Nucleotide Sequence Analysis of 77.7 kb of the Human V_β T-Cell Receptor Gene Locus: Direct Primer-Walking Using Cosmid Template DNAs

JERRY L. SLIGHTOM,*,†,[1] DAVID R. SIEMIENIAK,*,† LEANG C. SIEU,* BEN F. KOOP,‡ AND LEROY HOOD§

*Molecular Biology Unit 7242, The Upjohn Company, Kalamazoo, Michigan 49007; †Human Genome Center, The University of Michigan, Ann Arbor, Michigan 48109-0650; ‡Department of Biology, Center for Environmental Health, University of Victoria, Box 1700, Victoria, British Columbia V8W 2Y2, Canada; and §Department of Molecular Biotechnology, University of Washington, 4909 25th Avenue N.E., Seattle, Washington 98105

Received July 29, 1993; revised November 30, 1993

The nucleotide sequence of 77.7 kb from the human T-cell receptor β-chain locus was determined directly from three overlapping cosmid clones using the primer-walking approach. Computer-aided analyses of this sequence reveal the presence of at least 11 genic regions that are closely related to the human T-cell receptor β variable region (TCRBV) gene family. These include five germline sequences that have previously been determined, V_β21.2, V_β8.1, V_β8.2, V_β8.3, and V_β16, and four whose sequences have partially been determined at the mRNA level, V_β6, V_β23, V_β12.2, V_β24. The two remaining V_β Tcr-related sequences have eluded discovery by cDNA and RT-PCR cloning and genomic blot hybridization methods. These two V_β Tcr-related genes lack >75% nucleotide sequence identity with any other V_β Tcr gene member and therefore, by convention, are referred to as new subfamily members V_β25 and V_β26. This lack of shared identity with other subfamily members explains why they were not detected by hybridization. The promoter regions of these V_β Tcr genes contain the conserved Tcr decamer element located between 80 and 110 bp 5′ of the translation start site, generally near a putative TATAA promoter element. Our sequence analysis also reveals that a 3.3-kb duplication unit was involved in the recombination event that produced the closely related V_β8.1 and 8.2 gene subfamily members. This sequenced region of the V_β locus contains an average number of repetitive DNA elements (21 *Alu*, three L1, three *MER*, and three retrovirus-related elements). © 1994 Academic Press, Inc.

INTRODUCTION

The mammalian immune response is dependent on the interaction of various hematopoietic cell types, among which T lymphocytes play a central role as they synthesize and express surface receptors that recognize foreign antigens. The T-cell receptor (Tcr) molecules consist of two types of heterodimeric polypeptides: the α/β Tcr are the molecules that recognize most foreign antigens, while the γ/δ Tcr are expressed by a minor population of T cells. The Tcr are encoded as a family of genes that have an organization similar to that of the immunoglobulin heavy and light chain genes. The polypeptide chains of the Tcr (α, β, δ, and γ) are divided into a variable (V) region that recognizes antigens and a constant (C) region that anchors the T-cell receptor to the T-cell membrane. The variable regions are encoded by a multiplicity of distinct gene segments: variable (V), diversity (D), and joining (J) for β and δ chains and V and J for α and γ chains (see reviews by Lai *et al.*, 1989; Hunkapiller and Hood, 1989; Davis, 1990). During lymphocyte differentiation, these gene segments undergo DNA rearrangements, mediated by adjacent DNA rearrangement signals, to form a contiguous V gene (e.g., $V_\alpha J_\alpha$ and $V_\beta D_\beta J_\beta$) that is spliced together during transcription to its specific C gene region. The use of these different processes involved in the maturation of a T cell greatly increases T-cell gene diversity.

In this report, we are concerned with the investigation of the human Tcr variable-region β-chain (TCRB). The functional V_β gene locus[2] maps to chromosome 7 (Barker *et al.*, 1984; Caccia *et al.*, 1984), although an orphan locus maps to chromosome 9 (Robinson *et al.*, 1993). At present, the exact number of V_β genes encoded in this locus is unknown; however, statistical estimates suggest that humans may have between 60 (Concannon *et al.*, 1986) and 100 (Kimura *et al.*, 1986) functional V_β gene segments. At the time of this report, about 57 distinct V_β gene segments had been identified by cDNA and RT-PCR analyses, and these sequences fall into 24 subfamilies (Toyonaga and Mak, 1987; Ferradini *et al.*, 1991; Robinson, 1991; Li *et al.*, 1991; Gomolka *et al.*, 1993). Lai

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under Accession No. U03115.

[1] To whom correspondence should be addressed. Telephone: (616) 385-5304. Fax: (616) 385-4599.

[2] The HGMW-approved symbols for the genes discussed in this paper are as follows: V_β21.2, TCRBV21S2; V_β8.1, TCRBV8S1; V_β8.2, TCRBV8S2; V_β8.3, TCRBV8S3; V_β16, TCRBV16S1; V_β6, TCRBV6S1; V_β23, TCRBV23S1; V_β12.2, TCRBV12S2; V_β24, TCRBV24S1; V_β25, TCRBV25S1; and V_β26, TCRBV26S1.

0888-7543/94 $6.00

et al. (1988) reported the construction of a physical map of the human TCRBV locus in which they identified the location of 40 of these V_β gene segments using Southern blot analysis of large DNA fragments separated by field-inversion gel electrophoresis. This analysis and that by Robinson *et al.* (1993) indicate that the Tcr V_β gene locus extends over about 605 to 655 kb of chromosome 7 depending upon which insertion/deletion-related polymorphisms locus is mapped. Lai *et al.* (1988) also reported the cloning of about half of this locus in a series of overlapping cosmid clones.

Although the present gene mapping and RT-PCR cloning approaches have proven to be very successful in the investigation of the number and organization of the Tcr V_β gene family, it will no doubt be difficult to depend on these methods to complete this task due to the complex expression patterns of the Tcr V_β genes and because of the presence of pseudo-V_β Tcr genes. Completion of this task can be achieved only by obtaining the complete nucleotide sequence of the V_β Tcr gene cluster. We report here an initial part of this task; we have obtained the nucleotide sequence of a continuous 77.7-kb stretch of this cluster, which was obtained from three overlapping cosmid clones, H7.1, H12.18, and H130.1, originally isolated by Lai *et al.* (1988). Several V_β Tcr gene segments have already been sequenced from this cluster, which includes three members of the V_β8 Tcr gene subfamily, V_β8.1 and 8.2 (a recently duplicated gene pair), V_β8.3 (Siu *et al.*, 1986), and V_β21.2 (Wilson *et al.*, 1990). Two other V_β Tcr gene segments, V_β6 and V_β12-related Tcr genes, were mapped to this cluster by Lai *et al.* (1988); however, the identification of other V_β Tcr gene segments that might be present in this region of the V_β Tcr gene cluster remained unknown until now.

The size of the sequencing project described here approaches those that have generally been considered large scale, i.e., greater than 50 kb; however, this scale is rapidly being expanded due to the increased emphasis on obtaining the complete sequence of the genomes of many organisms, especially that of human. Most large-scale sequencing projects utilize the random or "shotgun" sequencing strategy, which uses large numbers of small subclones and universal oligomer primer(s) to randomly obtain bits of sequence information from a λ or cosmid size insert. After enough bits of sequence to statistically cover the insert are obtained, this insert sequence is reconstructed by assembly of the overlapping sequences. This methodology has the advantage of being applicable to the high throughput mode of automated instruments that utilize nonradioisotopic DNA sequencing methods (Edwards *et al.*, 1989; Legouis *et al.*, 1991; Martin-Gallardo *et al.*, 1992; McCombie *et al.*, 1992; Sulston *et al.*, 1992; Wilson *et al.*, 1992; Koop *et al.*, 1994; Beck *et al.*, 1992). For the present sequencing project, we investigated the feasibility of using the more traditional primer-walking methodology (using radioisotope labeling), but with the addition of two specific challenges; first to sequence cosmid template DNA directly (avoiding the use of a large number of subclones) and second to obtain sequence reactions whose readings extend in the range of 800 bp (reducing the number of reactions needed). We have previously demonstrated that the standard dideoxy-chain termination sequencing method (Sanger *et al.*, 1977) using T7 polymerase is capable of obtaining sequence information directly from cosmid template and that sequence readings in the range of 800 bp can routinely be obtained (Siemieniak *et al.*, 1991).

MATERIALS AND METHODS

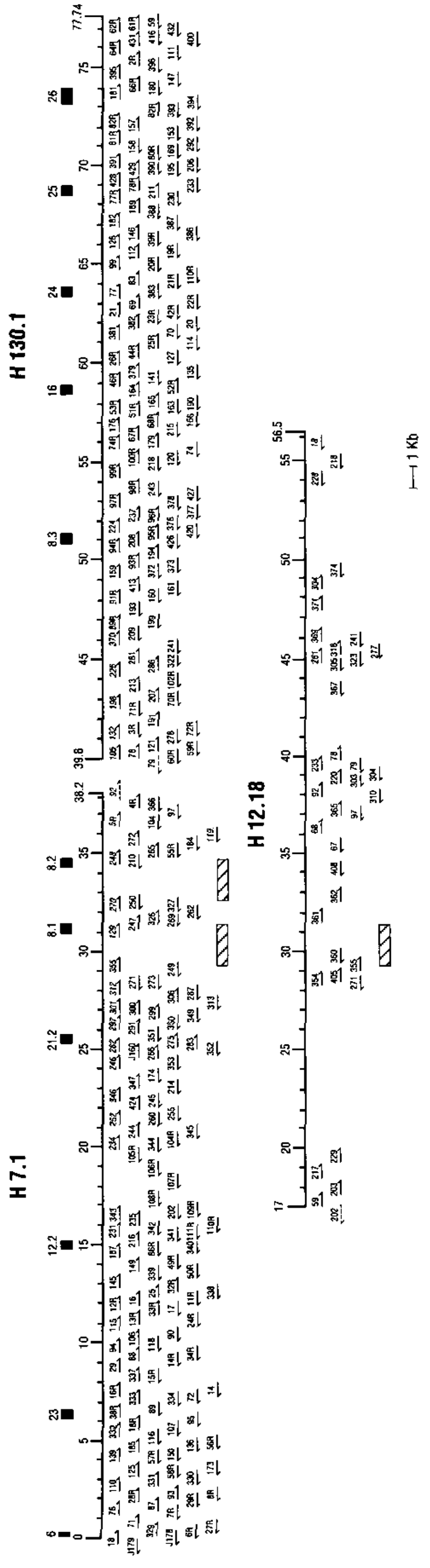
Reagents and equipment. T7 polymerase (Sequenase version 2) and Sequenase sequencing kits were purchased from U.S. Biochemicals. Deoxynucleotides (dNTPs) and dideoxynucleotides (ddNTPs) were obtained from Pharmacia, and [α-^{32}P]dATP and [α-^{32}P]dCTP (sp act, >3000 Ci/mmol, 10 mCi/ml) were purchased from Amersham. The [γ-^{32}P]ATP (sp act, >9000 Ci/mmol) used for end-labeling of restriction enzyme fragments for chemical sequencing was purchased from ICN. Chemicals used for chemical sequencing were obtained from the vendors recommended by Maxam and Gilbert (1980). X-ray role film, 20 cm × 25 m (Kodak XAR-351) was purchased from Kalamazoo X-ray and the DNA sequencing gel stands and safety cabinets (described by Slightom *et al.*, 1991) were purchased from Fotodyne, Inc.

Oligonucleotides. Oligomer primers (usually 20 nucleotides in length) were synthesized on an Applied Biosystems DNA synthesizer (Model 380B) using phosphoramidite chemistry. The synthesized oligomer primers were deprotected in NH_4OH at 55°C, and the NH_4OH was removed by evaporation in a Savant Speedvac. Dried oligomer primers were resuspended in sterile double-distilled H_2O, concentrations measured (OD_{260}), and diluted to a final concentration of 10 μg/ml in sterile double-distilled H_2O and used without any further purification or analysis.

Cosmid DNA preparation. Since large amounts of each cosmid template DNA (H7.1, H12.18, and H130.1) were needed, a standard large plasmid DNA preparation method was used. This method included the use of two CsCl centrifugation steps to ensure removal of protein and RNA contaminates. This cosmid DNA isolation method utilizes the basic alkaline extraction method described by Birnboim and Doly (1979) followed by the use of 7 *M* NH_4OAc for the neutralization step (Morelle, 1989). A detailed description of the method, for the preparation of cosmid DNA from 1-liter growths, has been reported previously (Slightom *et al.*, 1991). The two CsCl ethidium-bromide gradients were done using a fixed-angle rotor (Ti70.1 or equivalent) to ensure purity. Vertical-angle rotors were avoided because they did not provide adequate separation of the plasmid or cosmid DNA band from small RNA molecules.

DNA sequencing reactions. The DNA sequencing strategy used for this project utilized the chemical sequencing method (Maxam and Gilbert, 1980) to obtain sequence initiation points within the cosmid inserts and the DNA chain termination (Sanger *et al.*, 1977) method to obtain closure of sequence contigs. The chemical sequencing reactions were done essentially as described by Maxam and Gilbert (1980). About 10 μg of cosmid DNA was subjected to digestion by a restriction enzyme (either *Bam*HI or *Eco*RI) followed by 5'-end-labeling with [γ-^{32}P]ATP as described by Slightom *et al.* (1991) and digestion with the corresponding second restriction enzyme (*Eco*RI or *Bam*HI) to obtain asymmetric ^{32}P-labeled DNA fragments. Enzymatic DNA sequencing reactions were done using T7 polymerase, using both [α-^{32}P]dATP and [α-^{32}P]dCTP, as described by Siemieniak *et al.* (1991). The ratio of ddNTP:dNTP was reduced to 1:30, which allowed many of these sequence reactions to be read beyond 800 bp from the oligomer primer. Sequencing errors were minimized by obtaining sequence information from both DNA strands (see Fig. 1) and by proofreading the sequencing films back into the assembled sequence twice by different individuals.

Computer-aided DNA sequence analyses. The nucleotide sequence was searched for potential coding regions using GRAIL (Gene Recognition and Analysis Internet Link, Oak Ridge, TN; Uberbacher and Mural, 1991) and specifically for V_β Tcr genes using the INHERIT



(Applied Biosystems, Foster City, CA) computer packages. The locations of repetitive DNA elements [SINEs, LINEs, MERs, $(TG)_n$, etc.] were identified by search routines in the INHERIT and GCG (Devereux *et al.*, 1984) (Madison, WI) DNA analysis computer packages. A computer program designed by Jurka and Milosavljevic (1991) was used for *Alu* classification. Database searches (GenBank, r74) were done using the FASTA (Pearson and Lipman, 1988) program available within the GCG computer package.

RESULTS AND DISCUSSION

DNA Sequencing Strategy

The nucleotide sequence of these cosmid clones was determined using the strategy described by Siemieniak *et al.* (1991), which mostly uses the dideoxy-chain termination method (Sanger *et al.*, 1977) to extend sequence information for cosmid/insert boundaries and from sequence initiation points (IPs) located within the cosmid insert. Internal IPs were obtained using the chemical sequencing method (see Materials and Methods), from previously sequenced regions (Siu *et al.*, 1986; Wilson *et al.*, 1990), and from the overlapping boundaries of the cosmid clones (Fig. 1). Since cosmids H7.1 and H130.1 span the major portion of this cloned V$_\beta$ Tcr gene contig (Lai *et al.*, 1988; Fig. 1) we attempted to obtain at least five IPs for each of these cosmids. From the final sequence, we determined the locations of all IPs used, including those obtained by the chemical sequencing method. A total of seven IPs were used to primer-walk cosmid H7.1 [at the following locations: 0 kb (5′-end), 8.0 kb (*Bam*HI-*Eco*RI fragment chemical sequence), 12 kb (*Bam*HI-*Eco*RI fragment chemical sequence), 17 kb (overlap with cosmid H12.18), 26 kb (V$_\beta$21), 31 kb (V$_\beta$8.1), and 38.2 kb (3′-end)], and six IPs were used for cosmid H130.1 [39.8 kb (5′-end), 40.5 kb (*Bam*HI-*Eco*RI fragment chemical sequence), 51 kb (V$_\beta$8.3), 56.5 kb (overlap with cosmid H12.18), 62 kb (*Bam*HI-*Eco*RI fragment chemical sequence), and 77.7 (3′-end)]. The location of each cosmid vector pTL5-insert junction with respect to the final sequence is listed in Table 1. Only the 5′ pTL-insert junction of cosmid H130.1 could not be sequenced using a primer from the flanking cosmid re-

FIG. 1. Organization of cosmid clones H7.1, H12.18, and H130.1 and a summary of the strategy used to determine their nucleotide sequence. Cosmid clones H7.1, H12.18, and H130.1 were isolated and mapped as described by Lai *et al.* (1988), and the nucleotide sequence of these clones was determined using the primer-walking method as described by Siemieniak *et al.* (1991). The numbered lines (in kb) correspond with the size of the individual cosmid cloned inserts, starting from the 5′-end of cosmid clone H7.1 (0 kb) to the 3′-end of cosmid clone H130.1 (77.7 kb), and they indicated the regions of cosmid clone 12.18 that overlap with either H7.1 or H130.1. The locations of V$_\beta$ Tcr germline genes are indicated by black boxes and corresponding numbers above clones H7.1 and 130.1. The primer-walking sequencing strategy is shown by arrows below the cosmid clone maps, where the oligomer primer number identifies each arrow, and the direction and distance sequenced correspond to the direction and length of each arrow. The hashed boxes shown below the V$_\beta$8.1 and 8.2 Tcr genes show the regions that were previously sequenced from subclones of either cosmid H7.1 or H12.18 (Siu *et al.*, 1986; W. Funkhouser, B.F.K., manuscript in preparation).

TABLE 1

Cosmid Cloned Insert Junctions and Organization with Respect to Vector pTL5

Cosmid	Nucleotide positions[a]		pTL5 junction oligomers[b]	Insert orientation with respect to pTL5
	5′ end	3′ end		
H7.1	0	38,183	18 (yes) 59 (yes)	5′ to 3′ (N)[c]
H12.18	17,016	56,554	59 (yes) 18 (yes)	3′ to 5′ (U)
H130.1	39,789	77,743	18 (no) 59 (yes)	5′ to 3′ (N)

[a] All cosmid insert junctions were at *Sau*3AI restriction enzyme sites, which is consistent with their construction (Lai *et al.*, 1988). Nucleotide positions include the *Sau*3AI recognition site.

[b] Oligomers 18 and 59 flank the 5′ and 3′ junctions, respectively, of cosmid pTL5. These oligomers were used to determine insert nucleotide sequences adjacent to the vector cloning site.

[c] Orientation of inserts cloned in cosmid vector pTL5. The symbol (N) indicates inserts that have the same 5′ to 3′ orientation as the vector, while the symbol (U) indicates that the insert has the opposite orientation.

gion, which suggests that this clone may contain multiple copies of this pTL5 vector junction.

The composite sequencing strategy (Fig. 1) for these cosmids shows that both DNA strands were sequenced (with the inclusion of previously sequenced regions, hashed line regions in Fig. 1, and overlapping regions from cosmid 12.18). Comparison of the overlapping sequences from cosmids H7.1 and H130.1 with those derived from cosmid H12.18 also revealed the location for 24 putative polymorphic sites in about 16 kb of overlapping sequence reads (0.15% polymorphism), as these cosmid clones were derived from DNA isolated from more than one individual (Lai *et al.*, 1988). These putative polymorphic sites are listed in Table 2 along with the nucleotide difference found at each site. Lai *et al.* (1988) characterized this cosmid contig by the presence of a *Sfi*I site that they mapped about 10 kb 5′ of the V_β8.1 gene; we have located this *Sfi*I site between positions 23966 and 23978, or about 7 kb 5′ of the V_β8.1 translation initiation site.

TABLE 2

Polymorphic Nucleotide Positions Determined from Sequencing Overlapping Regions of Cosmid Clones H7.1, H12.18, and H130.1

Nucleotide position	Cosmid clones	Polymorphic difference
17,267	H7.1/H12.18	C/–
17,307	H7.1/H12.18	C/–
17,439	H7.1/H12.18	T/–
17,464	H7.1/H12.18	–/A
17,473	H7.1/H12.18	T/C
19,090	H7.1/H12.18	T/C
19,694	H7.1/H12.18	T/A
29,688	H7.1/H12.18	C/T
29,993	H7.1/H12.18	T/C
35,487	H7.1/H12.18	C/T
35,516	H7.1/H12.18	T/C
35,603	H7.1/H12.18	A/G
35,744	H7.1/H12.18	T/C
35,781	H7.1/H12.18	A/G
36,289	H7.1/H12.18	A/T
36,683	H7.1/H12.18	A/G
36,689	H7.1/H12.18	C/T
38,008	H7.1/H12.18	G/A
38,025	H7.1/H12.18	G/A
44,619	H130.1/H12.18	A/T
45,470	H130.1/H12.18	T/C
45,697	H130.1/H12.18	A/G
55,215	H130.1/H12.18	C/T
56,420	H130.1/H12.18	A/T

The sequence strategy used and depicted in Fig. 1 was straightforward; it did not require the use of any special computer-aided program (other than the GCG program, ASSEMBLE) to manage or to assemble the individual sequencing runs into the final composite sequence. In addition, very few sequencing procedural problems were encountered in this project. We anticipated that the presence of repetitive elements (SINEs and LINEs) could interfere with the direct primer-walking strategy; however, as described below the density of these elements in this region of the V_β Tcr gene cluster is not high. The 290-bp *Alu* elements did not present any problem because they could easily be sequenced across, and the accidental use of a primer derived from an *Alu* element sequence was reduced by screening new oligomer primers against a consensus *Alu* sequence. A minor sequencing problem was encountered between positions 13,060 and 13,547, which contain an unusual stretch of simple sequence DNA consisting of about 387 bp of alternating purine and pyrimidine bases followed by about 100 purine bases. With the identification of the 5′ and 3′ boundaries for this region, flanking oligomer primers were designed and this region was sequenced across on both DNA strands.

We did, however, encounter one major sequencing problem directly attributed to sequencing cosmid size templates that even our long sequence readings could not resolve. This problem involved our efforts to sequence across the duplicated V_β8.1 and 8.2 Tcr gene regions. Because this is a recent duplication event (see below), the duplication units (about 3.3 kb each) have nearly identical sequences, diverging by an average of 6%, while divergence in the genic regions is much less, about 2%. The sequence of about 0.8 kb from each duplication unit, which included the V_β8.1 and 8.2 Tcr genes, was previously determined from cosmid H7.1 subclones (Siu *et al.*, 1986), and more recently the length of sequence information from each duplication unit was extended to about 2.1 kb (W. Funkhouser, B.F.K., manuscript in preparation). This sequence information was used to locate divergent sequence regions that were used to design oligomer primers that could potentially prime unique sequence readings using cosmid template. This strategy was successful at the duplication unit boundaries (where divergence is greater) and we were able to sequence into and out of each duplication unit. This information was important in determining the precise duplication unit junctions. However, as we approached the V_β8.1 and 8.2 Tcr genes proper, the design of primers became more difficult (containing fewer mismatches) due to the higher degree of shared identity, which resulted in multiple primed sequence readings. Since the complete sequences of these V_β8 Tcr genic regions were already known, we discontinued our attempt to directly sequence the complete duplication units using cosmid DNAs. About 700 bp from each duplication unit was not sequenced directly from cosmids H7.1 or H12.18 (Fig. 1).

Another major problem encountered in this project was that about 25% of the primers used (122 primers) failed to primer readable sequencing reactions. The exact reason for each of these failures was not determined; however, since these primers were selected from the ends of "raw" sequence readings, most failures were probably the result of reading errors incorporated into the synthesized primers. A check of the failed primer sequences with the final sequences indicated that this was true for about half of the failed oligomer primers. A second reason for primer failure could be that these primers were not subjected to any quality control, except that obtained at the level of the synthesis instrument. Thus, the quality of these primers was subject to sequence entry errors and DNA synthesis instrument errors. It was decided from the onset of this project that it would be more efficient not to spend the resources to improve the quality control process for oligomer synthesis. However, the fact that this lack of quality control for the selection and synthesis of oligomer primers increases the work load by about 25% indicates that these are areas for improvements. This process could be improved by the addition of several simple steps, which include the use of a computer-aided program to improve the oligomer primer selection process (ensure selection is from an accurate reading and that the best primer composition is used), on-line transfer of oligomer sequence information from the selection program to the synthesis instrument, and improved quality of DNA synthesis.

A summary of the materials used and sequencing efficiency achieved for this sequencing project is listed in Table 3. We were able to complete this project using sequence information obtained from 360 primer-directed sequencing reactions, which had an average sequence reading length of 680 bp. However, if only the sequence readings that were designed to be read to their maximal length are included, the average reading length is increased to 762 bp. In many cases, we were able to obtain sequence readings that extended beyond 900 bp, and for several reactions readings extending beyond 1000 bp were obtained. Clearly, a major strength of the T7 polymerase sequence reaction and the gel methodology used (Siemieniak *et al.*, 1991; Slightom *et al.*, 1991) is the ability to obtain long sequence readings, which greatly reduced the amount of materials and time-consuming tasks associated with the DNA sequencing operation (number of oligomer primers, number of DNA sequencing reactions, amount of cosmid DNA used, number of gels, and the number of manually read sequencing ladders). The compete nucleotide sequence determined directly from these cosmid clones was 76,393 bp, which does not include the 1350 bp of the duplicated V_β8.1 and 8.2 Tcr gene regions (see above). This complete sequence was obtained on both DNA strands from reading a total of 244,367 bp (Table 3), which is a sequencing redundancy of 3.2-fold, slightly above the expected minimum redundancy level of 2.5-fold, but well below the 5- to 6-fold coverage found for sequencing project that predominately used the random sequencing strategy (Edwards *et al.*, 1989; Legouis *et al.*, 1991; Martin-Gallardo *et al.*, 1992; McCombie *et al.*, 1992; Sulston *et al.*, 1992; Wilson *et al.*, 1992; Beck *et al.*, 1992). Clearly, the major advantage of the direct primer-walking strategy is the reduced number of subclones and sequence reactions needed to obtain the 77.7 kb of nucleotide sequence. However, a distinct disadvantage is that it suffers from a low throughput due to the time needed for film exposure, manual reading, and synthesis of new oligomer primers after each series of primer-walking steps.

TABLE 3

Summary of Sequencing Project: Materials Used and Sequence Efficiency

Total No. of oligomer primers used	482
No. of readable sequences generated	360
No. of oligomer primer failures	122
Overall primer efficiency	75%
Total amount of cosmid DNAs used	1.5 mg
Total number of bp read by primer-walking	244,357 bp
Gross No. bp read/primer	680 bp
Net No. bp read/primer[a]	762 bp
Finished bp sequenced by primer-walking	76,393[b]
Sequence redundancy	3.2-fold

[a] This calculation excludes reactions not intended to achieve maximum reading length.

[b] This number excludes the 1350 bp previously by Siu *et al.* (1986).

Search for V_β Tcr Genes

Analysis of the variability and number of V_β Tcr genes by cDNA and RT-PCR methods has been very successful in identifying 57 unique V_β Tcr genes, a large percentage of the total estimated genes (Concannon *et al.*, 1986; Kimura *et al.*, 1987; Robinson, 1991). However, relatively few of the human V_β Tcr genes have been investigated at the germline level; those that have include five members of the V_β8 gene subfamily (Siu *et al.*, 1986), three members of the V_β21 gene subfamily (Wilson *et al.*, 1990), a V_β16 gene (Smith *et al.*, 1987), and several members of the V_β5, 6, and 13 subfamilies (Li *et al.*, 1991). The organization of the V_β Tcr genes appears to be conserved, consisting of two exons separated by a short intron. The size of the first exon is difficult to determine since it contains the 5′-untranslated region that can vary considerably in length (Anderson *et al.*, 1988) and a coding region that varies between 49 to 79 bp (Siu *et al.*, 1986; Wilson *et al.*, 1990; Smith *et al.*, 1987; Li *et al.*, 1991). The size of the second exon is more conserved, between 290 and 293 bp in length. However, identification of V_β Tcr gene subfamily members in a newly sequenced region is difficult because the subfamily members show considerable divergence from each other and because their intron-interrupted coding regions are relatively small. Two computer-aided search routines, the Oak Ridge National Laboratory GRAIL server (Uberbacher and Mural, 1991) and the INHERIT package were used for this analysis. The latter is a more direct search because it utilizes sequence information from related DNA elements or genes in a search routine based on the sequence alignment algorithm of Smith and Waterman (1981). In the INHERIT search, a total of 57 human- and mouse-derived V_β Tcr cDNA sequences were used in dot-matrix analysis with the 77.7-kb sequence (Fig. 2).

GRAIL analysis of this 77.7-kb region revealed the locations of 21 potential coding regions, of the nearly 40 found by the open reading frame analysis, which includes 8 of the 11 V_β Tcr gene exon 2 regions (Fig. 2). The dot-matrix result from the INHERIT analysis clearly revealed the locations of the 5 known V_β Tcr genes and suggested locations for at least 6 other family members without displaying much interference from other potential coding regions (Fig. 2). The regions suggested to encode additional V_β Tcr genes were further analyzed using pairwise alignment programs BESTFIT (based on the algorithm of Needleman and Wunsch, 1970) and FASTA (Pearson and Lipman, 1988) from the GCG computer package (see below). A summary of the V_β Tcr genes identified in this region of the human V_β Tcr cluster is presented in Table 4 along with the sequence of potential heptamer and nonamer recombination signal sequences (Sakano *et al.*, 1979). This analysis identified the locations of the known V_β Tcr genes (8.1, 8.2, 8.3, 16, and 21.2) and the locations of two other V_β Tcr genes mapped to this region (exon 2 of V_β6 and the complete V_β12 gene) (Fig. 3). Most importantly, this analysis located genes encoding V_β23 and V_β24 and two additional genes that appear to represent new members of the human V_β Tcr gene family, as they do not share >75% sequence identity with any known V_β Tcr gene subfamily member. We refer to these new human V_β Tcr gene subfamily members as V_β25 and V_β26 (Figs. 3H and 3I, respectively); however, the functionality of these genes is questionable (see below). The INHERIT dot-matrix analysis also suggests the existence of other sequence regions that share more distant identities with Tcr gene parts, most notably two regions between V_β24 and 25 (positions 66456 to 66526 and 66677 to 66711) that share identity with Tcr exon 1 and exon 2 of V_β12.2, respectively. However, the existence of an additional V_β Tcr gene in this region is questionable since this region does not contain any of the conserved molecular components (see above) associated with a V_β Tcr gene. At present, we do not know whether these V_β Tcr-related sequences represent a proto-Tcr gene or if their existence is only coincidental.

General Features of V_β Tcr Gene Flanking and Untranslated DNA Regions

The identification of consensus promoter elements (CCAAT and TATAA) in the 5′-flanking sequences of these V_β Tcr genes is difficult because they show considerable divergence. Functional mapping of the transcriptional start site of 14 murine V_β Tcr genes by Anderson *et al.* (1988) revealed that the location of this site, with respect to the translation start site, also varies considerably and that many of these genes utilized multiple transcriptional start sites. Many of the murine V_β Tcr gene transcriptional start sites were found 70 to 120 bp 5′ of the translation start site; however, others were located at distances ranging between 260 and 750 bp 5′ of the translation start site (Anderson *et al.*, 1988). The 5′-flanking regions of the human V_β8 Tcr gene family has been studied previously by Siu *et al.* (1986), and putative CCAAT and TATAA promoter elements were assigned; however, none of these exactly matches its respective consensus sequence. Our analysis of the 5′-flanking regions of the other human V_β Tcr genes located in this sequenced region shows a similar degree of diversity in the location and sequence of potential CCAAT and TATAA promoter elements. Figures 3A to 3I indicate the locations for potential promoter elements that most closely match (generally differing by only one mismatch) the consensus promoter element sequences. Whether any of these potential promoter elements play a role in regulating the expression of a particular V_β Tcr gene remains to be determined.

Although the sequences of these V_β Tcr 5′-flanking DNA regions are diverse, they must function in a similar manner to direct the expression of the rearranged V_β Tcr gene. A similar degree of diversity is also found for the immunoglobulin V-region gene promoters, but a conserved octamer sequence element was identified about 90 to 160 bp 5′ of the translation initiation site, for the light-chain encoding genes (ATTTGCAT) and heavy-chain encoding genes (ATGCAAAT). These elements appear to determine the tissue specificity of immunoglobulin transcription (Parslow *et al.*, 1984; Falkner and Zachau, 1984; Bergman *et al.*, 1984; Mason *et al.*, 1985). Sequence elements that match either of these immunoglobulin V-region octamer elements were not found in the 5′-flanking regions of the murine V_β Tcr genes described by Anderson *et al.* (1988) nor are they found in the 5′-flanking regions of the human V_β Tcr described in this report. Exact matches to these immunoglobulin promoter elements were found (six matches) within the 77.7-kb sequence; however, none of these sequences is located in the appropriate 5′ promoter location of the V_β Tcr genes (data not shown).

Anderson *et al.* (1988) did report finding that most of the murine and the human V_β8.1 genes share a con-

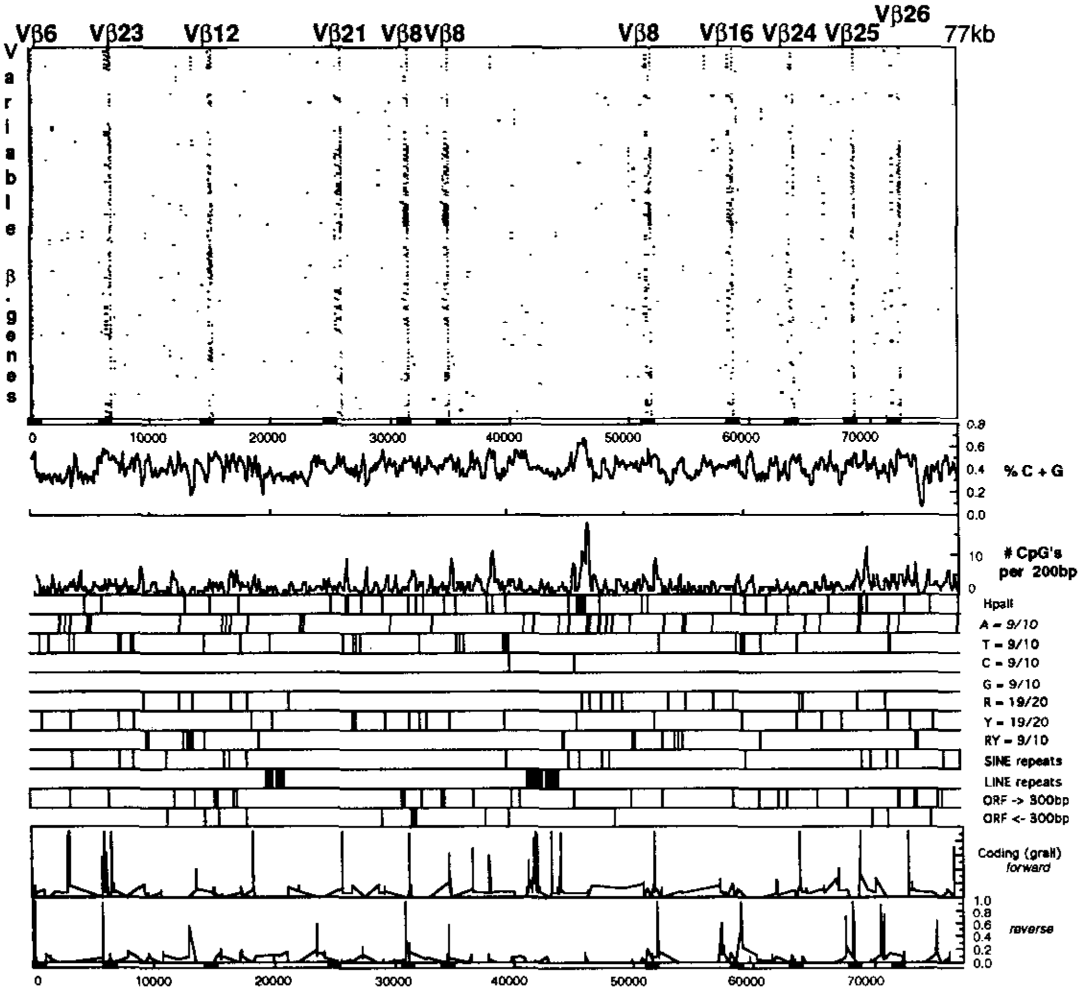


FIG. 2. *A summary of features found in the 77.7-kb region of the human TCRBV gene cluster. The top portion represents the identity* search pattern obtained from the INHERIT scan of 57 human and mouse V_β Tcr gene sequences linked in tandem on the vertical axis against the 77.7-kb sequence. The vertical dotted lines correspond to the location of V_β Tcr-related sequences whose identities were determined by more detailed analyses (see text). The assigned V_β Tcr gene number is presented above the corresponding vertical dot line. The lines shown below the map correspond to various searches of this sequenced region. The first line indicates the % G + C, the second line indicates the number of CpG dinucleotides found in a 200-bp window, and the third line shows the location of *Hpa*II sites. The lines marked A, T, C, G, R, Y, and RY mark the positions of simple repeats, where greater than 9 (19) positions out of 10 (20) correspond to the respective base specificity (R, purines; Y, pyrimidines; RY, alternating purines and pyrimidines). The positions of SINEs and LINEs are indicated below the simple repeats. The next *line shows the positions of open reading frames greater than 300 bp in length (all three reading frames are combined, but each DNA strand is* presented separately). The two graphs show the peak position of coding regions identified using GRAIL.

served decamer element with the consensus sequence AGTGAYRTCA, which is generally located close to a putative TATAA element. This finding has been further confirmed, as similar Tcr decamer elements are present, 80 to 106 bp 5′ of the translation start site, in human $V_\beta5$, 6, and 13 Tcr gene subfamily members (Li *et al.*, 1991). Using a DNase I footprint analysis of this region of the human $V_\beta8.1$ Tcr gene, Royer and Reinherz (1987) found that this Tcr decamer element does bind nuclear binding proteins. A search of this 77.7-kb sequence revealed *the presence of seven copies of this Tcr decamer*, and interestingly, all of these are located in the 5′-flanking region of a V_β Tcr gene. However, unlike its position in previous reports, this Tcr decamer is found in both orientations with respect to V_β Tcr gene transcription. Sequence motifs that exactly match this decamer sequence were found between 72 and 106 bp 5′ of the translation start site of $V_\beta23$, 21 (opposite strand), 8.1 and 8.2 (both strands), 8.3 (opposite strand), and 16 (opposite strand). Sequences that closely match the consensus Tcr decamer element were found in similar locations of the remaining V_β Tcr genes; these locations are indicated in Figs. 3A to 3I. One additional observation is that several of the V_β Tcr genes contain an exact match to the sequence TGATGTCACTG. Exact matches to this sequence are found near the Tcr decamer element of $V_\beta23$, 21, 8.1, 8.2, 8.3, and 16 (Figs. 3A and 3C–3F).

Other than the presence of these Tcr decamer-related *elements, the nucleotide sequences of murine and hu*man V_β Tcr genes are very diverse, suggesting that the mechanism that controls their expression may be more complex than that seen for other genes. This is most likely the case, since expression control should also involve the relocation of other enhancer elements following the rearrangement events that bring together the full genetic complement of the V_β Tcr gene. In addition, ex-

TABLE 4

Locations of T-Cell Receptor V_β-Type Genes within 77.7-kb Sequenced Region

V_β gene	Coding Exon 1	Intron 1	Exon 2	Heptamer[a]	Nonamer
6	—	—	1/253	(5)-CACAGCA	(23)-TCACAAACC
23	6305/6383	6384/6493	6494/6783	(6)-CACAGAC	(22)-GTACCCAAA
12.2	14717/14765	14766/14871	14872/15161	(5)-CACAGTG	(22)-CACGTAAAC
21.2	25365/25413	25414/25504	25505/25797	(5)-CACAGTG	(23)-GCAGAAAAC
8.1	30967/31015	31016/31115	31116/31408	(5)-CACAGCG	(23)-GCAGAAAAC
8.2	34290/34338	34339/34438	34439/34731	(5)-CACAGCG	(23)-GCAGAAACC
8.3	51481/51529	51530/51629	51630/51922	(5)-CACAGCG	(23)-GCAGAAACC
16	58424/58472	58473/58558	58559/58851	(5)-CACAGTG	(22)-TGCAAAACC
24	63499/63547	63548/63672	63673/63962	(5)-CACAGAG	(21)-GTTCATAAA
25	68508/68556	68557/68663	68664/68956	(5)-CACAATG	(21)-GACACAGAC
26	72124/72172	72173/72563[b]	72564/72853	(5)-CACAGCA	(22)-GTGCAAACC
			Consensus	CACAGYG	GYNNAAA$^{A}/_{C}$C

[a] Numbers in parentheses refer to the number of basepairs 3′ of the Tcr coding region for the heptamer element and from the end of the heptamer for the nonamer element.

[b] The intron of V_β26 is 290 bp longer than expected due to the insertion of an *Alu* element.

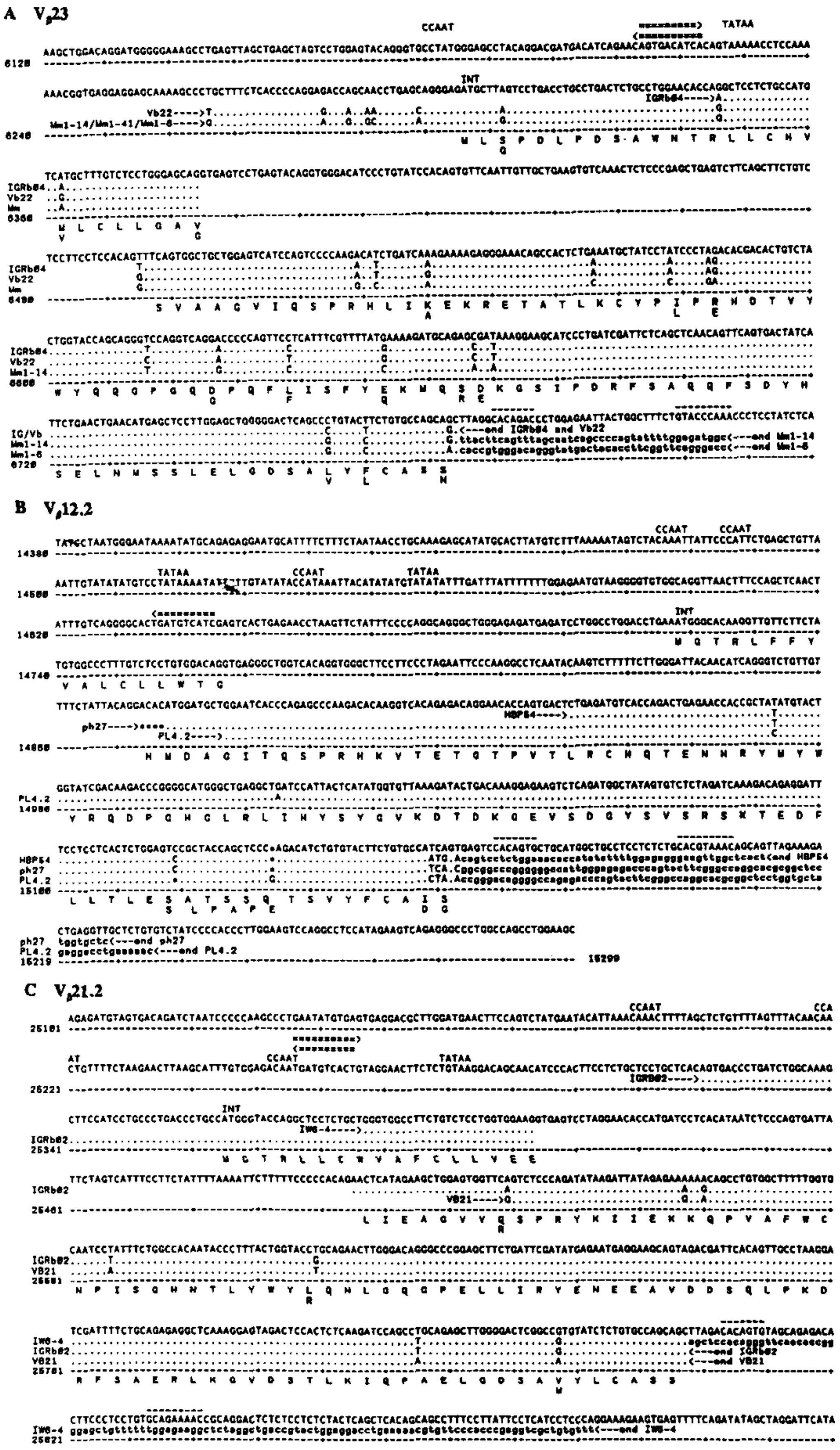
pression of a V_β Tcr gene may not be dependent on the use of strong promoter elements located in the V_β 5′-flanking DNA region, which could explain why this region is not well conserved. This is consistent with the mapping of a strong transcriptional enhancer in a region 5.5 kb 3′ of the murine C_β gene (Krimpenfort *et al.*, 1988; McDougall *et al.*, 1988; Gottschalk and Leiden, 1990).

V_β Tcr Gene Organization and Subfamily Assignments

The 11 V_β Tcr genes located in this 77.7-kb region of the V_β Tcr cluster all have the same 5′ to 3′ orientation with respect to the C_β gene regions (Li *et al.*, 1988; Wilson *et al.*, 1988; Robinson *et al.*, 1993). Assignment of these V_β Tcr genes to a specific V_β gene subfamily is relatively straightforward since by convention each subfamily differs by >25% at the nucleotide sequence level. The evolution of the V_β Tcr gene cluster has been the result of many gene duplication events, some of which have involved only a single V_β Tcr gene segment and others of which may have involved multiple V_β Tcr gene segments (mega duplications). The result is that members of the same gene subfamily may be widely distributed across the V_β Tcr gene cluster. Some of the subfamily members are nearly identical, indicating recent duplication events; such is the case with V_β8.1 and 8.2 gene segments (see below). Other duplication events may be much more ancient. For example, Lai *et al.* (1988) mapped six loci that share a high degree of identity and belong to the V_β6 subfamily, but these genes are spread across nearly 250 kb of the V_β Tcr gene cluster. The assignment of a specific V_β gene sequence, derived from sequencing a cDNA or RT-PCR product, to a specific germline gene location within the cluster is difficult due to the presence of polymorphic sites and/or nucleotide errors incorporated as part of the cDNA or RT-PCR synthesis, cloning, and/or sequencing processes. Such assignments are also made more complicated since the 3′-ends of cDNA clones are likely to include nucleotide sequence modifications (chew backs and base insertions) that are part of the N-region diversity that occurs during the joining of the V and D gene segments (see reviews by Hunkapiller and Hood, 1989; Davis, 1990).

The FASTA computer program (Pearson and Lipman, 1988) was used to search GenBank (r74) with the coding region of each V_β Tcr gene listed in Table 4. FASTA GenBank searches were used to identify which

FIG. 3. The nucleotide sequence of the V_β germline genes located in this 77.7-kb region of the V_β cluster in comparison with the most related cDNA and RT-PCR clone sequences found in GenBank. The nucleotide sequence of each V_β Tcr gene is presented (frames A to I) starting about 200 bp 5′ of the translation start site (ATG codon) and extending beyond the nonamer recombination signal. The nucleotide sequences are numbered to correspond with the continuous 77.7-kb sequence (GenBank Accession No. U03115), and the deduced amino acid sequence of each V_β Tcr gene is presented below the counting line. The nucleotide sequences derived from the most closely related cDNA or RT-PCR clones (Table 5) are presented below each germline Tcr gene sequence; however, for these latter sequences, only the differences (starting and ending positions and potential polymorphic sites) are presented. Dots indicate sequence positions that are identical with the germline V_β Tcr gene shown in each respective frame (in some cases the sequences of the identical regions of cDNA and RT-PCR clones were grouped to conserve space). Nucleotide sequence differences that result in amino acid replacements are indicated by the presence of the corresponding amino acid below the counting line. The germline and cDNA and RT-PCR clone sequences for V_β8.1 and 8.2 have been combined to conserve space (frame D). The germline sequence of V_β8.1 is shown on the top line, followed by the most closely related cDNA and RT-PCR clone sequences, which are then followed by the germline sequence of V_β8.2 and its most closely related cDNA and RT-PCR clone sequences. Each V_β8 germline sequence has its own respective nucleotide sequence numbering system, which can be identified by the 8.1 or 8.2 designation added to the end of each respective germline sequence line. The locations of potential promoter elements, CCAAT and TATAA, are indicated above the sequence, and the locations of potential Tcr decamer (and related TGATGTCACTG) elements on either one or both DNA strands are indicated by the direction of double-dashed-line arrows. Recombination heptamer and nonamer signal sequences (also presented in Table 4) are identified by single overlines. Lowercase sequence shown near the 3′-end of some cDNA sequences corresponds to regions of the rearranged V_β Tcr gene that may be the result of N-region diversity; no effort was made to maximize the alignment of the diversity region of these cDNA-derived sequences with the germline sequence.


A $V_\beta 23$
CCAAT
TATAA
INT
IGRb84
Vb22
Wm1-14/Wm1-41/Wm1-6
6128
6248
6368
6488
6608
6728
<---end IGRb84 and Vb22
<---end Wm1-14
<---end Wm1-6
B $V_\beta 12.2$
CCAAT
TATAA
INT
ph27
PL4.2
HBP54
14388
14508
14628
14748
14868
14988
15108
15219
15299
<---end HBP54
<---end ph27
<---end PL4.2
C $V_\beta 21.2$
CCAAT
TATAA
INT
IGRb82
IW6-4
VB21
25101
25221
25341
25461
25581
25701
25821
<---end IGRb82
<---end VB21
<---end IW6-4

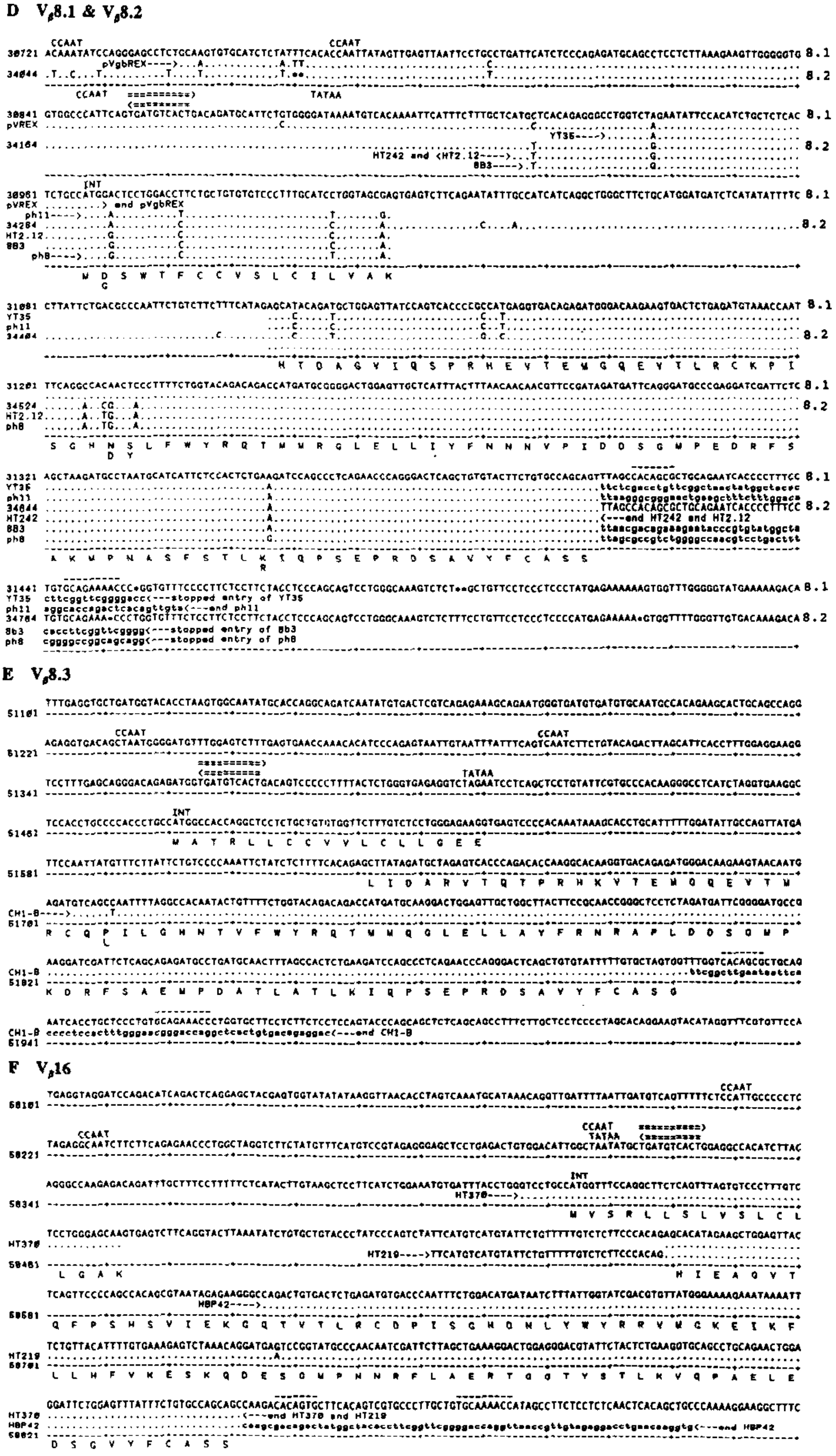


FIG. 3—*Continued*

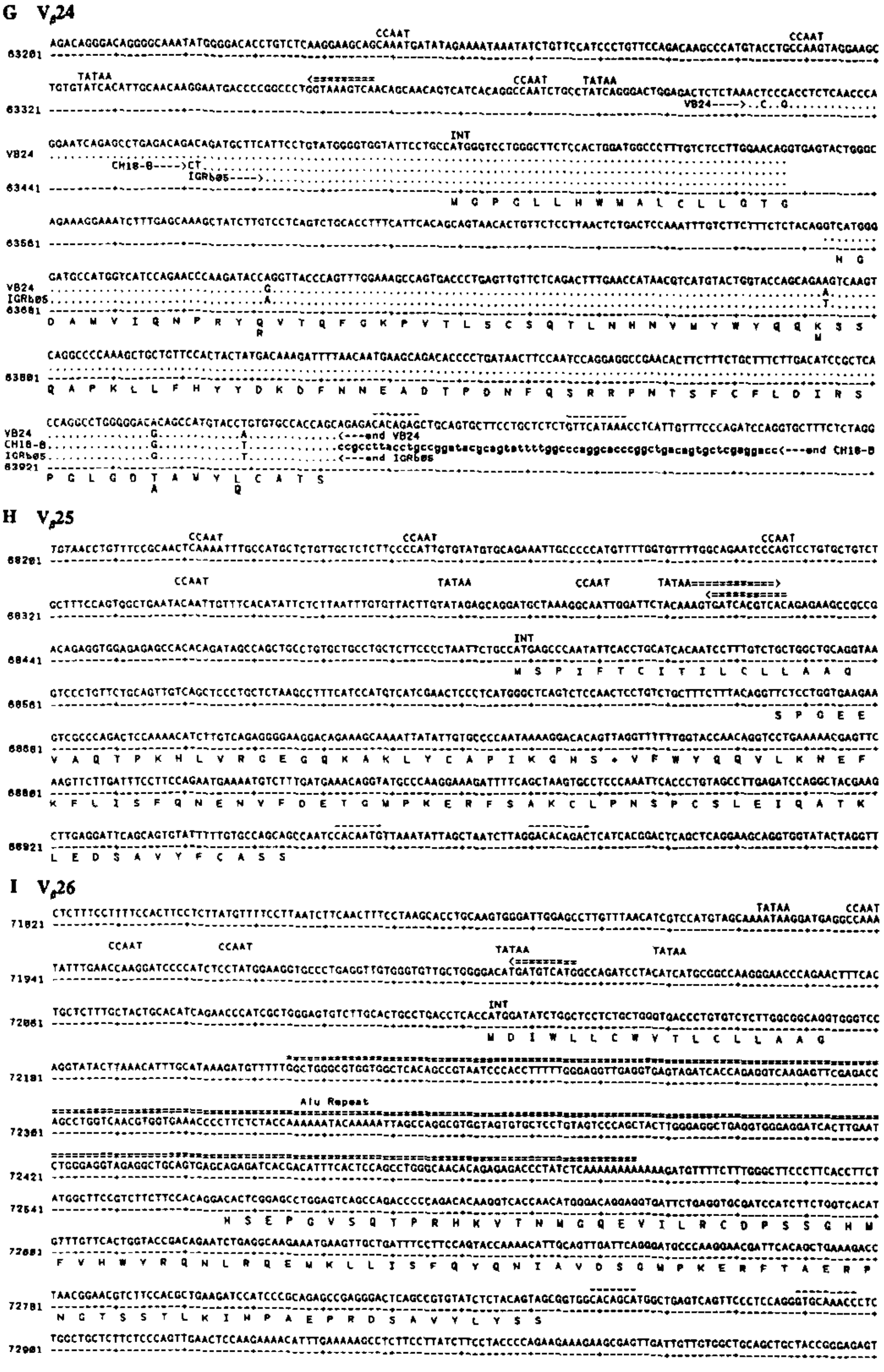


FIG. 3—*Continued*

of these V_β Tcr cDNA and RT-PCR sequences shared the highest degree of identity with each of the complete V_β Tcr germline sequences presented in Figs. 3A to 3I. A summary of this analysis is listed in Table 5, and the alignments for the most related sequences are shown in Figs. 3A–3I along with the location and identity of potential polymorphic sites. Table 5 also includes the official T-cell receptor variable gene segment designations according to Clark *et al.* (1993).

The sequence of this 77.7-kb region of the V_β Tcr cluster starts within exon 2 of a $V_\beta 6$ gene (Table 4). The available sequence of this exon (253 bp) shares a high degree of identity with the sequences determined from many $V_\beta 6$ cDNA clones; cDNA clones L17Tiβ (Leiden *et*

TABLE 5

Cross-Referencing of Sequenced Human T-Cell Receptor V_β cDNA and Germline Clones

V_β gene	Clone(s)	GenBank No(s).	%Identity[a]	DNA type	Ref(s).
6	H7.1	New	—	Germline	This paper
(TCRBV6S1)[b]	L17Tiβ	M15564	100/253	mRNA	Leiden *et al.* (1986)
	IGRb10	X58805	100/253	mRNA	Ferradini *et al.* (1991)
	L17β	M13552	100/253	mRNA	Leiden and Strominger (1986)
	HBVT11	M27386	100/253	mRNA	Kimura *et al.* (1987)
	HBVT116	M27385	99.2/253	mRNA	Kimura *et al.* (1987)
	ph 22	M14261	97.8/253	mRNA	Tillinghast *et al.* (1986)
23	H7.1	New	—	Germline	This paper
(TCRBV23S1)	IGRb04	X58799	99.7/328	mRNA	Ferradini *et al.* (1991)
	HT183	X57613	99.5/381	mRNA	Plaza *et al.* (1991)
	Vβ22	M62378	98.8/409	mRNA	Robinson (1991)
	Mm1-14	M60530	95.3/408	mRNA	Levinson *et al.* (1992)
	Mm1-41	M60545	95.3/408	mRNA	Levinson *et al.* (1992)
	Mm1-6	M60535	95.1/408	mRNA	Levinson *et al.* (1992)
12.2	H7.1	New	—	Germline	This paper
(TCRBV12S2)	ph27	M14268	98.7/315	mRNA	Tillinghast *et al.* (1986)
	HBP54	X04935	97.7/222	mRNA	Kimura *et al.* (1986)
	PL4.2	M13862	97.5/237	mRNA	Concannon *et al.* (1986)
	IGRb13	X58808	87.7/293	mRNA	Ferradini *et al.* (1991)
	KT2	M64352	86.5/400	mRNA	Boitel *et al.* (1992)
	Mm13-13	M60539	85.8/359	mRNA	Levinson *et al.* (1992)
	Mm13-2A2	M60531	84.8/407	mRNA	Levinson *et al.* (1992)
	Mm19-94	M60540	84.7/360	mRNA	Levinson *et al.* (1992)
21.2	H7.1	New	—	Germline	This paper
(TCRBV21S2)	H7.1	M33234	100/1667[c]	Germline	Wilson *et al.* (1990)
	IW6-4	X56665	100/320	mRNA	Hansen *et al.* (1991)
	IGRb02	X58797	99.2/387	mRNA	Ferridini *et al.* (1991)
	Vβ21	M62377	98.5/268	mRNA	Robinson (1991)
	IGRb01	X58796	89.3/365	mRNA	Ferridini *et al.* (1991)
	TCRBV21.3	M33235	88.2/455	Germline	Wilson *et al.* (1990)
	HT-11	X57724	88.1/311	mRNA	Plaza et *al.* (1991)
	H18.1(21.1)	M33233	86.0/1261	Germline	Wilson *et al.* (1990)
8.1	H7.1/H12.18	New	—	Germline	This paper
(TCRBV8S1)	H7.1(8.1)	X07192/Y00349	100/775	Germline	Siu *et al.* (1986)
	YT35	K01571/X00437	100/342	mRNA	Yanagi *et al.* (1984)
	4D8	K02885/X01417	100/222	mRNA	Sims *et al.* (1984)
	ph11	M14265	99.4/345	mRNA	Tillinghast *et al.* (1986)
	PL3.3	M13858/M16307	98.8/255	mRNA	Concannon *et al.* (1986)
	8B3	M81773	97.6/509	mRNA	Toyonaga *et al.* (1985)
	HT242	X57720	97.5/394	mRNA	Plaza *et al.* (1991)
	p8H7.1B5(8.2)	K02546	97.1/519	Germline	Siu *et al.* (1984)
	HT2.12	X57619	97.0/394	mRNA	Plaza *et al.* (1991)
	HBP41(8.3)	X04925	96.8/281	mRNA	Kimura *et al.* (1986)
	ph8	M14264	96.5/347	mRNA	Tillinghast *et al.* (1986)
	H7.1(8.2)	X007222	96.0/772	Germline	Siu *et al.* (1986)
	H7.1(8.1/8.2)	Duplication	93.8/3349	Germline	This paper
	Mm8-91	M60548	93.5/400	mRNA	Levinson *et al.* (1992)
	Mm8-1A8	M60547	93.1/319	mRNA	Levinson *et al.* (1992)
8.2	H7.1	New	—	Germline	This paper
(TCRBV8S2)	H7.1(8.2)	X07222	100/772[c]	Germline	Siu *et al.* (1986)
	HT242	X57720	100/342	mRNA	Plaza *et al.* (1991)
	PL3.3	M13858/M16307	100/255	mRNA	Concannon *et al.* (1986)
	p8H7.1B5(8.2)	K02546	99.8/519	Germline	Siu *et al.* (1984)
	8B3	M81773	99.7/492	mRNA	Toyonaga *et al.* (1985)
	HBP41(8.3)	X04925	99.6/281	mRNA	Kimura *et al.* (1986)
	HT2.12	X57619	99.5/394	mRNA	Plaza *et al.* (1991)
	ph8	M14264	99.1/342	mRNA	Tillinghast *et al.* (1986)
	4D8	K02885/X01417	98.6/222	mRNA	Sims *et al.* (1984)
	YT35	K01571/X00437	97.7/342	mRNA	Yanagi *et al.* (1984)
	ph11	M14265	97.1/345	mRNA	Tillinghast *et al.* (1986)
	H7.1(8.1)	X07192/Y00349	97.3/775	Germline	Siu *et al.* (1986)
	H7.1(8.2/8.1)	Duplication	93.8/3349	Germline	This paper

TABLE 5—*Continued*

V$_\beta$ gene	Clone(s)	GenBank No(s).	%Identity[a]	DNA type	Ref(s).
8.2	Mm8-91	M60548	93.0/400	mRNA	Levinson *et al.* (1992)
(continued)	Mm8-1A8	M60547	92.2/319	mRNA	Levinson *et al.* (1992)
8.3	H130.1	New	—	Germline	This paper
(TCRBV8S3)	λgt7-4.4	X07223	100/741	Germline	Siu *et al.* (1986)
	CH1-B	M73463	99.5/218	mRNA	Lunardi *et al.* (1992)
	4D8	K02885/X01417	83.8/222	mRNA	Sims *et al.* (1984)
	λgt7-8.5	X06936	81.6/737	Germline	Siu *et al.* (1986)
	ph11	M14265	80.5/344	mRNA	Tillinghast *et al.* (1986)
	YT35	K01571/X00437	80.4/373	mRNA	Yanagi *et al.* (1984)
	ph8	M14264	80.1/346	mRNA	Tillinghast *et al.* (1986)
	HT242	X57720	79.2/394	mRNA	Plaza *et al.* (1991)
	8B3	M81773	79.0/400	mRNA	Toyonaga *et al.* (1985)
	HT2.12	X57619	78.9/394	mRNA	Plaza *et al.* (1991)
	H7.1(8.1)	New	78.5/741	Germline	This paper
	H7.1(8.2)	New	78.3/741	Germline	This paper
	H18.1(ψ8.4)	X07224	76.8/741	Germline	Siu *et al.* (1986)
16	H130.1	New	—	Germline	This paper
(TCRBV16S1)		X06154/Y00349	100/720	Germline	Smith *et al.* (1987)
	HT370	X57723	100/350	mRNA	Plaza *et al.* (1991)
	HBP42	X04933	100/237	mRNA	Kimura *et al.* (1986)
	HT219	X57722	99.3/330	mRNA	Plaza *et al.* (1991)
24	H130.1	New	—	Germline	This paper
(TCRBV24S1)	CH18-β	M73464	99.7/375	mRNA	Lunarde *et al.* (1992)
	HT77	X57725	99.7/367	mRNA	Plaza *et al.* (1991)
	IGRb05	X58800	99.7/366	mRNA	Ferradini *et al.* (1991)
	HT1.8	X57726	99.5/367	mRNA	Plaza *et al.* (1991)
	Vβ24	M62376	99.5/375	mRNA	Robinson (1991)
25	H130.1	New	—	Germline	This paper
(TCRBV25S1)	HBVT72	M27390	69.7/297	mRNA	Kimura *et al.* (1987)
	PL3.9	M13860/M16309	68.6/312	mRNA	Concannon *et al.* (1986)
	244	M87323	68.5/340	mRNA	Chen *et al.* (1992)
	Vβ-ALT12-1	M11956	67.5/332	mRNA	Ikuta *et al.* (1985)
	2Q29	M22007	66.2/340	mRNA	Wade *et al.* (1988)
26	H130.1	New	—	Germline	This paper
(TCRBV26S1)	p8H7.1B5(8.2)	K02546	71.9/288	Germline	Siu *et al.* (1984)
	H7.1(8.2)	X07222	71.9/288	Germline	Siu *et al.* (1986)
	HT242	X57720	71.6/335	mRNA	Plaza *et al.* (1991)
	8B3	M81773	71.3/335	mRNA	Toyonaga *et al.* (1985)

[a] List includes cDNA, RT-PCR, and germline clone sequences that share highest degree of identity. The first number is the % identity and the second number is the length of the sequence comparison.

[b] Official T-cell receptor variable gene segment designations according to Clark *et al.* (1993).

[c] Assuming corrections in previous V$_\beta$21 germline sequence, M33234 (Wilson *et al.*, 1990), since sequences were derived from the same cosmid clone H7.1.

al., 1986), IGRb10 (Ferradini *et al.*, 1991), L17β (Leiden and Strominger, 1986), and HBVT11 (Kimura *et al.*, 1987) are identical to this V$_\beta$6 Tcr germline sequence. It appears that the several cloned cDNAs that contain identical V$_\beta$6 subfamily members have been given separate subfamily numbers; for example, subfamily members V$_\beta$6.5, 6.8, and 6.9 share 100% nucleotide sequence identity (Toyonaga and Mak, 1987). In fact, each of these V$_\beta$6 subfamily members also shares 100% identity with the V$_\beta$6 gene described here; thus, this V$_\beta$6 gene could be assigned to any of these subfamily members. The V$_\beta$ Tcr subfamily numbering system is being updated (Clark *et al.*, 1993) and can effectively and accurately be done once the nucleotide sequence of the complete V$_\beta$ locus is determined. The correct subfamily nomenclature for this V$_\beta$6 Tcr gene is TCRBV6S1. The high frequency with which V$_\beta$6 transcripts have been cloned suggests either that this gene is highly expressed or that it is represented in more than one identical, or nearly identical, gene copy.

The V$_\beta$23 Tcr gene subfamily was first classified by the sequence of cDNA clone IGRb04 (Ferradini *et al.*, 1991) and it has subsequently been identified in cDNA clone HT183 (Plaza *et al.*, 1991) and cDNA clone VB22 (Robinson, 1991). The first complete germline sequence of a V$_\beta$23 gene is shown in Fig. 3A along with these cDNA-derived sequences. The exon 1 coding region of the V$_\beta$23 gene is unusual, as it is 30 bp longer than that found for the other V$_\beta$ Tcr genes (Table 4). None of these V$_\beta$23 cDNA sequences shares complete identity with this V$_\beta$23 gene sequence (IGRb04 differs only at the first nucleotide position) (Table 5; Fig. 3A). This small degree of divergence (not more than 1.2%) could be due to polymorphisms or possible errors in the sequence (PCR or sequence generated). However, we cannot rule out the possibility that other members of the $V_\beta23$ gene subfamily exist. The next most related sequences are three V_β Tcr cDNA sequences derived from rhesus monkey (*Macaca mulatta*), clones Mm1-14, -41, and -6 (Levinson *et al.*, 1992), which share better than 95% identity (Table 5; Fig. 3A). This high degree of identity strongly suggests that these cDNA clones were derived from an orthologous member of the rhesus $V_\beta23$ gene subfamily.

The first germline sequence of a $V_\beta12$ Tcr gene subfamily member is shown in Fig. 3B. Genomic mapping (Lai *et al.*, 1988) suggests that there may be at least two $V_\beta12$ genomic locations and our GenBank search found five related cDNA derived sequences. Of these $V_\beta12$ cDNA sequences, three share a high degree of identity with this germline $V_\beta12$ sequence, greater than 97% (Table 5; Fig. 3B); these include clones ph27 (Tillinghast *et al.*, 1986), HBP54 (Kimura *et al.*, 1986), and PL4.2 (Concannon *et al.*, 1986). From the $V_\beta12$ subfamily numbering system suggested by Toyonaga and Mak (1987), this $V_\beta12$ germline gene is most related to subfamily member $V_\beta12.2$. The remaining two cDNA cloned sequences are identical to each other, clones IGRb13 (Ferrandini *et al.*, 1991) and KT2 (Boitel *et al.*, 1992); however, they share only about 87% identity with this $V_\beta12.2$ gene. The sequences of cDNA clones HBP54 and ph27 would share 100% identity with this $V_\beta12.2$ gene except for the difference of five nucleotides at the 3′ end of the HBP54 sequence and the 5′ 39 bp of ph27 (Fig. 3B). These high degrees of divergence located at the ends of these cDNA sequences may be indicative of cDNA cloning artifacts and N-region diversity due to the joining of the V and D gene segments, respectively. Three cDNA sequences derived from rhesus monkey are orthologous members of the $V_\beta12$ Tcr subfamily; they share their highest degree of identity (greater than 95%) with the $V_\beta12.3$ subfamily member (data not shown).

Three members of the human $V_\beta21$ Tcr gene subfamily have been subjected to previous extensive analyses; they were first detected by hybridization with a mouse $V_\beta13$ Tcr gene probe, followed by their isolation and nucleotide sequence analysis (Wilson *et al.*, 1990). The individual $V_\beta21$ Tcr genes were designated on the basis of their respective genomic *Eco*RI fragment sizes: $V_\beta21.1$ (6.6 kb), $V_\beta21.2$ (2.6 kb), and $V_\beta21.3$ (1.6 kb). $V_\beta21.1$ and $V_\beta21.2$ Tcr genes have been mapped to cosmid clones H18.1 and H7.1, respectively (Wilson *et al.*, 1990). Analysis of the complete sequence of cosmid H7.1 reveals the 1667-bp sequence determined by Wilson *et al.* (1990) and the presence of the $V_\beta21.2$ Tcr gene. The overlapping sequences did show some differences; however, these were resolved after careful rechecking and the correct sequence is included in the 77.7 kb that we submitted to GenBank. Even though these $V_\beta21$ Tcr gene subfamily members share greater than 86% nucleotide sequence identity, their genomic locations are spread across about 200 kb of the V_β Tcr locus (Wilson *et al.*, 1990), which suggests that they are most likely not the product of local gene duplication events, but could be the result of a large gene duplication event(s). The comparative nucleotide sequence analysis by Wilson *et al.* (1990), confirmed here, showed that all three $V_\beta21$ genes are complete and should be functional; however, they were only able to find transcripts for $V_\beta21.1$ and $V_\beta21.3$. Our FASTA search of GenBank revealed three recent cDNA sequence entries that share a high degree of identity, greater than 98% (Table 5 and Fig. 3C), with $V_\beta21.2$. In fact, the sequence of one cDNA clone, IW6-4 (Hansen *et al.*, 1991), shares 100% identity with the $V_\beta21.2$ gene sequence over a total of 320 bp. The small number of differences (potential polymorphic sites) among the other cDNA sequences is indicated in Fig. 3C.

The $V_\beta8$ Tcr gene subfamily has been extensively studied by Siu *et al.* (1986), who sequenced 775 bp from the $V_\beta8.1$ and 772 bp from the $V_\beta8.2$ Tcr gene regions, both of which were isolated from cosmid clone H7.1. Both $V_\beta8$ genes are expressed as cDNA sequences that share 100% identity with each gene that has been determined (Fig. 3D). As pointed out by Siu *et al.* (1986), the $V_\beta8.1$ and 8.2 Tcr genes are closely linked, separated by about 3 kb, and they share nearly identical sequences, which suggests that they arose from a recent gene duplication event. These suggestions were confirmed in our analysis because we identified the boundaries of each duplication unit. The $V_\beta8.1$ duplication unit starts at position 29,229 and extends to position 32,539, after which the $V_\beta8.2$ duplication unit begins at position 32,540 and ends at position 35,851. The lengths of these duplication units are nearly identical, and over their complete length they share 93.8% identity. The age of this duplication event can be estimated from the divergence level of the noncoding DNAs, which is 6.67% or, corrected for multiple substitutions (Hayashida and Miyata *et al.*, 1983), is calculated to be 6.99%. If we assume that these noncoding DNAs are evolving as neutral DNA at an average branch rate of about 0.13% change/million years since last sharing a common catarrhine ancestor (Bailey *et al.*, 1991; M. Goodman, pers. comm., July 1993, Wayne State University, Detroit, MI), we can calculate that this $V_\beta8$ duplication event occurred about 27 million years ago (MYA) (6.99/0.13 = 53.7 MYA for both branches, or 27 MYA for each branch). This estimate is supported by analysis of the orthologous $V_\beta8.1$ and 8.2 gene regions from several primate species, which finds this duplication present in species that last shared a common ancestor with humans in the stem of the catarrhine branch (W. Funkhouser, B.F.K., manuscript in preparation). The mechanism responsible for this duplication event appears not to involve flanking repetitive elements (SINEs or LINEs), which is unlike the primate fetal globin gene duplication event as it involved a crossover between flanking L1 elements (Fitch *et al.*, 1991). However, a member of the *MER26* repetitive element family is located near the 5′ boundary of each duplication unit (see below), and its role, if any, in this duplication event is currently unknown. It is conceivable that this $V_\beta8$ duplication event is the result of an unequal crossover between unrelated sequences.

The $V_\beta 8.3$ gene is located about 16.8 kb 3′ of $V_\beta 8.2$ and it was previously cloned, λgt7-4.4, and sequenced by Siu *et al.* (1986). The 741-bp sequence determined by Siu *et al.* (1986) is identical to that which we determined from cosmid 130.1 (Table 5). Comparing the sequences of the $V_\beta 8$ Tcr genes, Siu *et al.* (1986) speculated that the $V_\beta 8.3$ gene may be a pseudogene due to a deletion in its potential CCAAT element. However, our GenBank search found a recently published $V_\beta 8$-related Tcr cDNA clone sequence, CH1β (Lunardi *et al.*, 1992), that shares better than 99% identity over 218 bp with $V_\beta 8.3$, differing by only one potential polymorphic site (Table 5; Fig. 3E). The sequence of this transcript shares only 78% identity with other members of the $V_\beta 8$ Tcr gene subfamily (Table 5); thus, the $V_\beta 8.3$ gene appears to at least produce a mRNA transcript.

The complete genomic sequence of the $V_\beta 16$ Tcr gene, 720 bp, was determined by Smith *et al.* (1987) from clone $\lambda V_\beta 16$, and this sequence is identical with the one we obtained from the $V_\beta 16$ Tcr gene in cosmid 130.1 (Table 5). The FASTA GenBank search found several sequence matches; two of these cDNA clones, HT370 (Plaza *et al.*, 1991) and HBP42 (Kimura *et al.*, 1986), share 100% identity with the $V_\beta 16$ Tcr gene sequence. The remaining cDNA clone, HT219 (Plaza *et al.*, 1991), shows a somewhat usual pattern of shared identity in that its 5′ end 37 bp are derived from the intron region while the remaining sequence differs only at one position (Table 5; Fig. 3F). It appears that the mRNA transcript contained in clone HT219 was partially spliced when its cDNA copy was synthesized.

The existence of the $V_\beta 24$ Tcr gene locus was first suggested by Ferradini *et al.* (1991) from the sequence of cDNA clone IGRb05, and the finding of this gene in cosmid 130.1 is the first description of a $V_\beta 24$ germline sequence. Additional $V_\beta 24$ cDNA clone sequences have been described within the past few years, and interestingly, all of these share better than 99% identity with this $V_\beta 24$ gene sequence (Table 5; Fig. 3G). These other cDNA clones include $V_\beta 24$ (Robinson, 1991), HT77 and HT1.8 (Plaza *et al.*, 1991), and CH18β (Lunardi *et al.*, 1992). Because the number of differences among these sequences is small, they may represent either transcripts from a closely related $V_\beta 24$ gene(s) or potential polymorphic sites. The fact that $V_\beta 24$ transcripts have been frequently cloned indicates that it is either a highly expressed V_β Tcr gene or that it may be represented in nearly identical multiple copies.

The dot plot shown in Fig. 2 suggests the presence of two additional complete V_β Tcr-like genes, one located between positions 68,508 and 68,956 and another located between positions 72,124 and 72,843. Detailed analysis of this regions reveals a typical V_β Tcr gene structure for the first gene (a 49-bp exon 1, a 107-bp intron, and a 293-bp exon 2; Fig. 3H) and a similar structure for the second gene (a 49-bp exon 1, a somewhat larger intron of 391 bp, and a 290-bp exon 3; Fig. 3I). The FASTA GenBank search did not find any sequences that share better than 72% identity with these V_β Tcr-like genes (Table 5); thus, they are referred to as $V_\beta 25$ and $V_\beta 26$ Tcr genes, respectively. The $V_\beta 25$ Tcr-like gene could encode a V_β Tcr-like polypeptide (Fig. 3H), except that its reading frame contains a termination codon, and thus this gene is most likely not functional, a pseudogene. It remains to be determined whether the pseudogene status of the $V_\beta 25$ gene is predominant in the human population and other primate species. The larger intron of $V_\beta 26$ Tcr gene is due to the insertion of a 290-bp *Alu* element 43 bp downstream from the intron donor splice junction (position 72,215) (Fig. 3I). The presence of this *Alu* element and the lack of any evidence from cDNA and RT-PCR cloning experiments that identify a transcript from this gene lead us to speculate that this gene may also be a pseudogene. This *Alu* element could alter the ability of this gene to be expressed, and if expressed, the presence of this *Alu* element could produce a transcript incapable of translation or if translated, the resulting protein product may not be related to any V_β Tcr polypeptide. In addition, this *Alu* element could contain sequences that share identity with intron splice acceptor sequences, which could greatly alter the coding region of any transcript product. A scan of this *Alu* element sequence for the presence of other potential splice acceptor sequences that closely matches the consensus sequences (YYYYYYYACAG, determined from the functional V_β Tcr genes; data not shown) finds a close match at positions 72,459 to 72,469 (TTTCACTCCAG). This *Alu* element is a member of the *Alu*-J subfamily (Jurka and Milosavljevic, 1991), the more ancient *Alu* subfamily estimated to have evolved about 55 MYA (Labuda and Striker, 1989). Thus, retroposition of this *Alu* element into the intron of the a proto-$V_\beta 26$ Tcr gene could have occurred early in primate evolution. The timing of this *Alu* retroposition event remains to be determined, as well as whether this $V_\beta 26$ Tcr gene can be transcribed.

Analysis of Repetitive DNA Elements

SINEs. The most prominent class of repetitive DNA elements in primate is the *Alu* element, which is derived from the 7SL RNA gene (Daniels *et al.*, 1983; Li *et al.*, 1982; Ullu *et al.*, 1982), and as many as one million copies may be present in the human genome (Deininger and Schmid, 1979). On average, the distribution of these elements should be about one every 4 kb (Hwu *et al.*, 1986), a density of 0.25/kb. However, results from recent large-scale sequencing projects that sample differ human chromosomal regions reveal a considerable amount of variation in *Alu* distribution. A 67-kb region of the X chromosome that encodes the Kallmann gene has a low *Alu* density, about 0.1/kb (Legouis *et al.*, 1991), while a 90-kb region of the HLA class III gene locus on chromosome 6 has clusters of *Alu* elements with densities in the range of 2/kb, which complicated the DNA sequencing process (Iris *et al.*, 1993). A search of this 77.7-kb region of the V_β Tcr cluster for *Alu* elements reveals a total of 21 *Alu*-related sequences, which include complete and half-*Alu* elements (Fig. 2; Table 6). The orientation of these *Alu* elements with respect to the V_β Tcr gene appears to be biased, with 14 having the same and 7 having

TABLE 6

Location of Repetitive Element Family Members

Type	No.	Position	Length	Orientation	Identity (%)	Class
Alu	1	3484/3772	290	←	86.2	Sx
	2	7501/7650	148	←	78.0	J
	3	8654/8942	290	←	89.9	Sb
	4	11357/11686	290	→	86.9	Sx
	5	16004/16293 (DR)	290	→	85.9	Sq
	6	16582/16867	288	→	88.2	Sc
	7	17843/18135	290	→	84.0	Sx
	8	39556/39850 (DR)	290	←	86.4	Sp
	9	44689/44977	290	→	88.6	Sb
	10	45442/45704	260	←	81.7	Sx
	11	47446/47743	289	→	78.9	J
	12	48507/48797	290	→	85.2	Sq
	13	59599/59887 (DR)	289	←	89.6	Sx
	14	59895/60184	289	←	78.0	J
	15	69176/69464	280	→	89.2	Sb
	16	69996/70285 (DR)	289	→	84.6	Sp
	17	71380/71540	160	→	88.1	Sx
	18	72215/72505 (DR)	290	→	83.5	J
	19	76008/76298	290	→	84.5	Sp
	20	77278/77447	170	→	70.6	J
	21	77507/77743 (+)	234 (+)	→	89.3	Sb
LINE	1	20097/20939	843	←	65.0	
	2	41297/43315	2019	→	75.7	
	3	43316/44688	1373	→	81.7	
MER26	1	29358/29506	130	←	70.0	
	2	32687/32813	130	←	67.8	
MER32		61004/61160	152	→	75.6	
LTR1		37904/38479	544	→	63.4	
LTR5		46193/46544	352	→	90.9	
OFR		18318/18677	329	←	74.5	

the opposite orientation. Five of these *Alu* elements are flanked by direct repeats (Table 6). The overall *Alu* density in this sequence is about 0.26/kb, near the average density predicted for the human genome. However, these repeats are not randomly distributed; they do appear in clusters of different densities, 0.56/kb between positions 7502 and 18,136, 1/kb between 44,689 and 48,797, and 1.2/kb between 69,176 and 72,505, up to a high of 1.8/kb between positions 76,007 and 77,742. *Alu* elements have been further classified according to their evolutionary origin into two distinct subfamilies referred to as *Alu*-J (old), which is more similar to 7SL DNA, and *Alu*-S (new) (Jurka and Smith, 1988). The *Alu*-S family can be further divided into five distinct subfamilies referred to as *Alu*-Sx, -Sq, -Sp, -Sc, and -Sb (Jurka and Milosavljevic, 1991). Analysis of these 21 *Alu* elements shows that only 5 are of the *Alu*-J family, with the others distributed among the five *Alu*-S subfamilies (Table 6). These *Alu* elements will be important comparative markers for mapping the evolutionary events that occurred in the V_β Tcr gene cluster during the decent of primates.

LINEs. The family of long-interspersed repetitive DNA sequences (LINEs) is possibly present at greater than 10,000 copies per mammalian genome (Burton *et al.*, 1986; Fanning and Singer, 1987). The sizes of the longest LINEs found in primate species are in the range of 6 to 7 kb; however, in many cases, only a short part of a LINE may be found as a result of deletions or insertions that have occurred during or after the original insertion event (Tagle *et al.*, 1992). Only three regions that contain partial LINEs were found in this part of the V_β Tcr gene cluster (Fig. 2; Table 6). The region between positions 19,393 and 21,124 shows a complex pattern of multiple LINE insertions (in both directions), with the best match (65%) being between positions 20,097 and 20,939 (Table 6). The arrangement of the two adjacent LINEs (positions 41,297 and 43,316) suggests that these elements may be the result of a single LINE insertion event followed by several deletion events that removed about 990 bp of 5′, 1.4 kb of center, and 630 bp of 3′ from the originally inserted LINE. The 3′ end of this LINE may have been deleted by the insertion of *Alu* element 9 (Table 6) which overlaps 3′ end of this LINE.

MERs and other repetitive elements. In addition to from SINEs and LINEs, the human genome contains an assortment of other repetitive element families, which may be represented at lower copy numbers, in the range of 200 to 10,000 per haploid genome. Jurka *et al.* (1992) reported a collection of 53 prototypic sequences representing known families of repetitive element found (at that time) in the human genome. These include elements that are referred to as medium reiteration frequency repeats (MER) (Jurka, 1989), which currently include 33 families (Jurka *et al.*, 1993), the THR and OFR components of transposable element THE-1 (Paulson *et al.*, 1985; Misra *et al.*, 1987), and the endogenous human retrovirus long-terminal repeat sequence elements referred to as LTR1 through LTR10 (Jurka *et al.*, 1992). This 77.7-kb sequence region was searched using the up-to-date list of the human prototypic repetitive DNA elements, which revealed the presence of repetitive elements related to LTR1, LTR5, OFR, MER26 (two locations), and MER32 (J. Jurka, pers. comm., 25 March 1993). The locations of these elements are listed in Table 6 along with the degree of identity that they share with their corresponding prototypic element. The only significant features associated with any of these elements is the high degree of identity shared between the LTR5 element and its prototype sequences (90.5%) and the locations of the two MER26 elements; each is located very near the 5′-ends of a respective duplication unit of the V_β8.1 and 8.2 Tcr genes (see above). At present, it is difficult to determine whether these MER26 elements in each V_β8 duplication unit played any role in this duplication event or whether they are the result of a single MER26 element being carried along as part of the duplication event. Each of these MER26 elements is located about 145 bp 3′ of the purposed duplication unit junctions, and if they were involved in the event, they would be expected to be located at the junction boundaries. In addition, we would also expect to find a third MER26-related element located at the 3′-end of the second duplication (none was found). However, as suggested above, this duplication event occurred approximately 27 MYA, and since then other events that altered these boundary sequences may have occurred. Further examination of the ends of orthologous duplication regions from other primate species may shed some light on whether these MER26 elements were involved in facilitating this duplication event.

CONCLUSIONS

Having the complete nucleotide sequence of this region of the TCRBV cluster clearly shows the value of this type of information, in that we have been able to determine the germline sequences and locations of additional TCRBV genes (V_β6, 23, 12.2, and 24) and we have located two additional V_β Tcr-related genic regions, that may or may not be functional (V_β25 and 26). From our analysis, we have also located potential polymorphic sites between the individual cosmid clones described here and among the available cDNA-derived sequences found in GenBank. These polymorphic sites could prove useful in the identification of specific V_β Tcr genes that are associated with autoimmune diseases. This nucleotide sequence information will also be valuable in the design of experiments to test the function of putative promoter elements, CCAAT, TATAA, and the Tcr decamer element associated with each TCRBV gene. Even though this region of the V_β Tcr cluster contains an average density of repetitive elements, in particular, *Alu* elements, it will be interesting to determine whether any diseases related to defective V_β Tcr genes involve interferences arising from repetitive elements altering the Tcr gene rearrangement mechanism(s). It is most exciting to have located two additional V_β Tcr genes, V_β25 and 26, and if they are found to be nonfunctional in the human population, it will be most interesting to determine whether they are functional in other primate or mammalian species. If they are functional, it will be extremely interesting to learn their function and the consequence that their loss has played in the evolution of the human species.

The experience that we gained from doing this project strongly indicates that a primer-directed walking approach, which has the goal of obtaining sequence information from large DNA templates (cosmid size) and long sequencing readings, is a feasible strategy for large-scale sequencing projects. The major advantages are the need for few subclones, the reduced number of sequencing reactions, and the ease (straightforwardness) of assembling the final sequence. Major disadvantages that we observed include difficulties in sequencing across large duplication regions (larger than 1 kb per duplication unit) that share >95% identity, low throughput, and lack of automated steps. The successful sequencing of large duplication regions will require some subcloning and mapping efforts. Throughput could be greatly enhanced by increasing the number of IPs, and these IPs should be obtained by methods that are independent of sequence composition or available six-base restriction enzyme recognition sites. More IPs could be obtained by increasing the coverage of overlapping clones (cosmids or λ), since each overlap vector–insert junction potentially represents an addition IP.

A primer-directed strategy would be of even greater benefit if it could be adapted to automated instruments, which could be done with improved detection sensitivity and gel technology. Additional ways to improve throughput include having a dedicated online oligonucleotide synthesis facility or having access to prepared libraries of short oligomer primers (Studier, 1989; Siemieniak and Slightom, 1990; Kieleczawa *et al.*, 1993). Indeed, it is conceivable that the throughput rate could match that achieved by the M13 random sequence strategy provided that a method could be developed for the rapid identification of all members of an oligomer library that have the potential to prime a sequencing reaction within a particular cloned insert. Such development would allow a primer-direct sequencing approach to be used in a random mode similar to the M13 strategy, but without the need to generate numerous subclones.

ACKNOWLEDGMENTS

The authors thank E. Lai for cosmid clones H7.1, H12.18, and H130.1 and for sharing his knowledge of these clones during the sequencing process, and N. Hatzenbuhler and D. Riley for synthesizing the many oligomer primers used. We thank J. Jurka (Linus Pauling Institute of Science and Medicine) for help in classifying the type of *Alu* elements and for help in searching for other repetitive DNA elements using his updated file of prototypic sequences for the human genome. We also thank J. Bock (The Upjohn Company) for help in rechecking some of the DNA regions that were difficult to sequence, and M. Goodman (Wayne State University) for helpful discussions concerning primate evolution. We also extend a special thanks to the reviewers for their analysis of our sequence data and helpful comments. This research was partially supported by grants from DOE (DE FG03-91 ER61182) and NIH- (RO1HG00356) to L.H.

REFERENCES

Anderson, S. J., Chou, H. S., and Lou, D. Y. (1988). A conserved sequence in the T-cell receptor β-chain promoter region. *Proc. Natl. Acad. Sci. USA* **85:** 3551–3554.

Bailey, W. J., Fitch, D. H. A., Tagle, D. A., Czelusniak, J., Slightom, J. L., and Goodman, M. (1991). Molecular evolution of the $\psi\eta$-globin gene locus: Gibbon phylogeny and hominoid slowdown. *Mol. Biol. Evol.* **8:** 155–184.

Barker, P. E., Ruddle, F. H., Royer, H.-D., Acuto, O., and Reinherz, E. L. (1984). Chromosomal location of human T-cell receptor gene $T_i\beta$. *Science* **226:** 348–349.

Beck, S., Kelly, A., Radley, E., Khurshid, F., Alderton, R. P., and Trowsdate, J. (1992). DNA sequence analysis of 66 kb of the human MHC class II region encoding a cluster of genes for antigen processing. *J. Mol. Biol.* **228:** 433–441.

Behlke, M. A., and Loh, D. Y. (1986). Alternative splicing of murine T-cell receptor β-chain transcripts. *Nature* **322:** 379–382.

Bergman, Y., Rice, D., Grosschedl, R., and Baltimore, D. (1984). Two regulatory elements for immunoglobulin k light chain gene expression. *Proc. Natl. Acad. Sci. USA* **81:** 7041–7045.

Birnboim, H. D., and Doly, J. (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Res.* **7:** 1513–1523.

Boitel, B., Ermonval, M., Panina-Bordignon, P., Mariuzza, R. A., Lanzavecchia, A., and Acuto, O. (1992). Preferential $V\beta$ gene usage and lack of junctional sequence conservation among human T cell receptors specific for a tetanus toxin-derived peptide: Evidence for a dominant role of a germline-encoded V region in antigen/major histocompatibility complex recognition. *J. Exp. Med.* **175:** 765–777.

Burton, F. H., Loeb, D. D., Voliva, C. F., Martin, S. L., Edgell, M. H., and Hutchison, C. A., III. (1986). Conservation throughout mammalia and extensive protein-encoding capacity of the highly repeated DNA long interspersed sequence one. *J. Mol. Biol.* **187:** 291–304.

Caccia, N., Kronenberg, M., Saxe, D., Haars, R., Bruns, G. A. P., Governman, J., Malissen, M., Willard, H., Yoshikai, Y., Simon, M., Hood, L., and Mak, T. W. (1984). The T cell receptor β chain genes are located on chromosome 6 in mice and chromosome 7 in humans. *Cell* **37:** 1091–1099.

Chen, Z. W., Yamamoto, H., Watkins, D. I., Levinson, G., and Letvin, N. L. (1992). Predominant use of a T-cell receptor $V\beta$ gene family in simian immunodeficiency virus gag-specific cytotoxic T lymphocytes in a rhesus monkey. *J. Virol.* **66:** 3913–3917.

Clark, S. P., Arden, B., and Mak, T. W. (1993). Human T-cell receptor variable gene segment families. *Immunogenetics,* in press.

Concannon, P., Pickering, L. A., Kung, P., and Hood, L. (1986). Diversity and structure of human T-cell receptor β-chain variable region genes. *Proc. Natl. Acad. Sci. USA* **83:** 6598–6602.

Daniels, G. R., Fox, G. M., Lowensteiner, D., Schmid, C. W., and Deininger, P. L. (1983). Species-specific homogeneity of the primate Alu family of repeated DNA sequences. *Nucleic Acids Res.* **11:** 7579–7593.

Davis, M. M. (1990). T cell receptor gene diversity and selection. *Annu. Rev. Biochem.* **59:** 475–496.

Deininger, P. L., and Schmid, C. W. (1979). A study of the evolution of repeated DNA sequence in primates and the existence of a new class of repetitive sequences in primates. *J. Mol. Biol.* **127:** 437–460.

Devereux, J., Haeberli, P., and Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* **12:** 387–395.

Edward, A., Voss, H., Rice, P., Civitello, A., Stegemann, J., Schwager, C., Zimmermann, J., Erfle, H., Caskey, C. T., and Ansorge, W. (1990). Automated DNA sequencing of the human HPRT locus. *Genomics* **6:** 593–608.

Falkner, F. G., and Zachau, H. G. (1984). Correct transcription of an immunoglobulin k gene requires an upstream fragment containing conserved sequence elements. *Nature (London)* **310:** 71–74.

Fanning, T. G., and Singer, M. F. (1987). LINE-1: A mammalian transposable element. *Biochim. Biophys. Acta* **910:** 203–212.

Ferradini, L., Roman-Roman, S., Azocar, J., Michalaki, H., Triebel, F., and Hercend, T. (1991). Studies on the human T cell receptor α/β variable region genes. II. Identification of four additional V_β subfamilies. *Eur. J. Immunol.* **21:** 935–942.

Fitch, D. H. A., Bailey, W. J., Tagle, D. A., Goodman, M., Sieu, L., and Slightom, J. L. (1991). Duplication of the γ-globin gene mediated by L1 long interspersed repetitive elements in an early ancestor of simian primates. *Proc. Natl. Acad. Sci. USA* **88:** 7369–7400.

Gomolka, M., Epplen, C., Buitkamp, J., and Epplen, J. T. (1993). Novel members and germline polymorphisms in the human T-cell receptor Vb6 family. *Immunogenetics* **37:** 257–265.

Gottschalk, L. R., and Leiden, J. M. (1990). Identification and functional characterization of the human T-cell receptor β gene transcriptional enhancer: Common nuclear proteins interact with the transcriptional regulatory elements of the T-cell receptor α and β genes. *Mol. Cell. Biol.* **10:** 5486–5495.

Hansen, T., Qvigstad, E., Lundin, K. E. A., and Thorsby, E. (1991). Sequences of four previously undescribed human T-cell receptor β chain variable genes. *Tissue Antigens* **38:** 99–103.

Hayashida, H., and Miyata, T. (1983). Unusual evolutionary conservation and frequent DNA segment exchange in class I genes of the major histocompatibility complex. *Proc. Natl. Acad. Sci. USA* **80:** 2671–2675.

Hunkapiller, T., and Hood, L. (1989). Diversity of the immunoglobulin gene superfamily. *Adv. Immunol.* **44:** 1–63.

Hwu, H. R., Roberts, J. W., Davidson, E. H., and Britten, R. J. (1986). Insertion and/or deletion of many repeated DNA sequences in human and higher ape evolution. *Proc. Natl. Acad. Sci. USA* **83:** 3875–3879.

Ikuta, K., Ogura, T., Shimizu, A., and Honjo, T. (1985). Low frequency of somatic mutation in β-chain variable region genes of human T-cell receptors. *Proc. Natl. Acad. Sci. USA* **82:** 7701–7705.

Iris, F. J. M., Bougueleret, L., Prieur, S., Caterina, D., Gwenael, P., Perrot, V., Jurka, J., Rodriguez-Tome, P., Claverie, J. M., Dausset, J., and Cohen, D. (1993). Dense Alu clustering and a potential new member of the NF kB family within a 90 kilobase HLA class III segment. *Nature Genet.* **3:** 137–145.

Jurka, J. (1989). Novel families of interspersed repetitive elements from the human genome. *Nucleic Acids Res.* **18:** 137–141.

Jurka, J., and Milosavljevic, A. (1991). Reconstruction and analysis of human *Alu* Genes. *J. Mol. Evol.* **32:** 105–121.

Jurka, J., and Smith, T. (1988). A fundamental division in the *Alu* family of repeated sequences. *Proc. Natl. Acad. Sci. USA* **85:** 4775–4778.

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

Jurka, J., Kaplan, D. J., Duncan, C. H., Walichiewicz, J., Milosavljevic, A., Gayathri, M., and Solus, J. F. (1993). Identification and characterization of new human medium reiteration frequency repeats. *Nucleic Acids Res.* **21:** 1273–1279.

Kieleczawa, J., Dunn, J. J., and Studier, F. W. (1992). DNA sequencing by primer walking with strings of contiguous hexamers. *Sciences* **258:** 1787–1791.

Kimura, N., Toyonaga, B., Yoshikai, Y., Du, R.-P., and Mak, T. W. (1987). Sequences and repertoire of the human T cell receptor α and β chain variable region genes in thymocytes. *Eur. J. Immunol.* **17:** 375–383.

Kimura, N., Toyonaga, B., Yoshikai, Y., Triebel, F., Debre, P., Minden, M. D., and Mak, T. W. (1986). Sequences and diversity of human T cell receptor β chain variable region genes. *J. Exp. Med.* **164:** 739–750.

Koop, B. F., Rowen, L., Wang, K., Kuo, C. L., Seto, D., Lenstra, J. A., Howard, S., Shan, W., Deshpande, P., and Hood, L. (1994). The human T-cell receptor Cα/Cδ region: Organization, sequence, and evolution of 97.6 kb of DNA. *Genomics* **19:** 478–493.

Krimpenfort, P., de Jong, R., Uematsu, Y., Dembic, Z., Ryser, S., von Boehmer, H., Steinmetz, M., and Berns, A. (1988). Transcription of T cell receptor β-chain genes is controlled by a downstream regulatory element. *EMBO J.* **7:** 745–750.

Labuda, D., and Striker, G. (1989). Sequence conservation in *Alu* evolution. *Nucleic Acids Res.* **17:** 2477–2491.

Lai, E., Concannon, P., and Hood, L. (1988). Conserved organization of the human and murine T-cell receptor β-gene families. *Nature* **331:** 543–546.

Lai, E., Wilson, R. K., and Hood, L. E. (1989). Physical maps of the mouse and human immunoglobulin-like loci. *Adv. Immunol.* **46:** 1–59.

Legouis, R., Hardelin, J.-P., Levilliers, J., Claverie, J.-M., Compain, S., Wunderie, V., Millasseau, P., Paslier, D. L., Cohen, D., Caterina, D., Bougueleret, L., Delemarre-Van de Waal, H., Lutfalla, G., Weissenbach, J., and Petit, C. (1991). The candidate gene for the X-linked Kallmann syndrome encodes a protein related to adhesion molecules. *Cell* **67:** 423–435.

Leiden, J. M., Fraser, J. D., and Strominger, J. L. (1986). The complete primary structure of the T-cell receptor genes from an alloreactive cytotoxic human T-lymphocyte clone. *Immunogenetics* **24:** 17–23.

Leiden, J. M., and Strominger, J. L. (1986). Generation of diversity of the β chain of the human T-lymphocyte receptor for antigen. *Proc. Natl. Acad. Sci. USA* **83:** 4456–4460.

Levinson, G., Huges, A. L., and Letvin, N. L. (1992). Sequence and diversity of rhesus monkey T-cell receptor β chain genes. *Immunogenetics* **35:** 75–88.

Li, W. Y., Reddy, R., Henning, D., Epstein, P., and Busch, H. (1982). Nucleotide sequence of 7S RNA: Homology to Alu DNA and LA 4.5S RNA. *J. Biol. Chem.* **257:** 5136–5142.

Li, Y., Szabo, P., Robinson, M. A., Dong, B., and Posnett, D. N. (1990). Allelic variations in the human T cell receptor Vβ6.7 gene products. *J. Exp. Med.* **171:** 221–230.

Li, Y., Szabo, P., and Posnett, D. N. (1991). The genomic structure of human Vβ6 T-cell antigen receptor genes. *J. Exp. Med.* **174:** 1537–1547.

Lunardi, C., Marguerie, C., and So, A. K. (1992). Identification of novel human T-cell receptor Vβ gene segments by the anchored-polymerase chain reaction. *Immunogenetics* **36:** 314–318.

Martin-Gallardo, A., McCombie, W. R., Gocayne, J. D., FitzGeralds, M. G., Wallace, S., Lee, B. M. B., Lamerdin, J., Trapp, S., Kelly, J. M., Liu, L.-I., Dubnick, M., Johnson-Dow, L. A., Kerlavage, A. R., de Jong, P., Carrano, A., Fields, C., and Venter, J. C. (1992). Automated DNA sequencing and analysis of 106 kilobases from human chromosome 19q13.3 *Nature Genet.* **1:** 34–39.

Mason, J. O., Williams, G. T., and Neuberger, M. S. (1985). Transcription cell type specificity is conferred by an immunoglobulin V_H gene promoter that includes a functional consensus sequence. *Cell* **41:** 479–487.

Maxam, A., and Gilbert, W. (1980). Sequencing end-labelled DNA with base-specific chemical cleavage. *Methods Enzymol.* **68:** 499–560.

McCombie, W. R., Martin-Gallardo, A., Gocayne, J. D., FitzGerald, M., Dubnick, M., Kelly, J. M., Castilla, L., Liu, L. I., Wallace, S., Trapp, S., Tagle, D., Whaley, W. L., Cheng, S., Gusella, J., Frischauf, A.-M., Poustka, A., Lehrach, H., Collins, F. S., Kerlavage, A. R., Fields, C., and Venter, J. C. (1992). Expressed genes, *Alu* repeats and polymorphisms in cosmids sequenced from chromosome 4p16.3 *Nature Genet.* **1:** 348–353.

McDougall, S., Peterson, C. L., and Calame, K. (1988). A transcriptional enhancer 3′ of $C_{\beta2}$ in the T cell receptor β locus. *Science* **241:** 205–208.

Morelle, G. (1989). A plasmid extraction procedure on a miniprep scale. *BRL Focus* **11:** 7–8.

Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48:** 443–453.

Parslow, T. G., Blair, D. L., Murphy, W. J., and Granner, D. K. (1984). Structure of the 5′ ends of immunoglobulin genes: A novel conserved sequence. *Proc. Natl. Acad. Sci. USA* **81:** 2650–2654.

Pearson, W. R., and Lipman, D. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85:** 2444–2448.

Plaza, A., Kono, D. H., and Theofilopoulos, A. N. (1991). New human Vβ genes and polymorphic variants. *J. Immunol.* **147:** 4360–4365.

Robinson, M. A. (1991). The human T cell receptor β-chain gene complex contains at least 57 variable gene segments. *J. Immunol.* **146:** 4392–4397.

Robinson, M. A., Mitchell, M. P., Wei, S., Day, C. E., Zhao, T. M., and Concannon, P. (1993). Organization of human T-cell receptor β-chain genes: Clusters of V_β genes are present on chromosomes 7 and 9. *Proc. Natl. Acad. Sci. USA* **90:** 2433–2437.

Royer, H. D., and Reinherz, E. J. (1987). Multiple nuclear proteins bind upstream sequences in the promoter region of a T-cell receptor β-chain variable-region gene: Evidence for tissue specificity. *Proc. Natl. Acad. Sci. USA* **84:** 232–236.

Sakano, H., Huppi, K., Heinrich, G., and Tonegawa, S. (1979). Sequences at the somatic recombination sites of immunoglobulin light-chain genes. *Nature* **280:** 288–294.

Sanger, F. S., Nicklen, S., and Coulson, A. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74:** 5463–5467.

Siemieniak, D. R., Sieu, L. C., and Slightom, J. L. (1991). Strategy and methods for directly sequencing cosmid clones. *Anal. Biochem.* **192:** 441–448.

Siemieniak, D. R., and Slightom, J. L. (1990). A library of 3342 useful nonamer primers for genome sequencing. *Gene* **96:** 121–124.

Sims, J. E., Tunnacliffe, A., Smith, W. J., and Rabbitts, T. H. (1984). Complexity of human T-cell antigen receptor β-chain constant- and variable-region genes. *Nature* **312:** 541–545.

Siu, G., Clark, S. P., Yoshikai, Y., Malissen, M., Yanagi, Y., Strauss, E., Mak, T. W., and Hood, L. (1984). The human T cell antigen receptor is encoded by variable, diversity, and joining gene segments that rearrange to generate a complete V gene. *Cell* **37:** 393–401.

Siu, G., Strauss, E. C., Lai, E., and Hood, L. E. (1986). Analysis of a human Vβ gene subfamily. *J. Exp. Med.* **164:** 1600–1614.

Slightom, J. L., Siemieniak, D. R., and Sieu, L. C. (1991). DNA sequencing: Strategy and methods to directly sequence large DNA molecules. *In* "Phylogenetic Analysis of DNA Sequences" (M. M. Miyamoto and J. Cracraft, Eds.), pp. 18–44, Oxford Univ. Press, New York.

Smith, W. J., Tunnacliffe, A., and Rabbitts, T. H. (1987). Germline sequence of two human T-cell receptor Vβ genes: Vβ8.1 is transcribed from a TATA-box promoter. *Nucleic Acids Res.* **15:** 4991.

Smith, T., and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195–197.

Studier, F. W. (1989). A strategy for high-volume sequencing of cosmid DNAs: Random and directed priming with a library of oligonucleotides. *Proc. Natl. Acad. Sci. USA* **86:** 6917–6921.

Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qui, L., Dear, S., Coulson, A., Craxton, M., Durbin, R., Berks, M., Metzstein, M., Hawkins, T., Ainscough, R., and Waterston, R. (1992). The *C. elegans* genome sequencing project: A beginning. *Nature* **356:** 37–41.

Tagle, D. A., Stanhope, M. J., Siemieniak, D. R., Benson, P., Goodman, M., and Slightom, J. L. (1992). The β globin gene cluster of the prosimian primate *Galago crassicaudatus:* Nucleotide sequence de-termination of the 41-kb cluster and comparative sequence analyses. *Genomics* **13:** 741–760.

Tillinghast, J. P., Behlke, M. A., and Loh, D. Y. (1986). Structure and diversity of the human T-cell receptor β-chain variable region genes. *Science* **233:** 879–883.

Toyonaga, B., and Mak, T. W. (1987). Genes of the T-cell antigen receptor in normal and malignant T cells. *Annu. Rev. Immunol.* **5:** 585–620.

Toyonaga, B., Yoshikai, Y., Vadasz, V., Chin, B., and Mak, T. W. (1985). Organization and sequences of the diversity, joining, and constant region genes of the human T-cell receptor β chain. *Proc. Natl. Acad. Sci. USA* **82:** 8624–8628.

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

Ullu, E., Murphy, S., and Melli, M. (1982). Human 7SL RNA consists of a 140 nucleotide middle-repetitive sequence inserted in an Alu sequence. *Cell* **29:** 195–202.

Wade, T., Bill, J., Marrack, P. C., Palmer, E., and Kappler, J. W. (1988). Molecular basis for the nonexpression of Vβ17 in some strains of mice. *J. Immunol.* **141:** 2165–2167.

Wilson, R. K., Koop, B. F., Chen, C., Halloran, N., Sciammis, R., and Hood, L. (1992). Nucleotide sequence analysis of 95 kb near the 3′ end of the murine T-cell receptor α/δ chain locus: Strategy and methodology. *Genomics* **13:** 1198–1208.

Wilson, R. K., Lai, E., Concannon, P., Barth, R. K., and Hood, L. E. (1988). Structure, organization and polymorphism of murine and human T-cell receptor α and β chain gene families. *Immunol. Rev.* **101:** 149–172.

Wilson, R. K., Lai, E., Kim, L. D. H., and Hood, L. (1990). Sequence and expression of a novel human T-cell receptor β-chain variable gene segment subfamily. *Immunogenetics* **32:** 406–412.

Yanagi, Y., Yoshikai, Y., Leggett, K., Clark, S. P., Aleksander, I., and Mak, T. W. (1984). A human T cell-specific cDNA clone encodes a protein having extensive homology to immunoglobulin chains. *Nature* **308:** 145–149.