it particularly sensitive to the method by which the data are coded. The methodology used in constructing the data set used by Quentin (1988) is subject to criticism. This study directly addresses the methodology used in the Quentin (1988) study in both *Alu* sequence selection and alignment strategy. The results support the Willard, *et al.* (1987) theory of amplification/fixation, but require reconsideration of positive subfamily identification as a completely objective process.

BACKGROUND

Classifying DNA

One of the easiest and most informative methods to classify DNA segments (sequences) is by estimating the number of times a sequence (or one closely related to it) occurs in the genome, i.e. its copy number. **Single-copy** DNA sequences are those that, by definition, occur only once in the haploid genome. These are the sequences that are, in most cases, responsible for the production or regulation of a product; most of the regular expression protein coding genes fall into this category (Watson, *et al.*, 1987). **Infrequently-repetitive** DNA has less than 100 copies in the genome and has a Cot value (reannealing rate) of greater than 100. **Moderately-repetitive** DNA sequences are those that occur usually between 10 and 100,000 times in the genome and possess Cot values between 0.1 and 100. These sequences range from the essential histones and transfer RNAs (which need multiple copies due to the incredible demand for them in the cell), to multigene families, to processed pseudogenes. Much of the genome of higher eucaryotes, however, belongs to the class of DNA organization termed **highly-repetitive** DNA, or those occurring more than 100,000 times in the genome and having Cot values less than 0.1 (Singer and Berg, 1991). Although much of this type of DNA lies in centromeric and telomeric regions, a significant portion of it is dispersed throughout the genome. Much of the dispersed, repetitive DNA that is found in the genome owes its distribution to its **mobility** in the genome,[1] such as

[1]Some highly repetitive sequences, such as Variable Number Tandem Repeats (VNTRs), show no evidence of mobility in the genome; that is, in examination of their structure and nucleotide sequences, they show none of the qualities which are traditionally characteristic of mobile DNA sequences.

the mammalian family of Short Interspersed Nucleotide Elements (SINEs), the *Alu* sequence family.

Mobile Genetic Elements

There are three types of mobile genetic elements found in eucaryotes. Although each possesses unique characteristics, they all share some features in common. Some of these features include the presence of stable mutations in genes resulting from the **insertion** of the element, and the presence of terminal direct repeats flanking the sequence which arise from target site duplication (Watson, *et al.*, 1987) (see Figure 1). This by-product of insertion is essential for identification of these elements in the genome.

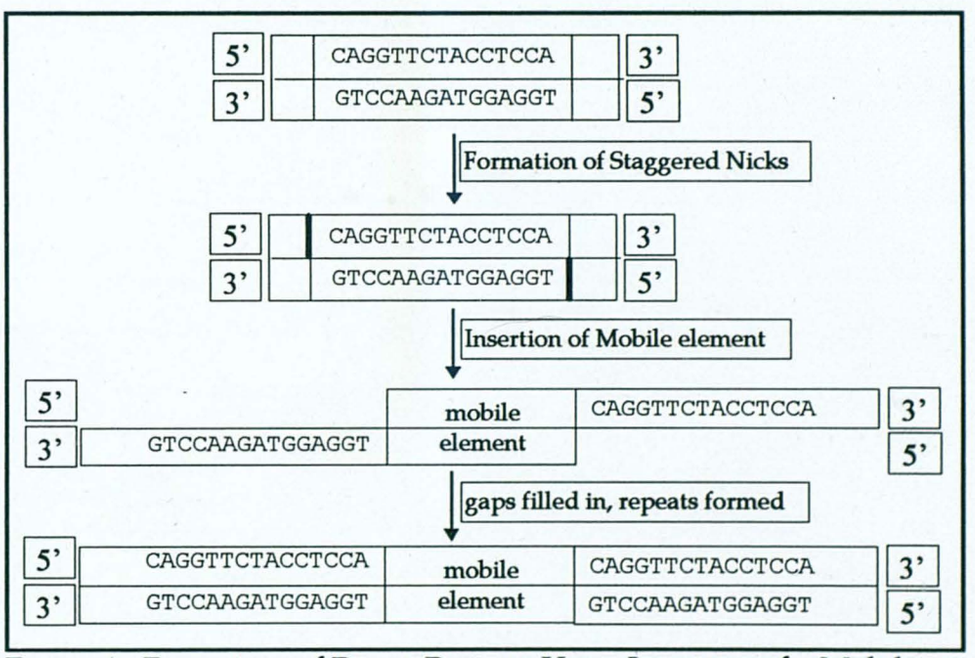


Figure 1. Formation of Direct Repeats Upon Insertion of a Mobile Element

Transposable elements are found in procaryotes as well as eucaryotes. They all possess the genes required for their transposition as well as specific DNA sequences that are inverted at either end of the sequence (Singer & Berg, 1991).

Examples of transposable elements include insertion sequences (IS) in *Escherichia coli*, controlling elements in maize, and P-elements in *Drosophila spp.*. **Retrotransposons** (class I & II), whose mobility depends upon transcription and reverse transcription, can encode multiple protein-coding genes, but always have, at least, the gene for reverse transcriptase (see below). They also possess long terminal repeats on either end of their central segment. Class I retrotransposons are differentiated from class II based on their short inverted repeats flanking the class I sequence and their resemblance to retroviral proviruses (both of which the class II elements lack) (Singer & Berg, 1991). **Retroposons**, or **retrogenes** (pseudogene-like elements), lack regions that code for any specific gene. This classification includes SINEs and LINEs (Long Interspersed Nucleotide Elements), which comprise the majority of this category and appear to belong to a larger classification of processed pseudogenes that are able to move about in the genome. Both LINEs and SINEs utilize retroposition as a means of mobility.

Retroposition

Retroposition appears to be the dominant mechanism for the formation of mammalian repetitive DNA sequence families (Deininger and Daniels 1986). The process of retroposition involves the transcription of DNA into RNA and the subsequent **reverse transcription** back into the DNA. The retroposition process appears to be passive, in that the enzymes required may be supplied by retrotransposons or retroviruses. Processed pseudogenes (i.e. non-coding copies of functional genes) could be copies of small RNAs, such as 7SL RNA or transfer-RNA, or of messenger-RNAs. Many retrogenes (retroposons) appear to be processed pseudogenes.

This process is initiated by a promoter sequence for an RNA polymerase (usually RNA polymerase III for SINEs and RNA polymerase II for LINEs) that is generally located within the sequence and facilitates the transposition of that sequence (Daniels and Deininger 1985). The

fact that the promoter is located within the *Alu* sequence is **internal** and is transposed along with it, allows each new copy to have the potential for its own retroposition (Deininger and Daniels 1986). SINEs, whose size generally ranges from under 100 to over 500 base pairs (bp), do not code for enzymes that advance the transposition process (usually reverse transcriptase) but utilize cellular mechanisms; while LINEs, sometimes up to 7 kilobase pairs (kb) may include open reading frames (ORFs) and retroposition enzymes (Deininger 1989). The internal RNA polymerase promoter has been instrumental in identifying dispersed repetitive element families and the location of the elements themselves throughout the genome (G. Daniels, private communication). The promoter for RNA polymerase III is encoded within SINEs due to their probable origin as class III (RNA polymerase III-transcribed) genes that encode small cytoplasmic RNAs. More generally, SINEs contain, in addition to the internal promoter, a 3′ A-rich terminus. (Deininger 1989) which is necessary for the internal insertion of *Alu* elements (discussed below).

7SL RNA and SINE Origins

The nucleotide sequence of the 7SL RNA gene (a class III gene found in all higher eucaryotes), regardless of the organism in which it is found, is highly conserved. The 7SL RNA is a small cytoplasmic RNA (scRNA) which is the only RNA part of the signal recognition particle (SRP). The SRP contains 6 proteins (between 10,000 and 75,000 Daltons) and is part of a vital reaction that directs proteins into the lumen of the endoplasmic reticulum for cellular transport. The sequence of the 7SL RNA is so highly conserved throughout the genomes of higher eucaryotes that 7SL RNA from amphibian or insect cells can reconstruct an SRP from mammalian proteins (Singer and Berg, 1991). There are only 3-4 functional copies of the 7SL RNA gene in primate genomes, but several hundred pseudogenes exist. The *Alu* family's, a SINE, monomeric units contain sequences that exhibit homology to both ends of the 7SL RNA gene; the central 155 base pairs of the 7SL RNA gene is absent in the monomeric unit (Li, *et al.* 1982 and Ullu, *et al.* 1982).

The discovery of the existence of homology[2] between the human *Alu* family and the 7SL RNA gene was the first step in deducing the origin of repeated DNA sequences. Several other mammalian SINE families have also been shown to exhibit homology to class III genes, including the *Galago crassicaudatus* Type II *Alu* subfamily which was derived from a methionine tRNA (Daniels and Deininger 1985); identified SINEs have been shown to be either dysfunctional 7SL or transfer-RNA. These genes undergo structural modifications that improve their efficaciousness in retroposition and amplification (Deininger and Daniels 1986), consequently, different SINEs have different structures.

Alu Sequences

The *Alu* family of repeated DNA sequences is the most ubiquitous family of all SINEs in primates. In humans, up to 500,000 *Alu* elements are estimated to occupy approximately 5% of the entire genome (Rinehart, *et al.* 1981. This repetitive element consists of two approximately 130 base pair (bp) head-to-tail repeated units (monomers) which show roughly 70% sequence identity to each other. The RNA polymerase III promoter is located in the left half of this sequence, while the right half has no apparent function in the transcription of the element (Fuhrman, *et al.* 1981) . This SINE shows a strong preference for insertion into small poly-A regions (Daniels and Deininger 1985) and into introns and flanking regions of class III genes. An additional, and unique, feature of the human *Alu* family is the insertion of a 31 bp segment in the right half of the sequence (Deininger, *et al.* 1981). Recent research has shown that the

---

[2]There is a distinction between several terms which needs to be made. *Homology* is the term that is used to describe to nucleotide sequences that have been determined to probably have a common ancestry. This determination is made through nucleotide sequence analysis and other methods, but the use of the term implicitly states that a relationship of this type exists. Use of the term *sequence identity*, or just *identity*, implies no such relationship, only that two, or more, sequences show a good deal (usually expressed as a percentage) of base pairs in common at similar (or analogous) positions. Oftentimes, sequences being analyzed must be brought into *alignment*, so that their analogous positions can be examined for similarity.

dimeric *Alu* segment has arisen in several stages. The first of these was the origination of the Fossil Alu Monomer (FAM) from the 7SL RNA and the subsequent diversion that the FAM underwent (Quentin, 1992b). The next stage in the evolution of the *Alu* family was the restructuring of the FAM into two monomeric segments, the Free Left Alu Monomer (FLAM) and the Free Right Alu Monomer (FRAM). The final stage began with the fusion of a FLAM and a FRAM into a dimeric *Alu* sequence (Quentin, 1992a).

The *Alu* family seems to be under no natural selective pressure, and accumulate mutations (insertions, deletions, transitions, and transversions) at a rate which approximates the neutral mutation rate (approximately 0.5% every million years), or that of nearby intergenic regions (Sawada, *et al.*, 1985). Sequences that arose at a roughly equivalent time in evolutionary history should possess approximately equivalent amounts of dissimilarity. However, even though *Alu* sequences are primate-specific SINEs, all *Alu* elements do not seem to be equally divergent from each other. *Alu* sequences can be grouped into subfamilies which show a marked degree of subfamily specificity. The previously noted homology of the human *Alu* sequence to the 7SL RNA, and the *Galago crassicaudatus* Type II *Alu* subfamily to the methionine-tRNA gene (also a class III gene), and the observation that many mammalian SINEs may be derived from amplified tRNA pseudogenes (Daniels and Deininger 1985), explains the existence of **taxon**-specific groups. It appears that most SINEs have arisen independently from separate class III genes (Deininger & Daniels 1986). In addition, several research groups have shown that subfamily organization of *Alu* sequences occurs not only between species, but **within** species, especially in humans (Bains 1986, Slagel, *et al.* 1987, Britten, *et al.* 1988, Jurka and Smith 1988, Quentin 1988). Subfamily organization of the *Alu* family began in the free monomer stage, as recent studies have shown the existence of subfamilies in the two aforementioned free monomer groups (Quentin, 1992a).

Alu Subfamily Organization

Various theories have attempted to explain the existence of subfamilies in the human dimeric *Alu* subfamily. Slagel, *et al.* (1987) proposed that subfamilies arose by one of two methods: either there were relatively few copies of *Alu* sequences that were capable of amplification when the family arose, or variant members appeared that were more capable of amplification. Willard, *et al.* (1988) proposed a new scheme for the appearance of subfamilies; that over the course of primate evolution, new *Alu* members arose from different progenitors (all 7SL RNAs) successively and were then amplified and fixed in their new locations in the genome (the amplification/fixation theory). The Willard *et al.* (1987) study argued that three subfamilies were formed by successive episodic bursts, originating from different progenitors. The three subfamilies found by this group all vary in the amount of similarity they possess (diverged, major, and conserved) which, given a neutral mutation rate, account for their amplification at successive times in primate evolution. The Willard, *et al.* (1988) alternative of successive waves of amplification, provides an explanation of how different subfamilies can exist, with significantly varying degrees of similarity, but without a progressive order from a single progenitor. Quentin (1988) found support for the amplification/fixation theory for subfamily formation via a statistical dimension reduction methodology. Correspondence analysis (see below) was performed on a set of available *Alu* elements in order to elicit latent subfamily structure. Presumably, a statistically insignificant amount of variation between the different resulting groups would support the amplification/fixation theory (or a multiple progenitor theory). Alternatively, if the subfamilies arose through the appearance of variant members capable of amplifying more effectively (as in punctuated equilibrium), there should be a logical progression in the mutational analysis, which is absent.

Repeated DNA sequences constitute a significant portion of the genomes of higher eucaryotes. Two major hypotheses regarding this phenomenon have been advanced. The first is that extensive proliferation of seemingly non-functional repeated sequences is one method of gene regulation at the translational level and is therefore maintained an optimized by natural selection. The second is that this DNA serves no other purpose than its own propagation. This unregulated promiscuity can lead to insertional mutations and deletions of entire genomic segments by homologous recombination. One fairly common occurrence of this type of mutation, with regards to *Alu*, is in the human low density lipoprotein (LDL) receptor gene which leads to hypercholesterolemia. Extensive deletions through homologous recombination in this gene the impair body's ability to regulate between the different forms of cholesterol and result in cholesterol buildup in the plasma (Singer and Berg, 1991).

Problems with Subfamily Identification

The analysis presented in the Quentin (1988) paper clearly illustrated the existence of the differing *Alu* subfamilies in primates and the variation that exists within and between them. The methodology that was used for the data coding in this study is, however, problematic.

First, in constructing the data set, the author aligned 168 human and non-human *Alu* elements to a consensus sequence constructed by a previous researcher on **human** *Alu* sequences (Kariya, *et al.* 1987). There are significant problems caused by using this type of data coding. The consensus sequence used to align the sequences is an average of all the human sequences in their native structure. Aligning human and nonhuman *Alu* sequences to an artificially constructed average introduces a systematic bias into the data set. Such bias could result in first, the possibility of newer sequences (i.e. those available for analysis after the consensus was constructed) being misaligned; and, second, because different subfamilies' *Alu* sequences have

arisen independently, their structural peculiarities may also lend themselves to inappropriate alignment when forced to the human consensus. *A global alignment* is the construction of an aligned data set where regions or segments of the whole sequence are simultaneously optimized for similarity (Ginsburg, 1994). A global alignment would allow the relationships between the sequences to be elicited while avoiding the systematic bias inherent in aligning sequences to a consensus sequence.

Second, Quentin eliminated "non-informative" data, which introduced a source of subjective error, and ran the analysis on the "informative" positions that remained. The elimination of "non-informative" sites, is, in part, necessary due to the extensive computations that are required to perform this analysis. Non-informative positions are those that would contribute very little, if at all, to subfamily identification because nearly all of the sequences have the same nucleotide at that location (Quentin 1988). In addition, certain regions of all *Alu* sequences, such as the highly variable regions of multiple CpG dinucleotides (Bains 1986) and central and terminal poly-A regions are certainly "non-informative." The potential, however, is that subjective elimination of non-informative sites could result in comparison errors. Such errors would be compounded due to the iterative nature of the correspondence analysis algorithm. Consequently, incorrect distances would be assigned to the resulting groupings and conclusions based on those distances could be erroneous. Objectivity in the elimination of non-informative sites could be enhanced by eliminating only those sites that possess statistically insignificant variance estimates.

In addition, some sequences that were included in the Quentin study were included inappropriately, for reasons discussed below; and, their inclusion may conceivably have miscontributed to an overall conception of dissimilarity.

DATA

The Quentin (1988) study was used as a point of departure for this study. As such, all of the original *Alu* sequences that could be located, or could be justifiably included, were included. The GenBank identifier, as well as the number of *Alu* sequences identified with it (identified as N) are listed in Table 1. Also included is a key to notes on exceptions and inconsistencies between these data and those given in the Quentin study (identified as I). These notes are discussed later in this section.

TABLE 1

LIST OF SEQUENCES INCLUDED IN THIS ANALYSIS

GENBANK ID	N	I	GENBANK ID	N	I	GENBANK ID	N	I
Human			**Human**			**Chimpanzee**		
HUMADAG	17	1	HUMHPARS	3		PTAZGLO	1	
HUMAGG	3		HUMIGVKA	2		PTGL01	1	
HUMALPPA	1		HUMINS2	1		PTRE123	2	
HSALUAGP	6	2	HUMLDLR18	2	7	CHPRSA	2	12
HUMAPOC2	4	3	HUMP5311	1		**Orangutan**		
HUMAPOE4	4		HUMPOMC	5	8	ORAHBBE	2	
HUMBLYM1	1		HUMRSA1	1		**Owl Monkey**		
HUMC1AIN1	1		HUMRSA16	1		ATBOWL1	1	
HUMCEA	1	4	HUMRSA27	1		ATHOWL1	1	
HUMFIXG	5		HUMRSAOLD	1		ATHOWL6	1	
HUMGAST2	1		HUMRSAP3	1		**Galago**		
HUMGHN	1		HUMRSKPA1	1		GCREG13	1	
HUMHBA4	5	5	HSREP10	9	9	GCREG19	1	
HUMHBB	6	6	HUMTKRA	12	10	GCREG9	1	
HUMHBB51	1		HUMTPA	22	11	**Gorilla**		
HUMHBBRT	2		HSMLVI2	1		GGALU	1	

The full set of data used in this study (Table 1) consisted of 138 human and nonhuman *Alu* sequences which were obtained from the GenBank and EMBL databanks using the NCBI program RETRIEVE via Internet electronic mail (e-mail) server.

Data Retrieval and Editing

Sequences were obtained using the RETRIEVE electronic mail server[3] for access to the GenBank and EMBL nucleic acid databanks. The returned sequences (retrieve files) were inspected for accuracy regarding *Alu* sequence content. Gene descriptions were used in most cases to confirm that the retrieved sequence was the one requested.[4] Once retrieved, the files were downloaded to a personal computer. The complete GenBank retrieve file is actually a record (or set of records, depending upon the ambiguity of the query) consisting of the nucleic acid sequence, as well as a number of fields that contain information about the sequence, such as the locus name, its GenBank Accession (identification) number, important features and their respective locations, and reference citations. Any of these fields may be specified in the query. The *Alu* sequence(s) was then extracted from the overall gene sequence using a text editor according to the information in the Features field of the record.

Unfortunately, the extraction of the *Alu* sequences from the genes was not always straightforward. The most frequently encountered obstruction was that the complement of the desired sequence was presented. The multiple sequence editor, ESEE (Cabot and Beckenbach,

---

[3]The Internet address for the server is RETRIEVE@NCBI.NLM.NIH.GOV. For information on how to use this server to obtain nucleic acid or protein sequence data, send an e-mail message to the above address with single word "help" in the body of the message. Full instructions for use and query structuring will be returned to the sender.

[4]While some GenBank loci identifications have varied over time, the descriptions of the sequences used in the Quentin (1988) study have stayed fairly consistent. The source descriptions of the *Alu* sequences in Quentin (1988) were matched to the current GenBank descriptions of the genes. In fact, these descriptions were sometimes used as string literals in the query used to retrieve some sequences.

1989)[5] provided a simple way to complement and reverse the *Alu* sequence to the correct orientation for analysis.

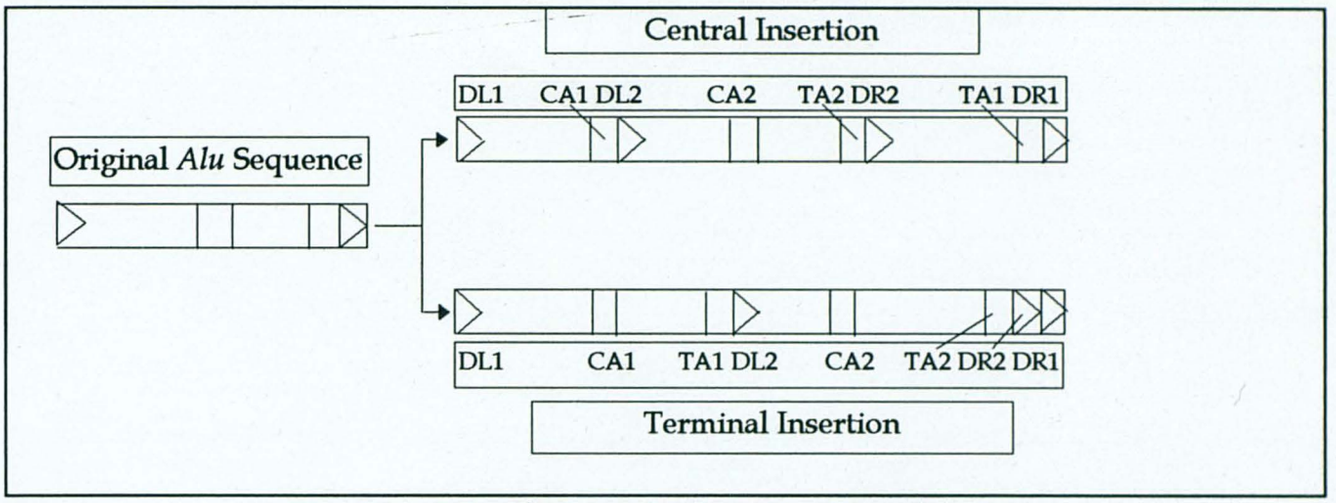


Figure 2. Alu Sequence Insertion Scenarios

Original *Alu* Sequence (Alu Sequence 1)
DL1 = direct repeat left, *Alu* Sequence 1
DR1 = direct repeat right, Alu Sequence 1
CA1 = central A-rich region, Alu Sequence 1
TA1 = terminal A-rich region, Alu Sequence 1

Inserted Alu Sequence (*Alu* Sequence 2)
DL2 = direct repeat left, Alu Sequence 2
DR2 = direct repeat right, Alu Sequence 2
CA2 = central A-rich region, Alu Sequence 2
TA2 = terminal A-rich region, Alu Sequence 2

More than once, *Alu* sequences were encountered that were significantly longer than what is considered "normal" for these sequences. Careful examination of these sequences was necessary since long sequences could contribute significantly to misalignment of the data set (Higgins, *et al.*, 1992). Specifically, these *Alu* sequences were examined for insertions by other *Alu* sequences. Given their preference for insertion into adenosine-rich (A-rich) regions, and the two A-rich regions that are characteristic of *Alu* sequences, it is not unlikely to find such an occurrence (G. Daniels, private communication). The two obvious possibilities for internal insertion are outlined in Figure 2. To extirpate one *Alu* sequence from the other, the sequences must be searched for direct repeats (figure 1). The program MACAW (Schuler, *et al.*, 1991) was

[5]The programs used for alignment (CLUSTAL V and ESEE 1.09) and editing (MACAW 1.0) are available from the software/msdos_software directory on the server at the European Molecular Biology Laboratory (EMBL). The Internet address for the ftp site is ftp.EMBL-Heidelberg.de and the E-mail address is Netserv@EMBL-Heidelberg.de. Programs can be obtained from either source. To have the programs mailed electronically, a UUDecoder and compiler are necessary. The decoder can be obtained from the server. A directory listing of all software available for this platform can be obtained by sending a message to the E-mail address above with the request DIR MSDOS_SOFTWARE. More general help can be obtained by sending a message with the word HELP as a query.

used to search these sequences for direct repeats, primarily in the regions following A-rich areas. To ensure thoroughness, the complement of the "long" sequence was searched as well. *Alu* sequences that were obtained this way, as well as other specific difficulties that were encountered, are documented below in the section "Data Inconsistencies."

Multiple Sequence Alignment and Data Selection

The edited sequences were then used as the input data set for a multiple sequence alignment program. There are basically two methods for computationally aligning multiple sequences. The first is to compare all possible sequence pairs looking for the best fit. This quickly becomes "computationally intractable" as sequences are added (Ginsburg, 1994). The alternative, the progressive pairwise approach, boasts a much faster alignment time at the cost of producing sub -optimal alignment. Due to this method's less-than-perfect alignment status, it is fairly standard practice for multiple sequence alignment programs to export the aligned data set to a multiple sequence editor for adjustment (Ginsburg, 1994).

The program that was used to align the *Alu* sequences in this study was Clustal V (Higgins, *et al.*, 1992) which is based on the Feng and Doolittle (1987) algorithm as modified by Higgins and Sharp (1989).[6] This program, available from the EMBL Netserv ftp site, derives a measure of

---

[6]CLUSTAL V is available in its C source code format and is ready for compilation on VAX/VMS (C), Decstation 2100 (Ultrix C), Sun (Gnu C), Macintosh (Think C), IBM-compatible PC (Turbo C). Minor modifications must be made to the CLUSTALV.h header file depending upon the platform being used and the memory requirements of the data set to be aligned. This study had a fairly large data set, and was therefore aligned on a Sun 670 mainframe computer. Changes needed to accommodate this data set are as follows:

```
#elif UNIX
#define DIRDELIM '/'
#define MAXLEN              700     /*(from 3000)*/
#define MAXN                200     /*(from 50)*/
#define FSIZE               3000    /*(from 15000)*/
#define LINELENGTH          60
#define GCG_LINELENGTH      50.
```

In the first section, the platform designation needs to be defined by uncommenting the line on which the choice lies. For this study, the lines

```
#define VMS 1                   /* VAX VMS */
```

similarity by comparing each pair of sequences with each other. This produces a similarity matrix, or preliminary alignment, which is then used to guide the final alignment. The original measure of similarity is deduced using the Wilbur and Lipman (1983) algorithm, and each alignment step is propagated by the Myers and Miller (1988) algorithm (Ginsburg, 1994).

Due to CLUSTAL V's use of the Feng and Doolittle (1987) algorithm, an alignment is produced which contains many gaps. These gaps represent both insertions and deletions in the multiple alignment. An insertion in one sequence at one position results in a gap at that position for every sequence which does not have an insertion at that position. Likewise, a deletion is also represented by a gap. The insertion of gaps is controllable to a certain extent by the modification of the operating parameter of gap penalties. Reducing this number increases the likelihood of gap insertions which may result in a better alignment. Reducing this number too much, however, could result in each sequence being aligned opposite a long terminal gap (Higgins, *et al.* 1992).

This, naturally, results in the overall lengthening of the sequences; the length for the completely aligned data set is 550 bp. Although these gaps are definitely informative with regards to subfamily identification, large insertions will make an unnecessarily large point of distinction in the analysis. Thus it is essential that as much of the original data set is preserved for input into the analysis as is feasible, while keeping the computation time to a workable level. The Quentin (1988) method for reducing the data set to a workable level consisted of identifying positions in the alignment which were different in over 70% of the cases (sequences) and extracting these "very frequent base changes" into a single, gap-free alignment for use in the analysis. This

---

/* #define UNIX 1 Ultrix/Decstation or Gnu C for Sun */
were changed to:
/*#define VMS 1 VAX VMS */
#define UNIX 1 /* Ultrix/Decstation or Gnu C for Sun */
to accommodate for the Sun 670 on which this program was run.

resulted in the reduction of the data set from 282 bp to 86 positions being used in the analysis. This being the case, only the most informative sites (as in Quentin (1988)) were used. A more conventional methodology was used in this present study to deem positions as "informative". Positions were selected from the final aligned output file based on a variance analysis conducted on all positions. Those locations that possessed a variance of greater than 1.90 (roughly 70% or more of the *Alu* sequences' bases were different) were included. This resulted in a final data set that consisted of 45 positions compared to the 86 positions from the Quentin (1988) study.

A cursory visual examination of the final output from the program CLUSTAL V showed that an optimal alignment was not produced. Adjustment of the alignment was made using the MS-DOS® Editor version 1.1 (Microsoft Corp.) on an IBM-compatible personal computer with an 80386DX microprocessor. This editor provided the advantage of unlimited file size[7] as well as line and column positions. The multiple sequence editor ESEE (Cabot and Beckenbach, 1989) also provided these and other features, but was limited to displaying approximately 10 sequences on a "page" at one time.

The final data set was the alignment of nucleotide positions of 138 *Alu* sequences, aligned so that regions of the entire set of sequences have the maximum number of base pairs in common. The alignment was constructed so that gaps (as mentioned above) were inserted into the sequences (identified as "-") to maximize regions of local similarity. Achieving an optimal alignment depends on the judgment of the aligner, the algorithm used in the computer program, and a subjective definition as to what constitutes "optimal ."

---

[7]Normally, this editor limits the number of characters on a line to 255. The Microsoft Windows® Notepad editor allows for unlimited line size, but limits text to 32K. This unfortunate limit prevented this editor from being used to finish aligning the data, as the full set was more than twice this size.

Data Coding

The data coding, which is necessary due to a methodological artifact of correspondence analysis, is a simple conversion of categorical data into sequential integers. The coding scheme is given in Table 2. Since the data are read by the algorithm as categorical, the actual value of each level is unimportant in the analysis (i.e. no actual math is performed on the values themselves). The only importance in the values is that they are distinct from each other (SPSS, Inc. 1990).

TABLE 2

DATA CODING SCHEME

Nucleotide	A	C	T	G	R,Y,N (X)	gap (-)
Code	1	2	3	4	5	6

The abbreviations for the nucleotides in Table 2 correspond to the current conventions for abbreviating nucleic acid data: adenosine (A), cytosine (C), thymine (T), guanine (G), purine (R), pyrimidine (Y), and unidentified nucleotide (N). The X designation is given to any non-A/C/T/G or non-gap position in CLUSTAL V when nucleic acid data is used.

Data Inconsistencies with Quentin (1988)

While every attempt was made to duplicate the data set used in the original study, much of the data that is currently accessible in GenBank and EMBL is inconsistent with the data specified as being included in the Quentin (1988) study. The inconsistencies encountered during the gathering phase of this study, as well as the exclusion of data that was considered inappropriate

for analysis, are listed and explained here. The numbers correspond to the numbers associated with the particular sequence under column "I" in Table 1.

1. HUMADAG

Twenty one *Alu* sequences were named as being included in Quentin (1988) from this gene (Human Adenosine Deaminase), two of which are greater than 540 nucleotides long. Close examination of these two long *Alu* sequences yielded the following conclusions and actions:

The first sequence's 3' terminus (as identified in GenBank) consisted of an approximately 100 nt long region of oligonucleotides $(TAAA)_n$. This sequence was truncated to facilitate alignment and analysis.

The second sequence (as identified in GenBank) contained 4 adenosine-rich areas, leading to the conclusion that one *Alu* sequence was inserted in another. Upon close examination, *Alu* "J" (as identified in GenBank) was found to contain a central insertion, as described above. The two *Alu* sequences were extracted from one another, and placed in the analysis as sequences "HUMADAGJa" and "HUMADAGJb". Their approximate locations were:

14,837 ... 14,971 Left half, *Alu* Ja
14,977 ... 14,989 direct repeat, *Alu* Jb
14,990 ... 15,226 *Alu* Jb
15,230 ... 15,242 direct repeat, *Alu* Jb
15,243 ... 15,386 Right half, *Alu* Ja

These numbers are approximate for the central positions (not terminal) above due to *Alu* Jb being on the complement strand; this was modified prior to numbering.

2. HSALUAGP

This gene contained all six of the *Alu* sequences reported by Quentin (1988), however sequence six was fragmented into two pieces (as identified in GenBank) and was reassembled for analysis.

3. HUMAPOCII

Five *Alu* sequences were identified by Quentin (1988) for analysis, while three *Alu* sequences are identified in GenBank. This discrepancy can be accounted for in that two of the three GenBank identified sequences are over 600 bp long. These were both found to contain right *Alu* sequence insertions and were separated. The resulting five sequences were then used in this study.

4. HUMCEA

This *Alu* sequence is described in GenBank as "truncated," however its length is greater than 240 bp and is included in this analysis.

5. HUMHBA4

Six *Alu* sequences were identified by Quentin (1988) for analysis, while five *Alu* sequences are identified in GenBank. *Alu* sequence 3 is approximately 600 bp long, and was found to contain a centrally inserted *Alu* segment. The fourth GenBank-described sequence is 248 bp long, but contains no A-rich 3' terminus. An A-rich region exists in this gene, however, directly 3' to the GenBank-described end of this *Alu* sequence. This region was included as part of "HUMHBA4E" to facilitate alignment. A-rich regions convey little information regarding the differences in subfamilies since they are characteristic of all *Alu* sequences (G.

Daniels, private communication; Quentin, 1988) and were not included in the final data matrix for analysis.

6. HUMHBB

Nine *Alu* sequences were identified by Quentin (1988) for analysis, while eight *Alu* sequences are identified in GenBank. The comment field of the GenBank entry for this gene contained references to two additional *Alu* sequences for which approximate locations were given. The first of these corresponded to a location that was given in the features field and the more precise features designation was used. The second referenced a location that corresponded to none of the features-identified *Alu* sequences. This region was extracted, along with the 300 bp flanking it, and searched for an *Alu* sequence. A 335 bp region, not including flanking direct repeats, was located and included in this analysis as "HUMHBBE".

7. HUMLDLR18

The first *Alu* sequence identified in GenBank possessed no A-rich 3'-terminus. An A-rich region, exists, however, directly 3' to the GenBank-denoted 3' end of this repeat. This sequence was extended an additional 12 bp to include a portion of this A-rich region to facilitate alignment, but was excluded from the analysis (see note 5, above). The second *Alu* sequence identified in GenBank for this gene is incomplete, and was excluded from analysis. Truncated or partial *Alu* sequences would not drastically alter the alignment of which they are a part, but their inclusion in an analysis would serve to distinguish missing segments as gaps in the sequence and therefore as differences or points of distinction. These truncated sequences would, therefore be unnecessarily discriminated in the analysis.

8. HUMPOMC

Six *Alu* sequences were identified by Quentin (1988) for analysis. While all are readily identifiable from their GenBank identifications, five of the *Alu* sequences are full length, and the sixth is approximately one-third normal length. The shortened sequence was excluded from analysis.

9. HSREP10 (HUMTBB5 in Quentin (1988))

Alu sequences 9 and 10 have GenBank identifications that overlap. Sequence 9 is truncated and was therefore excluded.

10. HUMTKRA

Fourteen *Alu* sequences were identified by Quentin (1988) for analysis, while thirteen *Alu* sequences are identified in GenBank. One of these is only 70 bp in length and was excluded from analysis. Position 6,421 (*Alu* sequence 8) is given in GenBank as "m." This is assumed to be a typographical error and was replaced with "n" for the purposes of analysis. This information is not available in the original paper.

11. HUMTPA

Twenty eight *Alu* sequences were identified by Quentin (1988) for analysis, while twenty two full *Alu* and six "half-*Alu*" sequences are identified in GenBank. The "half-*Alu*" sequences were excluded from analysis.

12. CHPRSA

The following list of base positions have been identified as the locations for the two *Alu* sequences noted in the description of this gene by GenBank, but for which locations are

absent. These positions have been communicated to the network administrator at NCBI and acknowledged.

repeat region 177 ... 183
/direct repeat Alu 1
misc. 184 ... 484
/complement Alu 1
repeat region 485 ... 491
/direct repeat Alu 1

repeat region 1259 ... 1268
/direct repeat Alu 2
misc. 1269 ... 1569
/Alu 2
repeat region 1570 ... 1579
/direct repeat Alu 2

13. Other Sequences

Other sequences included in the Quentin (1988) analysis have been excluded from this study for a variety of reasons. These sequences and their reasons for exclusion are listed here.

AGMRSASPC - only one of the three identified *Alu* sequences were locatable

HUMPAIB - this *Alu* sequence is truncated and was excluded

HUMLDLRFH - this *Alu* sequence is truncated and was excluded

HSIFNIN3 - this *Alu* sequence is truncated and was excluded

HUMMYCC - this *Alu* sequence is truncated and was excluded

HUMKIN10 - this *Alu* sequence is truncated and was excluded

GCRRSAGAg - this sequence has been unlocatable

AGMRSA - this sequence has been unlocatable

no name given - 5 *Alu* sequences from the Orangutan beta globin region have been unlocatable

14. Exclusions

The following sequences presented special difficulty in alignment. They were particularly prone to aligning opposite terminal gaps, even when stringent gap penalties were applied. Some of them contained long insertions, which made manual alignment particularly difficult. While exclusion of the sequences was not anticipated to drastically affect the analytical outcome of this study, the problems created by their inclusion outweigh the

suspicion generated by their exclusion. These sequences are: HSGAPDP, AGMRSASPC, HUMHBBE, HUMHBBH, HUMHBBI, HUMADAGB, and HUMRSA6.

The following six sequences were also excluded from the final analysis. The alignment resulted in a large number of gaps occupying the "informative" positions that were selected by the variance analysis. These sequences were then highly discriminated in the correspondence analysis and were shown as extreme outliers in the final object scores plot. AGMRSASAT, HUMADAGG, HUMADAGH, HUMAPOC2D, HS7SLRNA, and PTRE123C.

METHODOLOGY

Correspondence Analysis

The method used to analyze this data set was the same method used in Quentin's original study. *Analyse des correspondances*, or correspondence analysis, is analysis by dimension reduction to obtain graphical output. This method is similar, algebraically, to the ecological method of reciprocal averaging, but yields graphical, not numerical, results (Greenacre 1984). This method gained wide popularity in France in the 1960's, led by the statistician Jean-Paul Benzécri, but has only recently gained more worldwide acceptance. Its first use in the study of nucleic acid sequences was by Y. Quentin in 1988, which is the study that is currently under scrutiny.

Correspondence analysis, in its simplest terms, describes a set of multidimensional data (in this case, the number of dimensions is equal to the number of nucleotide positions) in a few, or optimally, 2, dimensions for easy display on a plane in Euclidean space. It can be thought of as principal components analysis of nominal (categorical) data (SPSS, Inc. 1990). In order to achieve the reduction of dimensions that is necessary for this two-dimensional display, a computer program identifies group centroids of one dimension, re-orients the points into a lower dimension, and uses the re-orientation as a point of reference for future iterations. The resulting distances between the points on the plane represent actual distances or differences in the data. Each sequence (known as an object or case) is represented by a point on the plane. Points that are closer together, regardless of their coordinate location, are identified as being more similar.

The major advantage to using this methodology is the graphical output that is produced. Greenacre (1984) comments:

> ...graphical displays provide the best summaries of data - a picture is worth a thousand numbers. A graphical description is more easily assimilated and interpreted than a numerical one and can assist all three functions [of statistics] ... summarizing a large mass of numerical data, simplifying the aspect of the data by appealing to our natural ability to absorb visual images, and (hopefully) providing a global view of the information, thereby stimulating possible explanations.

Most rectangular data sets can be reduced to one, two, or three dimension for easy representation and intelligibility (Greenacre 1984).

Analysis Procedure

The aligned data set was read into SPSS for Windows 6.0 in ASCII format and the HOMALS procedure (version 0.6) in the SPSS Categories module was used to conduct the multiple correspondence analysis. HOMALS is an acronym which stands for HOmogeneity analysis by Alternating Least Squares. In each iteration that is performed, all *Alu* sequences are examined at each nucleotide position and a similarity index is calculated. All sequences with similar categorical values for that position will be grouped closely (for a single position, they would be grouped on a single point) and sequences with differing categorical values would be grouped as far away as possible. The actual values for these points are termed object scores (the sequences in this type of analysis are called objects). When the next position is added, some of the sequences that were grouped as one point will still have the next position in common; some however, will not. The latter will, once again, be moved as far away from their original location as can be, but still as far from the sequences with whom they, as yet, have zero positions in common with, and so on for subsequent positions. When the end of the positions is reached, one iteration has been performed and the process begins again, as now, the first position is once again informative for discrimination. This entire process is repeated until a convergence point is

reached (hence the "alternating least squares" portion of the name). These points are then plotted on a two-dimensional graph. The sequences in this plot are closest to the other sequences with whom they have the most positions in common with, and farthest from those with whom they have the least in common with. Unlike most planar graphs, the axes do not actually correspond to any particular variable, but rather to discrimination dimensions and are therefore, for the sake of clarity, left unlabeled. Since multiple correspondence analysis replaces the categorical values of each sequence with a numerical (ordinal level) value (the object scores), the final output from this procedure can then be used in another procedure that require that type of data.

The groups which could be interpreted as subfamily identifications were identified through the use of a hierarchical cluster analysis of the final object scores. Hierarchical cluster analysis forms clusters by grouping the sequences in the data set into larger and larger groups until they are all part of the same cluster. This methodology, which attempts to identify relatively homogeneous groups of sequences based on their two-dimensional object scores, was chosen due to its lack of assumptions about the number of groups (clusters) that are contained in the data set (Norusis 1993). The **centroid clustering** method (which is the most similar to the correspondence analysis algorithm) of cluster formation was chosen to identify these groups. This method calculates the distance between two clusters as the distance between their means for all of the items. This method only calculated the centroids of the object scores obtained from the correspondence analysis. The results of these analyses can be seen in the next section[8].

[8]A simplified example of a complete alignment and analysis (correspondence analysis and cluster analysis) can be found in Appendix B.

RESULTS AND DISCUSSION

Empirical Expectations

Since the data coding methodology used by Quentin (1988) differs markedly from the methodology used in this study, the sequence selection and alignment strategies employed in this study were expected to significantly alter the final results, or at the least result in an outcome of a different magnitude from that found by of Quentin (1988).

Primary Analysis

The results of the correspondence analysis are shown in Figure 3. The points that are plotted on the graph are the actual object scores for the analysis.

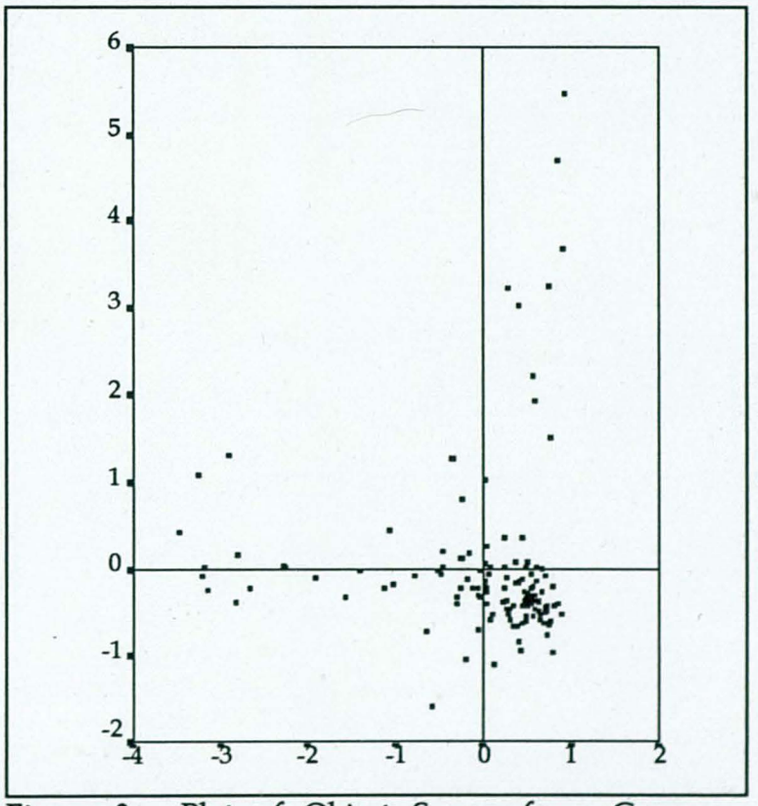


Figure 3. Plot of Object Scores from Correspondence Analysis

The examination of the object scores plot resulting from the correspondence analysis (Figure 3) and the dendrogram provided by the cluster analysis (Figure 4) revealed three major groupings (Figures 5 and 6). The identification of three distinct subfamilies is consistent with the results of Willard, *et* al. (1986) but is in contrast to the subfamily groupings found by Quentin (1988) (see Appendix C),[9] although there is a certain amount of visual similarity between the object scores plot from the Quentin (1988) and from this study.

The final subfamily identification can be interpreted in several ways. As the groupings that are drawn around the points are the result of a hierarchical cluster analysis (see Figure 4) there are differing levels at which a subfamily could legitimately be discriminated. Instead of arbitrarily deciding which level to identify with subfamily classification, a contour plot was drawn around the points for each level of the hierarchical cluster analysis (Figure 6). Figure 5 shows a schematic diagram of the Figure 6 plot, which lacks the spatial distinction made by the correspondence analysis, but shows the groupings differentiated by the cluster analysis more clearly, in addition to the levels at which the clusters discriminated (see below).

The dendrogram produced by the hierarchical cluster analysis shows a complete breakdown of cluster formation. The "rescaled distance cluster combine" in Figure 4 is a measure of the distance at which the clusters are combined (Norusis 1993) or the level at which they are discriminated. The dendrogram is traditionally read from left to right. The greater the horizontal distance, the more distant the differentiation between the groups. In this case, the

[9]Comparison of the subfamily groupings is, however, difficult because the method with which subfamilies were identified in Quentin (1988) is unclear. Quentin refers to the "automatic classification" consequence of the correspondence analysis procedure and indicates that such automatic classifications were the basis for the identified subgroups. However, no reference to this procedure could be found in either the biological or statistical literature. Consequently, the identified subfamily groups here are based on the complementary hierarchical cluster analysis procedure described above.

three cluster solution is the most visually and numerically distinct and is the most interpretable solution, but is subject to several subjective considerations (see below).

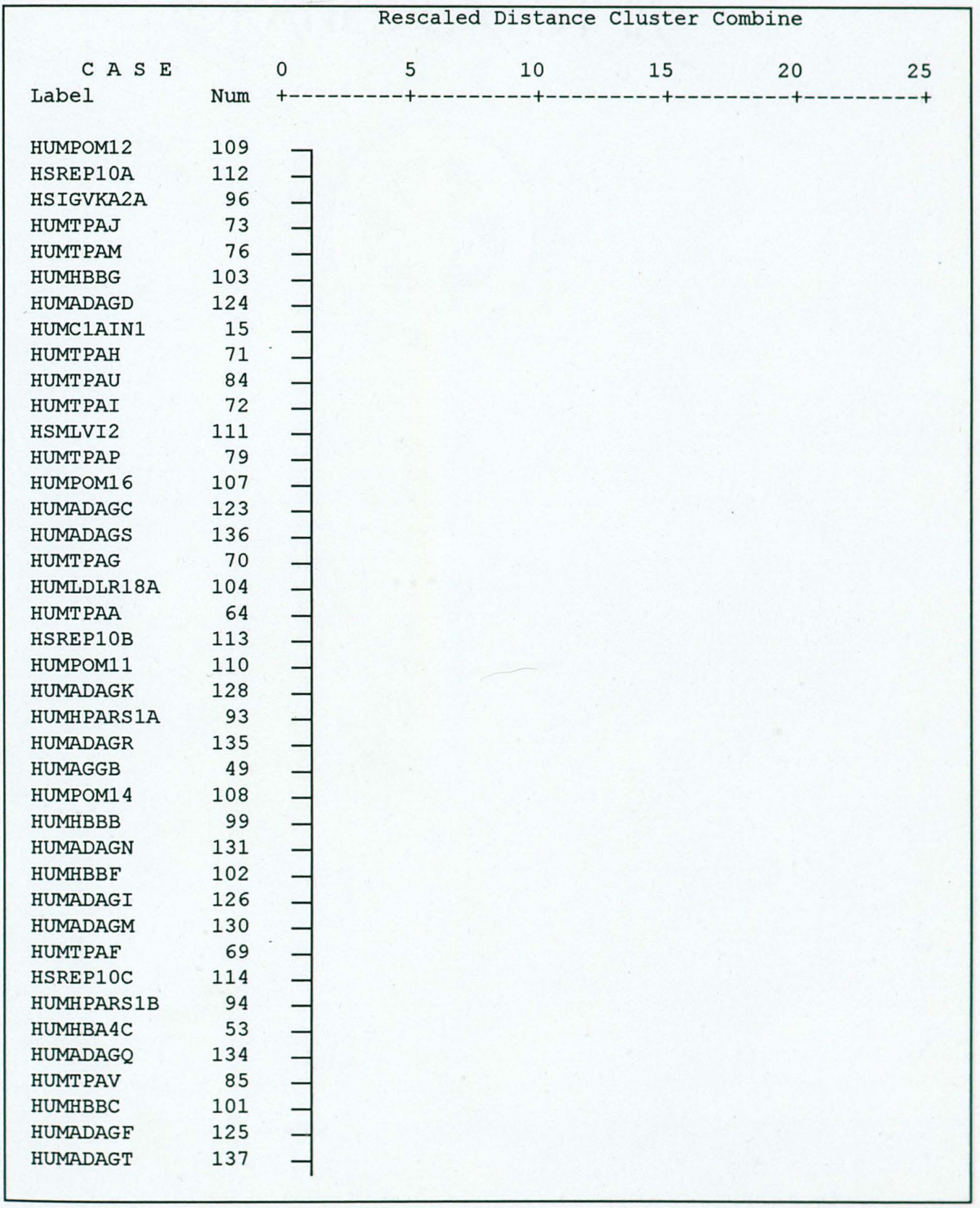


Figure 4. Hierarchical Cluster Analysis of Object Scores

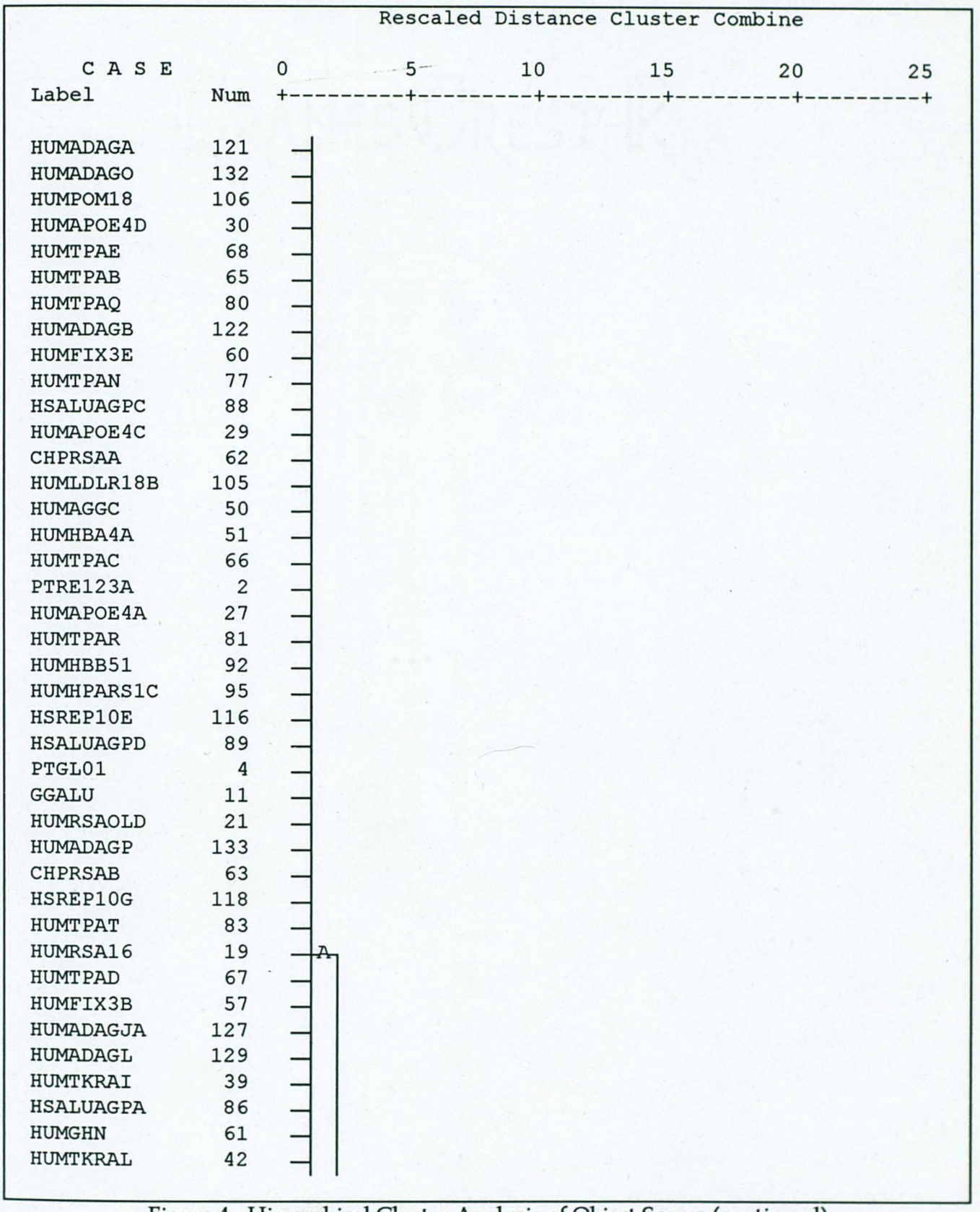


Figure 4. Hierarchical Cluster Analysis of Object Scores (continued)

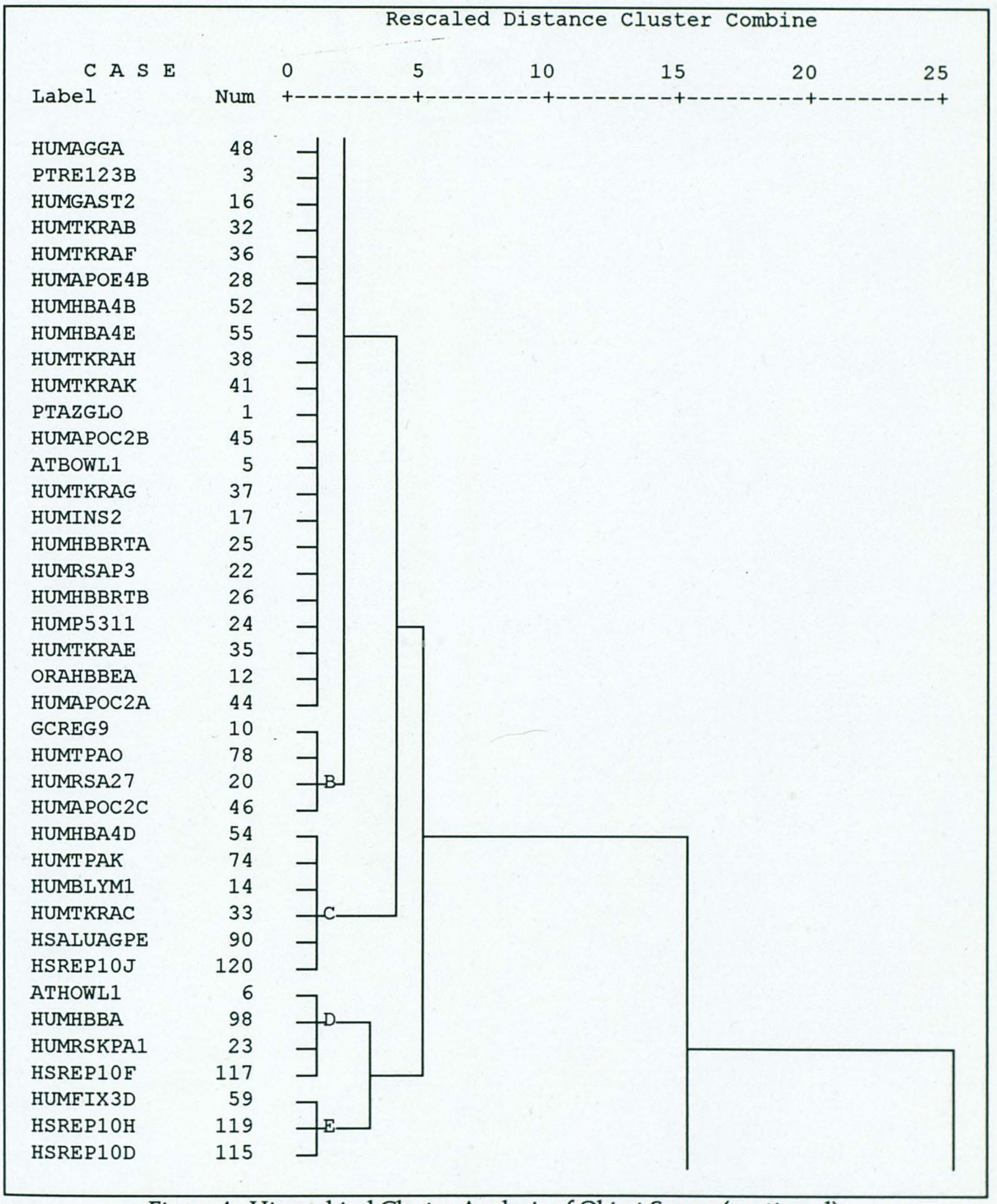


Figure 4. Hierarchical Cluster Analysis of Object Scores (continued)

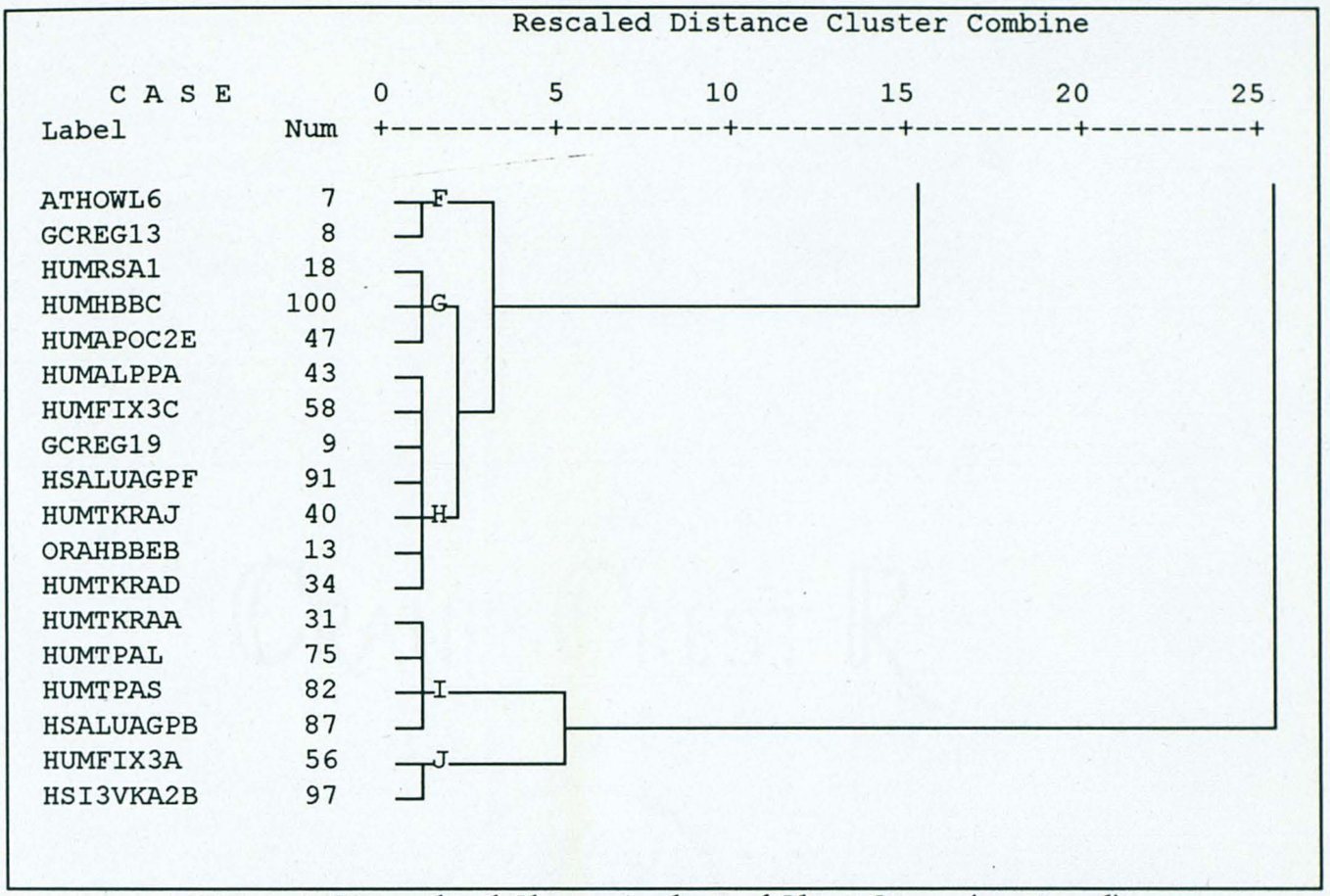


Figure 4. Hierarchical Cluster Analysis of Object Scores (continued)

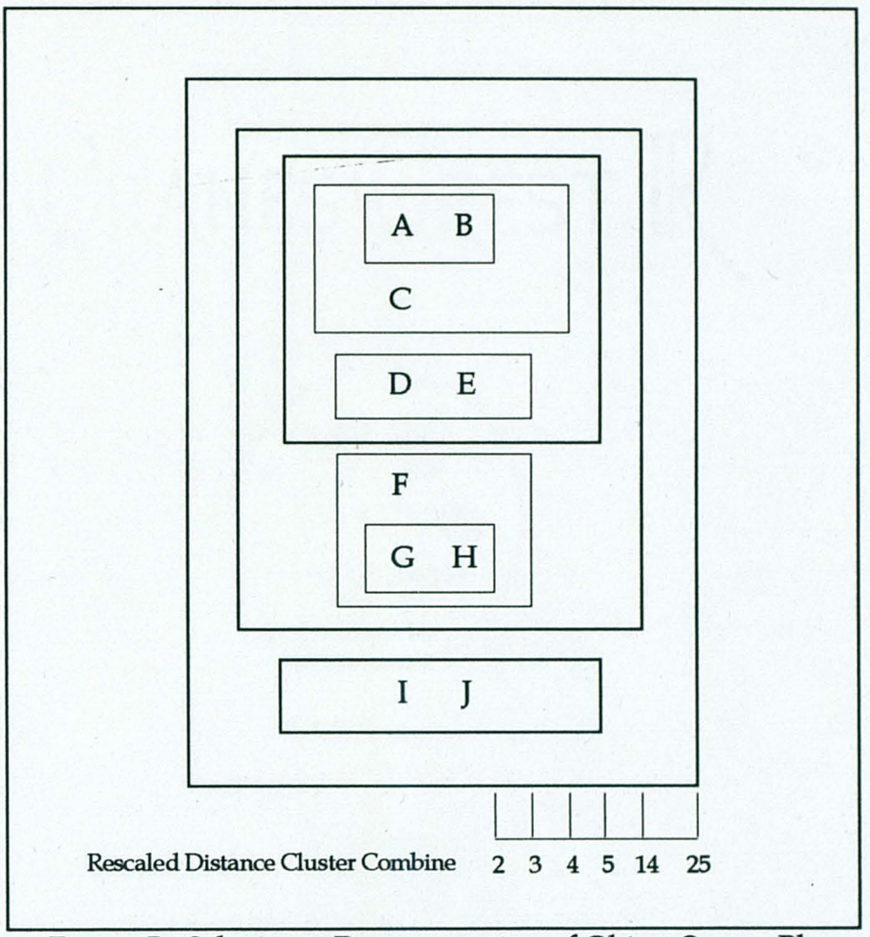


Figure 5. Schematic Representation of Object Scores Plot

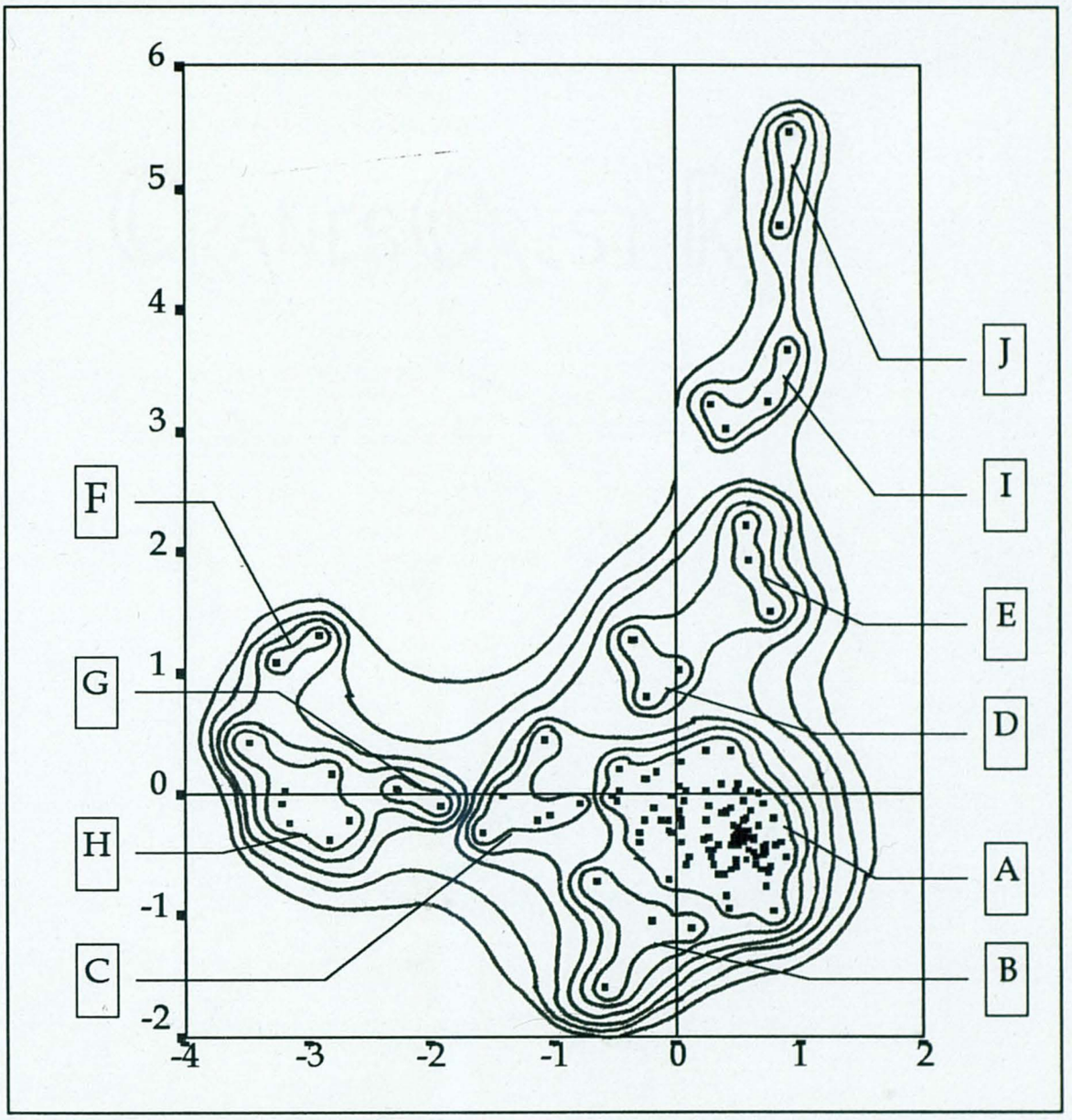


Figure 6. Plot of Object Scores with Hierarchical Cluster Analysis Delimitations

Non-Human Primate *Alu* Sequences

Figure 7 identifies all non-human species whose *Alu* sequences were included in the final analysis and their location on the object scores plot. All points that are not labeled are human *Alu* sequences. These points' lack of coherent pattern supports the theory of amplification/fixation. The Slagel, *et al.* (1987) theory of 'fitter' progenitors would be more appropriate if there had been species-specific clusters discernible in this plot. The lack of such

clusters indicates that the formation of these elements occurred prior to the formation of these taxa.

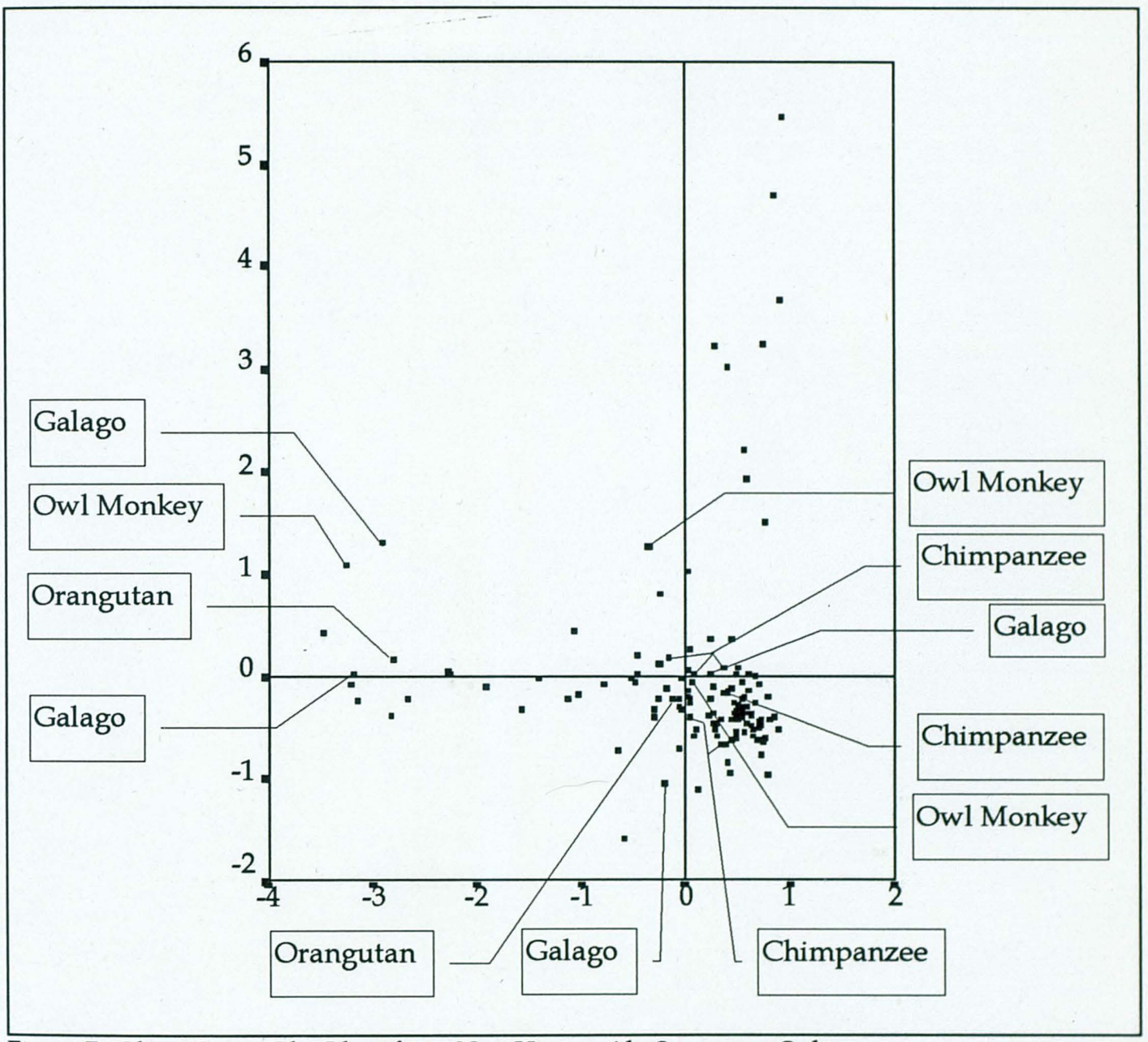


Figure 7. Object Scores Plot Identifying Non-Human Alu Sequences Only

Limitations of this Type of Study

The limitations of this study can be grouped into four categories: the subjectivity of the multiple sequence alignment prior to analysis, the selection of informative data points within the

sequences (prior to analysis), the sensitivity of the correspondence analysis to the alignment, and the subjectivity inherent in subfamily identification.

The CLUSTAL V algorithm (Higgins, *et al.*, 1992) used to align the data set is not guaranteed to produce an optimal alignment. An algorithm which could produce such an optimal alignment (by comparing all sequence pairs) for such a large data set is, currently, computationally intractable. Thus the sub-optimal alignment provided by CLUSTAL V must be adjusted manually. The manual insertion of additional gaps to better align a section of nucleotides involves a good deal of subjectivity, and alternative adjustment strategies may have resulted in outcomes of different magnitudes by changing the informative positions selected for analysis.

In addition, reduction of the data set to those positions which contribute most to the differences between sequences, while facilitating computational efficiency and promoting clarity in analysis, results in a loss of potentially relevant data. Because of the sensitivity of the alignment process to the sequences included, and the sensitivity of the correspondence analysis to the alignment, exclusion of such data may have biased the object scores and, therefore, the resulting correspondence plots and cluster analysis.

The method used to identify subfamilies from the object score plots was also highly subjective. The agglomerative hierarchical cluster analysis used to facilitate a more objective subfamily identification, while widely accepted and more conservative, still involves a great deal of subjectivity. Cluster analysis produces dendrograms which, when translated to the object scores plot can be used to produce a contour plot. Subfamilies could be chosen from almost any level of this contour plot; the three-cluster level was chosen based on its visual distinctiveness.

The above four points illustrate the enormous amount of flexibility in this study, and others like it. A trial run of these data included a large section of the terminal A-rich region. This inclusion affected not only the shape of the object scores plot, but also the object score-dependent cluster analysis which followed. In this particular example, clustering was far more distinct, and a five cluster solution was the most appropriate. Any of the factors of informative position selection criteria, selection of appropriate clustering methodology, and alignment judgment calls could easily have similar results.

Implications and Direction for Future Research

This study provides additional support for the Willard, *et al.* (1987) amplification/fixation theory of *Alu* subfamily formation by showing apparently independently arising subgroups with varying degrees of similarity; this is in direct opposition with Bains (1986) findings of equally divergent subfamilies. It is important to recognize, however, that the results of this study are incomplete. A superficial examination of the data and analysis could readily yield a three subfamily solution in this situation; however, the subjectivity of this analysis, compounded at nearly every step of the process, precludes almost any solution that possess internal validity.

Accepting the limitations of this analysis, a logical extension of this study would be the inclusion, alignment and analysis of more recently available data. The addition of newer data, especially from non-human primates, could possibly have enlightening effects on the underlying subfamily structure in human *Alus*.

APPENDICES

APPENDIX A

Global Alignment of Included Sequences

```
Sequence          10        20        30        40        50        60        70
PTAZGLO    ----------------GGCCGGGC-ACGGT-GGT----TCAC-GCTCATAATCCCAGCA-CTTTCGG-------AGG
PTRE123A   ----------------CCGACAGAC-ACGGT-GGC----TCAC-ACCTGTCATCCCAGCA-CTTTGGG-------AGG
PTRE123B   AGAAAAATAAATGCATAAGGCGGGGC-GCGGT-GGC----TCAC-GC-TGTAATCCCAGCA-CATTGGG-------AGG
PTGL01     ----------GTTGCGGGGGCTGGGT-GTGGT-GGC----TCAC-GCCTGTCATCCCAGCA-CTTTGGG-------AGA
ATBOWL1    ----------------AGGGCAGGC-GTGGT--GC----TCAT-GCCTGTAATCCCAGCA-CTTTGGA-------AGG
ATHOWL1    ----------------AGTCTGGGA--GCAT-GGC----TCAC-ATCTGTAATCCCAGCA-CTTTTGG-------AGG
ATHOWL6    ----------------------------------------------------------------------------
GCREG13    ----------------------GTA--CCAT-GGC----TCAC-TCATGTAATCCTCGCA-CTCTGAG-------AGG
GCREG19    ----------------------GTA---CAT--------TTAT--------------------CTGGG-------AGG
GCREG9     ------------TACTGTGCCAGGCCGGGGC-AGC----TGGCACCTTGTAATCCCAGCA-CTCCCGG-------AG-
GGALU      ----------------GGCCGGGC-GCGGT-GGC----TCAA-GCCTGTAATCCCAGCA-CTTTGGG-------AGG
ORAHBBEa   ----------------GGGCCAGGT-GTGGT-GGC----TCAC-ACCTATAATCCCAGCA-CTTCAGG-------AGG
ORAHBBEb   ----------------------------------------------AG--------GG-------AGG
humblym1   ----------------GCTGG---GTGGT-GGC----TCACG--CTGTAATCCTAGCA-CTGTGGG-------AGG
humc1ain1  ----------------GGCGGGGC-ACGGT-GGC----TCAT-GCCTGTAATCCCAGCA-CTTTGGG-------AGG
humgast2   ----------------CCGGGT-GCGGT-GGC----TCAT-GCCTGTAATCCCAGCA--TTTGGG-------AGG
humins2    ----------------GGCTGGGT-GCGGT-GGC----TCAT-GCCTATAATCCCAGCA-CTTTAGG-------AGG
humrsa1    ------------------------------T-GGC----TCAG-GCCTGTAATCCCAGCAAATTT-GG-------AGG
humrsa16   ----------------GGCTGGGT-GCAGT-GGC----TCAC-ACCTGTAATCTCAGCA-CTTTGGG-------AGG
humrsa27   ----------------GCTGGGT-ACGGT-GGC----TCAT-TCCTGTAATCCCAGCA-CTTTGGG-------AGC
humrsaold  ----------------GGCCAGAC-ATGGT-GGC----TCAT-GCCTGTAATCCCAGCA-CTTTGGG-------AGG
humrsap3   ----------------GGCTA----GGCGC-GGG----TTCACGCCTGTAATCCCAGCA-TTTTGGG-------AGG
humrskpa1  ----------------GGCT--CA--TGGT-GGC----TCAT-GCCTATAATCCCAACA-CTTTGGA--------GG
hump5311   ----------------GCTGGAC-GTGGT-GGC----TCAC-AATTGTAATCCCAGCA-CTCTGGG-------AGG
humhbbrta  ----------------GGCTGGAT-GCGGT-GGC----TCAG-GCTTGTAAACCCAGCAACTTTGGG-------AGG
humhbbrtb  ----------------GGCTGGGT-GTGGT-GGC----TCAC-GCCTGTGATCCCAGCA-CTTTCAG-------AGG
humapoe4a  ----------------GCTGGGC--CGGT-GGC----TCAC-CCCTGTAATCCCAGCA-CTTTGGG-------AGG
humapoe4b  ----------------GGCCGGGC-ATGGT-GGC----TCAG-GCCTGTAATCTCAGCA-CTTTGGG-------AGG
humapoe4c  ----------------GGCTGGGG-GCGGT-AGC----TCAC-GCCTGTAATCCCAGCA-CTTTGGG-------AGG
humapoe4d  ----------------GCCAGGT-GCGGT-GGC----TCAC-GCCTGTAATCCCAGCA-CTTTGGG-------AGA
HUMTKRAa   ----------------GCAGGGC-ATGGT-GGC----TAAC-ATCTGTAATCCCAGCA-CTTTGGG-------ATG
HUMTKRAb   ----------------GCCAGGT-GTGGT-GGC----CCAC-GCCTTTAATCCCAGCG-CTTTGGA-------AGG
HUMTKRAc   ----------------GGCCAGA-CACAG-CAG---CTCAT-GCCCGTAATC-------TTTGGGG-------AGG
HUMTKRAd   ----------------------------------------------------------CTCTGGG-------AGG
HUMTKRAe   ----------------GGCCAGGT-GTGGT-AGC----TCAT-GCCTGTAATCCCAGCG-CTTTGGG-------AGT
HUMTKRAf   ----------------GGCAAGAT-GCAGTGACT----CACA--CCTGTAATCTCAGC--CTGTGGG-------AGG
HUMTKRAg   ----------------GGCCGGGC-ACGGC-GGC----TCAT-GCCTGTAATTCCAACA-CTTTGGG-------AAG
HUMTKRAh   ----------------GACCAGGC-GTGGT--------TCAC-GCCTGTAATCCCAGCA-CTTTGGG-------A-G
HUMTKRAi   ----------------GGCCAGGC-GTGGT-GGC----TCAC-GC-TGTAATCCCAGCA-CTTTGGG-------AGG
HUMTKRAj   ----------------------------------------------------------------------------
HUMTKRAk   ----------------GGCTGGGC-A-GGT-GGC----TCAC-ACCTGTAATCCCAGCA-CTTTGGA-------AGG
HUMTKRAl   ----------------AGCCAGGC-GCAGT-GGC----TCAT-GCCTGTAATCCCAACA-CTTTGGG-------AGG
HUMALPPA   ----------------------------------------GTAATCCCAGCA-CTTTGGG-------AGG
HUMCEA     ----------------------------------------------------------------------------
HUMAPOC2A  -------GAAAGAAAAAGGGATAGGT-ACAAT-GGC----TCAT-GCCTGTAATCCCAGCA-CTCTGGG-------AGG
HUMAPOC2B  ------------CTGTAGGCTGGGG-CTGGT-GGC----TTAC-ACTTGTAATCCAAACG-CTTTGGG-------A-G
HUMAPOC2C  ----------------GGGCATG-GTGGC-TCA----CGCC----TGTAATTCCAGCA-TTTTTGG-------AGG
HUMAPOC2E  ---------------TC-ACTTGAG-----GT----------------------CAAGAA--GTTTGGA--------GG
HUMAGGa    ----------AAATGAC-GGCCGGGC-GCGGT---C----TCAC-GCCTGTAATCCCAGCA-CTTTGGG-------AGG
HUMAGGb    ------AAGGAAAAAGT-GGCCCGGC-TCAGT-GGC----TCAT-GCCTGTAATCCCAGCA-CAAAGAG-------AGG
HUMAGGc    -------AGAAGATTCCGGCCAGGC-GC--T-TAT----CTCAC-GCTTGTAATCC-AGCACTTTGGG--------AAG
HUMHBA4a   ----------------GCTGGGT-GTGGT-GGC----TCAC-GCCTGTCATCCCAGCA-CTTTGGG-------AGA
HUMHBA4b   ----------------------GGC-ACGGT-GGC----TCAC-ACCTGTAATCCCAGTA-CTTTGGG-------AGG
HUMHBA4c   ----------------GGCCAGGC-TCGGC-G-------CAC-GCCTGGAATCCCAGCA-CTTTGGG-------AGG
HUMHBA4d   ------------------------T--T-GGC----TCAC-GC-TGTAATCCCAGCA-CATTGGA-------AGG
HUMHBA4e   ----------------------GGC-ACGGT-GGC----TCAC-GCCTGTAATCCCAGCA-CTTTGGG-------AGG
HUMFIXGa   ----------------GGCCTGGC-ATGGT-GGC----TCAC-ACCTATAATCCCAGCA-CTTTCAG-------AGG
HUMFIXGb   ----------------GGCCAGGTGCAGT-GGC----TCAT-GCCAATAATCCCAGCA-CTTTGGG-------AGA
HUMFIXGc   ----------------------------------------------------------------------------
HUMFIXGd   -------GAGGGAAAT--GGCCGGGT-GCAGT-CG-----TCAC-GCCTGTAATCCCAGCA-CTTTGGG-------AGG
HUMFIXGe   ----------------GCTGGGC-CCAGT-GGC----TCAC-GCCTATAATCCCAGCA-CTTCTGG-------AGG
HUMGHN     ----------------GAAGGGAG--CAGT-GGT----TCAC-GCCTGTAATCCCAGCAA-TTTGGG-------AGG
CHPRSAa    ----------------CTGGGT-GTGGT-GGC----TCAC-GCCTGTGATCCCAGCA-CTTTCAG-------AGG
CHPRSAb    ----------------GCTGGAT-GCGGT-GGC----TCAG-GCTTGTAAACCCAGCA-CTTTGGG-------AGG
HUMTPAa    ----------------GCTGGGC-GCGGT-GGC----TCAC-ACCTATAATCCCAGCA-CTTTGGG-------AGG
HUMTPAb    ----------------CCAGGC-ATGTT-GGC----TCAT-GCCTGTAATCCCAGCA-CTTCGGG-------GGG
HUMTPAc    ----------------CTGGGC-ATGGT-GGC----TCAC-GCCTGTAATCCCAATA-CTTTGCG-------AGG
HUMTPAd    ----------------CTGGGC-ACAGT-GGC----TCTT-GCCTGTAATCCCAGCA-TTTTGTG-------AGG
HUMTPAe    ----------------CCGGGC-ACACA--GC----TCCT-GCCTGTAATCCCAGCA-CTTTGGG-------AGC
HUMTPAf    ----------------GCCAGGC-GTGTT-GGC----TCAC-GCCTGTAATCCTAGCA-CTTTGGG-------AGG
HUMTPAg    ----------------GCAAGGT-GTGGT-GGC----TCAC-ATCTGTAATCCCAACA-CTTTGGG-------AGG
HUMTPAh    ----------------GCCAGGC-ACAGT-AGT----TCAC-ACCTGTAATCCCAGCA-CTTTGGG-------AGG
HUMTPAi    ----------------GCTGGGC-GCGGT-GGC----TCAC-ACCT-TGATCCCAGCA-CTTTTGG------GAGG
HUMTPAj    ----------------GCGGGGC-GTGGT-GGC----TCAC-GCCTGTAATCCCACCA-CTTTGGG-------AGG
HUMTPAk    ------------------------GT-GAC----TCAC-ACCTGTAATCTCAGCA-CTTTGGG-------AGG
HUMTPAl    ----------------CCAGGT-GCAAT-GGC----TCAC-ATCTGTAACTCCAGCA-CTTTAGG-------AGG
HUMTPAm    ----------------GCTGGGC-ACGGT-GGC----TCAT-GCCTGTAACCCCAGCA-CTTTGGG-------AGA
```

```
Sequence          10        20        30        40        50        60        70
HUMTPAn   --------------------CCAAAC-ACGGT-GAC----TGAT-GCTGGTAATCCCAGCA-CTTTGGG-------TGG
HUMTPAo   ------------------AGGCTGG--AAGGT-GGT----GCAT-GCCTGTGATCCCAGCA-CTTTAGG-------AGG
HUMTPAp   --------------------GCTGGGC-GTGGT-GGC----TCAC-GC--GTAATCCCAGCA-CTTTGGG-------AAG
HUMTPAq   --------------------CCAGGC-ACAGT-GGC----TTAT-GCCTATAAACCCAGCA-CTCTGGG-------AGG
HUMTPAr   ------------------GGCCAGGT-GCAGT-GGC----TCGT-GTCTATAATCCCAGCA-CTTTGGG-------AGG
HUMTPAs   ------------------GCCGGGC-GTGTT-GGC----TCAC-ACCTGTACTCCCAGCA-CTTTGGG-------AGG
HUMTPAt   ------------------GCCGGGC-GCGGT-GGC----TCAC-GCCTGTAATCCCAGCA-CTTTGGG-------AGG
HUMTPAu   ------------------AGCCTGGT-GTGGT-GGT----TCAC-GCCTGTAATCCCAGCA-CTTTGGG-------AGG
HUMTPAv   ------------------GGTCAGGT-GCGG--GGC----TCAT-GCCTGTAATCCCAGCA-CTTTGAG-------AGG
HSALUAGPa ----------------GGGTGGGGC-ACCAC--GC----TCAC-GCCTGTAATCCCAGGA-CTTTGGG-------AGG
HSALUAGPb ----------------GGCTGGGA--GTGGT--GG----TCAC-GCCTATAATCC-AGCA-CTTTGGG-------AGG
HSALUAGPc ------------------CGGGCC--AGGCG-CCGTGGCTCAT-GCCTGTAATCCCAGTA-CTCTGGG-------AGG
HSALUAGPd ----------------GGCCAGGC--GCGGT-GGC----TCAC-GCCTGTAATCCCAGCA-CTTTGGG-------AGG
HSALUAGPe ------------------------------------------------GTAATCCCAGCA-CTTTGGG-------AGG
HSALUAGPf ------------------------------------------------------------CTTAATG-------GAC
HUMHBB51  ----------------GCCGGGC--GCAGT-GAC----TCAC-ACCTGCAATCCCAGCA-CTTTGGG-------AGG
HUMHPARS1a --------AGTTTCTAGGCCAGAC--ACGGT-GGC----TCAT-GCCTGTAATCCCAGCA-CTTTGGG-------AGG
HUMHPARS1b ----------------GGTCGGGC--GTGAT-GGC----TCAC-ACCTGTAATCTCAACA-CTTTGGG-------AGG
HUMHPARS1c --------AGTTTCTAGGCAGGGT--GCGGT-GGC----TCAC-ACCTGTAATCCCAACA-CTTTGGG-------AGG
HSIGVKA2a ---------------CT--GCTGGGT--GTGGT-GGC----TCAT-GCCTGTAATCCCAGCA-CT---------------
HSIGVKA2b ----------------GGCCGGGC--GCAGT--GC----TCAA-GCCTGTAATCCCAGCA-CTTTGGG-------AGG
HUMHBBa   ------------------GGT--GCGGT-GGC----TCAC-TCCTGTAATCACAGCA-GTTTGGG-------ACG
HUMHBBb   ----------------GGGCCAGGC--GTGGT-GGC----TCAC-ACCTATAATCCCAGCA-CTTCAGG-------AGG
HUMHBBc   ------------------------------------------------------ATCC----------GG-------AGG
HUMHBBd   ----------------AGGCTGGGA--GTGGT-GGC----TCAC-GCCTGTAATCCCAGAA-TTTTGGG-------AGG
HUMHBBf   ----------------GCTGGGT--GTGGT-GGC----TCAC-GCCTGTGATCCCAGTA-CATTCAG-------AGG
HUMHBBg   -------------CCTG-GCTGGAT--GCGGT-GGC----TCAG-GCTTGTAAACCCAGCA-CTTTGGG-------AGG
HUMLDLR18a ------------------CCGGGC--GCAGT-GGC---CTCAC-GCCTGTAATCCCAGCA-CTTTGGG-------AGG
HUMLDLR18b ----------------CAGACA--CAGGT-GCC----TCAC-GCCTGTAATCCCAGCA-CTTTGGG-------AGG
HUMPOMC8  ----------------GGGCCGGGC--ATGGT-GGC----TCAC-ACCTATAATCCCAGCA-CTTTGGG-------AGG
HUMPOMC6  ----------------AGACCAGGC--AGGAT-GGC----TCAT-GGCTGTAATCCCAGCA-CTTTTAG-------ATG
HUMPOMC4  ----------------GGCCGGGG--GCGGT-GGC----TCGC-GCCTGTAATGCCAGAA-CTTTGGG-------GGG
HUMPOMC2  ----------------GGCTGGGT--GTGGT-GGC----TAAC-ACCTGTAATCCCAGCA-CTTCAGG-------ATA
HUMPOMC1  ----------------GGCGGGGC--ACAGT-GGC----TCAC-GCCTGTAATCCCAGCA-CTCTGGG-------AGG
HSMLVI2   ----------------TGCCGGGC--GCGGT-GGC----TCAC-GCCTGTAATCCCAGCA-CTTTGGG-------AGG
HSREP10a  ----------------GGCCGGGC--AAGGT-GGC----TCAT-GCTTGTAATCC-AGCA-CTTTGGG-------A-G
HSREP10b  ----------------TCACTGGGC--ATGGT-GGT----TCAT-GCCTATAAACCCAGGA-CTTTGGG-------AGG
HSREP10c  ----------------GGGTGGGC--ACGGT-GGC----TCAC-ACCTGTAATCCCAGCA-CTTTGGG-------AGG
HSREP10d  ----------------CAGACCAGGC--GCGTT--GC----TCAT-GCCTGTAATCCCAGCA-CTTTGGG-------AAG
HSREP10e  ----------------GGCCAGGC--GTGGT-GGC----TCAC-GCTTGTAAACCCAGCA-CTTTGGG-------AGG
HSREP10f  ----------------GGCCAGGC--GCAGT-GGC----TCAT-GCCTGTAATCC-AGCA-CTTAGGG-------AGG
HSREP10g  ----------------GGCCAAGC--ATGGC---T----GCAC-GCCTGTAATCCCAGCA-CTTTAGG-------A-G
HSREP10h  ----------------GACTGGGC--ACAGT--GC----TCAT-GCCTGTAATCCTAACA-CTTTGGG-------AGG
HSREP10j  ----------------------GT--GCGGT-GGC----TCAC-ACCTGTAATCCCAGCA-CTTTGGG-------AGG
HUMADAGa  ----------------GCCTGGGC--GCAGT-GGT----TCAT-GCCTGTAATCCCAGGA-CTTTGGG-------AGG
HUMADAGb  ----------------GGCCAGGC--GTGGT-GGC----TCAT-GCCTGTAATCAC---G-CCTTGGG-------AAG
HUMADAGc  ----------------GCCCAGG---AGGTT-GGC----TCAC-GCCTGTAATCCCAGCA-CTTTGTG-------AGG
HUMADAGd  ----------------GGCGGGGT--GCAGT-GGC----TCAC-GCCTGTAATCCCAGCA-CTTTGGG-------AGG
HUMADAGf  ----------------GGCCGGGT--GCGGT-GGC----TCAT-GCCTGTAATCCCGGCA-CTTTGGG-------AGG
HUMADAGi  ----------------GGCAGGGT--AAGGC-AGC----TCAC-GCCTGTAATCATAACA-CTTTGGG-------AGG
HUMADAGja ----------------GGTCGGGC--GTGGT-GGT----TCAT-GCCTGTAATCCCAACA-CTTTGGG-------AGG
HUMADAGk  ----------------GGCTGGGC--ACGAT-GGT----TCAC-ACCTGTAACCCCAGCA-CTTTGGG-------AGG
HUMADAGl  ----------------GGCTGGGC--ATGGG-GGC----TCAC-CCCTGTAATCCAGCCA-CTTTGGG-------AGG
HUMADAGm  ----------------GGTCGGTT--GCGGT-GGC----ACAT-ACCTGTAATCCCAGCA-CTTTGGG-------AGG
HUMADAGn  ----------------GAGCCAGGC--ACAGT-GGC----TCAT-GCCTGTAATCCCAGCA-CTGTGGG-------AGG
HUMADAGo  ----------------AGAGCGTGC--ATGGT-G-C----TCAC-GCCTGTAATCCCAGCA-CTTTGGG-------AGG
HUMADAGp  ----------------GGCCGGGT--GCAGT-GGC----TCAT-ACCTGTAATCACAGTA-CTTTGGG-------AAG
HUMADAGq  ----------------GGCTGGGC--ATGGT-GGC----TCAC-GCCTGTAATCCCAGCA-CTTTGGC--CTCCCAG-
HUMADAGr  ----------------GGCCAGGC--GCAGT-GGC----TCAC-GCCTGTAATCCCAGCA-TTTTGGG-------AGG
HUMADAGs  ----------------GGCCAGGT--ACAGT-GGC----TCAC-GCCTGTATCCCCAGCA-CTTTGGG-------AGG
HUMADAGt  ----------------GGCCAGGC--GTGGT-GGT----TCAT-GCCTGTGATCCCAGCA-CTTTGGG-------AGG
```

```
Sequence  80        90        100       110       120       130       140       150
PTAZGLO     CCGAGGCAGGCAGATCACC---TGAGGT-C-AAGA--GTTCGACA-CCAGCC-TGGCCAAC--------ATGGTGAAAT
PTRE123A    CTGAGGCGGGAGGATCACC---TGAGGT-C-GGGA--GTTCGAGA-CCAGCC-TGACCAAT--------ATGGAGAAAC
PTRE123B    CTGAGGTGGGCAGATTCGC----GAGGT-C-AGGA--GATCGAGA-CCATCC-TCGCTAAT--------GCAGTGAAAC
PTGL01      CCGGGGA-GGGCAGATCACT--TGAGGT-C-AGGA--GTTTGAGA-CCAGCC-TGATCAAC--------ATGGTGAAAC
ATBOWL1     CCAAG-T-AGGCAGATCACC--CTAGAT-C-AGGA--GTTCAAGA-CCAGC--TGGCCAAC--------ATGGTGAAAC
ATHOWL1     CTAAGGT-GGGTGGACC-----CGGGGC-C-GGAA--ATGCAAGC-CCAGCC-TTGCCAAC--------ATGGTAAAAC
ATHOWL6     ---AGGT-GGCAGACCAC-----AAGG----AGGA--GTTTGAGA-CCAGCC-TGACCAAC--------ATGGTGAAAC
GCREG13     CTGAGGA-GGATGGATTGCT--TGAGCT-C-ACGA--GTTTGAGA-CCAGCCT-GAGCAAG--------AGTGAGACCC
GCREG19     CCAAGGC-AGGTAGACTGCT--TGAGCT-C-AAGA--GTTTGAGA-CAAGCT-TAAACAAG--------A----GCAAG
GCREG9      CCAAGGC-AGGTGGATTGCT--AGAGCC-C-AGGA--GTTTGAGA-CCAGCC-TGAGCAAG--------A----GCCAG
GGALU       CCGAGGC-GGGCGGATCAC----GAGGT-C-AGGA--GATCGAGA-CCATCC-TGGCTAAC--------ACAGTGAAAC
ORAHBBEa    CCAAGGC-GGGCAGATCATC--TGAGGT-C-AGGA--GTTCAAGA-CCAGCC-TGGCCAAC--------ATGGCAAAAC
ORAHBBEb    CAGAGGT-GGGCAGATCAT----GACGT-C-AAGA--GATCGAGA-CCATCC-TGGCAAAC--------ATGGTGAAAT
humblym1    CTGAGGC-AGGA-GGATTGC--TTGAAC-C-AAGG--TATTCAAG-AAGAGCATGGGCAAC--------ATGATGAGAC
humc1ain1   CCAAGGC-GGGCGGATCACCA-TGAGGT-C-AGGA--GTTTGAGA-CCAGTC-GGGCCAAC--------ATAGTGAAAC
humgast2    CCGAGGC-GGGTGGATCACC--TGAGGT-C-AGGA--GTTCGAGA-CCAGCC-TGGCCAAC--------ATGGTGAAAC
humins2     CTGAGGC-GGGCAGATCACC--TGAGGT-C-GGGA--GTTCAAGA-CCAGCC-TGACCAAC--------AGGGAGAAAC
humrsa1     CCAAGGC-AAGGGGATCAC----AAGGT-G-AAGA--GATCAAGA-CCATCC-TGGCCAAT--------ACAGTGAAAC
humrsa16    CTGAGGC-AGGAGGATTAC----GAGGT-C-AGGA--GATTGAGA-CCATCC-TGGCTAAC--------ACAGTGAAAC
humrsa27    C-GAGGC-AGGTGX-TCACT--TGAGGT-C-AGGA--GTTCGAGA-CCAGCC-TGACCAAC--------A-GGTGAAAC
humrsaold   CCGAAGC-AGGAGGATCAT--TTGAGCC-T-GGGA--GTTTGAGA-CCAGCC-TGGGCAAC--------ATAGCAGACC
humrsap3    CTGAGAC-GGGTGGAATCAT---GAGGT-C-AGGA--GATCGAGA-CCATCC-TGGCTAAC--------ATGGTGAAAC
humrskpa1   CTGAGGC-AGGAGGATCACT--TGAGCC-G-AGGA--GTTCAAGA-CCAGCC-TGGGCA-C--------ATAATGAGAT
hump5311    CTGAGAC-AGGTGGATCGCC--TGAGCC-C-AGGA--GTTTGAGA-CCAGCC-TGGGCAAC--------ACTGTGAGAC
humhbbrta   CCAAGGC-AGGCAGATCACT--TGAGGT-C-AGGA--GTTCAAGA-CCAGCC-TGACCAAC--------ATGGTGAAAC
humhbbrtb   CCGAGGA-GGGTGGATCACC--TGATGT-T-AGGA--GTTCGAGA-TCAGCC-TGACCAAC--------ATGGTGAAAC
humapoe4a   CCAAGGT-GGGAGGATCACT--TGAGCC-C-AGGA--GTTCAACA-CCAGCC-TGGGCAAC--------ATAGTGAGAC
humapoe4b   CC--GGC-GGGTGGATCACT--TG--GT-C-AGGA--GTTTGAGA-CCTGCC-TGGCCAAC--------ATGGTGAAAG
humapoe4c   CCACAGT-GGGCGAATCACT--TAAGGT-C-AGGA--GTTTGAGA-CCAGCC-TGGCCAAC--------ATGGTGAAAC
humapoe4d   TCGAAAC-GGGCAGATCACC--TGAGGT-C-AGGA--GTTCCAGA-CCAGCC-TGGCCAAC--------ATGGTGAAAC
HUMTKRAa    CTGAAGT-AGGAGGATTGCT--TGAGAC-C-AGGA--GTTCAAGA-CCAGCT-TGGGCAAC--------ATAGCAAGAC
HUMTKRAb    CTGAGGT-GCGTTGATCACT--TGAGGT-T-AGGA--GTTTGAGA-CCACCC-TGGTCAAC--------GTGGTGAAAC
HUMTKRAc    CCAAGAC-GGGAGGATCACT--TGAACT-C-AGGA--GTTCGAGA-CCAGCC-TGGGTAAC--------ACAGCGAGAC
HUMTKRAd    CTAAGGC-GGGTGGATCACT-TAAAAGT-T-AGGA--GTCTGAGA-CCAGCC-TGGCCAAC--------ATGGTGAAAC
HUMTKRAe    TCAAGGCGGGGCGGATCACC--TGAGGT-T-GGGA--GTTTGAGA-CCAGCT-TGACCAAC--------ATGGAGAAAC
HUMTKRAf    CCAAGGT-GGACAGATCACT--TGAGCC-C-AGGA--GTTGGAGA-CTCACC-TGGGCAAC--------ATAGTAAAAC
HUMTKRAg    CCAAGGC-GGGTGGATCACC--TGAGGT-C-AGGA--GTTCAAGA-CCAGCC-TAGCCAAC--------ATGGTGAACC
HUMTKRAh    CTGAGGC-ATGCGGATCACC--TGAGGT-C-AGGA--GTTCCAGA-CCAGC--TGGCCAAC--------ATGGTGAAAC
HUMTKRAi    CTGAGGC-AGGCAGATCACC--TGAGGT-T-AGGA--GTTCCAGA-CCAGCC-CGGTCAAC--------ATGATGAAAC
HUMTKRAj    ----------------------------------------A-CCAGCC-TGACCAAC--------ATGGAGAAAC
HUMTKRAk    CCTAGGC-GGGCGGATCACT--TGAGGT-C-AGGA--GTTTGAGA-CCAGCC-AGGCCAAC--------ATGGTGAAAC
HUMTKRAl    CCGAGGT-GGGTGAATCAC----GAGGT-C-AGGA--GATCGAGA-CCATCC-TGGCTAAC--------ACGGTGAAAC
HUMALPPA    CCGAGGT-GGGCGGATCAC----GAGGT-C-AGGA--GATGGAGA-CCATCC-TGGCTAAC--------ACGGTGAAAC
HUMCEA      ----------AAGACTCT-------GAC-C-AGA---GATCGAGA-CCATCC-TAGCCAAC--------ATCGTGAAAC
HUMAPOC2A   CCGAGGC-GGGTGGATCACT--TGAGGT-C-AGGA--GTTCGACA-CCAGCC-TTACCAGC--------ATGGTGAAAC
HUMAPOC2B   CCAAGGC-AAACGGATCACT--TGTGAT-C-AGGA--GTTGGAGA-CCAGCC-TGGCCAAC--------ATGGTGAAAC
HUMAPOC2C   CCAAGGC-AGGCAAATCACT--TGAGGT-T-AGGA--GTTCAAGA-CCAGCC-TGGCCAAC--------ATGGTGAAAC
HUMAPOC2E   CTAAGGC-GG-TGGATCACT--TGAGGGTC-AAGA--AGTTTGAGACC-AGC-CTACCAAC----------TGGCAAAA
HUMAGGa     CCAAGGC-GGGCGGATCAC----GAGGT-C-AGGA--GATCGAGA-CCATCC-TGGCTACC--------ACGGTGAAAC
HUMAGGb     CCAAGGT-GGGCGGATCTCC--CGAGGT-C-AGGA--GTTCAAGA-CCAACC-TGGCCAAA--------ACGGTGAAAC
HUMAGGc     CTGAGGC-GGACAGATCAC----GAGGT-C-AAGA--GATCAAGA-CCATCC-TGGACAAC--------ATGGTGAAAC
HUMHBA4a    CC-AGGA-GGGCAGATCACT--TGAGGT-C-AGGA--GTTTGAGA-CCAGCC-TGATCAAC--------ATGGTGAAAC
HUMHBA4b    CTGAGGC-GAGAGGATCACC--TGAGGT-C-GGGA--GTTTGAGA-CCAGCC-TGACCAAT--------ATGGAGAAAC
HUMHBA4c    CCGAGGT-GGGTGGGATCAGC-TGTGGT-C-GGGA--GTTCGAGA-CCAGGC-TGACCAAC--------ATGGAGAAAC
HUMHBA4d    CTGAGGC----TGGCAGATG--CGAGGT-C-AGGA--GATCGAGA-CCATCC-TCGCTAAT--------GCAGTGAAAC
HUMHBA4e    CCGAGGT-GGG-AGGATCACC-TGAGGT-C-GGGA--GTTTGAGA-CCACCC-TGATCAAC--------ATGTAGAAAC
HUMFIXGa    CCCAGGC-AGG-CAGATCACT-TGAGGT-C-AGGA--GTTCGACA-CCAGCC-TGGCCAACAAGGTGAAATGGTGAAAC
HUMFIXGb    CTGAGAC-GGGAGGATTGCT--TAAACC-C-AGGA--GTTTGAGA-CCAGCC-TGGGCAAC--------ACGGCGAAAC
HUMFIXGc    ----------------------------C-AGGA--GATCAAAA-CCATCC-TGGCTAAC--------ATAGTGAAAC
HUMFIXGd    CCAAGGC-GGGCGGATCAC----GAGGT-C-GAGA--GATCGAGA-CCATCC-TGGCCAAC--------ATGGTGAAAC
HUMFIXGe    CCAAGGT-GGGCGGATCACC--TGAGGT-T-AGGA--GTTTCAGG-CCAGCC-TGGCCAAC--------ATGGTGAAAC
HUMGHN      CCAAGGT-GGGTAGATCACC--TGAGAT-T-AGGA--GTTGGAGA-CCAGCC-TGGCCAAT--------ATGGTGAAAC
CHPRSAa     CCGAGGA-GGGTGGATCACC--TGATGT-T-AGGA--GTTCGAGA-TCAGCC-TGACCAAC--------ATGGTGAAAC
CHPRSAb     CCGAGGC-AGGCAGATCACT--TGAGGT-C-AGGA--GTTCAAGA-CCAGCC-TGACCAAC--------ATGGTGAAAC
HUMTPAa     CTGAGGC-AGGTGGATCAC----GAGGT-C-GGGG--GTTTGAGA-CCAGCC-TGACCAAC--------ATGGTGAAAC
HUMTPAb     CCGAGGT-GGGTGGATCACC--TGAGGT-C-ACGA--GTTCAAGA-CCAGTC-TGGCCAAC--------ATGGTGAAAC
HUMTPAc     CCAAGGC-GGGCAGATCACT--TGAGGT-C-AGGT--GTTCAAGA-CCAGCC-TGGCCAAC--------ATGGTGAAAA
HUMTPAd     CCATGGC-AAGAGGGTTGCT--TGAGGC-C-AGGA--GTTTGAGA-CTAGCC-TGGGCAAC--------AAAGCAAAAC
HUMTPAe     CCGAGGT-GGGCGGGTTGCT--TGAGCC-A-AGGA--GTTTGAAA-CCAGCC-CGGGTC---TTGAACATAGCGAAGAC
HUMTPAf     CCAAGGT-GGACAGATCACC--TGAGGT-T-GGGA--GTTCGAGA-CCAGCC-TGGCCAGC--------ATGCCGAAAC
HUMTPAg     CCAAGGA-GGGTGGGTGGCT--TGATCC-T-TGGA--GTTTGAGA-CCTCCC-TGGGTAAC--------ATGGCAAAAC
HUMTPAh     CCAAGGC-TGGCAGATCTCT--TGAGGT-C-AGGA--GTTCAAAA-CAAGCC-TGGCCAAC--------ATAGTGAAAC
HUMTPAi     CTGAGGA-AGGAGGATCATT--TGTGCC-C-AGGA--GTTCGAGA-CTAGCC-TGGACAAC--------ATAGAGAAAC
HUMTPAj     CCAAGGC-GAGCGGATCAC----GAGGT-C-AGGA--GATCGAGA-CCATCC-TGGCTAAC--------ACGGTGAAAC
HUMTPAk     CCGAGGT-GGGAGGATCGCT--TGAGCC-C-AGGA--GTTGGAGA-CCAGTC-TGGGCAAT--------ATAGTGAGAT
HUMTPAl     CGGATCC-AGGAGGATGGCT--TTAGCC-T-AGGA--GTTCAAGA-CCAGCA-TGGGCAAC--------ATAGTGAGAC
HUMTPAm     CCGAGGT-GGGTGGATCACT---TGAGT-C-AGGA--GTTCGAGA-CCAGCC-TGTTCAAT--------ATGGTGAAAC
```

```
Sequence  80        90        100       110       120       130       140       150
HUMTPAn     CCGAGGT-GGGTGAATCACT--TGAGCT-C-AGGA--GTTCGAGA-CCAGCC-TGGCCAAC--------ATGGTGAAAC
HUMTPAo     CCAAGGT-GGGAGGGTCGCT--GGAGCC-C-GGGA--GTTCAAGA-CCAATC-TGGGCAAAC-------ATAGCAAGTC
HUMTPAp     CCAAGGC-AGGTGGATCACC--TGAGGT-C-AGGA--GTTCAAGA-CCAGCC-TGGCCAAC--------ATGGTGAAAC
HUMTPAq     CTGAGGC-GGGCGGCTCACC--TGAGGT-C-AGGA--GTTTGAGA-CCAGCC-TGGCCAAC--------ATGGCGAAAC
HUMTPAr     CTGAGGC-GGGAGAACTGCT--TGAGCC-C-AGGA--GTTTGAGA-CCAGCC-TGGGCAAC--------AAAGTGAGAC
HUMTPAs     CTGAGGC-GGGCAGATCACC--TGATGT-T-AGGA--GTTCAAGA-CCAGCC-TGGCCAAC--------ATGGTGAAAC
HUMTPAt     CCGAGGC-GGGCGGATCAC----GAGGT-C-AGGA--GATCGAGA-CCATCC-CGGCTAAA--------ACGGTGAAAC
HUMTPAu     CTGAGGT-GGGAGGCTCACC--TGAGGT-C-AGGA--GTTCGAGA-CCAGCC-TGACCAAC--------ATGGAGAAAC
HUMTPAv     CTGAGGA-AGGCAGATCACC--TGAGGT-C-AGGA--GTTCGAGA-CCAGCC-TGGCCACC--------ATGGTGAAAC
HSALUAGPa   CCGAGGT-GGGCGGATCAC----AAGGT-C-AAGA--GATCGAGA-CAAGCC-TGGCCAAC--------ATGGAGAAAC
HSALUAGPb   CTGAGGC-AGGAGGATCACT--TGAGGC-C-AGAA--GTTCAAGA-CCAGTC-TGGGCAAC--------ATAGTGAGAC
HSALUAGPc   CCGAGGT-GGGCAGATTGCT--TAAGCT-C-AGGA--GTTCGAGA-CCAACC-TGGGCA-C--------ATGATGAGAC
HSALUAGPd   CTGAGGC-GGGCGGATCAT----GAGGT-C-AAGA--GATGGAGA-CCATCC-TGGCTAAC--------ATG-TGAAAC
HSALUAGPe   CCGAGAC-GGGTGAGTCACC--TGAGGT-C-GGGA--GTTCCAGA-CCAGCC-TGGCCAAC--------ATGA-GAAAC
HSALUAGPf   CACGATC-GGGAGAGTCACC--TGAGTC-C-TGGA--GTTCAAGA-CTAGCC-TGGGCAAC--------ATCAGTGAGA
HUMHBB51    CCAAGGA-GGGTGGATCAC----GAGGT-C-AAGA--GATGGAGA-CCATCC-TGGCCAAT--------ATGGTGAAAC
HUMHPARS1a  CCAAGGC-AGGCAGATCAC----GAGGT-C-AAGA--GATCGAGG-TCATCC-TGGCCAAC--------ATGGTGAAAG
HUMHPARS1b  CTAAGGC-GGGTGGATCACT--TGAGGG-T-GGGA--GTTTGAGA-TCAGCC-AGAGCAAC--------ATGGAGAAAC
HUMHPARS1c  CCGAGAC-AGGTGGATCAC----GAGGT-C-AAGA--GATCGAGA-TCATCC-GGGCCAAC--------ATGGTGAA-C
HSIGVKA2a   ---AGGC-GGGTGCATCACC--TGACAT-C-AGGA--GTTGAATA-CCAGCC-TGGCCAAC--------ATGGCAAAAC
HSIGVKA2b   CTGAAGC-AGGCAGATCAC----AAGGT-C-AGGA--GTTCGAGC-CCAGCC-TGGCCAAT--------ATGGTGAAAC
HUMHBBa     CCAAGGT-GGGCAGATCACT--TGAGGT-C-AGGA--GTTGGACA-CCAGCC-CAGCCAAC--------ATAGTGAAAC
HUMHBBb     CCAAGGC-GGGCAGATCATC--TGAGGT-C-AGGA--GTTCAAGA-CCAGCC-TGGCCAAC--------ATGGCGAAAC
HUMHBBc     CAGAGGT-GGGCAGATCA----TGAGGT-C-AAGA--GATCGAGA-CCATCC-TGGCAAAC--------ATGGTGAAAT
HUMHBBd     CCAAGGC-AGGCAGATCACC--TGAGGT-C-AAGA--GTTCAAGA-CCAACC-TGGCCAAC--------ATGGTGAAAT
HUMHBBf     CCGAGGA-GGGTGGATCACC--TGATGT-T-AGGA--GTTCGAGA-TCAGCC-TGACCAAC--------ATGGTGAAAC
HUMHBBg     CCGAGGC-AGGCAGATCACT--TGAGGT-C-AGGA--GTTCAAGA-CCAGCC-TGACCAAC--------ATGGTGAAAC
HUMLDLR18a  CCGAGGC-GGGTGGATCA----TGAGGT-C-AGGA--GATCGAGA-CCATCC-TGGCTAAC--------AAGGTGAAAC
HUMLDLR18b  CTGA-----GCTGGATCACT--TGAGTT-C-AGGA--GTTGGAGA-CCAGGC-CTGAGCAA-------CAAAGCGAGAT
HUMPOMC8    CCAAGGC-AGGCAGATCACA--TGAGGT-C-AGGA--GTTCAAGA-CCAGCC-TGGCCAAC-------TATGGTGAAAT
HUMPOMC6    CTAAGGT-GGGAGGATTGCT--TGAGCC-C-AGGG--ATTCAAGA-CAAGCC-TGGGCAAC--------ATAGTGAGAC
HUMPOMC4    CCAAGGC-GGGCGGATCACC--TGTTGT-C-GGGA--GTTTGAGA-CCAGCC-TGACCAAC--------ATGGAGAAAC
HUMPOMC2    CTGATGT-----GGATTACT--TGAGCC-C-AGGA--GCTTGAGA-CCAGCC-TAGGCAAC--------ATAGTGAAAC
HUMPOMC1    CTGAGTT-GGGCAGATCAC----GAGGT-C-AAGA--GATGGAGA-CATTCC-TGGCCAAG--------ATGGCGAAAC
HSMLVI2     CCGAGGC-GGGCGGATCAC----GAGGT-C-AGGA--GATCGAGA-CCATCC-CGGCTAAA--------ACGGTGAAAC
HSREP10a    CTGAGTT-GGGTGGATCGTT--TGAGTC-C-CGGA--GTTTGAGA-CCAGCA-TGGGCAAC--------ATAGAGAGAC
HSREP10b    CTGAGTC-AGGAAGAGAGCT--TGAGGT-C-AGGA--GTTCGAGA-CCAGCC-TGGGCACC--------ATGGTGAGAC
HSREP10c    CCGAGGC-GGGTGGATCACC--TGAGGT-C-GGGA--GTTTGAGA-CCAGCC-TGACTAAC--------ATGGAGAAAC
HSREP10d    CAGAGGC-AGGCAGAGTCACC-TGAGAG-G-TCAGGAATAGGAGA-CCAGCC-TGGCCAAC--------ATGGCGAAAC
HSREP10e    CCGAGGC-GGGCGGATAGC----GAGGT-C-AGGA--GATCAAGA-CTATCC-CGGCTAAC--------ACGGTGAGGC
HSREP10f    CCGAGGC-GGGCAGATTGC----GAGGT-T-AGGA--GATCGAGA-CCATCC-TGGCTAAC--------ACAGTGAAAC
HSREP10g    CCATGGC-AGGTGTATGCTC---CAGCT-C-AGGA--GTTCAAGA-ACAGCC-TGGGCAAC--------ATGTCGAAAC
HSREP10h    CAGAGGT-GGGAGGAGGTTCT-TGAGCC-C-AGCAG-CTTGGAGA-CTTAGC-CTGGCGGC-------CCTAGTGAGGC
HSREP10j    CCGAGGT-GGGCGGATCAC----GAGGT-C-AGGA--GATCGAGA-CCAACC-TGGCTAAC--------ACGGTGAAAC
HUMADAGa    CCAAGGC-AGGAGGATCACC--TGAGGT-C-GGGA--GTTTGAGA-CCAGCC-TGACCAAC--------ATGGTGAAAC
HUMADAGb    CCGAGGT-GGGTGGATCATG--TGAGGT-C-AGGT--GTTCGAGA-CCAGCC-TGGCCAAC--------ATGGTGAAAC
HUMADAGc    CTGAGGC-GGGAGGATCATT--TGAGTC-C-AGGA--GTTTGAGA-CTAGCC-TGGACAAC--------AAAACTAGAC
HUMADAGd    CAAAGGT-GGGTGGATCAT----GAGGC-C-AGGA--GTTCGAGA-CCAGCC-TGGCCAAC--------ATGGCAAAAC
HUMADAGf    CTGAGGT-GGGCGGATCACC--TGAAGT-C-GGGA--GTTCGAGA-CCAGCC-TGGCCAAC--------AAGGAGAAAC
HUMADAGi    CCTAGGT-GGGTGGATCACC--TGAGGT-C-AGGA--GTTCAAAA-CCAGCCT-GGCCAAC--------ATGGTGAAAC
HUMADAGja   CTGAGGT-GGGTGGATCACT--TGAGGT-T-AGGA--GTTCAACA-CCAGCCT-GGCCAAC--------ATGGTGAAAT
HUMADAGk    CCGAGGT-GGGCAGATCACT--TGAGGT-C-AGGA--GTTCAAGA-CCAGCCT-GGCCAAT--------ATGGTGAAAC
HUMADAGl    CCGAGGT-GGGTGGATCACC--TGAGGC-C-AGGA--GTTTGAGA-CCAGCCT-GGCCAAC--------ATGGTGAAAC
HUMADAGm    CTGAGGT-GGGTGGA-CACC--TGAGGT-C-AGGA--GTTCGAGA-CTAGCCT-GAGCAAC--------ATGGTGAAAC
HUMADAGn    CCAAGCT-GGGTGGATCACT--TGAGGT-C-AGGA--GTTCGAGA-CCAGCCT-GGCCAAC--------ATG-TGAAAC
HUMADAGo    CCAAGGC-GGGTGGATCACC--TGAGGT-C-AGGA--GTTCAAGA-CCAGCGT-GGCCAAC--------ATGG-GAAAC
HUMADAGp    CCAAGGCT-GGCGGATCACT--TAAGCC-C-AGGA--GTTTGAAA-CCAGCCT-GGGCAAC--------ACAGTGAAAC
HUMADAGq    -CAAGGT-CGGAGGATCAT----GAGGT-C-AGGA--GATTGAGA-CTATCCT-GGCCAAC--------ATGGTGAAAC
HUMADAGr    CCGAGGC-GGGCAGATCACC--TGAGGT-C-AGGA--ATTCGAGA-CCAGCCT-GGCAAAC--------ATAGTAAAAC
HUMADAGs    CTGAGGT-GGGCGGATCAT----GAGGT-C-AGGA--GATCGAGA-CCATCCT-GGCTAAC--------ACAGCGAAAC
HUMADAGt    C-GAGGC-AGGAGGATTGCT--TGAG-C-CTAGGA--GTTTGAGA-CCAGCCA-GGGCAAC--------ATAGTGAGAT
```

```
Sequence  160       170       180       190       200       210       220       230
PTAZGLO    TCCATCTGTACTAAAAATAAAAAAA----------------------------------------TTAGCCAGGC
PTRE123A   CCCAGTTATACTAAAAATACAAAA----------------------------------------TTAGCTGGG-
PTRE123B   TCTGTCTCTACTAAAAATAGAAAAAA----------------------------------------TTAGCCAGG-
PTGL01     TGTGTCTCTATTAAAAATACATAAA----------------------------------------TTAGCCAGG-
ATBOWL1    TC-GTCTCTACTAAAAATACAAAAA----------------------------------------TTAGCCGGG-
ATHOWL1    CACATGTCAAAAAAAAATACAAAAAG----------------------------------------TAGTTAAG-
ATHOWL6    T-TGTCTCTACTAAAAATATAAAAA----------------------------------------TTATCCAGG-
GCREG13    CCATCTCTAAAAA----------------------------------------TTAGAAGGG-
GCREG19    ACCCCATGTCTAACAAGTAGAAAAA----------------------------------------CTAGGCGGG-
GCREG9     AGTCTCTCTCTAAAA----------------------------------------CTAGC-----
GGALU      CCCGTCTCTACTAAAAAATACAAAAAA----------------------------------------TTAGCCGGG-
ORAHBBEa   CCCGTCTCTACTAAAAATACAAAAA----------------------------------------TTAGACTGG-
ORAHBBEb   CCCGTCTCTACTAAAAATAAAAAAA----------------------------------------TTAGCTGGG-
humblym1   CCTCGTCTCTACCAAAAATTTAAAAA----------------------------------------TTAGCTGGG-
humc1ain1  CACGTCTCTACTAAAAATACAAAAGA----------------------------------------TTAGCCTGGG
humgast2   CCCATCTCTACTAAAAATTACAAAAA----------------------------------------TTAGCCAGG-
humins2    CCCATCTTTACTAAAACTACAAAA----------------------------------------TTAGCTGGG-
humrsa1    GCTGTCACTACTAAAAATCCAAAAAA----------------------------------------TTAGXTAGG-
humrsa16   CCTGTCTCTA-TAAAAATACAAAAAA----------------------------------------GTAGCCGGG-
humrsa27   CCCGTCTCTACTAAAAAT----------
humrsaold  TC-ATCTCTACCAAAAAATAAAATAAAAA----------------------------------------TTAGCCAGG-
humrsap3   CCCGTCTCTACTAAAAATACAAACAACCAG----------------------------------------CCAGCCGAG-
humrskpa1  GCTGCCTCTAC-AGAAAATTTAAAAA----------------------------------------TTAGCTAGG-
hump5311   CCCATCTCTAC-AAAAC----------------------------------------ATGCAAAA-
humhbbrta  CCCATCTCTACTAAAAATACAAAA----------------------------------------TCAGCCGGGC
humhbbrtb  CCCCTATCTACTAAAAATAGAAAA----------------------------------------TTAGCCAGG-
humapoe4a  CCTGTCTCTACTAAAAATACAAAAA----------------------------------------TTAGCCAGG-
humapoe4b  CTCGTCTCTACTAAAAATACAAAAA----------------------------------------TTTAGCCGGG
humapoe4c  CCCATCTCTACTAAAAATACAAAAG----------------------------------------TTAGCCAGG-
humapoe4d  CCCGTCTCTACTGAAAATACAAAAAAG----------------------------------------TTAGTCGGG-
HUMTKRAa   CCCAGCTCTATAAAAATAAAATAAAA----------------------------------------TTAGCTG---
HUMTKRAb   CCCATCTCTACTGAAAAAAAAAACCAAAAA----------------------------------------TTATCTGGG-
HUMTKRAc   AC-----TGTGTGAAAAAAATGTAAAAA----------------------------------------TTAACTGGG-
HUMTKRAd   CCCATGTCTGCTAAAAATACAAAAA----------------------------------------TTAGCTGGG-
HUMTKRAe   CCCATCTCTACTAAAAATATAAAA----------------------------------------TTAGCTGGG-
HUMTKRAf   CCCATCTCTACCAAAAAAAAAAAGAAA----------------------------------------TTAGCCAGT-
HUMTKRAg   CTTGTCTCTAC-AAAAAT----------
HUMTKRAh   CCCATCTCTATTAAAAATACATAAA----------------------------------------TGAXTACG--
HUMTKRAi   CCCGTCTCTACTAAAAATACAAAAGG----------------------------------------TTAGCTGGG-
HUMTKRAj   TCCGTCTCTACTAAAAATACAAAAA----------------------------------------TTAGCCGGG-
HUMTKRAk   CCCATCTCTACTAAAAATACAAAAA----------------------------------------TTAGCCAAG-
HUMTKRAl   CCCGTCTCTACTGAAAATACAAAAAAA----------------------------------------TTAGCCAGG-
HUMALPPA   CCCTTATCTAT----------
HUMCEA     CCCATCTCTACTAAAAATACAAAAA----------------------------------------TGAGCTGGG-
HUMAPOC2A  CGCATCTATACTAAAAATACAAAAA----------------------------------------TAAGCCGGG-
HUMAPOC2B  CCCGTCTCTAGTAAAAACACAAAAA----------------------------------------TTAGACAGG-
HUMAPOC2C  CCCATCTCTACTAAAAATACAAAAA----------------------------------------TCAGGTGGG-
HUMAPOC2E  CCC-----TACTAAAAATACAAACA----------------------------------------TTAGCCTG--
HUMAGGa    CCCGTCTCTACTAAAAATACAAAAAA----------------------------------------TTAGCCGGG-
HUMAGGb    CCTGTCTCTACTAAAAACACAAAAA----------------------------------------TTAGCCAGG-
HUMAGGc    CTTGTCTCTACTAAAAATGTGAAAA----------------------------------------TTAGCTGGG-
HUMHBA4a   TGTGTCTCTATTAAAAATACATAAA----------------------------------------TTAGCCAGGA
HUMHBA4b   CCCAGTTATACTAAAAATACAAAA----------------------------------------TTAGCTGGG-
HUMHBA4c   CCCGTCTCTACTAAAAATACAAAAA----------------------------------------TTAGCTGG--
HUMHBA4d   TCTGTCTCTATTAAAAAAAGAAAAAA----------------------------------------TTAGCCACG-
HUMHBA4e   CCCATCTATACTAAAAATACAAAA----------------------------------------TCAGCCGGG-
HUMFIXGa   CCCGGCTTTACTAAAAATACAAAAA----------------------------------------TTAGCTGGG-
HUMFIXGb   CCCACCTCTAATTAAAAAAAATACAAAAA----------------------------------------TTAGCTAGG-
HUMFIXGc   CCCGTCTCTTCCAAAAATACAAAAAA----------------------------------------TTAGCCCGG-
HUMFIXGd   CCCATCTCTACTAAAAATACAAAAA----------------------------------------TTAACTGGG-
HUMFIXGe   CCTGTCTCTACTAAAAATAGAACAA----------------------------------------TTAGCTGGG-
HUMGHN     CCCGTCTCTACCAAAAAAACAAAAA----------------------------------------TTAGCTGAG-
CHPRSAa    CCCCTCTCTACTAAAAATAGAAAA----------------------------------------TTAGCCAGG-
CHPRSAb    CCCATCTCTACTAAAAATACAAAA----------------------------------------TCAGCCAGG-
HUMTPAa    CCCGTCTCTACTAAAA-TACAAAAAA----------------------------------------TTAGCTGGG-
HUMTPAb    CCTGTCTCTACTAAAAATACAAAAA----------------------------------------TTAACCAGG-
HUMTPAc    C----------
HUMTPAd    TCTGCTTCTACAAAAAAATTTTATAA----------------------------------------TTAGCTGGA-
HUMTPAe    TCTGTCTCTACAAAAAAATGAAAAAAAAAAAAAAA----------------------------------------TTAGCCAGA-
HUMTPAf    CCTGTCTCTACTAAAAATACAAAAA----------------------------------------TTAGCCAGG-
HUMTPAg    CCCATCTCAACAAAAAAATACAAAAAA----------------------------------------TTATTTGGG-
HUMTPAh    CCCATCTCTACTAAAAATACAAAAAAA----------------------------------------TTAGCTGGG-
HUMTPAi    CTTGTCTCTACACACACACACACACACACACACACACACACACTCTCTCTCTCTCTCTCTC----------TCAGCCAGG-
HUMTPAj    CCCGTCTCTACTAAAAATACAAAAA----------------------------------------TTAGCCGGG-
HUMTPAk    GCTGTCTCTACAAAAAATTTAAAAA----------------------------------------TTAGCCGGG-
HUMTPAl    ACCTCTGCCTCCCCCACCCCACCCCGCCTCTACAAAAAAATTTAAAAA----------TTAGCTGGG-
HUMTPAm    CCCGTCTCTACCAAAAATACAAAAA----------------------------------------TTAGCCGGG-
```

```
Sequence  160       170       180       190       200       210       220       230
HUMTPAn     CCTATCTCTACTAAAAATACAAAAA----------------------------------------------TGAGCCAGG-
HUMTPAo     CCCTGTCTCTACAAAAAATAAAAAAA----------------------------------------------TTAGCCAGA-
HUMTPAp     CCTGTCTCTACTAAAAATACAAAAA----------------------------------------------TTAGCCAGG-
HUMTPAq     CCTGTCTCTACTAAAAATACAAAAA----------------------------------------------TTAGCCGGA-
HUMTPAr     TCCATCTCTACAAAAAATACAAAAA----------------------------------------------TTAGTCAGG-
HUMTPAs     CCTGTCTCAACTAAAAATACAAAA-----------------------------------------------TTAGCGTAG-
HUMTPAt     CCCGTCTCTACTAAAACTACAAAAAA---------------------------------------------T-AGCCGGG-
HUMTPAu     CCCATCTCTACTAAAAATACAAAA-----------------------------------------------TTAGCCTGG-
HUMTPAv     CCTGCCTCTACTAAAAACACAAAAA----------------------------------------------TTAGCCAGG-
HSALUAGPa   TCCGTCTCTACTAAAAATACAAAAA----------------------------------------------TTAGCTGAG-
HSALUAGPb   TCTGTCTCTACAATTTTTTGTTTTTTAAATTTAA------------------------------------TTAGCCAAG-
HSALUAGPc   CCC-CATCTCTACAAAGTACAAAAAG---------------------------------------------TTAGCCAGG-
HSALUAGPd   CCCATCTCTACTAAAAATACAAAAACAAAA-------------------------------------------TTAGCTGGG-
HSALUAGPe   CCATCTCTACTGAAAATAACAACAACAAAAAAAA-------------------------------------TTAGCTGGG-
HSALUAGPf   CCCTATCTTTACACAAAAATATTTAAAAA----------------------------------------TGGTGCCA-
HUMHBB51    CCCGTGTCTACTAAAAATACAAAAA----------------------------------------------TTAGCTGGG-
HUMHPARS1a  CCCGTCTCTACTAAAAATACAAAAA----------------------------------------------TTAGGTGGG-
HUMHPARS1b  CTCGTCTCTACTAAAAATACAAAAAA--------------------------------------------TTAGCCAGG-
HUMHPARS1c  CACATCGCTACTAAAAATACAAAAA----------------------------------------------TTAGCTGGG-
HSIGVKA2a   CCTGTCTCTACTAAAAATACAAAAAATACAAAAGAAAA-------------------------------TTAGCTGGG-
HSIGVKA2b   CCTGTCTCTACTAAAAATACAAAAA----------------------------------------------TTAGCTGGG-
HUMHBBa     CTGGTCTCTGCTTAAAAATATAAAAA---------------------------------------------TTAGCTGGA-
HUMHBBb     CCCATCTCTACTAAAAATACAAAAA----------------------------------------------TTAGACTGG-
HUMHBBc     CCCGTCTCTACTAAAAATAAAAAAA----------------------------------------------TTAGCTGGG-
HUMHBBd     CCCATCTCTAC-AAAAATACAAAAA----------------------------------------------TTAGACAGG-
HUMHBBf     CCCCTATCTACTAAAAATAGAAAA-----------------------------------------------TTAGCCAGG-
HUMHBBg     CCCATCTCTACTAAAAATACAAAA-----------------------------------------------TCAGCCGGG-
HUMLDLR18a  CCCGTCTCTACTAAAAATACAAAAAA--------------------------------------------TTAGCCGGG-
HUMLDLR18b  CCCATCTCTAC-AAAAAC-CAAAAAGTTAAAAA-------------------------------------TCAGCTGGG-
HUMPOMC8    ACCATCTCTAGTATAAATAAATAAA----------------------------------------------TAAGCTGGG-
HUMPOMC6    CTCCACTTCTACAAAAAATAAAAAAG---------------------------------------------TTAGCCAGA-
HUMPOMC4    TCCGTCACTACTAAAAATACAAAAA----------------------------------------------TTAGCCGGG-
HUMPOMC2    CCTGTCTCTACCAAAAACAAAAACAAAAACAAAACAAAAAACCAAAACCCAAAA-----------------TTAGCTGGG-
HUMPOMC1    CCCGTCTCTACTAAAAATACAAAAA----------------------------------------------TTAGCCGGG-
HSMLVI2     CC-GTCTCTACTAAAAATAACAAAAAA--------------------------------------------TTAGCC-GG-
HSREP10a    CC-ACCTCTCTACAAAGAAAACTAAAAAA------------------------------------------TTAGCTGGG-
HSREP10b    CCTGTCTCTAC-AAAAAGAAAAAAATTTTAA----------------------------------------TTAGCTGAG-
HSREP10c    CCCATCTCTACTAAAAAT--AAAAAA---------------------------------------------TTAGCCCAGG
HSREP10d    CCCGTCACTATTAAAAATACAAAAAA---------------------------------------------TTAGCTAGG-
HSREP10e    CCCGTCTCTACTAAAAATAGAAAAAA---------------------------------------------TTAGCTGGA-
HSREP10f    CCTGTCTCTACTGAAAATACAAAAAA---------------------------------------------TTAGCCGGG-
HSREP10g    CCCACCTCTACAAAAAAAAGATACAAAAA------------------------------------------TTAGCCAGA-
HSREP10h    CCTGTCTCTAACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAATATATATATATATATATATATAAA-TTAGCCAGT-
HSREP10j    CCCGTCTCTACTAAAATTACAAAAAAAA-------------------------------------------TTAGCCGGG-
HUMADAGa    CCCGTCTCTACTAAAAAAAAAAAAAAAAATACAAAA-----------------------------------TTAGCCAGG-
HUMADAGb    CCCGTCTCTACTACAAATCCAAAAC-----------------------------------------------TTAGCAGGG-
HUMADAGc    CCCGTCTCTCCAAAAAATAAAAAAAG---------------------------------------------TTAGGGGAAG
HUMADAGd    CCCATTTCTACTAAAAATACAAAAAA---------------------------------------------TTAGCTGGG-
HUMADAGf    CCCGTCTCTACTAAAAATACAAAAAA---------------------------------------------TTAGCCGGG-
HUMADAGi    CCTGTCTCTACTAAAAGTACAAAAA----------------------------------------------TTAGCCGGG-
HUMADAGja   CCCCTCTCTATTAAAAATACAAAAAT-------------------------------------------------------
HUMADAGk    CCCATCTCTACTAAAAATACAAAAA----------------------------------------------TTAGCTGGG-
HUMADAGl    CCCATCTCTACTAAAAATGCAAAAAA---------------------------------------------TTAGCCAGG-
HUMADAGm    CCTGTCTCTACTAAAAATACAAAAA----------------------------------------------TTAACTGGG-
HUMADAGn    CCCATCTCCGCTAAAAATACAAAAA----------------------------------------------TTAGCTGGA-
HUMADAGo    CCCGTGTCTACTAAACATACAAAAAA---------------------------------------------TTAGCCAGG-
HUMADAGp    CCCTTCAAAACAAAAA-------------------------------------------------------TTCACCAGG-
HUMADAGq    CCCGTCTCTACCAAAAATACAAAAA----------------------------------------------TTAGTTGGG-
HUMADAGr    CCCATCTCTACTAAAAATACAAAACCA--------------------------------------------TTAGTTGGG-
HUMADAGs    GCTGTCTCTACTAAAAATACAAAAAA---------------------------------------------TTAGCCGGG-
HUMADAGt    CC-ATCTCTGC---AAATACAAAAAAAA-------------------------------------------GTAGTCAGG-
```

```
Sequence   240       250       260       270       280       290       300       310
PTAZGLO    -----CGCGGTGGTGCGTG----------------CCTG-TGATCCC--------AGCTACTCAGG--AGACTG-AAGC
PTRE123A   -----TGTC-TGCGCA-TG----------------CCTG-TAATCCC--------ACCTACTCGGG--AGGCTG-AGGC
PTRE123B   -----CGTAGTGGGGGGCG----------------CCTG-TAGTCCCAGCTACTCAGCTACTCGGG--AGGCTG-AGGG
PTGL01     -----AATGGCGGCCCATG----------------CCTG-TAATCCC--------AGCTACACGGG--AGACAG-AGGC
ATBOWL1    -----TGTGGTGGCAGGCA----------------CCTG-TAGTCCC--------ACCTACTTGGG--AGGCTG-AGGC
ATHOWL1    -----TGTGATGTTGTGTG----------------CCTG-TAGTCCC--------AGCTACTCCGG--AGGCTG-AGGC
ATHOWL6    -----CGTCATTGTGTGTT----------------CCTG-TAGTCCC--------AGCTGCTTGGG--AGGCTG-AGGC
GCREG13    -----CACTGTGGTAGGCA----------------CCTA-TAGTCCC--------AGCTACTTGGG--AGGCTG-AGGC
GCREG19    -----CATTATGGCAGGTG----------------TCTG-TAGTCCC--------AGCTAGTCAGA--AGGCTT-AGGC
GCREG9     -----CAGGTG------------------------CCTG-TAGTCCC--------AGCTACTAGGG--AAGCTG-AGGC
GGALU      -----CGTGGTGGCGGGCG----------------C-TG-TAGTCCC--------AGCTACGCGGG--AGGCTG-AGGC
ORAHBBEa   -----CATGGTGGCAGGTG----------------CCTG-TAATCCC--------AGCTACTCAGG--AGGCCA-AGGC
ORAHBBEb   -----CGTGGTGGCATGCG----------------CCTG-TAGTCCC--------AGCTACTCGGG--AGGCTG-AGGC
humblym1   -----CAGGGCATGATGAT---GCACAT--------CCC--TAATCCC--------AGCTACTTGAG--GGCTTG-AGGT
humc1ain1  -----TGTGGTGGTGTGCA----------------CCCTGTAACCCC--------AGCTAGTCAGG--AGGCTG-AGGC
humgast2   -----CATGGTGGCACGTG----------------CCTAT-AGTCCC--------AGATATTCTGG--AGGCTG-AGGC
humins2    -----TGTGGTGGCACATG----------------CCTGT-AATCCC--------AGATATTCGGG--AGGCTG-AGGC
humrsa1    -----CATGGXXGCAGCTGTGXAGTXCCAGTGT--GGTGT-AGTCCC--------AGCTACCTGGG--AGGCTG-AGGG
humrsa16   -----CATGGTGGCAGACG----------------CTGT-AGTCCC--------AGCTACTCAGG-----CTG-AAGC
humrsa27   ------------------------------------TGT-AATCTC--------AGCTACTTGGG--AGGCTG-AGGC
humrsaold  -----TGTGGTGXXGCATG----------------CXXGT-GGTGCC--------AGCTACTCAGA--AGGCTG-CAGT
humrsap3   -----TGTGGTGGTGGGCA----------------CCTGT-AGTCCC--------AGCTACTCAGGG-AGGCTG-AGGC
humrskpa1  -----CATGCTGGAATGTG----------------CCTAT-AGTCCC--------AGCTACCCAAG--AGACTG-ATGT
hump5311   -----GTTGGCTGGCCATGGTGGCATG--AA----CCTGT-GGTCCC--------AGCTACTCCGG--AGGCTG----A
humhbbrta  G----TGTGGTGCATGCCT----------------GCAG----TCCC--------AGCTATTCAGG--TGGCTG-AGGC
humhbbrtb  -----CATGGTGGTGGCAC-----ATG--------CCTGT-AGTGCC--------AGCTACTTGGG--AGGCTG-AGGC
humapoe4a  -----CATGGTGCCACACA----------------CCTGT-GCTCTC--------AGCTACTCAGG--AGGCTG-AGGC
humapoe4b  -----TGTGGTGGTGTGAG----------------CCTGT-AATCCC--------AGCTACTGAGG---------CAGC
humapoe4c  -----CGTGGTGGCGGGCG----------------CCTGT-AATCCT--------AGCTACTTGGG--AGGCTG-AGGC
humapoe4d  -----TGTGGTGGCATGTG----------------CCTGT-AGTCCC--------AGCTACTTGGG--AGGCTG-AGGC
HUMTKRAa   -----GGTGGAGGCGGGGGG----------ACG--CCTGT-AGTCCC--------AGCTACTTGGG--AGGCTG-AGGT
HUMTKRAb   -----CATGGTGGCAGCCT----------------CCTGT-AATCCC--------AGCTACTCGGG--AGGTTG-AGGC
HUMTKRAc   -----TGTGGTGGTGTGCG----------------CCTGT-AGTCCC--------ACCTGCTCAGG--GGGCTG-AGGT
HUMTKRAd   -----TGTGATGGTGTGTGG----------------CCAGT-AGTCCC--------AGCTACTCTTG--TGGCTG-AGGT
HUMTKRAe   -----CGTGGTGGCGCATG----------------CCTGT-AATCCC--------AGCTACTCGGG--AGGCTG-AGAC
HUMTKRAf   -----CATAGCAGCGCACA----------------CCTGT-GGTCCC--------TGCTACTCAGG--AGGCTG-AGGC
HUMTKRAg   -----CATGATGGTGGGTG----------------CCTGT-AATCCC--------ACCTACTTGGG--AGGC-G-AGGT
HUMTKRAh   -----CATG-TGGCACGCA----------------CCTGT-GATCCC--------AGCTACCCGGA--AGGATG-AGGC
HUMTKRAi   -----CATGGTGGCAGAAG----------------CCTGT-AATCCC--------AGCTACTCAGG--AGGCTG-AGGC
HUMTKRAj   -----CGTGGTGGCGCACTG----------------CCTGT-AATCCC--------AGCTACTTGGG--AGGCTG-AGGC
HUMTKRAk   -----CGTGGTGGCACACA----------------CCTGT-AATTCC--------AGCTACTTGGG--AGGCAG-AGGC
HUMTKRAl   -----CGTGGTGGTGGGCA----------------CCTGT-AGTCTC--------AGCTACTCGGG--AGGCTG-AGGC
HUMALPPA   ------------------------------GCG-CCTGT-AGTCCC--------AGCTACCCAGG--AGGCTG-AAGC
HUMCEA     -----CTTGGTGGCGCGC----------------ACCTGT-AGTCCC--------AGTTACTCGGG--AGGCTG-AGGC
HUMAPOC2A  -----TGTGGTAGCATATG----------------CCTGT-AATCCC--------AGCTATTCCAG--AGGCTG-AGAC
HUMAPOC2B  -----CATGGTGGTGCATG----------------CTTGT-ATTTCC--------AGTTACTCAGG--AGGCTG-AGGC
HUMAPOC2C  -----GCAAGTGGCT-CA----------------AGCCTGT-AATCCT--------AGCACTTTGGG--AGGCCA-AGGT
HUMAPOC2E  -----CACAGTG-CACATG----------------CCCAT-AATCCC--------AGCTACTCCAG--AGGCTG-AGGC
HUMAGGa    -----CGTAGTG-CGGGCG----------------CCTGT-AGTCCC--------AGCTACTTGGG--AGGCTG-AGGC
HUMAGGb    -----TGTGGTGGTGCATG----------------CCTGT-AATCTC--------AGCTATTTGGG--AGGCTG-AGGC
HUMAGGc    -----CGTGGTGGCACACA----------------CCTAT-AGTCCC--------AGCTACTCGGG--AGGCTG-AGGC
HUMHBA4a   ------ATGGCGGCCCATG----------------CCTGT-AATCCC--------AGCTACACGGG--AGACAG-AGGC
HUMHBA4b   -----TGTGGTGGCGCATG----------------CCTGT-AATCCT--------AGCTACTAGGA--AGGCTG-AGGC
HUMHBA4c   -----TGTGGTGGTGCACG----------------CCTGT-AATCCC--------AGCTACTCGGG--AGGTGG-AGGC
HUMHBA4d   -----ACTAGTGGGGGGCGC-------TGTAG---CCCAG-CTACTC--------AGCGACTCGGG--AGGCTG-AGGG
HUMHBA4e   -----CATGGTGGCCCATG----------------CCTGT-AAACCC--------ACCTACTCCGG--AGGCTG-AGGC
HUMFIXGa   -----CATGGTGGCGGGTG----------------CCTGT-AATCCC--------AGCTACTTGGG--AGGCTG-AGGC
HUMFIXGb   -----TGTGATGGCTCCCA----------------CCTGT-GCTCCC--------AGCTATTCAGG--AGGCTG-AGGT
HUMFIXGc   -----CGTGGTGGCAGGCG----------------CCTAT-AGTCCC--------AGCTACACGGG--AGGCTG-AGGC
HUMFIXGd   -----CATGGTGGCATGCG----------------CCTGT-AGTCCC--------AGGAGAATTGCTTGAACCT-GGGA
HUMFIXGe   -----CTTGGTAATGTGCA----------------CCTAT-AATCCC--------AGCTACTGGGG--AGGCTG-AGGC
HUMGHN     -----CCTGGTCATGCATG----------------CCTGG-AATCCC--------AACAACTCGGG--AGGCTG-AGGC
CHPRSAa    -----CATGGTGGTGGCAC----------------GCCTGT-AGTGCC--------AGCTACTTGGG--AGGCTG-AGGC
CHPRSAb    -----TGTG-TGGTGCATG----------------CCTGC-AGTCCC--------AGCTATTCAGG--TGGCTG-AGGC
HUMTPAa    -----CGTGGTGGCGGGCA----------------CCTGT-AATCTC--------AGCTACTCAGG--AGGCTG-AGGC
HUMTPAb    -----TGTGGTGGCAGGCG----------------CCTGT-AATCCC--------AGTTACTCAGA--AGGCTG-AGGC
HUMTPAc    -----TGTGGTGGCGGGCG----------------CCTGT-AATCCA--------AGCTATTTGGG--AGGCTG-AGGC
HUMTPAd    -----TATGGCAGCATGCC----------------CCTGT-AGTCCT--------AGCTACTTGGG--AGGCTG-AGAT
HUMTPAe    -----CATGGTGGCACGCA----------------CCTGT-AGTCCC--------AGCTACTTGAG--AAGCTG-AGGT
HUMTPAf    -----CATGGTGGCACACA----------------CTTGT-AATCCG--------AGCTACTCGGG--AGGCTG-AAGA
HUMTPAg    -----CATGGTGGTACATG----------------CCTGT-AGTCCC--------AGCTACTTGGG--AGGCTG-AGGT
HUMTPAh    -----CATGGTGGCGGGCG----------------CCTAT-AATCCC--------AGCTACTCAGG--AGGCTG-AGGC
HUMTPAi    -----CACAGTGGCACATG----------------CCTGA-AGTCCC--------AGCT-CT-GGG--AAGCTG-AGGC
HUMTPAj    -----CGTGGTGGCAGGCG----------------CCTGT-AGTCCC--------AGCTACTCAGG--AGGCTG-AGAC
HUMTPAk    -----TGTACTAGTATGCA----------------CCTGT-GGTCCC--------AGCTACTCAGG--AGGCTG-AGGC
HUMTPAl    -----TGTGGTGGTGTGCA----------------CCTGT-AGTCCC--------AGCTACTTGGG--AGGTTG-ACGC
HUMTPAm    -----CATGGTATCGGGCA----------------CCTGT-AATCCC--------AGCTACTCGGG--AGGCTG-AGGC
```

```
Sequence   240        250        260        270        280        290        300        310
HUMTPAn    -----CATGGTGGCAGGTG----------------CCTGT-AATCCC--------AGCTACTCAGG--AGGCTG-AGGC
HUMTPAo    -----CCTGGTA-----------------------TGT-AGTCCC--------AACTACTTGGG--AGGTTG-AGGC
HUMTPAp    -----CATGGTGGCAGGCA----------------CCTGT-AATCCC--------AGCTACTCGGG--AAGCTG-AGGC
HUMTPAq    -----TACGGTGGCAGGCA----------------CCTGT-AATCCC--------AGCTACTCAGG--AGGCTG-AGGC
HUMTPAr    -----TGTGGTGGCTTGTG----------------CCTAT-GGTCCC--------AGCTGCTTGGG--AGGCTC-AGGT
HUMTPAs    -----CATGGTGGCGCATG----------------CCTGT-AACCCC--------ACCTACTCGGG--AGGCTG-AGGT
HUMTPAt    -----CGTAGTGGCGGGCG----------------CCTGT-AGTCCT--------GGCTACTTGGG--AGGCTG-AGGC
HUMTPAu    -----CATGGTGGCACATG----------------CCTGT-AATCCC--------AGCTACTCAGG--AGGCTG-AGGC
HUMTPAv    -----TGTGGTGGTGCGTG----------------CCCAT-AAGCAC-------------------------------
HSALUAGPa  -----TGTGGTGGCATGCG----------------CCTGT-AGCTAC-------CAGCTACTCAGG--AGGCTG-AGGC
HSALUAGPb  -----TGTGGTGGTGCATG----------------CCTGT-AGTCCT--------AGCTACTCCAG--AGGCTG-AGGT
HSALUAGPc  -----TGTGGTGGCACGCA----------------CCTGT-GGTCCT--------AGCTACTCGGG--AGGCTG-AGCT
HSALUAGPd  -----CGTGGTGGCAGGCG----------------CCTGT-AGTCCC--------AGCTACTTGGG--AGGCTG-AGAT
HSALUAGPe  -----CGTTGTGGAGGGCA----------------CCAGT-AATCCC--------AGCGACTCAGG--AGGCTG-AGGC
HSALUAGPf  -----TGTACCTGAAAGCC----------------TAGCT-ACTCTGG----------ACTCTG---AGGCTG-GAAG
HUMHBB51   -----TGTGGTGACGCGCT----------------ACTGT-AGTCCC--------AGCTTCTCAGG--AGGCTG-TGGC
HUMHPARS1a -----CATAGTGGCGCACG----------------CCTGT-AGTCCC--------AGCTACCTGGG--AGGCTG-AGGC
HUMHPARS1b -----CATGGTGATATATG----------------CCTGT-AATCCC--------AGCTACTCGGG--AGGCTG-AGGC
HUMHPARS1c -----CGTGGTGGCACACG----------------CCTGT-AGTCCT--------AGCTACTTGGG--AGGCTG-AGGC
HSIGVKA2a  -----CGTGTTGGCATGGG----------------CCTGT-AATCCC--------AACTACTCAGG--AAGCTG-AGGC
HSIGVKA2b  -----CGTGGTGGCGGACG----------------CCTAT-ATTCCC--------AGCTACTAGGG--AGGCTG-AGGC
HUMHBBa    -----CGTGGTGACGGGAG----------------CCTGT-AATCC---------AACTACTTGGG--AGGCTG-AGGC
HUMHBBb    -----CATGGTGGCAGGTG----------------CCTGT-AATCCCC-------AGCTACTCTGG--AGGCCA-AGGC
HUMHBBc    -----CATGGTGGCATGCG----------------CCTGT-AGTCCC--------AGCTGCTCGGG--AGGCTG-AGGC
HUMHBBd    -----CATGATGGCAAGTG----------------CCTGT-AATCCC--------AGCTACTTGGG--AGGCTG-AGGA
HUMHBBf    -----CATGGTGGTGGCACATG-------------CCTGT-AATGCC--------AGCTACTTGGG--AGGCTG-AGGC
HUMHBBg    -----CGTGTGGTGCATG-----------------CCTGC-AGTCCC--------AGCTATTCAGG--TGGCTG-AGGC
HUMLDLR18a -----CGCGGTGGTGGGCA----------------CCTGT-AGTCCC--------AGCTACTCGGG---AGGCT-GAGG
HUMLDLR18b -----TATGGTGGCACGTG----------------CCTGT-GATCCC--------AGCTACTTGGG--AGGCTG-AGGC
HUMPOMC8   -----TGTGGTGGCGGGCG----------------CCTGT-AATCCC--------AGCTACTCTGG--AGCCTG-AGGC
HUMPOMC6   -----TGTGGTGGCATGTG----------------CCTGT-AGTCCT--------AGCTACTTGGG--AGGTT-------
HUMPOMC4   -----CGTGGTGGCGCATG----------------CCTGT-AATCCC--------AGCTACTCGGG--AGGCTG-AAGC
HUMPOMC2   -----CATGGTGGCATGCA----------------TCTGT-GGTTCC--------AGGTACTCAGG--AGGCTG-AAGC
HUMPOMC1   -----CGTGGTGGCGCACA----------------CCTGT-AGTCCC--------AGCTACTCGGG--TGGCTG-AGGC
HSMLVI2    -----CGTAGTGGCGGCG-----------------C-TGT-AGTCCC--------AGCTACTTGGG--AGGCTG-AGGC
HSREP10a   -----TGTGGTGGTGCATG----------------CCTGT-AGGTCC--------AGCTACTTGGG--AGACTG-AGGC
HSREP10b   -----TGAGGTGGTGTGCA----------------CCTGT-AGTCAC--------AGCTACTCGTG--AGGCTA-AGGT
HSREP10c   -----CATGGTGGCACATG----------------C-TGT-AATCCC--------AGCTACTCGGG--AGGCTG-AGGC
HSREP10d   -----TGTGGTGATGCATG----------------CCTGT-AGTCCC--------AGCTACTCG--------TG-AGGC
HSREP10e   -----TGTGGTGGTGGGTG----------------CCTGT-AGTCTC--------AGCTACTCGGG--AGGCTG-AGGC
HSREP10f   -----CGAGGTGCGGGCG----------------CCTGT-AGTCCC--------AGCTACTCGGG--AG-CTG-AGC-
HSREP10g   -----TCTGGTAGCATAAG----------------CCTGT-AGTCCC--------AGCTACTCAAA--AGGCTG-GGGC
HSREP10h   -----CATGGTGGTGGCACA----------------CCTGT-AGTACC--------AGTTACTAGGG--GGGCTC-AGGT
HSREP10j   -----AGTGGTGCGGGCG----------------CCTGT-AGTCCT--------AGCTACTCGGG--AGGCTA-GGCA
HUMADAGa   -----CGTGGTGGTGTATG----------------ACTGT-AATCCC--------AGCTACTCGGG--AGGCTG-AGGC
HUMADAGb   -----CACGGTGCCACGCG----------------CCTGT-AATCCC--------AGCTACTCGGG--ACGCTG-AGGC
HUMADAGc   TGTG-TGTGGTGGTGCACT----------------CCCGT-AGTCCC--------AGCTACTCAGG--AGGCTG-AGGC
HUMADAGd   -----CATGGTG-CGGGCG----------------CCTGT-AGTCCC--------ACGTACGCAGA--AGGCTG-AGGC
HUMADAGf   -----CATGGTGGTGCATG----------------CCTGT-AATTCC--------AGCTACTTGGG--AGGCTG-AGGC
HUMADAGi   -----CATGGTGGTGGGCG----------------CCTGT-AATCCC--------AGCTACTTGGG--AGGCTG-AGGC
HUMADAGja  ---------------GGTA----------------CCTAT-AATCCC--------AGCTACTCAGG--AGGCTG-AGGC
HUMADAGk   -----CATGGTGGTGGGTG----------------CCTGT-AGTCCC--------AACTACTCGGG--AGGCTG-AGGC
HUMADAGl   -----CGTGGTGGCGCACA----------------CCTGT-AATCCC--------AGCTACTTGGG--AGGCTG-AGAC
HUMADAGm   -----TGTGGTGGCACCTG----------------CCTAT-AGTCCC--------AGCTACTCCGG--AGGCTG-AGGC
HUMADAGn   -----TGTGGTGGCATGTG----------------CTTGT-AATCCC--------AGCTACTCAGG--AGGCTG-AGGC
HUMADAGo   -----CATGGTGGTGGGCA----------------TCTAT-AATCCC--------AGCTACTTGAG--AGGCTG-AGGC
HUMADAGp   -----CGTGGTGGCGTGCA----------------CCTGT-AGTCCC--------AGCTACTTGGG--AGGCTA-AGGT
HUMADAGq   -----TGTGGTGGTGTGTG----------------CCTGT-AATCCC--------AGCTACTCAGG--AGGCTG-AGGC
HUMADAGr   -----TGTGATGGTGGGTG----------------CCTGT-GATCCC--------AGCTACTTGGG--AGGCTA-AAGC
HUMADAGs   -----CGTGGTGGCATGCG----------------CCTGT-AATCCC--------AGCTACTTGGG--AGGCTG-AGGC
HUMADAGt   -----CATGGTGGTGCATG----------------CCTGT-AGTCCC--------AGCTACTTGGG--AGGCTG-AGGT
```

```
Sequence    320        330        340        350        360        370        380        390
PTAZGLO     AGGAGAA---TCACTTGA-A-CCCAGGAGGCA--GAGGTTG-CAGTGGGTCA------AAA-TCG--GCC---------
PTRE123A    AGGAAAA---TCGCTTGA-A-CCCGGGAGGTG--CAGGTTG-AGGTGAGCTG------AGA-TCA-CGCC---------
PTRE123B    AGGAGAA---TGGCGTGA-A-CCTGCGAGGTG--GAGCTTG-CAGTGAGCCG------AGA-TTG--GCC---------
PTGL01      AGGAGAA---TCGCTTGA-A-CCCAGGAGGTT--GAGGCTG-CAGTGAGCCAA-----AAA-TTG---CC---------
ATBOWL1     AGAAGAA---TCCCTTA--A-CCCG-GAGGCA--GAGGTTG-CAGTGAGCTG------AGA-CCA-CACC---------
ATHOWL1     AGGAAAA---TCACCTGA-A-TC---GAGGCA--GAGGTTG-CAGTGAGCTG------AGA-TCA-CACC---------
ATHOWL6     AGGAGAA---TCACTTGA-A-CCTGGGAGGTA--GAG-TTC---ATGAGCCG------AGA-TTG-CTCC---------
GCREG13     AAGAGAA---TTGCGTGAG--CCCAAGAGTTT--GAGATTG-CTGTGAGCTA------TGA-TGCCATGC---------
GCREG19     AGGAGGA---TTGCTTGAG--CTCAGGAGTTT--GAGGTTG-CTGTGAGCTA------TGA-TGA-TGCC---------
GCREG9      AAGAAAGAGCTTTGCTTGAG-CCCAAAAGTTT--GAGGTTG-CTGTGAGCTA------TGG-T----GCC---------
GGALU       AGGAGAA---TGGCGTGA-A-CCCGGGAGCG---GAGCTTG-CAGTGAGCCG------AGATCGC--GCC---------
ORAHBBEa    AGGAGAA---TCGCTTGA-A-TGCAGGAAGGG--GAGGTTG-CAGTGAGCCG------AAA-TCA-CGCC---------
ORAHBBEb    AGGAGAA---TCGTTTGA-A-CCCAGGAGGCG--AAGGTTG-CAATGAGCTG------AGA-TCG-TGCC---------
humblym1    GGCAGGA---TCACTTGA-A-CCTAGGAACATT-GAGGCTG-CAGTGAGCTA------TGA-TCT-TGCC---------
humc1ain1   AGGAGAA---TTGCATGA-A-CCCAGGAGGTG--GAGGTTG-CAGTGAGCTG------AGA-TCG-CGCC---------
humgast2    AGGAGAA---TCACTTGA-A-CCCGGGGGAGGCGGAGGTTA-TAGTGAGCCG------AGA-TCC-CACC---------
humins2     AGGAGAA---TCGCTTGA-A-CCTGGGAAGCA--GAGGTTG-CGCTGAGCCG------AGA-TGG-CACC---------
humrsa1     AGGACAA---TTGCTTGGAC-CCTGTGAGGCA--GAAGTTG-CAGTGAGCAA------AGA-XGG-CGCC---------
humrsa16    AGGAGAA---TGGCGTGA-A-CCCGGGAGGCA--GAGCTTG-CAGTGAGCCG------AGG-TCA-CACC---------
humrsa27    AGGAGAA---TCACTTGA---CXTGGGAGCCA--GAGGTTG-AAGTGAGCTG------AGA-TCA-GCT---------
humrsaold   GGGAGGA---GCACTTGA-G-TCXAGGAGGTG--GAAGCTG-CAGTGAGCCA------TGA-TGG-CACC---------
humrsap3    AGGAGAA---TGACTTGA-A-CCTGG-AGGTG--GAGCTTG-CAGTGAGCCAA-----CGA-TCG-CGCC---------
humrskpa1   GGGAGGA---TTGCTGGA-G-CCAGGTGGTAG---AGGCTG-CAGTGAGCCA------TGA-CTGG-TGC---------
hump5311    GGCAGGA---CTGCTCGA-G-CCGGGGAGGCA--AAGGCTG-CAGTAAGCCA------AGA-TCA-CGCC---------
humhbbrta   AGGAGAA---TTGCTTGA-A-CCCAGGAGGCA--GAGGTTG-CGGTGAGCCT------AGA-TTG-CACC---------
humhbbrtb   AGGAGAA---TCGCTTGA-A-CCTGGGAGGCA--GATGTTG-CAGTGAGCCT------GGA-TCA-TGCC---------
humapoe4a   AGGAGGA---TCGCTTGA-G-CCCAGAAGGTC--AAGGTTG-CAGTGAACCA------TGT-TCA-GGCC---------
humapoe4b   AGAA------TCGCTTGA-A-CCCAAGAGGCA--GAGGTTG-CAGTGAGCCA------AGA-TCG-TGCC---------
humapoe4c   AGGAGAA---TCGCTTGA-A-CCCGGGAGGCG--GAGGTTA-CAGTGAGCCG------AGA-TCT-CGCC---------
humapoe4d   AGGAGAA---TGGCGTGG-A-CCTGGGAGGCG--GAGCTTG-CAGTGAGCCG------AGA-TCC-CGCC---------
HUMTKRAa    GAGAGAA---TTGCTTGA-G-CCCAGGAGTTT--GAGACCA-GCCTGGGCA-------ACA-TAG-CAAG---------
HUMTKRAb    AGGAGAA---TCTCTTGA-A-CCCGGAAGGCA--GGGGTTG-CAGTGAGCTG------AGA-TCG-CTCC---------
HUMTKRAc    GGGACGA---TCACTTGA-G-CCCAAGGGTTT--GGGGTCA-C-GTAACAGT------AAG-CTA-TGAT---------
HUMTKRAd    GGGAGAA---TCGCTTGA------GACCCTT---GAGAATT-GGG--AGGT-------AGA-GAT-TGCAGGGAGCCGA
HUMTKRAe    AGGAGAGAA-TTGCTTGA-A-CCCAGGAAGCA--AAGTTTG-CGGTGAGATT------GTG-CCA-TCGT---------
HUMTKRAf    AGAAGGA-------TGGA-G-CCTGGGAGGTC--GAGGCTG-CAGTGAGTGG------TGA-TAG-CACC---------
HUMTKRAg    GGGAAAA---TCGCTTGA-A-TCCGGGTGGCA--GAGGTTG-CAGTGAGCCG------AGA-TCA-CACC---------
HUMTKRAh    AGAAC------TGCTTGA-A-CCCAAAAGGCG--GAGGTTG-CTGTGAGCTA------AGA-TCA-CGTC---------
HUMTKRAi    GGGAGAA---TCACTTGA-A-CCTGGGACATG--GAGGTTA-TAGTGAGCCG------AGA-CTG-CGCC---------
HUMTKRAj    AGGAGAA---TCGCTTGA-A-CCCGGTAGGCG--AAGGTTG-CAGTGAGCCA------AGA-TCG-CCCC---------
HUMTKRAk    ACAAGAA---TTGCTTGA-A-CCTGGGAGGCA--GAGGTTG-CAGTGAGCCA------AGA-TTA-TGCC---------
HUMTKRAl    AGGAGAG---TGGCGTGA-A-CCCAGGAGGCG--GAGCTTG-CAGTGAGCTG------AGA-TCA-CGCC---------
HUMALPPA    AGGATAA---TCGCTTGA-A-CCCGGGCGGCG--GAGATTG-CAGTGAGCCG------AGG-TCA-TGCC---------
HUMCEA      AGGAGAA---TCGCTTGA-A-CCCGGGAGGTG--GAGATTG-CAGTGAGCCC------AGA-TCG-CACC---------
HUMAPOC2A   AGGAGAA---TTGCTTGA-A-CCCAGGAAGCG--GAGGTTG-CAGTGAGCCA------GAT-TGT--GCC---------
HUMAPOC2B   TGGAGAA---TCGCTCAA-A-CCCGGAAGACA--GAGGTTG-CGGTGAGCCA------AAA-TTG-CGCC---------
HUMAPOC2C   GGGCGGA---TCACGAGG-T-CAGAAGTTCGA--GACCAGC-CTGGCCAGCA------TGG-TGAAACCC---------
HUMAPOC2E   ACACAGAA--TCGCTGGA-A-CCTGGGAGGCG--GAGGGTG-CAGTGAGCCG------AGA-TTG-CACC---------
HUMAGGa     AGGAGAA---TGGCGTGA-A-CCCGGGAGGCG--GAGCTTG-CAGTGAGCCG------AGA-TCG-CGCC---------
HUMAGGb     AGGAGAA---TCACTTGA-A-CCCAGGAGAGA--GAGGTTG-CAGTGAGCCG------AGA-TTG-TGCC---------
HUMAGGc     AGGAGAA---TCGCTTGA-A-CCTAGGAGGCG--GAGGTTG-CACTGAGCCG------AGA-TCA-CGCC---------
HUMHBA4a    AGGAGAA---TCGCTTGA-A-CCCAGGAGGTT--GAGGCTG-CAGTGAGCCA------AAA-CTT--GCC---------
HUMHBA4b    AGGAGAA---TCGCTTGA-A-CCCGGGAGGTG--GAGGTTG-AGGTGAGCCG------AGA-TCA-CGCC---------
HUMHBA4c    TGGAGAA---TCCTTTGA-A-CCAGGGAGTTG--GAGGTTG-CAGTGAGCCG------AGA-TCG-TGCC---------
HUMHBA4d    AGGAGAA---TGGCGTGA-A-CCTGCGAGGTG--GAGCTTG-CAGTGAGCCG------AGA-TTT--GCC---------
HUMHBA4e    AGGAGAA---TCATTTTA-A-CCAAGGAGGCA--GAGGTTG-CAGTGAGCTA------AGA-TCA-CACC---------
HUMFIXGa    AGGAGAA---TAGCTTGA-A-CCTGGGAGATG--GAGGTTG-CAGTGAGCTG------AGA-TCG-CACC---------
HUMFIXGb    GGGAGAA---TCACCTGA-G-CCTGGAAAGTC--GAGGCTG-CAGTGAATTG------TGA-TCA-CACC---------
HUMFIXGc    AGGAGAA---TGGCGTGA-A-CCGGGGAGGCG--GAGCTGG-CAGTGAGCCG------AGA-TCC-CGCC---------
HUMFIXGd    GGCGGAGG-TTGCAGTGA-G-CCAAGATCTCA--CCACTGC-TCTCCAGCCTGG----TGA-CAG-GGCA---------
HUMFIXGe    AGGAGAA---TCACTTGA-G-CCTGGGGAGCA--GGGGGTG-CGGGAGGTTG------CAG-TGA-GACA---------
HUMGHN      AGGAGAA---TCGCTTGA-A-CCCAGGAGGCG--GAGATTG-CAGTGAGCCA------AGA-TTG-TGCC---------
CHPRSAa     AGGAGAA---TCGCTTGA-A-CCTGGGAGGCT--GATGTTG-CAGTGAGCCT------GGA-TCA-TGCC---------
CHPRSAb     AGGAGAA---TTGCTTGA-A-CCCAGGAGGCG--GAGGTTG-CGGTGAGCCT------AGA-TTG-CACC---------
HUMTPAa     AGGAGAA---TTGCTTGA-A-CCT----GGTG--GAGGTTG-CAGTGAGCCG------AGA-TCA-CACC---------
HUMTPAb     AGGAGAA---TCACTTGA-A-CCTGGGAGGTG--GAGGTTG-CAGT-AGCCG------GGA-TCA-TGCC---------
HUMTPAc     AGGAAAA---TCGCTTGA-A-CCTGGGAGGTG--GAGGTTG-CAGTGAGCCA------AGA-TTG-TGCC---------
HUMTPAd     GGA-GGA---TCACTTGA-G-CCCAGGAGTTT--GAAGGTT-CAGTGACCTA------CGA-TCA-CACT---------
HUMTPAe     GAGAGGA---TCACTTGA-G-CCAGGGAGGTT--GAAACTG-CAGTGAGCTG------TGA-TCA-CGCC---------
HUMTPAf     AGGAGAA---TCGCTTAA-A-CCCAGGAGGCG--GAGGTTG-CAGTGAGCTG------AGA-TTG-CACC---------
HUMTPAg     GGGAGGA---TGGCCTGA-G-CCTGGGAGTTT--GAGGCTG-CAGTGAGCTG------TGA-TCA-TACC---------
HUMTPAh     AGGAGAA---TCGCTTGA-A-CCCGGGAGGCG--GAAGTTG-CAGTGAGCCA------AGA-TTG-CGCT---------
HUMTPAi     AGGAGGA---TCTCTTGA-G-CCTGGTGGGTC--AAGGCTG-CAGTGAACCA------TGT-TCA-TGCC---------
HUMTPAj     ATGAGAA---TGGCATGA-A-CCCGGGAGGCG--GA-GCTG-TAGTGAGCCG------AGA-TCA-CACC---------
HUMTPAk     GGGAGGA---TCGCTTAA-G-TTCAGGAGGTT--GGGACTT-CAGTGAGATA------TGA-TTA-CGCC---------
HUMTPAl     AGGAGGA---CTGCTTGA-G-CCTGGGAGGCA--GAGGTTG-CAGTGAGCCA------AGA-CAG-CACC---------
HUMTPAm     AGGAGAA---TCACTTGA-A-CCCAGGAGGTG--GAGGTTG-CAGTGAGCCG------AGA-TTG-TGCC---------
```

```
Sequence    320        330        340        350        360        370        380        390
HUMTPAn     AGGAGTA---TTGCTTGA-A-CCCAGGAAGTG--GGGGTTG-CAGTGAGCCG------AGA-TTG-TACC---------
HUMTPAo     AGAAGGA---TCACTTGA-G-CCCAGGAGTTG--GAGGCTG-CAGTAATCTA------CGA-TTA-TGCC---------
HUMTPAp     TGAAGA----TTGCCCAA-A-TCCAGGAAACG--GAGGTTG-CAGTGAGAGG------AGA-TG--CGCC---------
HUMTPAq     AGGAGAA---TTGCTTGA-A-CCTGGGAGATG--GAGGTTG-CAGTGAGCCG------AGA-TCA-TGCC---------
HUMTPAr     GGGAGGA---TCGCTTGA-G-CCCAGGAGGTT--GAGGCCA-CAGTGAGCAA------TGA-TTG-TGCC---------
HUMTPAs     AGAAGAA---TCGCTTGA-A-CCCAGGAGGTG--GAGATTG-CAGTGAGCCA------AGA-TCG-CGC----------
HUMTPAt     AGGAGAA---TGGCATGA-A-CCC-GGGAGGCG-GAGCTTG-CAGTGAGCCG------AGA-TCC-CGCC---------
HUMTPAu     AGGAGAA---TCGCTTGA-A-CCCGGGAGGCA--GAGGTTG-CAGTGAGCCG------AGA-TCG-CGCC---------
HUMTPAv     --GAGAA---TTGCTTGA-A-CCCCGGAGGTG--GAGGGTG-CAGTGAGCTG------AGA-TTG-CGCC---------
HSALUAGPa   AGGAGAA---TCGCTCAA-A-CCTGGGAGGTG--GAGCTTG-CAGTTAGCCG------AGA-TCT--GCC---------
HSALUAGPb   GGGAGGA---TTGCTTGA-G-CCCAGGAGTTT--GAGGCTG-CAGTTGAGCT------ATGATTCATACC---------
HSALUAGPc   GGGAGGA---TCTCTTGA-G-CCTGGGTGGCA--GAGGTTG-CAGTGAGCCA------AGA-TCC-CACC---------
HSALUAGPd   GGGAGAA---TGGCGTGA-A-CCTGGGAGGCA--GAGCTTG-CAGTGAGCTG------ATA-TGG-CGCC---------
HSALUAGPe   AGGAGAA---TCGCTTGA-A-CCCAGGAGGCA--CAGGTTG-CAGTGAGCCG------AGA-TTG-AGCC---------
HSALUAGPf   A---------TTGCTTGA-G-CCCAGGAGTTC--AAAGCTG-CAGTAAGCTG------TGA-TTT-TGCC---------
HUMHBB51    AGGAGAA---TCACTTGA-A-CCTGGGAGGTG--GACTTTG-CAGTGAGCTG------AGA-TTG-TGCT---------
HUMHPARS1a  AGGAGAA---TCACTTGA-A-CCTGGGAGGTG--GAGGTTG-CAGTGAGCCG------AGA-TCA-TGCC---------
HUMHPARS1b  AGGAGAA---TTGCTTGA-A-CCTCGGAGGTA--GAAGTTG-TGGTGAGCTG------AGA-TCA-TGCC---------
HUMHPARS1c  AGGAGAA---TTGCTTGA-A-CCGGGGAGGGG--GAGGTTG-CGATGAGCTG------AGA-TCG-CGCC---------
HSIGVKA2a   AGGAGAA---TCGCTTGA-A-CCCGGGAGGAG--GAGGTTG-CAGTGAGCTG------AGA-TAG-CATC---------
HSIGVKA2b   AGAAGAA---TCGCTTGA-A-CCCGGG----------------------------------------------------
HUMHBBa     AGGAGAA---TCGCTTGA-A-CCGGGGAGGT---GGAGTTT-GCACTGAGCAG-----AGA-TCA-TGCC---------
HUMHBBb     AGGAGAA---TTGCTTGA-A-TGCAGGAAGGG--GAGGTTG-CAGTGAGCCG------AAA-TCA-TGCC---------
HUMHBBc     AGGAGAA---TCGTTTGA-A-CCCAGGAGGCG--AAGGTTG-CAGTGAGCTG------AGA-TAG-TGCC---------
HUMHBBd     AGGAGAA---TTGCTTGA-A-CCTGGAAGGCA--GGAGTTG-CAGTGAGCCG------AGA-TCA-TACC---------
HUMHBBf     AGGAGAA---TCGCTTGA-A-CCTGGGAGGCA--GATGTTG-CAGTGAGCCT------GGA-TCA-TGCC---------
HUMHBBg     AGGAGAA---TTGCTTGA-A-CCCAGGAGGCG--GAGGTTG-CGGTGAGCCT------AGA-TTG-CACC---------
HUMLDLR18a  CAGGAGAA---TGGTGTG-AACCCGGGAAGCG--GAGCTTG-CAGTGAGCCG------AGA-TTG-CGCC---------
HUMLDLR18b  AGGAGGA---TCGCCTGA-G-CCCAGGAGGTG--GAGGTTG-CAGTGAGCCA------TGA-TCG-AGCC---------
HUMPOMC8    AGGAGAA---CCGCTTGA-A-CCTGGGAGGTA--GAGGTTG-CAGTGAGCCA------AGA-TCA-CACC---------
HUMPOMC6    ----------CACTTGA-G-GCCAGGAGTCT--GACGACA-CAGTAAGCTA------TGA-TCA-CACC---------
HUMPOMC4    AGGAGAA---TCGCTTGA-A-CCCGGGAGGCG--AAGGTTG-CGGTGAGCAG------AGA-TCA-CCCC---------
HUMPOMC2    ATGAGGA---TCACTTGA-G-CCCAGGAGGTG--GAGGCTG-CAGTGAGCAG------TGT-TCA-CACT---------
HUMPOMC1    AGGCGAA---TCGCTTGA-A-CCTGGGAGGCG--GAGGTTG-CAGTGAGCCA------AGA-TCG-CGCC---------
HSMLVI2     AGGAGAA---TGGCGTGA-A-CCCGGGAGGCG----GCTTG-CAGTGAGCCG------AGA-TCC-CGCC---------
HSREP10a    AAGAGGA---TCACGTGT-G-CCCAGAGGTG---GAGGTGA-CAGTGAGCCG------TGA-TCA-TGCC---------
HSREP10b    AGGATAA---TCACTTGA-G-CCTGG--------GAGGTTG-AGGCTGCCATGAGCTATGA-TGG-TGCC---------
HSREP10c    AAGAGAA---TTGCTTGA-A-CGGCGA-------GAGGTTG-TGGTGAGCCA------AGA-TCA-TGAC---------
HSREP10d    AGGAGAA---TCGCTTGA-A-TCTGGGAGGTG--GAGGTTG-CG------------------TC---------------
HSREP10e    AGGAGAA---TGGCGTGA-A-CCC---AG-----GAGGTGG-AGCTTGCAGTGAGCCGAGA-TCG-CGCC---------
HSREP10f    ----------------------------------------CT-GCAGTGAGCCGAGA-TCG-TGCC---------
HSREP10g    AGGAGGA---TCACTTGA---CCTGGGAAGTC--AAGGCTC-AGTGAGCCA-------TGA-TTG-CACC---------
HSREP10h    GGGAGAC---TTACTTGA-G-CCCAGGAGTTG--GAGGCTG-CAGTAGCTA-------TGA-TCA-CACC---------
HSREP10j    GGAGAA----TGGCGTGA-A-CCCAGGAGGCG--GAAGCT--TGGCAGTGA--GCCG-ACAT-TGT-GCC---------
HUMADAGa    AGGAGAA---TCGCTTGA-A-CCTGGGAGGTG--GAGGTTG-TGGTGAGCCA------AGA-TTA-CACC---------
HUMADAGb    CG-AGAA---CTGCTTAA-AATCCAGGAGGTG--GAGGTTG-CAGTGAGCCG------AGA-TTT-CGCC---------
HUMADAGc    GGGAGGA---TCGCTTGA-G-CCCAGGAGGTT--GAGGCTG-CAGTGAGGTG------TGA-TGG-TGCC---------
HUMADAGd    AGGAGAA---TGGCATAA-A-CCCGGGAGGTG--GAGCTTG-CAGTGAGCCG------AGA-TCT-CGCC---------
HUMADAGf    AGAAGAA---TCACTTGA-A-CCTGGGAGGCG--AAGGTTG-CAGTGAGCCG------AGG-TCG-TGCC---------
HUMADAGi    ACGAGAA---TCCCTTGA-A-CCTGGGAGGCA--GAGGCTG-CAATGAGCTG------AGA-TCT-TGCC---------
HUMADAGja   AGGAGAA---TTGCTTGA-A-CCTGGGGGCA---GAGATTC-CAGTGAGCCG------AGA-TTA-AATC---------
HUMADAGk    AGGAGAA---TCGCTTGA-A-CCCGGGAGGTG--GAGGTTT-CCGTGAGCTG------AGC-TGG-AGCC---------
HUMADAGl    AGGAGAA---TCACCTGA-A-CCTGGGAGGCA--GAGGCTG-CAGTGAGCCG------AGA-TTA-CGCC---------
HUMADAGm    ACAAGAA---TTGCTTGA-A-CCCAGGAGGTG--GAGGTTG-CAGTGAGAGG------AGA-TCA-CGTC---------
HUMADAGn    AGGAGAA---TCACTTGA-A-CCCAGGAGGTG--GAGGTTG-CAGTGAGCCG------AGA-TCA-TGCC---------
HUMADAGo    AGGAGAA---TCACTTGA-A-CCAGGGAGGTG--AAGGTTG-CAGTGAGTCG------AGG-TCG-TGCC---------
HUMADAGp    GGGAGGA---CCACCTGA-G-CCTGGAAGTC---AAGGCTA-CTGCGGGCCA------AGA-TTG-CACC---------
HUMADAGq    AGGAGAA---TCACTTGA-A-CCAGGGAGTCA--GAGGCTG-TGGTGAGCCG------AGA-TCA-TGCT---------
HUMADAGr    AGGAGAA---TCCCTTAA-A-C-TGGGAGGTG--GAGGTTG-CAGTGAGCTG------AGA-TCG-CACC---------
HUMADAGs    AGACGAA---TCACTTGA-A-CCCAGGAGGCA--GAGGCTG-CAGTGAGCTG------AGA-TGG-CGCC---------
HUMADAGt    GGGAGAA---TTGCTTGA-G-TCCAGGAAGTC--AAAGCTG-CAGTGAGCTG------TGA-TAA-TGCG---------
```

```
Sequence    400       410       420       430       440       450       460       470
PTAZGLO     ---------ACCACACTCCA---------------GTCTGGGAGACA-------GAGAAAGACTCCATCTCAGAAAC--
PTRE123A    ---------ACTGCACTCCA---------------GCCTGGGCAACAA------GAGCAAAACTCCATCTCAAAAAAT-
PTRE123B    ---------ACTGCA-TCCC---------------GCCTGGGTGACA-------GAGCGAGACTCCGTCTCAAAAATAA
PTGL01      ---------ACTGCACTCCA---------------GCCGGGGTGACAG------AGCAAGGG--CCTATCTCAAAAACA
ATBOWL1     ---------ACTGCAATCCA---------------GCCTGGGCGGCAG------AG-TAAGACTCCATCTCAAAAAAA-
ATHOWL1     ---------ACTGTACTCCA---------------GCCTCGGCGGCA----------GAAGACTC-GTCTCAAAAAAA-
ATHOWL6     ---------ACTGCA-TCCA---------------GCCTGGGCAGCA---------AGTGAGATTCATTTCAAAAAAA-
GCREG13     ---------ACTCTACCAAG---------------CCAAGGGTGACA------AA-GTGAGACTCTGTCTCAAAAAAAA
GCREG19     ---------ACACAGCACTCT---------------AGCTCGGGTGAC------AGAAAGGGACTCTGCCTCAAAAAAAA
GCREG9      ---------ACAGCACTCTA---------------AGAG--GGGACA------CAGTAAGAC-TTTGTCTCAGAAAAAA
GGALU       ---------ACTGCACTCCA---------------GCCTGGGCGACA------GAGCGAGAC-TCCGTCTCAAAAAAAA
ORAHBBEa    ---------ACTGCACTCCA---------------GCCTGGGCAATA------AGAGCGAAACTCCATCTCAAACAAAA
ORAHBBEb    ---------ATTGCACTCCA---------------GTCTGG-CAAC-------AGAGTGAGACTCCGTCTCAAAAAAAA
humblym1    ---------ACTGCACTCCA---------------TACTGCATGACA-------GAGCAAGACCCTGTCTCTTAGAAAA
humclain1   ---------ACTGCACTCCA---------------GCCTGGGCAACA------GTGCGAGAC-TCCATCTCAAAAAAAA
humgast2    ---------ACTGCACTCCA---------------GCCTGGGCAACA------AGAGTGAAACTCTGTCTAAAAAAAAA
humins2     ---------ATTGCACTCCA---------------GCCTGGGCAACG------AGAGCGAAACTCCGTCTCAAAAAAA-
humrsa1     ---------AXXGTACCCCA---------------CCCXGTGCAACA----GGACAAXGAGATTCTGTCTA-AAAAAAA
humrsa16    ---------ATTGCATTCCA---------------GCCTGGGCGAAA------GAGTGAGAC-TCTGTCTCAAAAAAAA
humrsa27    ---------ACTGCACTCC----------------GCXTGGCCAACA------AAGTTAGAC-TCCGCCTAAAAAAAAA
humrsaold   ---------ACTACACTCCA---------------GCCAGGGCAACA------GAGAGAGAC-TCTGTCTCAAAAAAGA
humrsap3    ---------ACTGTCATCAT---------------CAT-GGGTGACA------GAGAGAGACTCCCGTCTCAAAAAAAA
humrskpa1   ------------------------------------CTGGC-AAGA------GAGCAAGACAA---TCTCAAAAAGAA
hump5311    ---------ACTCCACTCCA---------------GCCTGGGCAAC-------AAAGCGAGACCCAGTCTCAAAGAAAA
humhbbrta   ---------ATTGCACTCTA---------------GCTTGGGCAATA------GGGATGAAACTCCATCTCAGAAGAGA
humhbbrtb   ---------ATTGCACTCCA---------------GCCTGGGCAACC------ATAGCAATGCTCCATCTCAGAAAAAA
humapoe4a   ---------GCTGCACTCCA---------------GCCTGGGTGACA-------GAGCAAGACCCTGTTTATAAATACA
humapoe4b   ---------ACTGCACTCTA---------------GCCTGGGTGACA-------GAGCCAGACTCCGTCTAAAAAAAAA
humapoe4c   ---------ACTGCACTCCA---------------GCCTCAGCAA---------GAGGGAGACTGT-CCTCAAAAAA--
humapoe4d   ---------ACTGCACTCCA---------------GCCTGGGCGACA-------GAGCGAGACTTCATCTCAAAAAAA-
HUMTKRAa    ------------------------------------------------------ATCCTACCTCTAGAAGAACA
HUMTKRAb    ---------ACTGCACTCTA---------------ACCTAGGCAACA-------GAGCGAGACTCCACCCCAAAAAGAA
HUMTKRAc    --TGA-GCCACTGCACTCCA---------------GCCTGGGTGACA-------GAGAGAGACCCTGTCCCAAAAACAA
HUMTKRAd    GATGGCGCCACTGCACTCCA---------------GCCTGGGTGACA-------GAGCAAGACTCTGTTTCATAAAAAA
HUMTKRAe    ---------ACTCCG--------------------GCCTGGGCAACA------AGAGCAAAACTCCGTCTCAAAAAAAA
HUMTKRAf    ---------ACTGCACTCCA---------------GCCCGGGCGACA-------AGGCCAGACCCTGTCTCAAAAAAAA
HUMTKRAg    ---------ATTGCACTCCA---------------GCCTGGGCGACA-------GAGTAAGACTCCGTCTCAAAAAAAA
HUMTKRAh    ---------ACTCCACTCCA---------------GCCAGGGCAACA-------GAGCAAGACTCCATCTCAAAAAAAA
HUMTKRAi    ---------ACTGCACTCCA---------------GCCTGGGCCACA-------GAGCAAGACT-GTTTTTAAAAACAA
HUMTKRAj    ---------ATTGCACTCCAA--------------GCCTGGGCAACA---ACAAGAGCAAAACTCAGTCTCAAAACAAA
HUMTKRAk    ---------ACTGCACTCCA---------------GCCTGGATGACA-------GAGCGAGACTCTGTCTCAAAAAAAA
HUMTKRAl    ---------ACTGCACTCCA---------------GCCTGGGCGACA-------GAGCGAGACTCCGTCTCAAAAAAAA
HUMALPPA    ---------ACTGCACTGCA---------------GCCTGGGCGACA-------GAGCGAGATTCTGCCTCAAAAATAA
HUMCEA      ---------ACTGCACTCCA---------------GTCTGG-CAACA-------GAGCAAGACTCCATCTCAAAAAGAA
HUMAPOC2A   ---------ACTGTACTCTA---------------GCCTGG-TGACA-------GAGCAAGACTCAGTCTTGGCGGGAA
HUMAPOC2B   ---------ATCGCACTCCA---------------GC-CTGGGCGAGA------GAACAAGACCTTGTCTCGGAAAAAA
HUMAPOC2C   ---------ATCTCTACTAAAAATACAAAATATTAGC-C-GGGCAT----GGTGGCAGGTGCTTGTGATTC---------
HUMAPOC2E   ---------ACTGCTGTCCA---------------GCCTCCCTGACA-------GAGTGAAACCCTGTCTCAAAAAAAA
HUMAGGa     ---------ACTGCACTCCA---------------GCCTTGGCGACA-------GAGCGAGACTCCGTCTCGGAAAAAA
HUMAGGb     ---------ATTGCACTTAA---------------GCCTGGGCGACA-------GAGTGAGACTCCATCTCAAAAAAAA
HUMAGGc     ---------ACTGCACTCCA---------------GCCTGGGCGACA-------GAGCGAGACTCCATCTCAAAAAAAA
HUMHBA4a    ---------ACTGCACTCCA---------------GCCGGGGTGTCA-------GAGCAAGGGCCTATCTCAAAAACAA
HUMHBA4b    ---------ATTGCACTCCA---------------GCCTGGGCAACAA------GAGCAAAACTCCGTCTCAAAAA---
HUMHBA4c    ---------ACAGCACTCTA---------------GCCTGA-CGACA-------CAGCGAGACTCTGTCTCAAAAAAAT
HUMHBA4d    ---------ACTGCACTCCA---------------GCCTGGGTCACA-------GAGCGAGACTCCGTCTCAAAAAAAA
HUMHBA4e    ---------ATTGCACTCCA---------------GCCTGGAAAACAA------CAGCGAAACTCCGCCTCAAAAAAAA
HUMFIXGa    --------------ACTGCA-----------------------------------------------------------
HUMFIXGb    ---------ACTGCACTTCA---------------GCCTGAGTGACA-------GAGTAAGACCCTATCTCAAAAAACA
HUMFIXGc    ---------ACTGCACTCCA---------------GCCTGGGCGACA-------GAGCGAGACTCCGTCTCAAAAAAAA
HUMFIXGd    ------------------------------------------------------AGACTCCGTCAAAAAAAAAA
HUMFIXGe    ---------AGATCGCACCA----GTGCACTCCCCATCCTGGGTGACA------GAGTGAGACTCTGTCTCAAAGAAAA
HUMGHN      ---------ACTGCACTCCA---------------GCTTGGTTCCC---------GAATAGACCCC-------------
CHPRSAa     ---------ATTGCACTCCA---------------GCCTGGGCAACCA------TAGCAATACTCCATCTTAGAAAAAA
CHPRSAb     ---------ATTGCACTCTA---------------GCTTGGGCAATAG------GGATGAAACTCCATCTCAGAAGAGA
HUMTPAa     ---------ACTGCACTCTA---------------GCCTGGGCGACA-------GAGCAAGACTCTGTCTCAAAAAAAA
HUMTPAb     ---------ATTGCACTCTA---------------GCCTGGGCAAAAA------GAGTGAAACTCCGTCTCAAAAAAAA
HUMTPAc     ---------ACTGCACCCCA---------------GCCT--GTGAAA-------AAGCGAGACTCCATCTTAAAAAAAA
HUMTPAd     ---------GCTGCACTCCA---------------GCCTGGGTGATG-------GAGTCAGACCTTGTCTCTAAATTAT
HUMTPAe     ---------ACTGCACTCCA---------------GTCTGGGTGACT-------GGGCGAGACCCTGTCTCAAAACAAA
HUMTPAf     ---------GTTGCAATCCA---------------GCCTGGGCAACA-------GAGTGAAACTCCATCTCAGAAAAAA
HUMTPAg     ---------ACTGCACTCCA---------------GCTCGA--GACA-------GAGCAAGATCCTGTCTCAAAAAAAA
HUMTPAh     ---------ACTGCA--------------------GCCTGGGCGACA-------GAGCAGGACTTCGTCTCAAAA----
HUMTPAi     ---------ACTGCACTCCA---------------GTCTGGATGACA-------GAGCGAGACCTAGTCTCAAAAAAAA
HUMTPAj     ---------ACTGCACTCCA---------------GCCTGGGCGACA-------GAGCGAGACTCCATCTTAAAAGAAA
HUMTPAk     ---------AATGCACTCCA---------------GCCTGGGTGACA-------AACTGAGATCCTGTCTCAAAAAAAA
HUMTPAl     ---------ATTGGACTCCA---------------GCCTGGGTAACA---------------GAGACCCTCGAAAAAAA
HUMTPAm     ---------ACTTCACTCCA---------------GCCTGGGCGACA-------GAGTGAGACTTTGTTTCAAAAATAA
```

```
Sequence     400       410       420       430       440       450       460       470
HUMTPAn    ---------ACTGCACTCCA---------------GCCTGGGCGACA-------GAGCAAGACTCCATCTCAAAATAAA
HUMTPAo    ---------ACTGCATTTCA---------------ACCTCAGTGACA-------GGGCAAGCCCTCACCTCTAAAACAA
HUMTPAp    ---------GCTGCACTCCA---------------GCCTGGGATAAA-------CAGCGAGACTCTGTCAAAAAAAAAA
HUMTPAq    ---------ACTGCACTCCA---------------GCCTGGGTGATA-------GAACAAAAAACTGTTTCAAAAAAAA
HUMTPAr    ---------ACTGCACTCCA---------------GCCTGGGTGACA---------GTGAGATCCTGCCTCAAAATAAA
HUMTPAs    ---------ATTGCACTCCA---------------GCCTGGGCAACA-------------------------AAAG
HUMTPAt    ---------ACTGCACTCCA---------------GCCTGGGCAACA-------GAGCGAGACTCCGTCTCAAAAAAAA
HUMTPAu    ---------ATCGCGCTTCA---------------GCCTGGGCGACAA------GAGCGGAACTCGATCTCAAGAAAAA
HUMTPAv    ---------ACTGCACTCCA---------------GCCTGGGTGGGTG---ACAGAGTGAGACTCTGTCTCAAAACAAA
HSALUAGPa  ---------ACTGCACTCCA---------------GCCTGG-TGACA---------GAGCAAGACTGTCTAAAAAACAA
HSALUAGPb  ---------ACTGCACTCCA---------------ACTTGAAGA-----------------------GCCTGAAAAAAAA
HSALUAGPc  ---------ACTACACTCCA---------------GCCTGGGCGACG------AGAGTGAGATGCCATTTCAGAACAAA
HSALUAGPd  ---------ACTGCACTCCA---------------GCCTGGTGACAG--------AGTGAGACTCCGTCTCAAGGAAAA
HSALUAGPe  ---------ATTGCACTCCA---------------GCCTGGGGGACG------GGAGTGAAACTCCGTATCAAAAAAAA
HSALUAGPf  ---------ACTACACTCCA---------------GCCTGTATGATG-----GAGCAAGAGACCCTGTCTCAAAAATAA
HUMHBB51   ---------ACTGCACTCCA---------------GCCTGGCGACAG--------AGTGAGACTCTATCTTAAAAAAAA
HUMHPARS1a ---------ACTGCACTGCA---------------GCCTGGTGACAA--------AGCGAGACTCCATCTCAAAAAAAA
HUMHPARS1b ---------ATTGTACTCCA---------------GCCTGGGCAACA------GGAGTGAAACTCTGTCTCAAAAAAAA
HUMHPARS1c ---------ACT-CACTCCA---------------TCCTGGTGACAG--------AGTGAGACTCTGTCTCACAAAAAA
HSIGVKA2a  ---------ATTGCACTCCA---------------GCCTGGGCAACA------AGAGTGAGACT-TCTCTAAAATAAAA
HSIGVKA2b  ------------------------------------------------------------------------------
HUMHBBa    ---------ATTGCACTCCA---------------GCCTCCAGAGCG------------AGACTCTGTCTAAAGAAAAA
HUMHBBb    ---------ACTGCACTCCA---------------GCCTGGGCAATA------AGAGCGAAACTCCATCTCAAACAAAA
HUMHBBc    ---------ATTGCACTCCA---------------GTCTGGCAACAG--------AGTGAGACTCCGTCTCAAAAAAAA
HUMHBBd    ---------ACTGCACTCCA---------------GCCTGGGTGACA-------GAACAAGACTCTGTCTCAAAAAAAA
HUMHBBf    ---------ATTGCACTCCA---------------GCCTGGGCAACC------ATAGCAATGCTCCATCTCAGAAAAAA
HUMHBBg    ---------ATTGCACTCTA---------------GCTTGGGCAATA------GGGATGAAACTCCATCTC---------
HUMLDLR18a ---------ACTGCAGTCCG---------------CAGTCTGGCCTGGGCGACAGAGCGAGACTCCGTCTCAAAAAAAA
HUMLDLR18b ---------ACTGCACTCCA---------------GCCTGGGCAACA-------GATGAAGACCCTATTT----------
HUMPOMC8   ---------ACTGCACTCCA---------------CACTCCAGCCAGGGCAACAGAGCAAGACTCCGTCTCAAAAAAAT
HUMPOMC6   ---------ATTGCACTCCA---------------GTCTGGGTAACA-------GAATGAGACCTTGTCTCAAAACAAA
HUMPOMC4   ---------TTTGCTCTCCA---------------GCATGGGCAACA------AGAGCGAAACTCCGTCTCAAAAAAAA
HUMPOMC2   ---------GCTGCATTCCA---------------GCCTGGGCAACA-------GAGTGAGACCCTAATTCAAAAACA-
HUMPOMC1   ---------ACTGCACTCCA---------------GCCTGGCGACAC--------AGCAAGACTCCATCTAAAAAAAAA
HSMLVI2    ---------ACTGCACTCCA---------------GCCTGGGCGACA-------GAGCGAGACTCCGTCTCAAAAAAAA
HSREP10a   ---------ACTGCACTCTA---------------GCCTGGGTGACA-GCAGGACAGACAGACCCTGTCTGAAAAAAAA
HSREP10b   ---------A-TGCCATCCA---------------GCCTGAGCGATG-------GTGTGAGACCCATCTCTAAAAAAGA
HSREP10c   ---------AGTGCACCCCA---------------GCGTGGGCAACA------AGAGTGAAACTTCGTCTCCAAAAAAA
HSREP10d   ---------ACTGCACTCCA---------------GCCTGGGCAACG-----------GAGACTCCGTCTCAAAAAAAA
HSREP10e   ---------ACTGCACTCCA---------------GCCTGGGCGACA-------GAGTGAGACTTTGTCTCAAAAAAAA
HSREP10f   ---------ACTGCATC-CA---------------GCCTGGGCGACA----------GCAAGCTCCGTCTCAAAAAAAA
HSREP10g   ---------TCTGTACTCCA---------------GCCTGGGCAACA-------TACCAAGACCCTGCCTGAAAATGAT
HSREP10h   ---------ATTCAGCCTGG---------------GAGACAGAGCAA-------------GGCCCTGTCTCTAAAACAA
HSREP10j   ---------ACTGCACTGCA---------------GCCTGGGCAACA--------GAGCGAGCTCCGTCTCAAAATAAT
HUMADAGa   ---------ATTGCATTCCA---------------GCCTGGGCAACA------AGAGCGAAACTCCATCTCATGAAAAA
HUMADAGb   ---------ACTGCACTCCA---------------GCCTGGGCGACA-------GAGCAAGAGTCCATCTCAAAAAAAC
HUMADAGc   ---------ACTGACCTTCA---------------GCCTGGGAGACA-------GAGCGAGACCCTGTCTCAAAAAAAA
HUMADAGd   ---------ATTGCACTCCA---------------GCCTGGGTGACA-------GAGTGAGACTCTGTCTCAAAAAAAA
HUMADAGf   ---------ATTGCACTCCA---------------GCCTGGGCAACA------AGAGCGAAACTCTGTCTAAAAAAAGA
HUMADAGi   ---------ACTGCACTCCA---------------GCCTGGGCAACA-------GAGCCAGACTCCATCTCAAAAAAAA
HUMADAGja  ---------ACTTTACTCCA---------------GCCTGGGTGAAA-------GTGCAAAACTCCACCTCAAAAAAAA
HUMADAGk   ---------ACTGCACTCCA---------------GCCTGGGCAACA-------GAGTGAAACTCCGTCTTAAAAAAAC
HUMADAGl   ---------ACTGCACTCCA---------------GCCTGGGCGACA------AGAGCGAAACTGCATCTCAAAAATAA
HUMADAGm   ---------ACTGCACTCCA---------------GCCTGGGAGACA-------GAGCGAGACTCCATCCGTCTCAAAA
HUMADAGn   ---------ACTGCACTCCA---------------GCCTGGGCGACA-------GAGCAAGACTCTATCTCAAAAGAAA
HUMADAGo   ---------ATTGCACTCCA---------------GCCTGGGCAACA------AGAGCAAGACTCCGTCTCAAAAAAAC
HUMADAGp   ---------ACTGCACTCCA---------------GCTTGGGTGACA-------GAGCAAGACCCTGTCTCAAAAAAAT
HUMADAGq   ---------ACTGCACTCCA---------------GCCTGG-TGACA-------GAGCAAGACTCTGTATCACAAAAAA
HUMADAGr   ---------ACTGCACTTCA---------------GCCTGGGCAACA-------GAGTGAGACTCTGTCTCAATAAATA
HUMADAGs   ---------ACTGCACTCCA---------------GCCTGGGCGACA-------GAGCAAGATTCTGTCTCAAAAAAAA
HUMADAGt   ---------ACTGCACTCCA---------------GCCTGGGTGACA-------GAGGGAGACCCTGTCTCAAAAAAAA
```

```
Sequence      480        490        500        510        520        530        540
PTAZGLO     ----------------------------------------------------------------------*
PTRE123A    ----------------------------------------------------------------------*
PTRE123B    ATAAATAAATAAATAAAAATAAAAATAAATAAATAAATAAATAAAT----------------------*
PTGL01      AAAACAAAGCAAAAAAAAAAAAAAAAAAAACCCCCAAAAACAAAAAAACACTTGGGGGATGTAGATAACGC*
ATBOWL1     ----------------------------------------------------------------------*
ATHOWL1     ----------------------------------------------------------------------*
ATHOWL6     ----------------------------------------------------------------------*
GCREG13     AAAA------------------------------------------------------------------*
GCREG19     AAAA------------------------------------------------------------------*
GCREG9      AAAATGTAC-------------------------------------------------------------*
GGALU       AAAAAAAAAAAAAA--------------------------------------------------------*
ORAHBBEa    ACAAAAACAAACAAAAACAAAAAT----------------------------------------------*
ORAHBBEb    AAAAAAAAAAAAAAAAAGAAAAAAGAAAAGAAAAAGAAAAAAGAAAAAGAAA----------------*
humblym1    AA--------------------------------------------------------------------*
humc1ain1   ----------------------------------------------------------------------*
humgast2    ----------------------------------------------------------------------*
humins2     ----------------------------------------------------------------------*
humrsa1     A---------------------------------------------------------------------*
humrsa16    ----------------------------------------------------------------------*
humrsa27    ----------------------------------------------------------------------*
humrsaold   A---------------------------------------------------------------------*
humrsap3    ----------------------------------------------------------------------*
humrskpa1   ----------------------------------------------------------------------*
hump5311    ----------------------------------------------------------------------*
humhbbrta   ----------------------------------------------------------------------*
humhbbrtb   ----------------------------------------------------------------------*
humapoe4a   ----------------------------------------------------------------------*
humapoe4b   ----------------------------------------------------------------------*
humapoe4c   ----------------------------------------------------------------------*
humapoe4d   ----------------------------------------------------------------------*
HUMTKRAa    AAAACAAAAC------------------------------------------------------------*
HUMTKRAb    AGAA------------------------------------------------------------------*
HUMTKRAc    ACAACAAACA------------------------------------------------------------*
HUMTKRAd    TAA-------------------------------------------------------------------*
HUMTKRAe    AAAAAA----------------------------------------------------------------*
HUMTKRAf    AAGGGGGA--------------------------------------------------------------*
HUMTKRAg    AAAAAAACCCAAACCCAC----------------------------------------------------*
HUMTKRAh    TAAAAATAAA------------------------------------------------------------*
HUMTKRAi    ACAAAAAAC-------------------------------------------------------------*
HUMTKRAj    ACAAAACAA-------------------------------------------------------------*
HUMTKRAk    TAGACAAA--------------------------------------------------------------*
HUMTKRAl    AAAA------------------------------------------------------------------*
HUMALPPA    AC--------------------------------------------------------------------*
HUMCEA      AAGAAAAGAAGACTCTGACC--------------------------------------------------*
HUMAPOC2A   AAAAGAATGGAAAAATTTAAAAAA----------------------------------------------*
HUMAPOC2B   AAAAAAAGAATTGTCGA-----------------------------------------------------*
HUMAPOC2C   ----------------------------------------------------------------------*
HUMAPOC2E   AAAAAATAGAGAAA--------------------------------------------------------*
HUMAGGa     ATAAAAACAAAAAATAAAAAAATG----------------------------------------------*
HUMAGGb     AAAAAAAAAAAAAACGAAAAGAAAAGAAAAAG--------------------------------------*
HUMAGGc     AAAAAAAGAAAGATCC------------------------------------------------------*
HUMHBA4a    AAACAAAGCAAAAAAAAAAAAAAAAAAACCCCCAAAA---------------------------------*
HUMHBA4b    ----------------------------------------------------------------------*
HUMHBA4c    AAATATATAAGGCCGGGCGCG-------------------------------------------------*
HUMHBA4d    AAAAAAAAATAATAATAATAATAAAAATAAAAAAAT----------------------------------*
HUMHBA4e    AAAA------------------------------------------------------------------*
HUMFIXGa    ----------------------------------------------------------------------*
HUMFIXGb    GAAAAAGAAA------------------------------------------------------------*
HUMFIXGc    AAAAAAAAA-------------------------------------------------------------*
HUMFIXGd    AAGAGAGG--------------------------------------------------------------*
HUMFIXGe    TAAATAAATAAATA--------------------------------------------------------*
HUMGHN      ----------------------------------------------------------------------*
CHPRSAa     AAAAAAAAAAAG----------------------------------------------------------*
CHPRSAb     AAAGAAGAAAAGA---------------------------------------------------------*
HUMTPAa     AAAAAAGAAAA-----------------------------------------------------------*
HUMTPAb     AA--------------------------------------------------------------------*
HUMTPAc     AAAAAAGAAAAG----------------------------------------------------------*
HUMTPAd     ATATATATATA-----------------------------------------------------------*
HUMTPAe     TAAATAAGAAA-----------------------------------------------------------*
HUMTPAf     AAAAAAAGTT------------------------------------------------------------*
HUMTPAg     AAAAAAAAAAA-----------------------------------------------------------*
HUMTPAh     ----------------------------------------------------------------------*
HUMTPAi     AAAAAAAAAAA-----------------------------------------------------------*
HUMTPAj     AAAAAAAAAGG-----------------------------------------------------------*
HUMTPAk     AAAAAAAAAAA-----------------------------------------------------------*
HUMTPAl     AAAAGAAAAAG-----------------------------------------------------------*
HUMTPAm     TAATAATAAAT-----------------------------------------------------------*
```

```
Sequence      480       490       500       510       520       530       540
HUMTPAn     ATAAAATAAAT----------------------------------------------------*
HUMTPAo     AACAAAACAAC----------------------------------------------------*
HUMTPAp     GCACCC---------------------------------------------------------*
HUMTPAq     GAAAAAAATG-----------------------------------------------------*
HUMTPAr     TACATAAATAT----------------------------------------------------*
HUMTPAs     ----------------------------------------------------------------*
HUMTPAt     AAAAAAAAAA-----------------------------------------------------*
HUMTPAu     AAAAGAAGAG-----------------------------------------------------*
HUMTPAv     AAAACAAAAAAAAACAAAAAAA-----------------------------------------*
HSALUAGPa   AAACAAAAGCAAAAACAAAAACCAAAAAAAAAGAAAGAAAC---------------------*
HSALUAGPb   AAAAAAA--------------------------------------------------------*
HSALUAGPc   AACAAAAACAAAAAA------------------------------------------------*
HSALUAGPd   CAAACAAACAAACAAAAAAA-------------------------------------------*
HSALUAGPe   AAAAAAATG-------------------------------------------------------*
HSALUAGPf   ATAAATAAATAAATAAATAAA------------------------------------------*
HUMHBB51    AAAAAAAAAAAAAA-------------------------------------------------*
HUMHPARS1a  AAAAAGTTTCTAG--------------------------------------------------*
HUMHPARS1b  CAAAAAACAAAAAA-------------------------------------------------*
HUMHPARS1c  AAAAGTTTCTAG---------------------------------------------------*
HSIGVKA2a   ATAATAATAATAATAAAAAA-------------------------------------------*
HSIGVKA2b   ----------------------------------------------------------------*
HUMHBBa     ACGAAAACAAACAAACAAACAAACAAACAAAACC-----------------------------*
HUMHBBb     ACAAAAAC-------------------------------------------------------*
HUMHBBc     AAAAAAGAAAGAAAGAAAAGAAAAAGAAAAAAGAAAAAATAA--------------------*
HUMHBBd     AAAAGAGAGATTCAA------------------------------------------------*
HUMHBBf     AA-------------------------------------------------------------*
HUMHBBg     ----------------------------------------------------------------*
HUMLDLR18a  CAAA-----------------------------------------------------------*
HUMLDLR18b  ----------------------------------------------------------------*
HUMPOMC8    AAATAAATAAATAA-------------------------------------------------*
HUMPOMC6    ACAAAATGAAACAAACAAACAAACAAACCCC--------------------------------*
HUMPOMC4    AAA------------------------------------------------------------*
HUMPOMC2    ----------------------------------------------------------------*
HUMPOMC1    AAAAAA---------------------------------------------------------*
HSMLVI2     AAAAAAAAAAAAAAAAAAA--------------------------------------------*
HSREP10a    AAAAAAAAA------------------------------------------------------*
HSREP10b    TCAAA----------------------------------------------------------*
HSREP10c    AAAAAAAATT-----------------------------------------------------*
HSREP10d    AAAAAAGC-------------------------------------------------------*
HSREP10e    AAAAAAATGGAATAAAAAAAA------------------------------------------*
HSREP10f    AAAAAAAAAAAAAAAGA----------------------------------------------*
HSREP10g    ATTATTATTAATAA-------------------------------------------------*
HSREP10h    ATAAATAATAACAATAAAAA-------------------------------------------*
HSREP10j    AATAATAATAATAATAATAATAAATAATAACAATT----------------------------*
HUMADAGa    AAAAAAAAA------------------------------------------------------*
HUMADAGb    AAAAACAAAAA----------------------------------------------------*
HUMADAGc    AGAGAAGAAAAAGAAAAGAAAAGAAA-------------------------------------*
HUMADAGd    AAAAAAAAAAAAAAA------------------------------------------------*
HUMADAGf    AAAAAAGAAAGAAA-------------------------------------------------*
HUMADAGi    AAAAACAACAACAACAAATAAATAAATGAATAAATA---------------------------*
HUMADAGja   AAAAAAA--------------------------------------------------------*
HUMADAGk    AAAAAACAAAA----------------------------------------------------*
HUMADAGl    AAAATAAGAATAAAATAAAA-------------------------------------------*
HUMADAGm    AAAAGAAAACGAAAA------------------------------------------------*
HUMADAGn    AAAAAAAA-------------------------------------------------------*
HUMADAGo    AAAACAAAAGAAAAAA-----------------------------------------------*
HUMADAGp    AAAAAAGAATAAA--------------------------------------------------*
HUMADAGq    AAAAAGAAAAAAAAAAGAAA-------------------------------------------*
HUMADAGr    AATAAATAAATAAATAA-ATAAATAAATAAA--------------------------------*
HUMADAGs    AAAAAAAAGATAA--------------------------------------------------*
HUMADAGt    AAAAAAGGAAGAAAGAAGAAAGAGAAAA-----------------------------------*
```

The 45 positions designated as informative through variance analysis using this alignment:

017, 019, 020, 021, 022, 023, 026, 058, 030, 043, 046, 052, 058, 061, 082, 088, 099, 102, 162, 177, 234, 244, 252, 254, 256, 280, 322, 336, 341, 348, 368, 381, 384, 438, 439, 442, 448, 449, 450, 451, 453, 454, 455, 456, 461.

APPENDIX B

An Example of a Five Sequence Alignment and Analysis

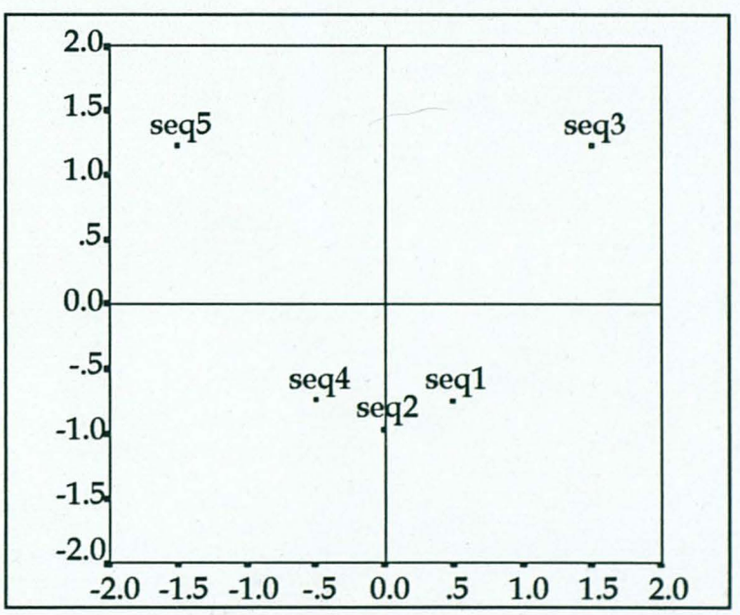
The following is a simplified example of the complete analysis that was performed in this study. The data here were artificially constructed in an attempt to clarify some of the terminology and procedures that may have been vague.

```
        123456
seq 1 AAAAA
seq 2 AAAAG
seq 3 AAGGA
seq 4 ACAAAG
seq 5 ACAG
```

The original five sequences are listed with their names

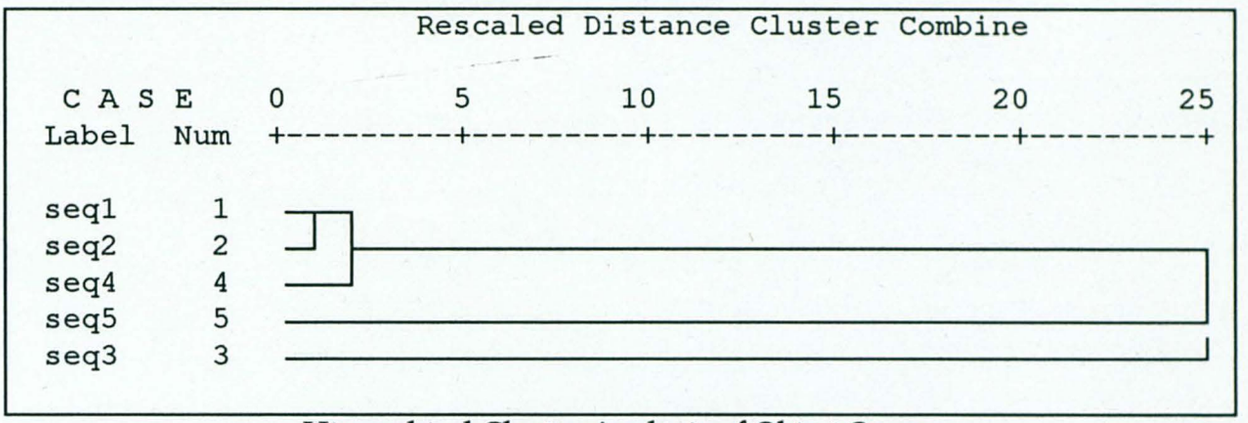
```
        123456
seq 1   A-AAAA
seq 2   A-AAAG
seq 3   A-AGGA
seq 4   ACAAAG
seq 5   AC--AG
```

The five sequences are then globally aligned. With a data set of this size, rapid manual alignment was possible; the data set from this study required computer alignment followed by manual optimization. An insertion of a C at position 2 in sequences 4 and 5 and a deletion of AA at positions 3 and 4 in sequence 5 can be observed.



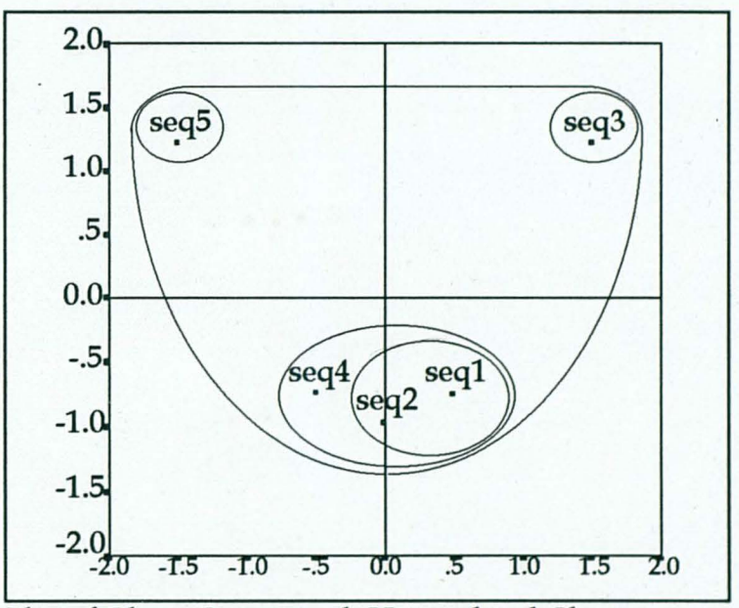
Plot of Object Scores

The correspondence analysis is performed and the resulting object scores are plotted on a two dimensional graph (above). In this instance, the individual object score points are labeled by their sequence name. In the main part of this study, labeling individual points would have resulted in an incomprehensible graph.



Hierarchical Cluster Analysis of Object Scores

The object scores are then used as new variables in a hierarchical cluster analysis, using the centroid method of computation. The resulting dendrogram (above) is then used to identify groupings on the object scores plot (below).



Plot of Object Scores with Hierarchical Cluster Analysis Delimitations

There are three groupings which can be observed. The obvious visual distinctiveness seen in the second object scores plot between the three groups is given support by the dendrogram which has the greatest distance between groups at the three cluster level.

This is the exact procedure, albeit simplified, that was followed in the main analysis of this study.

APPENDIX C

Plot of Correspondence Analysis Scores and Results of Automatic Classification from Quentin (1988)

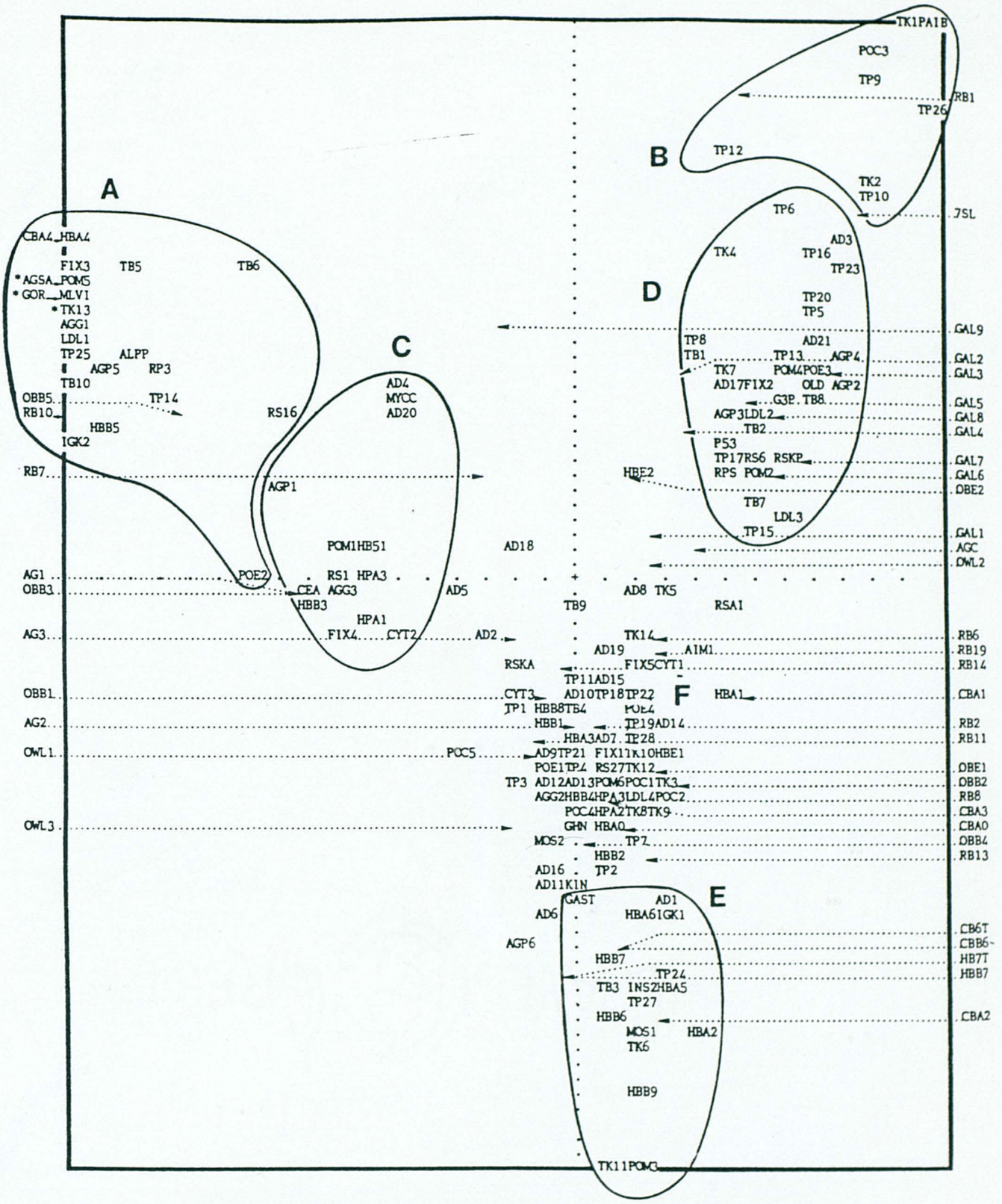


Fig. 4. Projection of the 168 Alu sequences on the first plane of the correspondence analysis. Sequences are labeled as in Table 1, and those that were at standard deviations of more than 2.3 from the center have been brought back to the edge of the figure. Lines are drawn around groups of sequences to indicate the results of the automatic classification. Additionally, the 10 Blur clones, the human 7SL RNA sequence, and primate sequences are added to this figure. The three sequences that may have been inserted recently are labeled with an asterisk.

Springer
Springer-Verlag New York
Publishers
175 Fifth Avenue
New York, New York 10010

July 25, 1994

William York
Philadelphia, PA 19104

Fax: 215-895-6975

Dear Mr. York,

You may have our permission for one time use only, in the English language, material in the manner and for the purpose as specified in your letter dated July 22, 1994, for use in your thesis.

This grant does not extend to use in any medium other than that specifically requested. It does not enable use of said material in a data base, video-disk, or other electronic storage or reproduction system with the exception of a University Microfilms edition.

Full citation must be made with full bibliographic reference as appropriate to the scholarly style of the printed work.

This permission does not extend to any copyrighted material from other sources which may be incorporated in the Work.

If you have any questions, please feel free to call me at 212-460-1505.

Sincerely,

Ian R. Gross
Rights & Permissions Administrator

REFERENCES

Altschul, S. F. (1993) A Protein Alignment Scoring System Sensitive at All Evolutionary Distances. *Journal of Molecular Evolution* 36: 290-300.

Altschul, S. F. (1991) Amino Acid Substitution Matrices from an Information Theoretic Perspective. *Journal of Molecular Biology* 219: 555-565.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology* 215: 403-410.

Bailey, A. D. and Shen, C. K. (1993) Sequential Insertion of Alu Family Repeats Into Specific Genomic Sites of Higher Primates. *Proceedings of the National Academy of Sciences of the United States of America* 15: 7205-7209.

Bains, W. (1986) The Multiple Origins of Human Alu Sequences. *Journal of Molecular Evolution* 23: 189-199.

Batzer, M. A. and Deininger, P. L. (1991) A Human-specific Subfamily of Alu Sequences. *Genomics* 9(3): 481-487.

Brini, A. T., Lee, G. M., Kinet, and J. P. (1993) Involvement of *Alu* Sequences in the Cell Specific Regulation of Transcription of the Gamma Chain of Fc and T cell Receptors. *Journal of Biological* Chemistry 268(2):1355-1361.

Britten, R J., Baron, W. F., Stout, D. B., and Davidson, E. H. (1988) Sources and Evolution of Human Alu Repeated Sequences. *Proceedings of the National Academy of Sciences of the United States of America* 85: 4770-4774.

Cabot, E. L. and Beckenbach, A. T. (1989) Simultaneous Editing of Multiple Nucleic Acid and Protein Sequences with ESEE. *Computer Applications in the Biosciences* 5 (3): 233-234.

Caccone, A. and Powell, J. (1989) DNA Divergence Among Hominoids. *Evolution* 43 (5): 925-942.

Chen, Z. W., McAdam, S. N., Hughes, A. L., Dogon, A. L., Letvin, N. L., and Watkins, D. I. (1992) Molecular Cloning of Orangutan and Gibbon MHC Class I cDNA: The HLA-A and -B Loci Diverged Over 30 Million Years Ago. *Journal of Immunology* 148(8): 2547-2554.

Daniels, G. R. and Deininger, P. L. (1991) Characterization of a Third Major SINE Family of Repetitive Sequences in the Galago Genome. *Nucleic Acids Research* 19 (7): 1649-1656.

Daniels, G. R. and Deininger, P. L. (1985) Repeat Sequence Families Derived from Mammalian tRNA Genes. *Nature* 317: 819-822.

Daniels, G. R. and Deininger, P. L. (1983) A Second Major Class of Alu Family Repeated DNA Sequences in a Primate Genome. *Nucleic Acids Research* 11(21): 7595-7610.

Daniels, G. R., Fox, G. M., Loewensteiner, D., Schmid, C. W., and Deininger, P. L. (1983) Species-specific Homogeneity of the Primate Alu Family of Repeated DNA Sequences. *Nucleic Acids Research* 11 (21): 7579-7593.

Deininger, P. L. (1989) SINEs: Short Interspersed Repeated DNA Elements in Higher Eucaryotes. In Douglas E. Berg and Martha M. Howe (eds.) *Mobile DNA,* Washington D. C.: American Society for Microbiology 619-636.

Deininger, P. L. and Daniels, G. R. (1986) The Recent Evolution of Mammalian Repetitive DNA Elements. *Trends in Genetics* 2 (3): 76-80.

Deininger, P. L., Jolly, D J., Rubin, C. M., Friedman, T., Schmid, C. W. (1981) Base Sequence Studies of 300 Nucleotide Renatured Repeated Human DNA Clones. *Journal of Molecular Biology* 151: 17-33.

Deininger, P. L. and Slagel, V. K. (1988) Recently Amplified Alu Family Members Share a common Parental Alu Sequence. *Molecular Cell Biology* 8 (10): 4566-4569.

Dugaiczyk, A (1993) The Emergence of New DNA Repeats and the Divergence of Primates. *Proceedings of the National Academy of Sciences of the United States of America* 90 (5): 1872-1876.

Ellis, N. A., Goodfellow, P. J., Pym, B., Smith, M., Palmer, M., Frischauf, A. M., and Goodfellow, P. N. (1989) The Pseudoautosomal Boundary in Man is Defined by an *Alu* Repeat Sequence Inseted on the Y Chromosome. *Nature* 337(6202): 81-84.

Farris, J. S. (1973) A Probability Model for Inferring Evolutionary Trees. *Systematic Zoololgy* 22: 250-256.

Felsenstein, J. (1985) Phylogenies and the Comparative Method. *American Naturalist* 125 (1): 1-15.

Felsenstein, J. (1978) Cases in Which Parsimony or Compatibility Methods Will be Positively Misleading. *Systematic Zoololgy* 27: 401-410.

Feng, D. and Doolittle, R. F. (1987) Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees. *Journal of Molecular Evolution* 25: 351-360.

Fuhrman, S. A., Deininger, P. L., LaPorte, P., Friedman, T., and Geiduschek. (1981) Analysis of Transcription of the Human *Alu* Family Ubiquitous Repeating Element by Eukaryotic RNA Polymerase III. *Nucleic Acids Research* 9: 6439-6456.

Ginsburg, M. (1994) Sequence Comparison. In Martin J. Bishop (ed.) *Guide to Human Genome Computing,* London: Academic Press, Inc. 215-248.

Greenacre, Michael (1984) *Theory and Applications of Correspondence Analysis*, London: Academic Press, Inc.

Higgins, D. G., Bleasby, A. J., and Fuchs, R. (1992) CLUSTAL V: Improved Software for Multiple Sequence Alignment. *Computer Applications in the Biosciences* 8: 189-191.

Higgins, D. G. and Sharp, P. M. (1989) Fast and Sensitive Multiple Sequence Alignments on a Microcomputer. *Computer Applications in the Biosciences* 5: 151-153.

Hixson, J. E. and Brown, W. M. (1986) A Comparison of the Small Ribosomal RNA Genes from the Mitochondrial DNA of the Great Apes and Humans: Sequence, Structure, Evolution, and Phylogenetic Implications. *Molecular Biology and Evolution* 3 (1): 1-18.

Hobbs, H. H., Lehrman, M. A., Yamamoto, T., Russell, D. W. (1985) Polymorphism and Evolution of *Alu* Sequences in the Human Low Density Lipoprotein Receptor Gene. *Proceedings of the National Academy of Sciences of the United States of America* 82 (22): 7651-7655.

Jagadeeswaran, P., Forget, B. G., and Weisman, S. M. (1981) Short Interspersed Repetitive DNA Elements in Eukaryotes: Transposable DNA Elements Generated by Reverse Transcription of RNA Pol III Transcripts? *Cell* 26: 141-142.

Jeffreys, A. J., Barrie, P. A., Harris, S., Fawcett, D. H., Nugent, Z. J. and Boyd, A. C. (1982) Isolation and Sequence Analysis of a Hybrid Delta-Globin Pseudogene from the Brown Lemur. *Journal of Molecular Biology* 156: 487-503.

Jurka, J. and Smith, T. (1988) A Fundamental Division in the Alu family of Repeated Sequences. ***Proceedings of the National Academy of Sciences of the United States of America*** 85: 4775-4778.

Jurka, J. and Zuckerkandl, E. (1991) Free Left Arms as Precursor Molecules in the Evolution of Alu Sequences. *Journal of Molecular Evolution* 33 (1): 49-56.

Kariya, Y., Kato, K., Hayashizaki, Y., Himeno, S., Tarui, S., and Kenichi, M. (1987) Revision of Consensus Sequence of Human *Alu* Repeat - A Review. *Gene* 53: 1-10.

Karlin, S. and Altschul, S. F. (1990) Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes. *Proceedings of the National Academy of Sciences of the United States of America* 87: 2264-2268.

Koop, B. F., Miyamoto, M. M., Embury, J. E., Goodman, M, Czelusniak, J., and Slightom, J. (1986) Nucleotide Sequence and Evolution of the Orangutan Epsilon-Globin Gene Region and Surrounding Alu Repeats. *Journal of Molecular Evolution* 24: 94-102.

Koop, B. F., Siemieniak, D., Slightom, J. L., Goodman, M., Dunbar, J., Wright, P. C., and Simons, E. L. (1989) Tarsius Delta- and Beta-Globin Genes: Conversions, Evolution, and Systematic Implications *Journal of Biological Chemistry* 264 (1): 68-79.

Koop, B. F., Tagle, D., Goodman, M., and Slightom, J. L. (1989) A Molecular View of Primate Phylogeny and Important Systematic and Evolutionary Questions. *Molecular Biology and Evolution* 6 (6): 580-612.

Lehrman, M. A., Schneider, W. J., Sudhof, T. C., Brown, M. S., Goldstein, J. L., and Russell, D. W. (1985) Mutation in LDL Receptor: Alu-Alu Recombination Deletes Exons Encoding Transmembrane and Cytoplasmic Domains. *Science* 227: 140-146.

Li, W.-Y., Reddy, R., Henning, D., Epstein, P. and Busch, H. (1982) Nucleotide Sequence of a 7S RNA (Homology to Alu DNA and 4.5S RNA). *Journal of Biological Chemistry* 257: 5136-5142.

Maeda, N., Wu, C., Bliska, J., and Reneke, J. (1988) Molecular Evolution of Intergenic DNA in Higher Primates: Patterns of DNA Changes, Molecular Clock, and Evolution of Repetitive Sequences. *Molecular Biology and Evolution* 5 (1): 1-20.

Myers, E. W. and Miller, W. (1988) Optimal Alignments in Linear Space. *Computer Applications in the Biosciences* 4:11-17.

Norusis, M. J. (1993) *SPSS for Windows Professional Statistics Release 6.0.* Chicago, IL: SPSS, Inc..

Quentin, Y. (1992a) Fusion of a Free Left Alu Monomer and a Free Right Alu Monomer at the Origin of the Alu Family in the Primate Genome. *Nucleic Acids Research* 20 (3): 487-493.

Quentin, Y. (1992b) Origin of the Alu Family: A Family of Alu-like Monomers gave Birth to the Left and Right Arms of the Alu Elements. *Nucleic Acids Research* 20 (13): 3397-3401.

Quentin, Y. (1988) The Alu Family Developed through Successive Waves of Fixation Closely Connected with Primate Lineage History. *Journal of Molecular Evolution* 27: 194-202.

Rinehart, F. P., Ritch, T. G., Deininger, P. L., and Schmid, C. W. (1981) Renaturation Rate Studies of a Single Family of Interspersed Repeated DNA Sequences in Human Deoxyribonucleic Acid. *Biochemistry* 20: 3003-3010.

Ryan, S. C. and Dugaiczyk, A. (1989) Newly Arisen DNA Repeats in Primate Phylogeny. *Proceedings of the National Academy of Sciences of the United States of America* 86: 9360-9364.

Sawada, I., Willard, C., Shen, C., Chapman, B., Wilson, A. C., and Schmid, C. W. (1985) Evolution of *Alu* Family Repeats Since the Divergence of Human and Chimpanzee. *Journal of Molecular Evolution* 22: 316-322.

Schuler, G. D., Altschul, S. F., and Lipman, D. J. (1991) A Workbench for Multiple Alignment Construction and Analysis. *Proteins: Structure, Function, and Genetics* 9: 180-190.

Shaw, J. P., Marks, J., and Shen, C. K. (1991) The Adult Alpha-Globin Locus of Old World Monkdeys: An Abrupt Breakdown of Sequence Similarity to Human is Defined by and *Alu* Family Repeat Insertion Site. *Journal of Molecular Evolution* 33(6): 506-513.

Shen, M. R., Batzer, M. A., Deininger, P. L. (1991) Evolution of the Master Alu Gene(s). *Journal of Molecular Evolution* 33 (4): 311-320.

Singer, M. And Berg, P. (1991) *Genes and Genomes: A Changing Perspective* Mill Valley, CA: University Science Books.

Sinnett, D., Richer, C., Deragon, J. M., and Labuda, D. (1991) Alu RNA secondary Structure Consists of Two Independent 7SL RNA-like Folding Units. *Journal of Biological Chemistry* 266 (14): 8675-8678.

Slagel, V., Flemington, E., Traina-Dorge, V., Bradshaw, H., and Deininger, P. (1987) Clustering and Subfamily Relationships of the Alu Family in the Human Genome. *Molecular Biology and Evolution* 4 (1): 19-29.

Smouse, P. E. and Li, W. (1987) Likelihood Analysis of Mitochondrial Restriction-Cleavage Patterns for the Human-Chimpanzee-Gorilla Trichotomy. *Evolution* 41(6): 1162-1176.

SPSS, Inc. (1990) *SPSS Categories* Chicago, IL: SPSS, Inc..

States, D. J., States, D. J., and Altschul, S. F. (1991) Improved Sensitivity of Nucleic Acid Database Searches Using Application-Specific Scoring Matrices. *Methods: A Companion to Methods in Enzymology* 3 (1): 66-70.

Tagle, D. A., Slightom, J. L., Jones, R. T., and Goodman, M. (1991) Concerted Evolution Led to High Expression of a Prosimian Primate Delta-Globin Gene Locus. *Journal of Biological Chemistry* 266 (12): 7469-7480.

Tagle, D. A., Stanhope, M. J., Siemieniak, D. R., Benson, P. Goodman M., and Slightom, J. L. (1992) The Beta-globin Gene Cluster of the Prosimian Primate Galago crassicaudatus: Nucleotide Sequence Determination of the 41-kb Cluster and Comparative Sequence Analyses. *Genomics* 13 (3): 741-760.

Trabuchet, G., Cheboune, Y., Savatier, P., Lachuer, J., Faure, C., Verdier, G., and Nigon, V. M. (1987) Recent Insertion of an Alu Sequence in the Beta-Globin Gene Cluster of the Gorilla. *Journal of Molecular Evolution* 25: 288-291.

Ullu, E., Murphy, S., and Melli, M. (1982) Human 7S RNA Consists of a 140 Nucleotide Middle Repetitive Sequence Inserted in an Alu Sequence. *Cell* 29: 195-202.

Watson, J. D., Hopkins, N. H., Roberts, J. W., Steitz, J. A., and Weiner, A. M. (1987) *Molecular Biology of the Gene.* Menlo Park, CA: The Benjamin/Cummings Publishing Company, Inc..

Wilbur, W. J. and Lipman, D. J. (1983) Rapid Similarity Search of Nucleic Acid and Protein Data Banks. *Proceedings of the National Academy of Sciences of the United States of America* 80:726-730.

Willard, C., Nguyen, H. T., and Schmid, C. W. (1987) Existence of at Least Three Distinct Alu Subfamilies. *Journal of Molecular Evolution* 26: 180-186.

Williams, S. A. and Goodman, M. (1989) A Statistical Test That Supports a Human/Chimpanzee Clade Based on Noncoding DNA Sequence Data. *Molecular Biology and Evolution* 6 (4): 325-330.