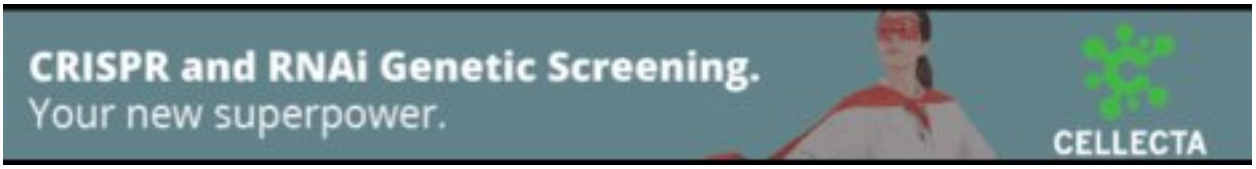
A gene-rich cluster between the CD4 and triosephosphate isomerase genes at human chromosome 12p13.

M A Ansari-Lari, D M Muzny, J Lu, et al.

Genome Res. 1996 6: 314-326
Access the most recent version at doi:10.1101/gr.6.4.314



RESEARCH

A Gene-rich Cluster between the *CD4* and Triosephosphate Isomerase Genes at Human Chromosome 12p13

M. Ali Ansari-Lari,[1] Donna M. Muzny, Jing Lu, Fei Lu, Caroline E. Lilley, Sophie Spanos, Tracy Malley, and Richard A. Gibbs

Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030

The genomic sequence of the human *CD4* gene and its neighboring region, located at chromosome 12p13, was generated using the large-scale shotgun sequencing strategy. A total of 117 kb of genomic sequence and ~11 kb of cDNA sequence were obtained. Six genes, including *CD4*, triosephosphate isomerase, B3 subunit of G proteins (*GNB3*), and ubiquitin isopeptidase T (*ISOT*), with known functions, and two new genes with unknown functions were identified. Using a battery of strategies, the exon/intron boundaries, splice variants, and tissue expression patterns of the genes were determined. Various computer software was utilized for analyses of the DNA and amino acid sequences. The results of the analyses and sequence-based strategies for gene identification are discussed.

The human T4 receptor (CD4) is a primary component of the immune recognition process (for review, see Janeway 1992) and is the major receptor for the human immunodeficiency virus type 1 (HIV-1) (Dalgleish et al. 1984; Klatzmann et al. 1984). The *CD4* gene is expressed in a subset of T cells, B cells, macrophages, granulocytes, and in specific regions of the brain (Maddon et al. 1987). The *CD4* locus maps to 12p12-pter and contains 10 exons that encode a 458 amino-acid mature protein (Maddon et al. 1985; Isobe et al. 1986). The fine structure of the locus has been determined by many physical studies and by generation of a 14-kb sequence encompassing exons 1–3 (Edwards and Gibbs 1992). Linkage disequilibrium between two polymorphic markers developed in this region (Edwards et al. 1991; Edwards and Gibbs 1992) have been used for evolutionary studies pertaining to the origins of modern human (Tishkoff et al. 1996). The nearest known gene is the glycolytic enzyme triosephosphate isomerase (TPI), which maps to 12p13. *TPI* is a housekeeping gene and is required for cell growth and maintenance (Brown et al. 1985). TPI deficiency can cause nonspherocytic hemolytic anemia with or without neuromuscular impairment (Maquat et al. 1985).

To further the understanding of the physical, genetic, and evolutionary history of the *CD4* locus and its neighboring region, we undertook a large-scale sequence analysis of genomic clones from this area. Six genes were identified within an 80-kb segment spanning the *CD4* and *TPI* loci (GenBank accession no. U47924; also see Table 1). Using a battery of PCR, cloning, hybridization, sequencing, and data base searching tools, the messages encoded by each gene, their splice variants, and their expression patterns have been analyzed. The combination of known and implied functional assignments of these genes suggests a region of potentially great importance.

[1]Corresponding author.
E-MAIL ma029926@bcm.tmc.edu; FAX (713) 798-5741.

RESULTS

Generation of the Sequence

Minimally overlapping cosmid clones from this region were isolated by screening a chromosome-12-specific cosmid library by hybridization and PCR. A schematic representation of the clones is shown in Figure 1A. Shotgun sequencing libraries were generated for the individual cosmids and the λ clones. For the majority of the regions, the random phase of sequencing resulted in multiple sequence reads from both strands. Direct and random reverse sequencing along with map–gap closure were employed to generate accurate, contiguous, bidirectional sequence of the region. More details on the sequencing strategies are described in Richards et al. (1994).

6:314–326 ©1996 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/96 $5.00

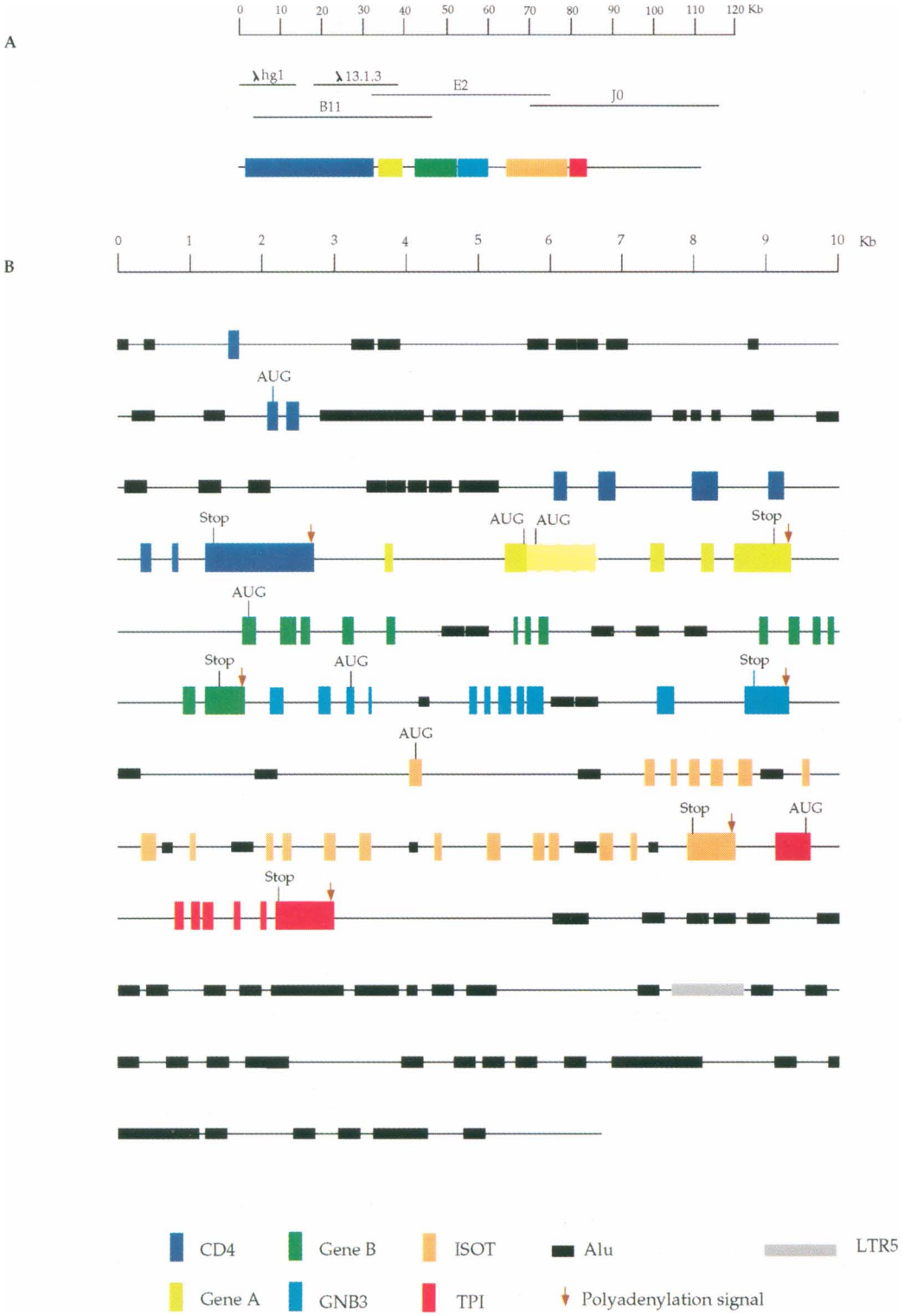


Figure 1 Schematic representation of the six genes on chromosome 12p13. The drawings are to scale (± 100 bp). (*A*) Map of the genomic clones that were used for sequencing. B11, E2, and J0 are cosmids. The relative position of the genes with respect to the clones are indicated. (*B*) Exons are represented by boxes and introns by straight lines. For gene *A*, the light yellow box represents an alternatively spliced exon (see Fig. 4). The polyadenylation signal is AAUAAA, except for gene *B* (AUUAAA).

Table 1. Structural and Functional Information of the Six Genes Located on Chromosome 12p13

Gene ID and spliced variants	Number of exons	ORF[a]	Size (kb) of transcripts by Northern	Length of cDNA (bp)	Accession no.	Exons not predicted by GRAIL
CD4						
hematopoeitic	10	458	3	3064	M12807; I09237	1(p); 2; 5(p); 10
brain	N.D.	N.D.	1.8; 3.0	N.D.		N.D.
Gene *A*						
1	5	304	1.7	1514	U47925	1; 5(p)
2	5	588	2.5	2516	U47928	1; 5(p)
3	I.C.	I.C.	3.4	I.C.	U47929	I.C.
Gene *B*	14	551 (543)		2099	U47926	6; 7; 9; 14(p)
GNB3						
1	11	340	2	1928	M31328; U47930	1; 2; 3; 11(p)
2	I.C.	I.C.	3.4–3.7	I.C.	U47931	I.C.
ISOT						
1	20	858	3.3	3181	U47927	15(p); 20(p)
2	20	835	N.D.	3080*	U35116	20(p)
TPI	7	284 (249)		1835	M10036	1(p); 3; 7(p)

The completed sequence was edited to the established community standard of 99.99% accuracy. Furthermore, as an indication of the sequence accuracy, comparison of our genomic sequence with the published cDNA sequence (1517 bp; GenBank accession no. M31328) for the B3 subunit of G proteins (*GNB3*) (Levine et al. 1990), and the edited portion of the cosmid vector sequence (3412 bp) with the published Lawrist 16 vector sequence showed a 100% sequence match. In addition, the overlapping segment (1698 bp) between E2 and J0 cosmids, which was edited independently, showed 100% sequence match.

Analyses of the Sequence

Five genes adjacent to *CD4* were identified: Gene *A*, gene *B*, *GNB3*, ubiquitin isopeptidase T (*ISOT*), and *TPI*. The exon/intron organization of the genes in this region is represented in Figure 1B.

Table 1. (*Continued*)

5′ exons identified by LA–PCR	EST matches	Related information
N.D.	R59028, H68199, R84400, R92959 R67367, R58980, R66647, R84399 R92957, H68200, H61295, H48407 T95138, H61299, T95042, T72416 R59963, R46412, R92387, R98829 R46327	interacts with T-cell receptor during antigen recognition; signal transducton pathway; major receptor for HIV
N.D.		
N.C.	H56593, H56592, R54138, R84482 R87327, T09310, R54086	sequence obtained from cDNA clone HIBBS19 (accession no. T09310); one Asp-rich and one Glu-rich domain; one putative transmembrane domain
N.C.	H56593, H56592, R54138, R84482 R87327, T09310, R87358, R54086 R88866, R12610	seven putative transmembrane domains; one Asp and one Glu-rich domain
identified		
N.C.	None	two Glu-rich domains; one leucine zipper pattern
2.5	H92898, T29297, H9784	one of the subunits of heterotrimeric G proteins; involved in signal transduction pathways
identified		
1	T11767, H15620, T35460, T08021 U25904, T08022, H15561, H41228 T35369, T11766, T24496, T08021	member of ubiquitin carboxyterminal hydrolase family; 3′ end of the sequence was obtained from cDNA clone HIBAA66 (accession no. T08021)
N.D.		
N.D.	>50 EST matches has been reported	interconversion of dihydroacetone phosphate and glyceraldehyde 3-phosphate

[a](ORF) The longest open reading frame from the first Met. For gene *B,* the number in the parentheses is based on the first Met with the preferred Kozak sequence context. For *TPI,* the number in the parentheses represents the known amino acid sequence. (I.C.) Incomplete information; (N.D.) not determined; (N.C.) did not change the existing data; (p); portion of the indicated exon that was not predicted by GRAIL. (*) The published cDNA sequence for ISOT-2 was edited by removing the sequence of the vector. GNB3(2) and A-3 splice variant were identified by LA–PCR.

Five of the six genes were identified first by similarity to expressed sequence tags (ESTs) (Table 1). Exons from all six genes were also identified by the GRAIL program (Uberbacher et al. 1991). Based on the GRAIL predictions, several primers were designed for reverse transcription PCR (RT–PCR), modified ligation-anchored PCR (LA–PCR) (Ansari-Lari et al. 1996), and 3′ RACE (rapid amplification of cDNA ends) (Frohman et al. 1988), followed by cloning and sequencing (Table 2). Not all of the exons predicted by GRAIL were detected by RT–PCR and sequencing. Further-

Table 2. Oligonucleotide List

Oligonucleotide ID	Sequence	Application
R595	CTACTTTCCCAACTATGAGGCTGG	PCR probe for isolation of E2 cosmid
R664	GTTCACTGTGAACCCTCAAGAGG	
R807	CAGTTCTGGCAGTGCCATCTTC	PCR probe for isolation of J0 cosmid
R995	CGAGTTGACTACATCATGCAGC	
2246	TTTCTGTGGGCTCAGGTCCCTAC	For PCR of CD4 gap
R730	CCATGAGTACAGCATGTCTCTCC	
R596	GTGCCTGCTTCTTGGCGTTGCTG	RT-PCR of gene A-2
R597	GAGTTCGGTCTCCATGGCAGTGCT	
R796	CTCCTACCATCGCATGTGGATG	RT-PCR of gene A-2
R836	GGCAGAGGTGTCTCATCGATG	
R1046	B-TCCTCAGCTTTCATACTGGG	LA-PCR of gene A (outer primer)
R595	CTACTTTCCCAACTATGAGGCTGG	LA-PCR of gene A (inner primer)
R618	CTGAGGAGGATGGAGATGAC	PCR probe for northern of gene A
R836	GGCAGAGGTGTCTCATCGATG	
R1045	B-CGAACTCCCGACATTCGTC	LA-PCR of gene B (outer primer)
R915	TGTACTTAGCCATGTCCTCC	LA-PCR of gene B (inner primer)
R801	GCTAAGCTGCTTCTGGAGGTG	3'-RACE of gene B (outer primer; with universal primer)
R802	CACGCAGACAACTGCGTCCTG	3'-RACE of gene B (inner primer; with universal primer)
R1017	TGTAAAACGACGGCCAGTTTTTTTTTTTTTTT	For 3'-RACE (Oligo-dT15 with complementary universal tail)
R798	CAGCACACACCTTCTTTGTAGC	RT-PCR of gene B
R832	GACTCCTGCAGCAGTTCTTTGG	
R803	CTGTAGTCCCGATAGGTGTAGG	PCR probe for northern of gene B (with R798)
R913	B-GCATAGGCACAGGTCATGACC	LA-PCR of GNB3 (outer primer)
R872	ACTCGTCCCACCACCTCTAG	LA-PCR of GNB3 (inner primer)
R871	CATAGGCACAGGTCATGACC	PCR probe for northern of GNB3
R937	GAGATGGAGCAACTGCGTCAG	
R804	GGGAAACAGTATGTGGAGAGAC	RT-PCR of ISOT (with R807)
R1044	B-TGGGTCTCCAGTGCCTGTAG	LA-PCR of ISOT (outer primer)
R1047	CCAAAGCCCAGAAACGTG	LA-PCR of ISOT (inner primer)
R805	TGATGAGGATGACATGGTCCTG	PCR probe for northern of ISOT (with R807)

All of the sequences are in 5′ → 3′ orientation. (B) Biotin group at the 5′ end.

more, some of the exons (coding and noncoding) identified by RT–PCR and LA–PCR were not predicted by GRAIL (Table 1). The repetitive elements were identified by CENSOR (Jurka et al. 1995) and PASAP (Jurka et al. 1992) software.

The sequence of the cDNAs corresponding to variant forms of gene *A*, gene *B*, and *ISOT* were determined. LA–PCR was used to obtain the sequence from 5′ ends of transcripts of gene *A*, gene *B*, *GNB3*, and *ISOT*. An array of computer programs, including BLAST (Altschul et al. 1990), FASTA (Pearson 1990), and BEAUTY (Worley et al. 1995) were utilized for the prediction of possible functions associated with these genes. The results of the analyses are summarized in Table 1.

CD4

The sequences corresponding to the 5′ and 3′ ends of the *CD4* gene were obtained from λ clones hg1 and 13.1.3, respectively. A gap remained between exons 3 and 4 that was not represented in the clones, and therefore the gap was cloned by a long-range PCR strategy (Cheng et al. 1994). The sequence of the gap was obtained by a shotgun library strategy, several subcloning steps, and a nested deletion series. The restriction map of the cloned gap was in agreement with the restriction map predicted from the obtained sequence. The amplification of the gap sequence from eight different individuals and analysis of the fragments by agarose gel electrophoresis did not reveal any polymorphism in this region (data not shown). The presence of more subtle polymorphism within the gap, however, has not been fully excluded.

Comparison of the silencer sequence of the mouse *CD4* gene (Sawada et al. 1994) and the genomic sequence of the human *CD4* gene indicates the presence of a region in intron 1 of human *CD4* with 75% sequence similarity (Fig. 2). This illustrates that cross-species sequence comparison can be a powerful strategy for the determination of regulatory elements. However, of the nine nuclear protein-binding sites identified on the mouse sequence (Sawada et al. 1994), only

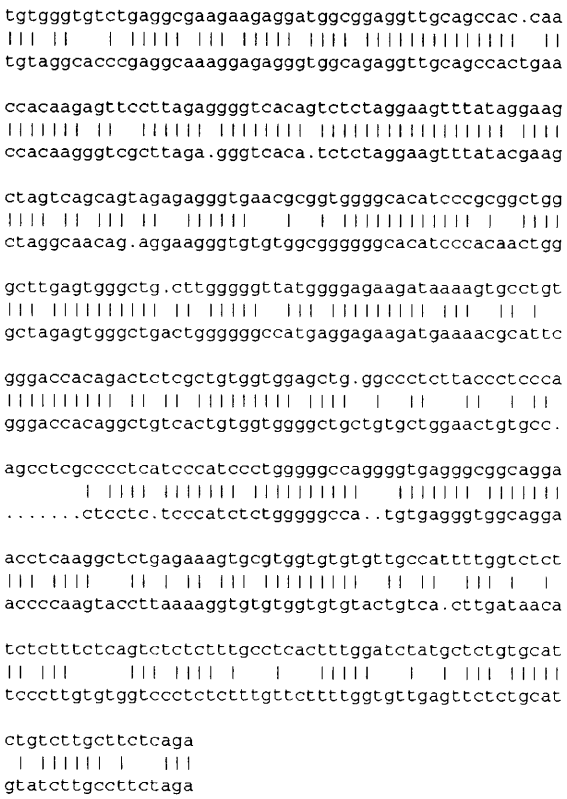


Figure 2 Identification of the hypothetical silencer of the human *CD4* gene. The upper sequence is human; the lower sequence is mouse. In human and mouse, the silencer segment is located at intron 1 of the *CD4* gene.

one was a perfect sequence match. Whether this indicates lack of conservation of the sequence for these sites or the requirement of another conserved domain in this region for the function of the silencer is not known.

Gene *A*

Three alternatively spliced forms (A-1, A-2, and A-3) were identified by RT–PCR and LA–PCR, and subsequently detected by Northern blot analysis (Fig. 3). Gene *A* appears to be mainly expressed in the brain with the A-2 (2.5 kb) variant as the predominant transcript. A low level of expression of the 3.4-kb transcript (A-3 variant) is also detected in liver. The splice variants are represented schematically in Figure 4A (see also Table 1). Exons 3 and 4 are shared by all three forms. Exon 2 is spliced at different positions in all three forms. Form A-1 does not show similarity to any known sequence. Form A-2 shows very weak similarity to the members of G-protein-coupled receptor superfamily. TMbase software (Hofmann and Stoffel 1993) identified seven significant transmembrane domains, as presented in Figure 5. Form A-3 was identified by LA–PCR strategy for determination of the 5′ end of gene B. In this form, exon 4 of gene A is spliced to exon 1 of gene *B* by skipping exon 5 of gene *A* [exon 5 contains a poly(A) signal]. The open reading frame (ORF) is maintained upon splicing to gene *B*. This form does not show similarity to any known sequence. The exact 5′ and 3′ end of this alternate form has not been obtained. All three forms maintain the same translation frame in exons 3 and 4.

Gene *B*

The cDNA corresponding to gene *B* was obtained by RT–PCR. The conceptual protein corresponding to the gene *B* does not show strong similarity to any known protein but does show a very weak similarity (16.5% identity) to the rat synaptonemal complex protein Sc 65 (Chen et al. 1992). Gene *B* appears to be a distinct gene as opposed to another alternatively spliced form of gene *A*. The first exon of this gene starts at a different position compared with the exon that is present in gene A-3 (see Fig. 4). However, the possibility of genes *A* and *B* being one gene with a collection of alternative transcripts has not been excluded. Tissue expression analysis of human adult tissues indicated a presence of a very faint band of ~2.1 kb in heart, lung, ovary, and skeletal muscle (data not shown). In addition, a higher, faint band of ~3.4 kb was observed in brain, liver, and other tissues mentioned containing the 2.1-kb band. The 3.4-kb band likely corresponds to the alternatively spliced form of gene *A* (A-3) observed in liver and brain (Fig. 3).

GNB3

Two major alternatively spliced forms of this gene have been identified by tissue expression analysis (2.0 and 3.3–3.7 kb, depending on the tissue), and both splice forms are expressed in all tissues analyzed (Fig. 3). A previously published cDNA sequence of 1545 bp (Levine et al. 1990) is identical to the corresponding exons of our genomic sequence. Using LA–PCR, we have extended the sequence at the 5′ end to a total of 1950 bp. This sequence likely corresponds to the 2.0-kb transcript. LA–PCR also identified a different splicing pattern at the 5′ end as indicated in Figure 4B.

ISOT

Tissue expression analysis indicated a high level of expression of a 3.3-kb transcript in brain, and

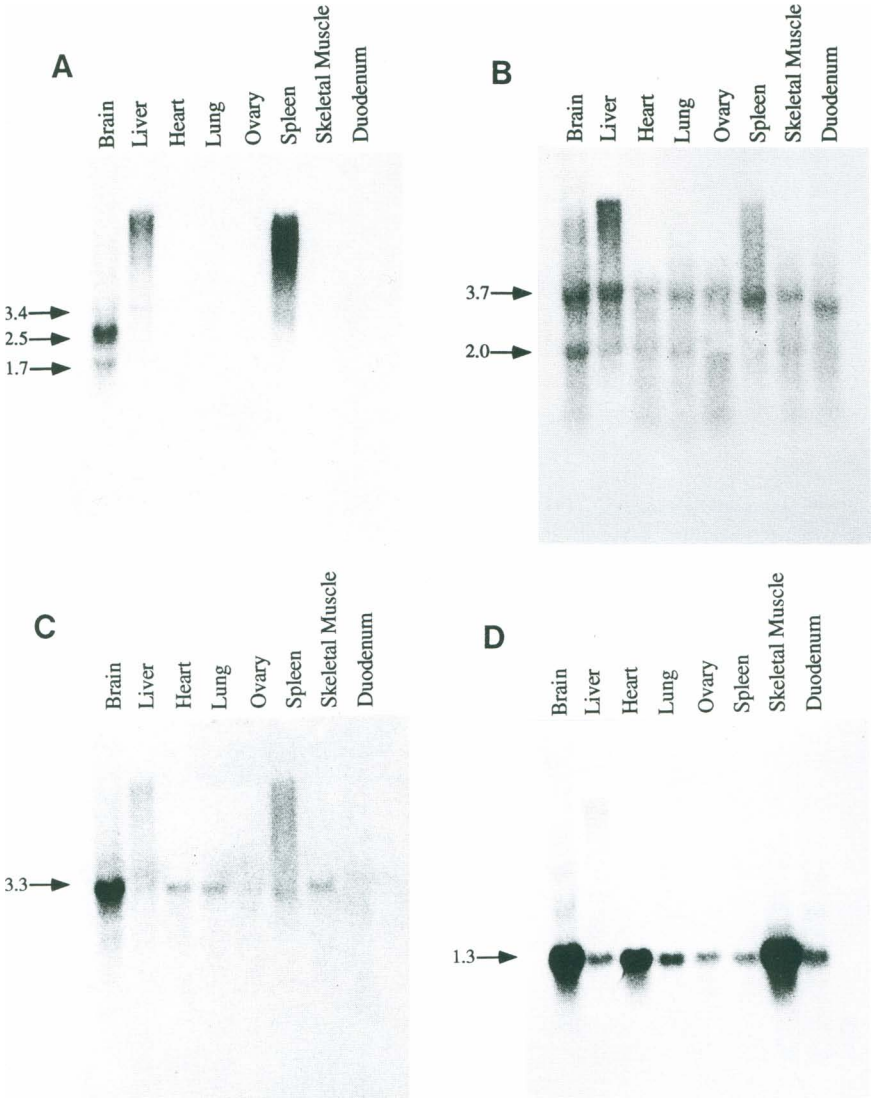


Figure 3 Tissue expression patterns of gene A (*A*), GNB3 (*B*), ISOT (*C*), and glyceraldehyde 3-phosphate dehydrogenase (*D*).

a low level of expression in heart, lung, spleen, and skeletal muscle (Fig. 3). The 5′ half of this gene was isolated by RT–PCR and LA–PCR. BLAST and BEAUTY searches show that this gene is a member of the ubiquitin carboxy-terminal hydrolase family. One form of this gene has been reported recently as the human *ISOT* (Wilkinson et al. 1995). The reported sequence is 3102 bp long. The comparison of this sequence to our cDNA sequence indicates alternative splicing in exon 15. Furthermore, there are 4 nucleotide differences between the two sequences. Two of these differences lead to Lys → Arg and Gly → Asp substitutions. Our cDNA and genomic sequences are in full agreement. The comparison of the amino acids for the two sequences is shown in Figure 6.

TPI

The comparison of our genomic DNA sequence for *TPI* (5029 bp) with the previously published *TPI1* sequence (4995 bp; accession no. X69723) shows 99.34% identity. This region is very GC rich (59%), especially at the 5′ end of the sequence (65% for the first 2000 bp). Some of the differences may be attributed to natural polymorphism, but other differences may be the result of variation in editing and/or sequence quality be-

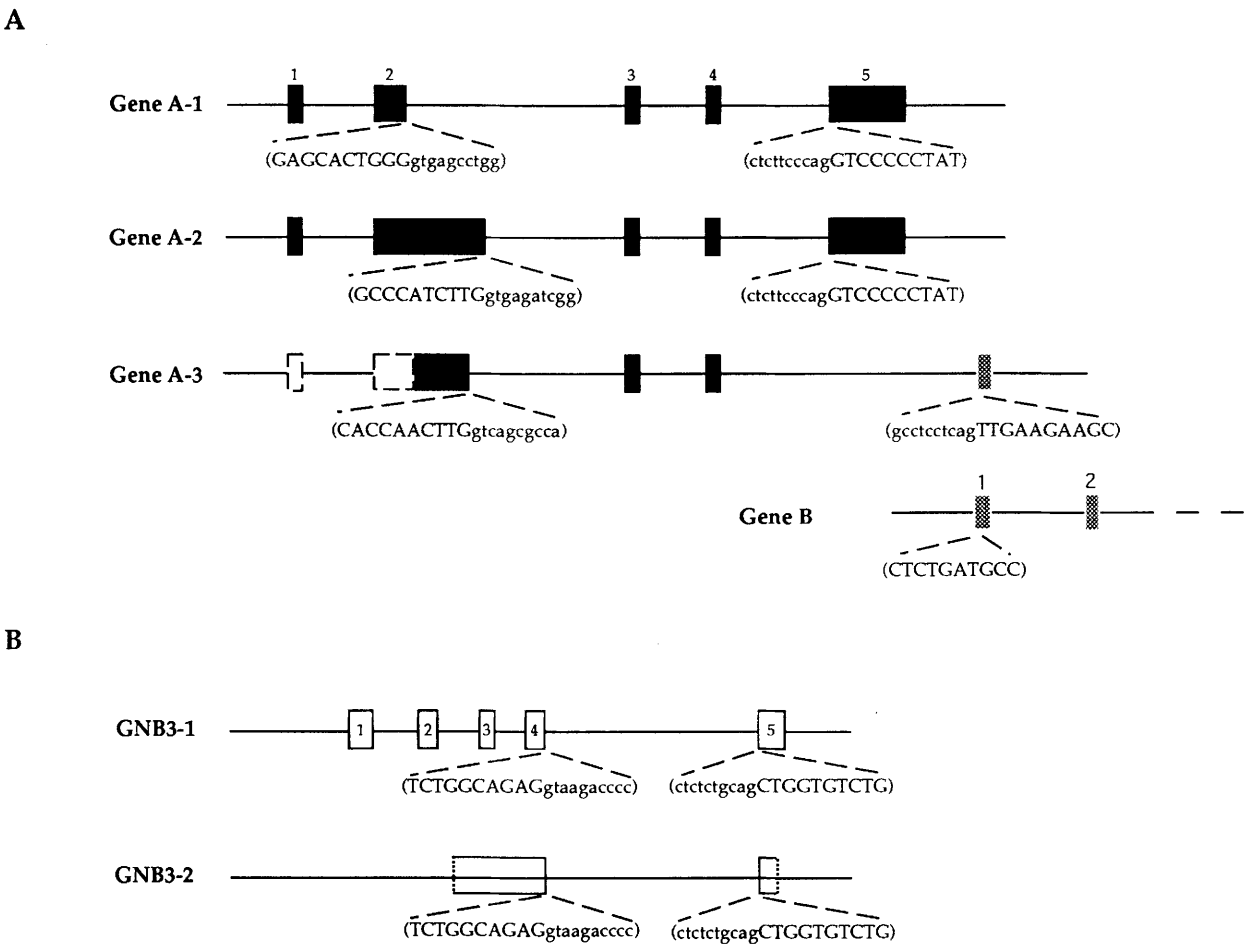


Figure 4 Splice variants. The drawings are not to scale. The sequence for variable splice junctions are indicated in parentheses. (*A*) Three alternatively spliced forms of gene *A*. Boxes represent exons. The stippled boxes are exons of gene *B*. The lowercase sequence is intronic; the uppercase sequence is exonic. For gene A-3, the broken-line boxes represent the likely splicing pattern at its 5′ end. The GenBank accession nos. for A-1, A-2, A-3, and gene *B* are U47925, U47928, U47929, and U47926, respectively. (*B*) Alternative splicing pattern of the *GNB3* gene. Exons 1–5 of *GNB3*(1) are shown. The segment representing portion of *GNB3*(2) was isolated by LA–PCR. Full-length GNB3(2) transcript has not been obtained. The GenBank accession nos. for *GNB3*(1) and *GNB3*(2) are U47930 and U47931, respectively.

tween the two groups. The sequence differences, however, do not result in any amino acid variation in the TPI protein.

DISCUSSION

In this study 117 kb of genomic sequence on 12p13 was determined, and the sequence was analyzed using an array of molecular and computational tools. Six genes, bounded by *CD4* and *TPI*, span 80 kb of the genomic interval. This represents one of the most gene-packed regions on a human chromosome reported to date, with a relatively short stretch of intergenic sequence separating the ends of the genes. The genomic and cDNA sequences were examined in detail to determine exon/intron boundaries, 5′- and 3′-untranslated regions, repetitive elements, and functional or structural features associated with each gene.

CD4 belongs to the large immunoglobulin supergene family but is an unusual member structurally, owing to the insertion of an intron (intron 3) within a usually contiguous segment coding for the variable domain (Littman and Gettner 1987). One unusual feature of this intron is its extreme Alu richness. Alu elements make up 5% of the human genome, and hence their distribution is estimated at one Alu element every 5000 bp (for review, see Deininger 1989). There are 22 full Alu elements and 7 half-Alu elements

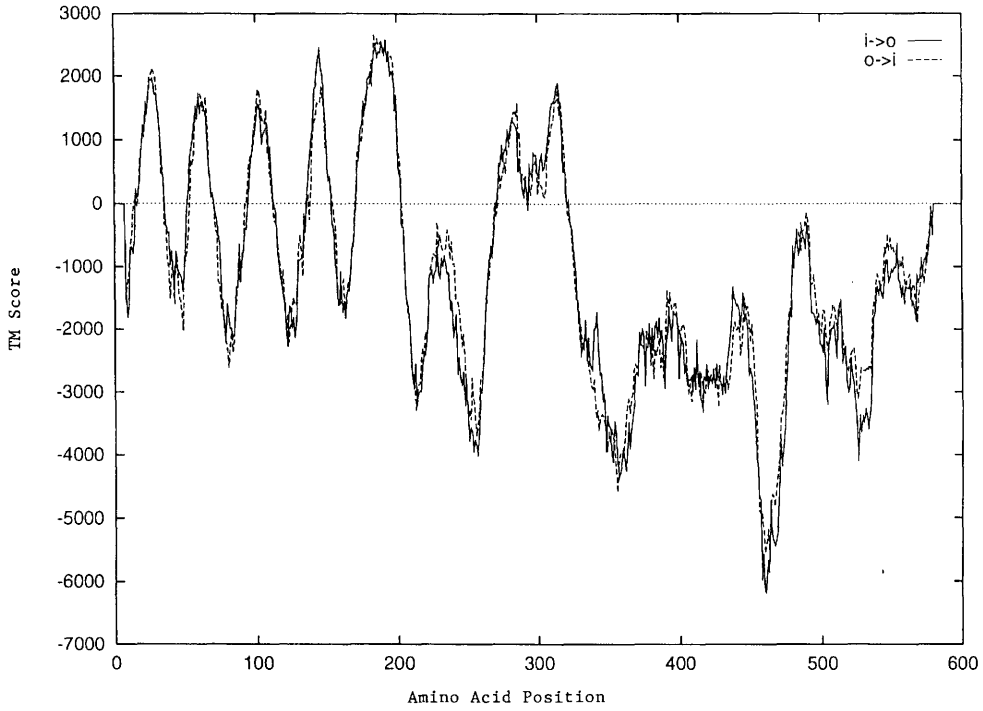


Figure 5 Hypothetical configuration of transmembrane (TM) domains for the gene *A-2* based on TMbase program. (i → o) Inside to outside; (o → i) outside to inside. Inside means the cytoplasmic face; outside means the lumenal face of the membrane (depending on the organelle). The prediction parameter is TM–helix length between 17 and 33. The i → o and o → i predictions are in very close agreement. The amino terminus is predicted to be at the outside. Only TM scores >500 are considered significant.

in the intron 3 or, on average, 1 Alu every 530 bp. Considering the unusual interruption of the V-like domain of *CD4* by the highly Alu-rich intron, the propagation of Alu elements might have had a major role in the evolution of this locus. Intron 1 of *CD4* is also relatively Alu-rich with approximately one Alu element every 1300 bp.

One of the genes in this region is *GNB3*. The G proteins are heterotrimeric proteins composed of α, β, and γ subunits (for review, see Rens-Domiano and Hamm 1995). They have a central role in numerous hormonal and sensory signal transduction pathways. Currently, 18 G-α, 5 G-β, and 7 G-γ subunits have been identified. In this study the identification of the 5′-untranslated region of *GNB3* represents an initial step toward determination of transcriptional regulation of this gene.

ISOT is a member of ubiquitin carboxy-terminal hydrolase gene family, which are ATP-independent proteases (Baker et al. 1992). Deubiquitinating enzymes are conserved among eukaryotes and have properties of thiol proteases. There are three highly conserved residues (Cys, His, His) that constitute the putative active site of this family of enzymes. Recently, the sequence of a cDNA for human ISOT (Wilkinson et al. 1995) has been reported that appears to be an alternatively spliced form of the cDNA reported here. This protein has been suggested to have a major role in disassembly of polyubiquitin chains after release of the chain from conjugated proteins or degradation intermediates (Wilkinson et al. 1995). Other deubiquitinating enzymes have been suggested to have a more central role in the degradation of ubiquitin-conjugated proteins (Papa and Hoschstrasser 1993). Whether the two alternatively spliced forms of human ISOT have different functions is not known.

Of the six genes identified here, the functional significance of genes *A* and *B* awaits further analysis. Although three of the genes show variable tissue expression pattern, all six genes are expressed in the brain. Whether the product from all six genes or a subset of them are involved in a coordinated or concerted functional pathway is not yet known.

This study has tested the potential of several techniques to be integrated into a systematic

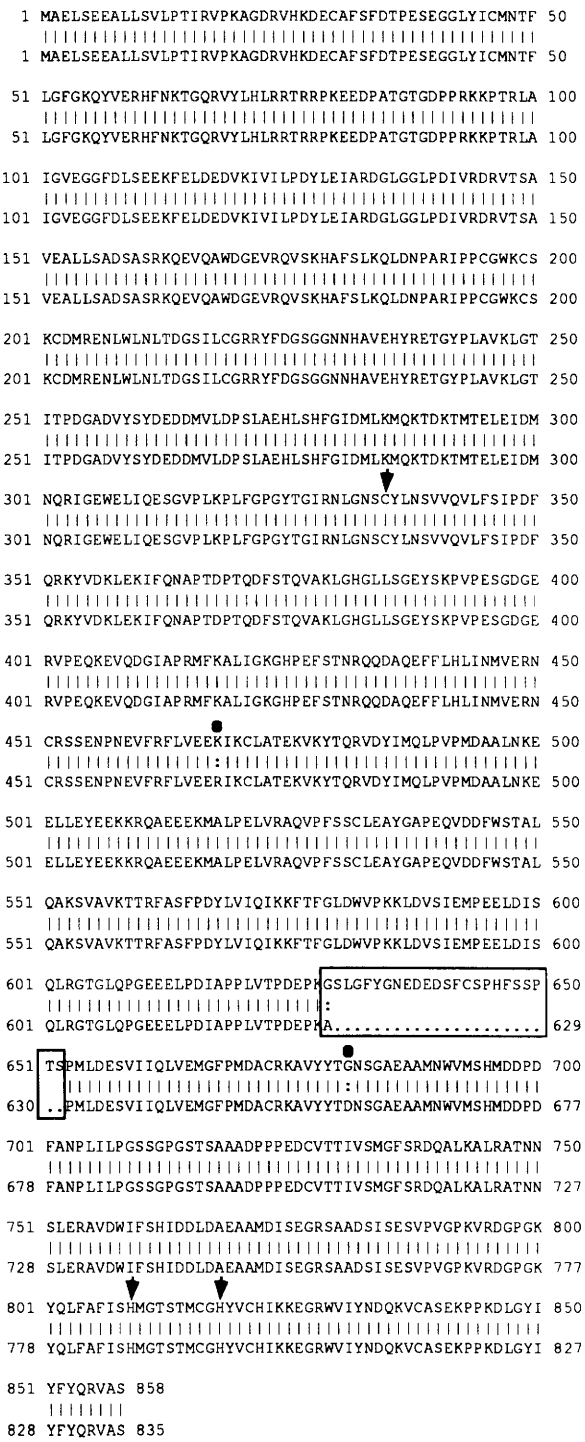


Figure 6 Amino acid comparison of the two alternatively spliced forms of ISOT. The *top* sequence was determined by our group (ISOT-1; GenBank accession no. U47927); the *bottom* sequence (ISOT-2; GenBank accession no. U35116) was reported by Wilkinson et al. (1995). The boxed segment represents the alternatively spliced region. (●) Amino acid differences between the two sequences. Arrows show the Cys and His of the active site of the enzyme.

analysis of completed genomic sequence. Fine detailed analyses of the sequence by RT–PCR, LA–PCR, and 3′ RACE were required for accurately defining exon/intron boundaries and possible spliced variants; however, these type of analyses are very time consuming and in general are not amenable to automation. In contrast, analysis with extensive EST data base searches and computer software for exon predictions were found to be very effective approaches for initial gene identification, and these can be easily automated. The continuing flow of EST sequence information into public data bases would improve this process. Based on this work and other large-scale genome sequencing projects in our laboratory, we predict ~80% of the genes could be identified by the current available public EST data base. With the predicted explosion of sequencing information in the coming years, EST data base searches and computer software for exon predictions should be the strategy of choice for rapid sequence analysis.

METHODS

Isolation of Cosmids

A human chromosome 12-specific cosmid library, LL12NC01, was kindly provided by Dr. Jeffery C. Gingrich (Human Genome Center; Biology and Biotechnology Research Program; Lawrence Livermore National Laboratory, Livermore, CA). The λ clones were gifts from Dr. Dan R. Littman (University of California, San Francisco). Library plating and screening were performed according to standard protocols (Sambrook et al. 1989). A Rediprime kit (Amersham Life Science, Arlington Heights, IL) and NICK columns (Pharmacia) were used for labeling and purification of the probes, respectively. The positive clones from the primary screen were mapped by PCR to identify minimally overlapping clones that were plated for secondary screening to isolate individual positive colonies. Cosmid DNA was isolated by equilibrium centrifugation in CsCl gradients essentially as described (Sambrook et al. 1989).

Sequencing of λ and Cosmid Clones

A shotgun sequencing library was generated for each individual λ and cosmid clone using one of the M13 adaptor-based strategies as described (Povinelli and Gibbs 1993; Andersson et al. 1994). Sequence templates were prepared as described (Kristensen et al. 1987). Clones were sequenced using random and directed sequencing strategies as described (Civitello et al. 1993). Directed reverse template preparation and sequencing was performed using a modified asymmetric PCR protocol (Muzny et al. 1994). Sequencing reagent kits were purchased from Perkin Elmer, and sequence reactions were electrophoresed on ABI 370, 373A, or 373 sequencers.

RT–PCR, LA–PCR, and 3′ RACE

Total RNA was isolated from white blood cells (WBC) or from HT-1080 human fibrosarcoma cells by guanidinium thiocyanate extraction followed by CsCl centrifugation as described (Sambrook et al. 1989). mRNA was isolated using oligo(dT)–cellulose type 7 Redi-column (Pharmacia) according to the manufacturer's protocol.

For RT–PCR and modified LA–PCR reactions (Ansari-Lari et al. 1996), single-strand cDNA synthesis was performed using random hexamer and oligo(dT_{15}). For 3′-RACE reaction (Frohman et al. 1988), oligo(dT_{15}) containing a universal tail was used to generate the first-strand cDNA. For LA–PCR and 3′ RACE, a nested PCR approach was used. The primers used for various strategies are indicated in Table 2.

Sequencing of ESTs, RT–PCR, LA–PCR, and 3′-RACE Products

RT-PCR and LA-PCR products were either cloned into pBluescript II (SK−) (Stratagene) or in pGEM-T vector (Promega). Two of the ESTS (Table 1), and the majority of RT–PCR products were sequenced using a cDNA–concatenation sequencing strategy (Andersson et al. 1994). Some of the RT–PCR products and all of the LA–PCR, and 3′-RACE products were sequenced using dye-terminator sequencing, or solid-phase sequencing (Gibbs et al. 1990) strategies.

Sequence Assembly

Sequence reads were edited using the SEQPREP software developed by the Molecular Biology Computational Resource Center at Baylor College of Medicine. Lambda clones were assembled by SAM software developed by the Molecular Biology Computational Resource Center at Baylor College of Medicine. Cosmid sequence assembly was performed using Staden XDAP and XGAP software (Dear and Staden 1991). Gap closure was performed as described (Richards et al. 1994). cDNA sequence editing and assembly were performed using Sequencher for Macintosh, version 3.0 (Gibbs and Cockerill 1995).

Tissue Expression Analysis

An adult human total RNA blot was purchased from BIOS Laboratories (New Haven, CT). Hybridization was preformed using Speed Hyb solution (BIOS Laboratories) with probes generated by RT–PCR indicated in Table 2. Washes were performed essentially according to the manufacturer's protocol.

Computer Analysis Program

An array of computer software, including GRAIL (Uberbacher et al. 1991), BLAST (Altschul et al. 1990), BEAUTY (Worley et al. 1995), FASTA (Pearson 1990), TMbase (Hofmann and Stoffel 1993), GCG (sequence analysis software package, v. 8, Genetics Computer Group, Madison, WI), and CENSOR (Jurka et al. 1995) were used for analysis of the sequence. Some of these programs have been assembled into a World Wide Web page by the Baylor College of Medicine genome informatic core (http://kiwi.imgen.bcm.tmc.edu:8088/search-launcher/launcher.html).

ACKNOWLEDGMENTS

We thank Dr. Christine Povinelli, Mark Tristan, and Frederico Mattioli for the sequencing and initial analysis of λ clones, Binhai Zheng for assistance in isolation of E2 cosmid, Dr. Bjorn Andersson and Meredith Wentland for assistance in construction of E2 cosmid sequencing library, and Kecia Rowland and Kimberly Edwards for help in sequencing of cDNA clones. We especially thank Dr. Randall Smith for helpful discussions during the course of this study. This work was supported in part by grants ROI HG00823 and P30 HG00210 from the National Center for Human Genome Research. The sequence data described in this paper have been submitted to the GenBank data library under accession nos. U47924–U47931.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Andersson, B., C.M. Povinelli, M.A. Wentland, Y. Shen, D.M. Muzny, and R.A. Gibbs. 1994. Adaptor-based uracil DNA glycosylase cloning simplifies shotgun library construction for large-scale sequencing. *Anal. Biochem.* **218:** 300–308.

Ansari-Lari, M.A., S.N. Jones, K.M. Timms, and R.A. Gibbs. 1996. Improved ligation-anchored PCR strategy for identification of 5′ ends of transcripts. *BioTechniques* (in press).

Baker, R.T., J.W. Tobias, and A. Varshavsky. 1992. Ubiquitin-specific proteases of *Saccharomyces cerevisiae*. *J. Biol. Chem.* **267:** 23364–23375.

Brown, J.R., I.O. Darr, J.R. Krug, and L.E. Maquat. 1985. Characterization of the functional gene and several processed pseudogenes in the human triosephosphate isomerase gene family. *Mol. Cell Biol.* **5:** 1694–1706.

Chen, Q., R.E. Pearlman, and P.B. Moens. 1992. Isolation and characterization of a cDNA encoding a synaptonemal complex protein. *Biochem. Cell Biol.* **70:** 1030–1038.

Cheng, S., C. Fockler, W.M. Barnes, and R. Higuchi. 1994. Effective amplification of long targets from cloned inserts and human genomic DNA. *Proc. Natl. Acad. Sci.* **91:** 5695–5699.

Civitello, A.B., S. Richards, and R.A. Gibbs. 1993. A simple protocol for the automation of DNA cycle

sequencing reactions and polymerase chain reactions. *J. DNA Seq.* **3:** 17–23.

Dalgleish, A.G., P.C. Beverley, P.R. Clapham, D.H. Crawford, M.F. Greaves, and R.A. Weiss. 1984. The *CD4* (T4) antigen is essential component of the receptor for the AIDS retrovirus. *Nature* **312:** 763–767.

Dear, S. and R. Staden. 1991. A sequence and editing program for efficient management of large projects. *Nucleic Acids Res.* **19:** 3907–3911.

Deininger, P.L. 1989. SINEs: Short interspersed repeated DNA elements in higher eucaryotes. In *Mobile DNA* (ed. D.E. Berg, and M.M. Howe), pp. 619–636. American Society for Microbiology, Washington, D.C.

Edwards, M.C. and R.A. Gibbs. 1992. A human dimorphism resulting from loss of an Alu. *Genomics* **14:** 590–597.

Edwards, M.C., P.R. Clemens, M. Tristan, A. Pizzuti, and R.A. Gibbs. 1991. Pentanucleotide repeat length polymorphism at the human CD4 locus. *Nucleic Acids Res.* **19:** 4791.

Frohman, M.A., M.K. Dush, and G.R. Martin. 1988. Rapid production of full-length cDNAs from rare transcripts: Amplification using a single gene-specific oligonucleotide primer. *Proc. Natl. Acad. Sci.* **85:** 8998–9002.

Gibbs, R.A. and M. Cockerill. 1995. Working on the assembly line. *Trends Biochem. Sci.* **20:**162–163.

Gibbs, R.A., P.-N. Nguyen, A. Edwards, A.B. Civitello, and C.T. Caskey. 1990. Multiplex DNA deletion detection and exon sequencing of the hypoxanthine phosphoribosyltransferase gene in Lesch-Nyhan families. *Genomics* **7:** 235–244.

Hofmann, K. and W. Stoffel. 1993. TMBASE-A database of membrane spanning protein segments. *Biol. Chem. Hoppe-Seyler* **374:** 166.

Isobe, M., K. Huebner, P.J. Maddon, D.R. Littman, R. Axel, and C.M. Croce. 1986. The gene encoding the T-cell surface protein T4 is located on human chromosome 12. *Proc. Natl. Acad. Sci.* **83:** 4399–4402.

Janeway, C.A. Jr. 1992. The T-cell receptor as a multicomponent signaling machine: CD4/CD8 coreceptors and CD45 in T-cell activation. *Annu. Rev. Immunol.* **10:** 645–674.

Jurka, J., J. Walichiewicz, and A. Milosavljevic. 1992. Prototypic sequences for human repetitive DNA. *J. Mol. Evol.* **35:** 286–291.

Jurka, J., P. Klonowski, V. Dagman, and P. Pelton. 1995. CENSOR—A program for identification and elimination of repetitive elements from DNA sequences. *Comput. & Chem.* **20:** 119–122.

Klatzmann, D., E. Champagne, S. Chamaret, J. Gurest, D. Guetard, T. Hercend, J.C. Gluckman, and L. Montagnier. 1984. T-lymphocyte T4 molecule behaves as receptor for human retrovirus LAV. *Nature* **312:** 767–778.

Kristensen, T., H. Voss, and W. Ansorge. 1987. A simple and rapid preparation of M13 sequencing templates for manual and automated dideoxy sequencing. *Nucleic Acids Res.* **15:** 5507–5516.

Levine, M.A., P.M. Smallwood, P.T. Moen Jr., L.J. Helman, and T.G. Ahn. 1990. Molecular cloning of B3 subunit, a third form of the G protein B-subunit polypeptide. *Proc. Natl. Acad. Sci.* **87:** 2329–2333.

Littman, D.R. and S.N. Gettner. 1987. Unusual intron in the immunoglobulin domain of the newly isolated murine *CD4* (L3T4) gene. *Nature* **325:** 453–455.

Maddon, P.J., D.R. Littman, M. Godfrey, D.E. Maddon, L. Chess, and R. Axel. 1985. The isolation and nucleotide sequence of a cDNA encoding the T cell surface protein T4: A new member of the immunoglobulin gene family. *Cell* **42:** 93–104.

Maddon, P.J., S.M. Molineaux, D.E. Maddon, K.A. Zimmerman, M. Godfrey, F.W. Alt, L. Chess, and R. Axel. 1987. Structure and expression of the human and mouse T4 genes. *Proc. Natl. Acad. Sci.* **84:** 9155–9159.

Maquat, L.E., R. Chilcote, and P.M. Ryan. 1985. Human triosephosphate isomerase cDNA and protein structure. *J. Biol. Chem.* **260:** 3748–3753.

Muzny, D.M., S. Richards, Y. Shen, and R.A. Gibbs. 1994. PCR based strategies for gap closure in large-scale sequencing projects. In *Automated DNA sequencing and analysis* (ed. M.D. Adams, C. Fields, and J.C. Venter), pp. 182–190. Academic Press, San Diego, CA.

Papa, F.R. and M. Hoschstrasser. 1993. The yeast *DOA4* gene encodes a deubiquitinating enzyme related to a product of the human *tre*-2 oncogene. *Nature* **366:** 313–319.

Pearson, W.R. 1990. Rapid and sensitive sequence comparisons with FASTP and FASTA. *Methods Enzymol.* **183:** 63–98.

Povinelli, C.M. and R.A. Gibbs. 1993. Large-scale sequencing library production: An adaptor-based strategy. *Anal. Biochem.* **210:** 16–26.

Rens-Domiano, S. and H.E. Hamm. 1995. Structural and functional relationships of heterotrimeric G-proteins. *FASEB J.* **9:** 1159–1166.

Richards, S., D.M. Muzny, A.B. Civitello, F. Lu, and R.A. Gibbs. 1994. Sequence map gaps and directed reverse sequencing for the completion of large sequencing projects. In *Automated DNA sequencing and analysis* (ed. M.D. Adams, C. Fields, and J.C. Venter), pp. 191–198. Academic Press, San Diego, CA.

Sambrook, J., E.F. Fritsch, and T. Maniatis. 1989. *Molecular cloning: A laboratory manual,* 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sawada, S., J.D. Scarborough, N. Killeen, and D.R. Littman. 1994. A lineage-specific transcriptional silencer regulates *CD4* gene expression during T lymphocyte development. *Cell* **77:** 917–929.

Tishkoff, S.A., E. Dietzsch, W. Speed, A.J. Pakstis, J.R. Kidd, K. Cheung, B. Bonne-Tamir, A.S. Santachiara-Benerecetti, P. Moral, M. Krings, S. Paabo, E. Watson, N. Risch, T. Jenkins, and K.K. Kidd. 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271:** 1380–1387.

Uberbacher, E.C. and R.J. Mural. 1991. Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci.* **88:** 11261–11265.

Wilkinson, K.D., V.L. Tashayev, L.B. O'Connor, C.N. Larsen, E. Kasperek, and C.M. Pickart. 1995. Metabolism of the polyubiquitin degradation signal: Structure, mechanism, and role of isopeptidase T. *Biochemistry* **34:** 14535–14546.

Worley, K.C., B.A. Wiese, and R.F. Smith. 1995. BEAUTY: An enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results. *Genome Res.* **5:** 173–184.

Received Janury 17, 1996; accepted in revised form March 7, 1996.