The Complete 685-Kilobase DNA Sequence of the Human β T Cell Receptor Locus

Lee Rowen, Ben F. Koop, Leroy Hood

The human β T cell receptor (TCR) locus, comprising a complex family of genes, has been sequenced. The locus contains two types of coding elements—TCR elements (65 variable gene segments and two clusters of diversity, joining, and constant segments) and eight trypsinogen genes—that constitute 4.6 percent of the DNA. Genome-wide interspersed repeats and locus-specific repeats span 30 and 47 percent, respectively, of the 685-kilobase sequence. A comparison of the germline variable elements with their approximately 300 complementary DNA counterparts reveals marked differential patterns of variable gene expression, the importance of exonuclease activity in generating TCR diversity, and the predominant tendency for only functional variable elements to be present in complementary DNA libraries.

Many important features of higher organisms (such as olfaction, transcriptional regulation, and immune recognition) are encoded by multigene families (*1*). However, the genomic organization of large multigene families remains largely uncharacterized. To explore the organization, biology, and evolution of a complex multigene family, we have sequenced 685 kb of DNA encompassing a family of human immune recognition genes, the β TCR locus. Two general technical advances made this achievement possible: (i) The tools of large-scale DNA sequencing have reached throughputs that permit DNA in the megabase range to be sequenced in reasonable times (*2, 3*). (ii) The computational tools for assembly and analysis of DNA sequence information have progressed to the point at which megabase blocks of sequence can be comprehensively analyzed (*3, 4*).

We chose the human β TCR locus for analysis because it is of appropriate size for current large-scale DNA sequencing techniques (~0.7 Mb) and because of the vital role of this gene family in immunity. TCR molecules are expressed in a quantized manner—one type of receptor for each T cell. They interact, in conjunction with molecules encoded by the major histocompatibility complex (MHC), with the myriad peptides derived from foreign proteins (antigens) and initiate appropriate steps to destroy or eliminate foreign viruses, bacteria, and parasites, as well as cancer cells. The classic TCR is a heterodimer composed of α and β chains, each of which is divided into variable (V) and constant (C) regions that recognize peptides and are attached to the T cell, respectively. The V_β regions are encoded by a multiplicity of V_β, D_β (diversity), and J_β (joining) gene segments, one each of which undergoes rearrangement during T cell differentiation to create a contiguous V_β gene. Transcripts of a V_β gene are spliced to those of a C_β coding region to generate a β messenger RNA (mRNA) (*5*). TCR rearrangements occur in the thymus in two developmental stages: (i) D to J → DJ (pre–T cell), and (ii) V to DJ → VDJ (T cell). DNA rearrangement signals—a highly conserved heptamer followed by a variable spacer [12 or 23 nucleotides (nt)] and an AT-rich nonamer—reside at the joining boundaries of the V, D, and J gene segments and, together with the RAG1 and RAG2 enzymes, mediate the DNA joining process (*6*). TCR diversity arises by several mechanisms: (i) the germline multiplicity of V, D, and J elements; (ii) combinatorial joining of the V, D, and J gene segments; (iii) the addition of nongermline nucleotides at the gene segment junctions by terminal deoxynucleotidyl transferase (N diversity); and (iv) the combinatorial heterodimeric associations of any one of many α and many β chains.

The β TCR elements constitute a multigene family located on human chromosome 7. The structure and partial organization of the β TCR family have been characterized to date through complementary DNA (cDNA) analysis of ~270 different β transcripts (*7*) and the analysis of a few selected germline or chromosomal regions of this locus (*8–10*). The V_β gene segments have been divided, primarily by cDNA analyses, into 26 subfamilies whose individual members exhibit ≥75% sequence homology at the DNA level. Before our studies, knowledge of the organization of the β locus was limited: 5′-(unknown number and order of V_β elements)-($D_\beta1$—$J_\beta1.1$–1.6—$C_\beta1$—$D_\beta2$—$J_\beta2.1$–2.7—$C_\beta2$—$V_\beta20$)-3′. The complete DNA sequence of the β TCR locus has thus provided new insights into the organization, evolution, and diversification of this gene family, as well as into the general architecture of the largest stretch of a human chromosome analyzed to date.

Genes of the β TCR Locus

A schematic representation of the sequence features of 685 kb of DNA spanning the β TCR locus is shown in Fig. 1 (*11*). Two categories of genes lie within this region: those encoding TCR elements and other genes (Fig. 1A). We used several gene-finding methods and similarity analyses to identify two types of non-TCR genes (*12*). A dopamine-β-hydroxylase–like gene lies at the 5′ end of this sequence. Eight trypsinogen genes (see below) are divided into two clusters, three immediately 3′ to the dopamine-β-hydroxylase–like gene and five immediately 5′ to the $D_\beta1$ gene segment. TCR elements were identified by dot-matrix comparison of the germline sequence with previously characterized cDNAs. This analysis revealed a total of 65 V_β gene segments, six of which are new. All but one of these lie between the dopamine-β-hydroxylase–like gene and the $D_\beta1$ gene segment. The duplicated DJC clusters, separated by 2.5 kb, each contain one D_β and six or seven J_β gene segments, as well as a C_β gene. At the 3′ side of the $C_\beta2$ gene are an enhancer element and the 65th V_β gene segment lying in an inverted translation reading frame with respect to the 81 other TCR elements. In total, the coding regions represent 4.6% of this sequence.

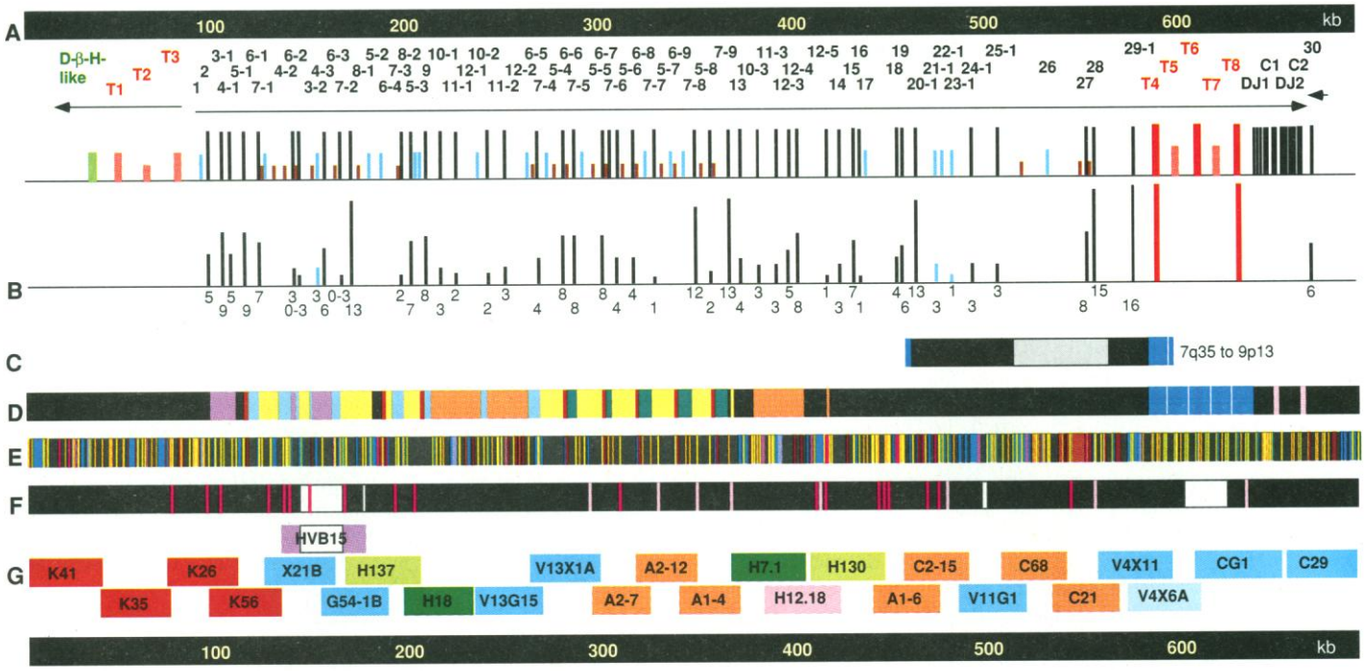
Historically, the V gene segments and subfamilies have been named according to their order of discovery. A revised nomenclature was published recently (*13, 14*) that largely preserves the conventional names. Now that the complete genomic sequence is available, we propose that a more logical nomenclature assigns the V subfamilies consecutive numbers, starting at the 5′ end of the locus. The individual subfamily members are then numbered sequentially after the subfamily designation. The proposed new system is presented in Fig. 1A, and Fig. 1H presents a translation between the proposed new system and a shorthand version of the recently revised nomenclature used by immunologists (*13, 14*). These two nomenclatures will be used in conjunction throughout the text,

L. Rowen and L. Hood are in the Department of Molecular Biotechnology, University of Washington, Seattle, WA 98195–7730, USA. B. F. Koop is at the Center for Environmental Health, Department of Biology, University of Victoria, Victoria, British Columbia, Canada V8W2Y2.

with the proposed new designation in parentheses after the shorthand conventional designation—for example, V_β4S1 (29-1).

Structures of the V_β Gene Segments

The V_β gene segment is composed of five elements: (i) a promoter region, (ii) a first exon coding for a signal peptide, (iii) an intron with 5′ and 3′ RNA splicing signals, (iv) a second exon encoding the V element, and (v) a DNA rearrangement signal sequence (*15*) (Fig. 2). In some V_β gene subfamilies (5, 6, 7, and 12), a conserved decamer with the consensus sequence AGTGAYRTCA (Y = C or T; R = A or G) has been shown to interact with nuclear binding proteins (*10, 16*). Our analysis confirms the presence of the decamer in many V gene subfamilies and further shows that the decamer is a subcomponent of a larger 14-nt palindrome (CAGTGAYRTCACTG) located ~80 to 120 base pairs (bp) 5′ of the transcriptional start site. Sequences matching at least 10 of the 14 nt have been identified in 42 of the 65 V gene segments



H

1	27S1	4-3	7S2	10-1	12S2	7-5	6S9	5-8	5S4	16	25S1	26	28S1
2	22S1	6-3	13S2	11-1	21S1	5-5	5S3	7-9	6S4	17	26S1	27	14S1
3-1	9S1	7-2	6S5	12-1	8S4	6-7	13S8	13	23S1	18	18S1	28	3S1
4-1	7S1	8-1	30S1	10-2	12S3	7-6	6S3	10-3	12S1	19	17S1	29-1	4S1
5-1	5S1	5-2	31S1	11-2	21S3	5-6	5S2	11-3	21S2	20-1	2S1	30	30S1
6-1	13S3	6-4	13S5	12-2	8S5	6-8	13S7	12-3	8S1	21-1	10S1		
7-1	6S7	7-3	6S1	6-5	13S1	7-7	6S6	12-4	8S2	22-1	29S1		
4-2	7S3	8-2	32S1	7-4	6S8	5-7	5S7	12-5	8S3	23-1	19S1		
6-2	13S2	5-3	5S5	5-4	5S6	6-9	13S4	14	16S1	24-1	15S1		
3-2	9S2	9	1S1	6-6	13S6	7-8	6S2	15	24S1	25-1	11S1		

Fig. 1. Schematic representation of the human β TCR locus. (**A**) Genes. A dopamine-β-hydroxylase–like pseudogene (D-β-H-like), showing ~40% amino acid similarity to human dopamine-β-hydroxylase, was identified at the 5′ end of the sequence by Xgrail (version 1.2), MpSRCH, and BlastX. The eight trypsinogen genes (T1 to T8) were identified by Xgrail, BlastX, or dot-matrix analyses with cDNAs. T4 and T8 have been previously identified as cDNAs for trypsinogen 1 and 2, respectively (*25*). Arrows indicate the direction of transcription. Functional genes (black) are indicated by full-height vertical lines, pseudogenes (blue) by half-height lines, and relics (maroon) by quarter-height lines. (**B**) Complementary DNAs. The number of instances of cDNAs assigned to a given gene, based on a search of the primate database of GenBank release 88, is indicated. (**C**) Translocation. Preliminary sequence data indicate that a duplication and chromosomal translocation (*34*) include ≥60 kb of DNA 5′ of V_β14S1 (27) and 40 kb of DNA 3′ of V_β3S1 (28). The gap between the two contigs is ~75 kb. The translocated region contains at least the V orphon genes V_β2S2 (20-2), V_β10S2 (21-2), V_β29S2 (22-2), V_β19S2 (23-2), V_β15S2 (24-2), V_β11S2 (25-2), and V_β4S2 (29-2), as well as the functional trypsinogen 3 (T9) gene. (**D**) Homology units. Eight types of locus-specific repeats have been identified by dot-matrix analysis and multiple sequence alignments. Color codes for homology units: A (purple), B (red), C (light blue), D (yellow), E (orange), F (green), G (dark blue), and H (pink). (**E**) Genome-wide repeats. With the use of cross_match and an extensive database of genome-wide interspersed repeats, ~30% of the sequence was divided into five broad categories: (i) SINEs Alu and MIR (yellow and orange); (ii) LINEs (blue); (iii) MaLR (maroon), retroviral (lavender), and other (red) LTRs; (iv) DNA transposons (green); and (v) unclassified repeats (pink). (**F**) Polymorphisms. Microsatellite repeats of more than eight consecutive units were examined for polymorphism. Those in dark pink were found to be polymorphic; those in light pink were not (*45*). Areas blocked in white represent insertion-deletion polymorphisms. (**G**) Cosmids. Cosmids H7.1, H12.18, and H130 were sequenced by a primer-walking method (*8*). All others were sequenced by the high-redundancy shotgun method (*11*). The two gaps between cosmids H18 and V13G15 and between CG1 and C29 were filled by sequencing PCR products prepared from genomic DNA. Colors indicate that cosmids were derived from different libraries (for example, K41 and X21B) or represent different haplotypes within a given library (for example, V4X11 and V4X6A). Overlapping cosmids of different colors indicate that the DNA is derived from different haplotypes and, hence, that the sequence variations detected in the overlapping regions might be polymorphic. These variations are annotated as such in the L36092 database entry. (**H**) Nomenclature. A translation between our proposed new nomenclature and the shorthand version of the conventional nomenclature (*13, 14*) is shown, starting at the 5′ end.

(Fig. 2); thus, even though all V_β genes exhibit similar expression patterns in T cells, 23 V_β gene segments do not have a motif similar to the 14-nt sequence.

The V region is composed of a leader peptide, which usually comprises 16 amino acids and is hydrophobic, and a V element, generally consisting of 99 amino acids with specific highly conserved residues (such as Gln^{25}, Cys^{42}, Cys^{101}, and Trp^{53}). As expected, genealogical analysis shows a clustering of the members of individual subfamilies. With the criterion of ≥75% sequence similarity, there are 30 V_β subfamilies in the β locus ranging in size from one to nine V members. The subfamilies are evolutionarily relevant in that they represent the most recent gene duplications of V homology units (see below).

The 65 V gene segments can readily be divided into functional genes and pseudogenes by computational analysis at the DNA level. Pseudogene lesions may arise from errors in gene expression (failures in transcription, RNA splicing, translation, or DNA rearrangement) or protein structure (premature termination, improper reading frame translations, or substitutions compromising protein function). By these criteria, 46 V gene segments appear functional and 19 represent pseudogenes (Fig. 1A and Table 1).

Dot-matrix analyses identified the 65 V gene segments described above. In the course

V nomenclature		Promoter	Init. codon	5′ splice	Intron 1 (bp)	3′ splice	ORF	cDNA	Recombination signal
		TCAGTGAYRTCACTGA		GT		AG			CACAGTG ACAAAAACC
27S1	1	ATAGTGACACCAACAG	+	TCAGGTGAGTCCTGGGC	117	TATTTTCATAGGCTC	-	-	CACAGCCC. TGCAGAGTCACCGCCTCCCTGTGCACAAACCTCCTGGA
22S1	2	CCAGTAATTCTGCCAG	+	GCAGGTCGATGCTTAGA	88	TTTTCTCACAGGACT	+	+	CACAGCCT. TGCAAAGACAACTCCAGCCTGTGCAAAATCCCTCACAG
9S1	3-1	GCTGTGACCTCACTGG	+	GCAGGTGAGTCCCGGGC	116	TTCCTTTGCAGGTCC	+	+	CACAGCCC. TGCAAAGTCACTGCATCCCTGTGCACAAACCTCCCGGC
9S2	3-2	GCTGTGACCTCAATGG	+	GCAGGTGAGTCCTGGGC	116	TTCCTTTACAGATCC	-	+	CACAGCCT. TACAGAGCCACTGCATCCCTGTGCACAAACCTCCCGGC
7S1	4-1	GCAGTGACATCACAAG	+	GCAGGTGAGTTAGGTTG	110	TCCTCCTACAGTTCC	+	+	CACAGCCT. TGCAGAGTCACCGCTTTCCTGTGCAGAAACCTTCGGGG
7S3	4-2	GCAGTGACATCACAGG	+	GCGGGTGAGTTGGGTTC	110	TCCCCCCGCAGTCCC	+	+	CACAGCCT. TGCAGAGTCACCGCTTTCCTGTGTAGAAACCTTAGGGG
7S2	4-3	GCAGTGACATCACAGG	+	GCGGGTGAGTTGGGTTC	110	TCCCCCCACAGTCCC	+	+	CACAGCCT. TGCAGAGTCACCGCTTTCCTGTGCAGAAACCTTAGGGG
5S1	5-1	ACAGTGACATCACAGT	+	GCAGGTGAGTCCCTGCA	127	CACAGGCCCAGTAAA	+	+	CACAGCCC. TACAAAGCCAACCACATTCTGTGCACAAACCTCCCTGG
31S1	5-2	ACAGTGACATCACAGG	+	GCAGGTGAATCCTGTGG	109	TTTTTCCTCAGGCCC	-	-	TGCAGGCC. TGCAGAGCCAAGAACATTCTGTGTACAAACATCCCTGC
5S5	5-3	ACAGTGACATCACAGG	+	GCAGATGAGTCCTGTGC	124	TTTTCCCACAGGCCC	-	-	CACAGCCC. TGCAGAGTCACTGGAACTCTGTGCACTAATCTCTCTGC
5S6	5-4	ACAGTGACATCACTGT	+	GCAGGTGAGTCCTGGGC	122	TTTTCCCACAGGCTC	+	+	CACAGCCC. GGCAGAGTCTCTGACATTCTATACATAAACTTCCTGCC
5S3	5-5	ACAGTGACATCACAGG	+	GCAGGTGAGTCCTGGGC	122	TTTTCCCACAGGCCC	+	+	CACAGCCC. GGCAGAATCACTGACATTCTGTATATAAACTTCCTGCC
5S2	5-6	ACAGTGACATCACAGG	+	GCAGGTGAGTCCTGGGC	123	TTTTCCCACAGGCTT	+	+	CACAGCCC. ATCAGAGTCACTGACGTTCTGTATATAAACTTCCTGCC
5S7	5-7	ACAGTGACATCACAGC	+	GAAGGTGAGTCCCGGGC	124	TTTTCCCACAGGCCC	-	-	CACAGCCC. AGCAGAGTCACTGACATTCTGTATATAAACTTCCGCCT
5S4	5-8	ACAGCGACATCACAGG	+	ACAGGTGAGTCCTGGGC	124	TTTTCCTGCAGGCCC	+	+	CACAGCCC. AGCAGAATCACTGACATTCTGTATATAAACTTCCTGCC
13S3	6-1		+	GCAAGTAAGTCCTGGGC	89	CTTCTCTGCAGGTCC	+	+	CACAGCGC. TGCATGGCCGTCTCCTCTCTGCACATAAAGGCAGGGAG
13S2	6-2		+	GCAGGTGGGTCCTGGGC	89	CTTCTCTGCAGGTCC	+	+	CACAGTGC. TGCACGGCTGTCTCCTCTCTGCACAGAAAGGCAAGGGA
13S2	6-3		+	GCAGGTGGGTCCTGGGC	89	CTTCTCTGCAGGTCC	+	+	CACAGTGC. TGCACGGCTGTCTCCTCTCTGCACAGAAAGGCAAGGGA
13S5	6-4	AGAATGAGGTCTCAGG	+	GCAGGTAAGTCCTGGCA	91	TGTAGGTCCAGTGTT	+	+	CACAGTGC. TGCACAGCCATCTCCTCTCTGTACATAAATGCAGGGGA
13S1	6-5		+	GCAGGTGGGTCCTGGGC	92	CTTCTCTGCAGGTCC	+	+	CACAGCGC. TACAAGGCCGTCTCCTCTCTGCACATAAAGGCAGGGAG
13S6	6-6		+	GCAGGTGGGTCCCGGAC	89	CTTCTCTGCAGGTCC	+	+	CACAGCGC. TACAAGACCATCTCCTCTCTGCACATAAAGGCAGGGTG
13S8	6-7		+	GCAGGTGGGTCCTGGGC	89	CTTCCCTGCAGGTCC	+	-	CACAGCGC. TGCAAGGCTGTCTCCTCTCTGCACATAAAGGCAAGGGA
13S7	6-8		+	GCAGGTGGGTCCTGGGC	88	CTTCTCTGCAGGTCC	+	-	CACAGCGC. TACAAGGTTGTCTCCTCTCTGCACATAAAGGCAGGGAG
13S4	6-9		+	CCAGGTGGGTCCTGGCA	94	CTTCTCTGCAGGTCC	+	-	CACAGCGC. TGCAAGCCTGTCTCCTCTCTGCACATAAAGGCACAGAG
6S7	7-1	GATGTGATGTCACTAT	+	GCAGGTGATTCCTCAGA	142	TGTCTCCACAGATCA	-	-	CACAGCAC. TACTGCTCCAGTGTCAGCTTGGTTCCCTAGGAAATGGG
6S5	7-2	GAAGTGATGTCACTGT	+	GCAGGTGAGTCCTCAGA	150	TGTCTCCACAGATCA	+	+	CACAGCAT. GGCACAGTCGCCTCCTTCCTGCTCATAAACCTCATCCT
6S1	7-3	GAAGTGATGTCACTGT	+	GCAGGTGAGTCCTCAGA	110	TGTCTCCGCAGATCA	+	+	CACAGCAT. GACACAATCGCCTCCTTCCTGCTCATAAACCTCCTCCT
6S8	7-4	GAAGTGATGTCACTGT	+	ACAGGTGAGTCCTTAGA	115	TATTTCCACAGATCA	+	-	CACAGCGT. GGCACAGTCGCCTCCTTTCTGCTCACAAACCTCATCCT
6S9	7-5	GAAGTGATGTTACTGT	+	ACAGGTGAGTCCTCAGA	127	TATTTCCACAGATCA	-	-	CACAGCGT. GGCACAGTCGCGTCCTTCCTGCTCACAAACCTCATCCT
6S3	7-6	GAAGTGATGTCACTGT	+	ACAGGTGAGTCCTCAGA	146	TATTTCCACAGATCA	+	+	CACAGTGT. GGCATAGTCGCCTCCTTCCTGTTCACAAACCTCATCCT
6S6	7-7	GAAGTGATGTCACTGT	+	ACAGGTGAGTCCTCAAA	155	TGTTTCCACAGATCA	+	+	CACAGCAT. GGCACAGTCGCCTCCTTCCTGTTCACAAACCTCATCCT
6S2	7-8	GAAGTGATGTCACTGT	+	ACAGGTGAGTCCTCAGA	137	TATTTCCACAGATCA	+	+	CACAGCAT. GGCACAGTTGCCTCCTTCCTGTTCACAAACCTCATCCT
6S4	7-9	TCAGTGATGTCACTGT	+	GCAGGTGAGTCCTCAGA	126	TATCTCCACAGATCA	+	+	CACAGCAT. GGCACAGTCGCCTCCTTCCTGCTCACAAACCCTCAGGC
30S1	8-1		+	CTAGGTGAACTCTGTAG	492	CTCTCCCACAGCCCC	-	-	CGCAGCCC. TGCACAGCCAACTGCTCTCTGCACAAAAAAGGGCAGCC
32S1	8-2	GGGGTGACACCACAGG	+	GCAGATAAGTTCTGGAA	136	CTCTCCAGCAGGCTC	-	-	CGCAGCCC. TGCACAGCCAGCTGCCCTCTGCACAAAAAGGGCAGTCA
1S1	9	GTGGTGACATCACAGG	+	GCAGGTGAGTCCTGGGC	132	CACAGGCCCAGTGGA	+	+	CACAGCCC. TGCATGAGCATCAGCCTTCTGTGCAATAACATTCCTGC
12S2	10-1	ACAGTGACATCATCAA	+	GCAGGTGAGGGCTGGTC	106	ACAGGACACAGGGAT	+	+	CACAGTGC. TGCACAGCTGCCTCCTCTCTGCACATAAAGGGCAGTTA
12S3	10-2	ACAGTGACATCATCAA	+	GCAGGTGAGGGCTGGTC	106	CTCTATTACAGGACA	+	+	CACAGTGC. TGCACAGCTGCCTCCTCTCTGCACGGAAACGGCAGTTA
12S1	10-3	GCACTGATGTCATCGA	+	ACAGGTGAGGGCTGGTC	106	TTCTATTACAGGACA	+	+	CACAGTGC. TGCATGGCTGCCTCCTCTCTGCACGTAAACAGCAGTTA
21S1	11-1	ACAATGATGTTACTGT	+	GCAGGTGTGTCCTAAGA	100	TTCCCCAACAGAACT	+	+	CACAGCGT. TGCAGAGACTT. TCTCTCCTGTGCACAAAACTCCAGGGC
21S3	11-2	GCAATGATGTCACTGT	+	GCAGGTGAGTCCTAGGA	92	TTCCTTCTCAGAACT	+	+	CACAGTGT. AGCAGAGACACTTCCCTCCTGTGCAGAAAACCAGAAAA
21S2	11-3	ACAATGATGTCACTGT	+	GAAGGTGAGTCCTAGGA	91	TTCCCCCACAGAACT	+	+	CACAGTGT. AGCAGAGACACTTCCCTCCTGTGCAGAAAACCGCAGGA
8S4	12-1	TCAGTGACGTCACTGA	+	GCGAGTGAGTTCTCCGA	99	TTCTTTTATAGCACA	-	-	CACAGCAC. TGCAGAATCTCCCCATCTCTGTGCAGAAACCCTGGTGC
8S5	12-2	TCAGTGTTGTCACTAA	+	GCGAGTGAGTCCTCATG	99	TTCTTTTATAGCATG	-	-	CACAGCGC. TGCAGAATCTCCCCCTCCCTGTGCAGAAACTCTGGTGC
8S1	12-3	TCAGTGATGTCACTGA	+	GCGAGTGAGTCTTCAGA	100	TTCTTTCATAGAGCA	+	+	CACAGCGC. TGCAGAATCACCCCTTTCCTGTGCAGAAAACCCGGTGT
8S2	12-4	TCAGTGATGTCACTGA	+	GCAAGTGAGTCTTCAGA	100	TTCCTTCATAGAGCA	+	+	CACAGCGC. TGCAGAATCACCCCTTTCCTGTGCAGAAACCCTGGTGT
8S3	12-5	ATGGTGATGTCACTGA	+	GAAGGTGAGTCCCCACA	100	TCTTTTCACAGAGCT	+	+	CACAGCGC. TGCAGAATCACCTGCTCCCTGTGCAGAAACCCTGGTGC
23S1	13	ACAGTGACATCACAGT	+	GCAGGTGAGTCCTGAGT	110	TTCCTCCACAGTTTC	+	+	CACAGACCCTGGAGAATTACTGGCTTTCTGTACCCAAACCCTCCTAT
16S1	14	ATGCTGATGTCACTGG	+	GCAAGTGAGTCTTCAGG	86	TCTTCCCACAGAGCA	+	+	CACAGTGC. TTCACAGTCGTGCCCTTGCTGTGCAAAACCATAGCCTT
24S1	15		+	ACAGGTGAGTACTGGGC	126	TTTCTCTACAGGTCA	+	+	CACAGAGC. TGCAGTGCTTCCTGCTCTCTGTTCATAAACCTCATTGT
25S1	16		+	GCAGGTAAGTCCCTGTT	107	TTTCTTTACAGGTTC	+	+	CACAATGT. TAAATATTAGCTAATCTTAGGACACAGACTCATCACGG
26S1	17		+	GCAGGTGGGTCCAGGTA	391	TTCTTCCACAGGACA	-	-	CACAGCAT. GGCTGAGTCAGTTCCCTCCAGGGTGCAAACCCTCTGGC
18S1	18		+	GCAGGTGAGTTTTCAGC	272	TCTCCTCTGAGGTCT	+	+	CACATTGA. TGCAGAGCCACATCCTCTCAGTCCACAAACATCCTCCA
17S1	19		+	GCAAGTGAGTCCTGGGT	132	TTCTTTTCCAGACAC	+	+	CACAGTGA. AGCACGGATGTCGCCTCTCTGTGCATAAATGTGCCCAG
2S1	20		+	CCAGGTATAAGCCTCCT	338	TGTCTCAGCAGGCTC	+	+	CACAGCGC. CAGGAGGGGATCAGACACCGCGGCAAGAACCCCTGCAG
10S1	21		+	CCAGGTAAGGCCCTGTT	113	TTCCTCCACAGGCTC	-	+	CACAGTGC. CGAATGTTAGCCCTTCTTAGAACACAAACTCATTATGG
29S1	22		+	GCAGGTGAGCTCAGCAC	132	ACATCTATCAGATGC	-	-	CACAATGG. AAGCACAACCATTGTCTCTCTGTGCGGAAATGTGTCCT
19S1	23		+	GCAGATGCGCCCTGAGC	150	TTCTCTTACAGACTC	-	+	CACAGCAC. TGAAATGTCAGTTCCTCTTAGCACACAAACTTGTCACA
15S1	24	ACAATGACATCACTTC	+	ACAGGTGAGTTTGGAAC	132	ATTTCCCACAGGGTC	+	+	CACAGTGC. TTCTTGGCCACCTGCTCTCTACACAGAAAGACAGACAC
11S1	25		+	GCAGGTAAGTCATAGAC	126	TCTTCCCATAGGCCT	+	+	CACAGTGC. TACATGGATACCGACACTCCGCACAGAAAGGGTCGCCT
28S1	26		+	GCAGGTGAGTCCTGGGC	141	CCCCCCAACAGGCCT	-	-	CATAGCAC. TGCATAGCCATATCCTCTCCACACAAAAAAGGTGCAT
14S1	27		+	GCAGGTGAGTCCCAGAA	129	TTCCTTTACAGGCCC	+	+	CACAGTGT. TGCACAGCCAGCTGCTCTCTGCACAAAAACAGAGGGTA
3S1	28	ACAATGACATCACAGA	+	GTAGGTGAGTCCTGGGA	137	GTCTCTCACAGGCCT	+	+	CACAGCGC. AGCACAGCTGCATCCTCTCTGCACAAAAAGAGCGGACG
4S1	29		+	CTAGGTATGAGCCTTGT	279	TCTGGAACTAGGCTC	+	+	CACAGTGC. GGGGCACAGATCAAAGATCTGAGCAAGAACCTCAGCTC
20S1	30		+	TTTGGTGGGTGCCCTCC	366	CTTTCCCACAGGGGT	+	+	CACACTGA. GCTGGGTGGGGCAGACATCTGTGCAAAAACCCCACCCT
Human V	70%	AGTGAYRTCAC G		GTGAGTCCT		CAG			CACAG GCA AG C C CTG CA AAA
	80%	A TGAYRTCAC G		GTG GT C		CAG			CACAG CA CTG CA AAA
	90%	TGAYRTCA		GT G		AG			CACAG AA
Human J	70%			GTAAGT					CAC GT C ACAAAAAC
	80%			GTAAGT					CAC CA AAAC
	90%			GT GT					CAC CA A
				Ig-TCR	70%				CACAGTG ACAAAAACC
					80%				CACAGT AC AAAC
					90%				CAC AA

Fig. 2. Features of the human V_β gene segments. A conserved 14-nt promoter sequence was identified ~80 to 120 bp upstream of exon 1. Boundaries and splice sites of exons 1 and 2 were identified by dot-matrix comparisons with similar cDNA sequences. Consensus sequences determined at the 70%, 80%, and 90% levels were obtained from all the promoters, intron 1 5′ splice sites, intron 1 3′ splice sites, and recombination signals of the TCR V_β gene segments. These consensus sequences are compared to consensus sequences (also at 70%, 80%, and 90% levels) obtained from human TCRα J gene segments and from immunoglobulin (Ig) variable gene segment rearrangement signals (*17*). For each gene, the size of the intron and the presence (+) or absence (–) of the correct initiation (init.) codon, an open reading frame (ORF) in exons 1 and 2, and a corresponding cDNA identified in GenBank are indicated.

of applying a new similarity search program [cross_match, a modified Smith-Waterman algorithm (*4*)], we identified 22 additional sequences with limited local similarity to V gene segments, each with several major lesions in one or more basic components (Fig. 1A). We term these V elements "relics" because they will presumably never regain functionality. The boundary between relics and some of the more damaged V pseudogenes is somewhat arbitrary. Translated sequences of relics are virtually impossible to align effectively in multisequence comparisons because of extensive insertions and deletions in the exons. Although relics provide no functional information, they contribute to a dynamic view of the evolutionary changes occurring in this multigene locus.

The RNA splicing signals for the introns of the V gene segments are conventional: 5′ GT and 3′ AG. The sizes of the introns correlate well with evolutionary proximity; that is, members of the same V subfamily generally share introns of similar size (Fig. 2). The range of intron sizes extends from 86 to 492 nt.

The DNA rearrangement signals of the β TCR locus are similar to those of their human and mouse immunoglobulin and J_α gene segment counterparts (*17*) (Fig. 2). All show the classical heptamer-spacer-nonamer structure. In only 2 of the 65 V gene segments did the spacer not comprise 23 bp, and in these instances it differed by only 1 bp. The possibility that sequence variation in the DNA rearrangement signals affects the efficiency of V_β to $D_\beta J_\beta$ rearrangements is suggested by the reduced expression of a V gene segment whose only sequence variation occurs in the 23-bp spacer portion of the recombination signal (*18*). This issue can now be investigated further in in vitro rearrangement systems and in appropriately constructed transgenic mice.

Comparison of TCR Germline and cDNA Sequences

The availability of the β germ line and >267 partial or complete V_β cDNA sequences makes possible an analysis of aspects of the phenotypic expression of V gene segments, including V, D, and J usage, junctional diversity, and the identification of pseudogene candidates. The 267 distinct β cDNAs present in the GenBank release 88 primate database (those with >50 bp at the 3′ end of the V gene segment) are encoded by 48 different V_β elements. Some V_β and J_β gene segments are expressed more frequently than others (Fig. 1B) (*19*). This differential V or J gene segment expression could arise from selection (in the thymus or periphery) or from differential transcription (perhaps as a result of differing promoter strengths), RNA stabilities, or DNA rearrangement probabilities. Indeed, V_β gene segments that contain [for example, V_β6S5 (7-2)] or lack [for example, V_β4S1 (29-1)] the 14-nt promoter palindrome can be highly expressed. With regard to the question of whether the V_β profiles of expression change from one individual to the next in the thymus and periphery, a variety of data suggests that the V_β expression profiles in the periphery are relatively fixed (*20*), whereas more limited data suggest that those in the thymus are also fixed but differ from those of the periphery (*21*). Thus, it appears that intrinsic as well as selective mechanisms may influence these ratios of V gene segment expression. An analysis of the promoter regions shows that those within V subfamilies are similar, even though V elements within a subfamily may be expressed in different amounts—for example, V_β6S5 (7-2) and V_β6S6 (7-7) (Fig. 1B). Likewise, the heptamer-nonamer sequences of these two V elements are identical, with the exception of a 1-bp substitution. In these examples, neither the promoter nor the DNA rearrangement signal appears to explain the markedly different amounts of expression.

Several events occur in the generation of junctional diversity (6): The boundaries of the gene segments to be joined are cleaved and modified by exonucleases, after which nongermline (N) nucleotides are added and the resulting junctional region sequences are ligated together (Fig. 3). Palindromic (P) nucleotides can arise from palindromic pairing of germline nucleotides at the ends of the gene segments. Few palindromic nucleotides were observed at the N region ends, which suggested that this mechanism for generating diversity is relatively unimportant. Evaluation of exonuclease activity at the 3′ end of the V gene segment (E_V) and N nucleotide addition at the $V_\beta D_\beta$ junction (N_V) was previously impossible because the germline 3′ boundaries of most of the V_β gene segments had not been established. We analyzed the junctional diversity from 193 β cDNAs sequenced across this region (Fig. 3) and showed that an average of 3.6 and 4.6 N nucleotides is added to the

Table 1. Pseudogene classification. Errors in splice sites and recombination signals were identified on the basis of features of the genomic sequence. The exons were joined from the genomic sequence and translated to identify lesions in the proteins. V_β13S4 (6-9) and V_β6S8 (7-4) appear to be pseudogenes by structural modeling (*23*). U, unknown amino acid.

V gene segment	Defects	Reported cDNA
27S1 (1)	Stop codon in exon 1; Cys → Arg in exon 2	No
9S2 (3-2)	Cys → stop in exon 2	Yes
31S1 (5-2)	Frameshift in exon 2; heptamer violates consensus	No
5S5 (5-3)	GT → AT in splice donor	No
5S7 (5-7)	Trp → Ser in exon 2	No
13S8 (6-7)	Trp → Arg in exon 2	No
13S4 (6-9)	Leu → Arg in exon 2	No
6S7 (7-1)	Cys → Tyr in exon 2; missing spacer and nonamer	No
6S8 (7-4)	Leu → Arg in exon 2	No
6S9 (7-5)	Frameshift in exon 2	No
30S1 (8-1)	Frameshifts in exon 2	No
32S1 (8-2)	Frameshifts in exon 2	No
8S4 (12-1)	U → stop in exon 2	No
8S5 (12-2)	U → stop in exon 2	No
26S1 (17)	Cys → Tyr in exon 2	No
10S1 (21-1)	Frameshift in leader	Yes
29S1 (22-1)	Frameshift in exon 2	No
19S1 (23-1)	GT → AT in splice donor	Yes (includes intron)
28S1 (26)	Cys → Tyr in exon 2	No

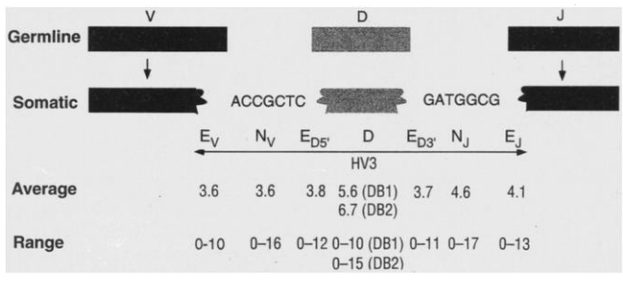


Fig. 3. Features of junctional diversity. One hundred and ninety-three cDNAs were identified from the primate database of GenBank release 88 and compared with the germline sequence. In the analysis of exonuclease activity and N nucleotide addition, the maximal uniquely identifiable extent of germline sequence for the V, D, and J regions was counted as germline sequence (rather than as N addition). Thirty-four cDNAs for which a unique germline D assignment could not be made were not included in the N-D-N analysis. E, nucleotides deleted by exonuclease; HV3, hypervariable region 3.

VD and DJ junctions, respectively. An average of 3.6, 3.8, 3.7, and 4.1 nt is removed from the V, D (5′ and 3′), and J ends, respectively (Fig. 3). There is no apparent correlation between the extent of N addition or nucleotide removal and specific V, D, or J elements (although, in some instances, there are insufficient cDNAs in the analysis to detect such a correlation if one exists). $D_\beta1$ gene segments may join either to $J_\beta1$ or $J_\beta2$ gene segments. One, and possibly two, $D_\beta2$ gene segments joined to a $J_\beta1.5$ gene segment, perhaps representing a nonhomologous recombination event between paternal and maternal chromosomes. The N regions on the 5′ side of the D element are enriched in C (35% C, 19% A, 23% G, 23% T), whereas the N regions on the 3′ side of D are enriched in G (35% G, 22% A, 22% C, 21% T). Given the preference of terminal deoxynucleotidyl transferase for deoxyguanosine triphosphate as substrate (*22*) and the 5′ → 3′ direction of polymerization, these data suggest that N addition proceeds from both ends of the D gene segment. The D1 and D2 gene segments are used in all three reading frames, although frames 1 and 2 appear to be used 2.5 times as often as frame 3 for both D segments.

These data emphasize the importance of nucleotide removal in generating diversity in the third hypervariable (HV3) region (defined as the number of nucleotides between the germline V and germline J portions of the cDNA). The range of HV3 size variation is 0 to 28 nt. Nucleotide removal thus allows the V and J germline gene segments to join to the size-variable HV3 sequences at many different positions, greatly enhancing diversity in this region. The HV3 region plays a central role in the recognition of MHC-peptide complexes.

Analysis of cDNAs also allows reexploration of the pseudogene question. Any V element expressed as a cDNA (mRNA) contains functional signals for gene expression (transcription, RNA splicing, and DNA rearrangement). Of the 48 V elements that are expressed as cDNAs (Fig. 1B), only 3 [$V_\beta9S2$ (3-2), $V_\beta10S1$ (21-1), and $V_\beta19S1$ (23-1)] are clearly pseudogenes. Both β loci are rearranged in virtually all T cells; although one rearrangement is almost always nonfunctional, there is no apparent reason to expect that both rearranged V gene segments would not be expressed. It is possible that transcripts from most nonfunctional rearrangements are rapidly turned over. All apparently functional V gene segments except three [$V_\beta13S7$ (6-8), $V_\beta13S4$ (6-9), and $V_\beta6S8$ (7-4)] appear to be expressed. Because these three gene segments appear functional by the one-dimensional criteria typically used in identifying pseudogenes, they were modeled in three dimensions (*23*). The models reveal that highly conserved hydrophobic leucine residues in $V_\beta13S4$ (6-9) and $V_\beta6S8$ (7-4) have been converted to hydrophilic arginines (Leu-Arg-Leu-Ile → Leu-Arg-Arg-Ile and Gly-Leu-Pro → Gly-Arg-Pro, respectively). In both instances, the leucine residues are buried in the hydrophobic face between the α and β subunits, and the charged arginine residues would completely disrupt this interface (*23*). The third apparently functional V_β element not expressed at the cDNA level [$V_\beta13S7$ (6-8)] may also be a pseudogene (a glycine substitution not previously detected in functional V elements may compromise the first hypervariable region). Hence, this detailed cDNA analysis provides a powerful tool for identifying putative pseudogenes by virtue of their lack of expression at the mRNA level.

The cDNA analyses are also useful for establishing exon-intron boundaries (Fig. 2) and for identifying new genes within a given locus (such as the dopamine-β-hydroxylase–like and trypsinogen genes). The importance of comparative germline-cDNA analysis is increasing as the number of available cDNA sequences (also being collected as expressed sequence tags) is increasing (*24*).

Boundaries of the Human β TCR Locus

We believe that we have sequenced all of the functional β TCR coding elements. First, and most important, all 267 β cDNA sequences previously identified are encoded by these elements lying within the sequenced families. Hence, any unidentified functional TCR elements would have to be rarely expressed. Second, the 5′ dopamine-β-hydroxylase–like pseudogene does not appear to have a function related to immune recognition and, accordingly, probably represents the 5′ boundary of the β TCR family. On the 3′ side, we have not yet reached a non-TCR gene, but because we have extended the 3′ and 5′ boundaries well beyond the last TCR elements it is likely that the entire β family is included in this sequence.

Intercalation of Trypsinogen Genes in the β TCR Locus

A dot-matrix analysis of the entire 685-kb DNA segment against itself revealed five tandemly arrayed 10-kb locus-specific repeats (homology units) at the 3′ end of the locus, in the 70-kb region between the $V_\beta4S1$ (29-1) and $D_\beta1$ elements (Fig. 4). These repeats exhibit 90 to 91% overall nucleotide similarity, and embedded within each is a trypsinogen gene. Alignment of pancreatic trypsinogen cDNAs with the germline sequences shows that these trypsinogen genes contain five exons that span ~3.6 kb. Further analyses revealed two pseudo trypsinogen genes and one relic trypsinogen gene at the 5′ end of the sequence, all in inverted transcriptional orientation. The eight trypsinogen genes are denoted T1 through T8 from 5′ to 3′ (Fig. 1A).

Analysis of the trypsinogen cDNAs in the databases reveals that the pancreas contains three mRNAs, but only two correspond to trypsinogen genes in the β TCR locus [T4 is denoted trypsinogen 1 and T8 is trypsinogen 2 (*25*)]. The third pancreatic cDNA [identified independently as trypsinogen 3 and 4 (*26*, *27*)], although closely related to the others, is distinct from the third apparently functional trypsinogen gene

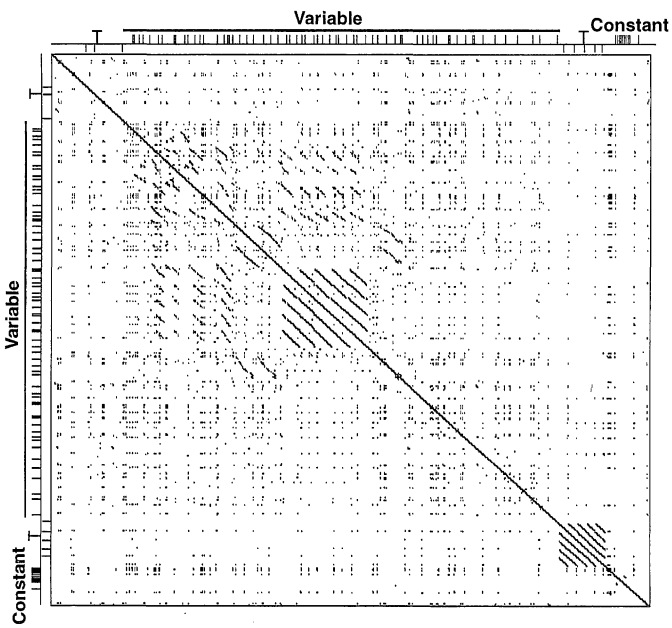


Fig. 4. Dot-matrix analysis of the β TCR locus sequence versus itself. The 685-kb sequence was plotted against itself with the Inherit analysis program (Applied Biosystems) and the following parameters: length, 50; offset, 50; match, 42. With the exception of the main diagonal, diagonal lines indicate internal repeats (homology units) in the sequence. The dots in the background indicate the common genome-wide interspersed repeats, for example, Alu. T, trypsinogen genes.

(T6) in this locus. The T6 gene is deleted in a common insertion-deletion polymorphism (see below); if it is functional, its function is apparently not essential. Comparison of the three-dimensional structures of T4, T6, and T8 suggests that the three residues responsible for catalysis (His^{57}, Asp^{102}, and Ser^{195}) and the nature of the substrate binding pocket are all highly conserved (*28*). However, variations in surface charge and shape distributions suggest that, if T6 has a distinct expression pattern, it might interact differently with other protein components. Indeed, polymerase chain reaction (PCR) analyses of pancreas, thymus, and liver suggest that T6 may be expressed in minute amounts in the thymus (*29*). Other researchers have shown that trypsinogen is expressed in the brain (*27*). The roles of these trypsinogen isozymes in nonpancreatic tissues are unknown.

The intercalation of the trypsinogen genes in the β locus has been conserved in mouse (*30*) and chicken (*31*), which diverged from humans ~65 million and 350 million years ago, respectively. This long-term conservation of intercalated gene organization may reflect shared functional or regulatory constraints, as has been postulated for genes in the MHC (such as class I, II, and III genes) that share similar long-term organizational relations (*32*).

A β Locus Translocation from Chromosome 7 to Chromosome 9

The chromosomal location of the gene corresponding to the third pancreatic trypsinogen cDNA was investigated by fluorescence in situ hybridization with human metaphase chromosomes and a cosmid clone containing three trypsinogen genes. Strong hybridization to chromosome 7 and weaker hybridization to chromosome 9 were observed (*33*). We have isolated and partially sequenced four cosmid clones from the chromosome 9 region. This region represents a duplication and translocation of a DNA segment from the 3′ end of the β locus that includes at least seven V_β elements (*34*) and a functional trypsinogen gene denoted T9 (Fig. 1C). Three features of this translocated fragment are of interest: (i) Portions of a 10-kb trypsinogen homology unit constitute the 5′ and 3′ boundaries of the translocated DNA segment, and they lie in an inverted orientation with respect to one another. These sequences may reflect a DNA retroposon-like mechanism of duplication. (ii) The translocated chromosome 9 DNA segment is complex in structure, with non–chromosome 7 DNA interposed between duplicated chromosome 7 regions. Hence, the duplication and translocation mechanism is complex. (iii) On the basis of preliminary sequence data, the T9 gene appears to encode the third trypsinogen mRNA present in the pancreas. Therefore, the genetic elements for tissue-specific and developmentally specific expression are contained within the translocated homology unit.

This translocation event depicts a mechanism for the origin of a multigene family. Because the terminally inverted sequences that may have played a role in the translocation event did not encompass D, J, and C together with V elements, the V gene segments on chromosome 9 are pseudogenes. In this regard, if the inverted V elements at the 3′ ends of the α/δ and β TCR families were to mediate a translocation event, all of the requisite TCR elements would be included in the translocation, which may be how the antibody and TCR gene families all arose from a common ancestor.

In one sense, multigene families have no necessary boundaries because they can duplicate and translocate to other chromosomes. It is apparent that the three antibody gene families and the three TCR gene families arose from a single common ancestor. The basic immunoglobulin homology unit representing the fundamental evolutionary unit of the six immune receptor rearranging gene families has duplicated many times to give rise to the >200 distinct members of the immunoglobulin gene superfamily (*32*). Indeed, complex multigene families (such as olfactory receptor and MHC-like genes) rarely remain confined to a single chromosomal locus.

Dynamic History of the β TCR Locus

The dot-matrix analysis of the β locus sequence against itself reveals the extraordinary complexity of this region of the genome (Fig. 4). Approximately 47% of the β locus is composed of locus-specific repeats (homology units) that have been duplicated between 2 and 10 times and are indicated by the multiplicity of diagonal lines in the dot plot (Fig. 4). A detailed analysis of the dot matrix shows eight major locus-specific repeats scattered across the multigene family (homology units A to H) (Fig. 1D). Some of the homology units (such as G) are tandemly arrayed, and others (such as D) are dispersed. The eight homology unit groups contain, respectively, (A) V gene subfamilies 3 and 4; (B) V gene subfamily 5; (C) a V_β relic; (D) V gene subfamilies 6, 7, and 8; (E) V gene subfamilies 10, 11, and 12; (F) a V_β relic; (G) the five 3′ trypsinogen genes; and (H) the $D_\beta1J_\beta1C_\beta1$ and $D_\beta2J_\beta2C_\beta2$ segments. The sequence reveals that the individual genes or gene segments belonging to multimembered gene subfamilies (such as V subfamily 6 or trypsinogen genes T4 to T9) are, for the most part, embedded in long sequence repeats. Evolutionary flux in subfamily membership is the outcome of duplications and deletions of homology units. The length of the homology unit groups, based on a multisequence alignment of the individual repeat blocks, varies from 0.7 kb (B) to 31.1 kb (E). Some of the homology units (such as H) have diverged significantly, whereas others (such as F and G) remain very similar. Homology units A and E show an overall range of divergence of 82 to 94% and 74 to 93%, respectively. The blocks within the complex homology unit D containing V subfamilies 6, 7, and 8 show overall divergence rates of 62 to 98%, which suggests that this locus-specific repeat embodies both the oldest and the newest of the evolutionary changes in V subfamily membership brought about by duplications. Homology unit H has retained similarity only in the constant gene segments, in which gene conversion has likely occurred. Indeed, variation of the percentage divergence among the individual sequences within a homology unit suggests that, over time, gene conversion among the sequences has been a frequent event.

The wide range of sequence variation among the locus-specific repeats and the large differences in the number of copies of each repeat indicate the dynamic nature of the evolution of this locus. However, genome-wide interspersed repeats were rarely present at the boundaries of the homology units, thus eliminating obvious possible sites for the nucleation of nonhomologous recombination (unequal crossover).

Sequence Variations and Polymorphisms

Sequence variations generally arise by one of three mechanisms: (i) single-base substitutions, (ii) small insertions or deletions, or (iii) the duplication or deletion of one or more members of tandem arrays of simple or more complex repeats. Polymorphisms are sequence variations with a frequency of ≥1% in the population. A variety of single-base polymorphisms (for example, restriction fragment length polymorphisms) has been identified from cDNA and limited germline analyses (*35*). Each of the 26 cosmid clones sequenced (Fig. 1G), with the exception of the first and last, showed up to 21.5 kb of overlap with its two neighbors. When clones were derived from the same haplotype (chromosome), overlap differences provided an estimate of the rate of sequence errors (one error per 5 kb over 60 kb of overlap). When clones were derived from different haplotypes, the overlap differences provided an estimate of natural sequence variation (1 out of 474 nt, on

average, was affected by substitutions or small insertions over a total of 129 kb of overlap). The fraction of actual polymorphisms as opposed to infrequent mutations is unknown. In some overlap comparisons, the rate of sequence variation is low (1 nt in 9.5 kb), whereas in others it is very high (63 nt in 21 kb). These differences could arise from regional differences in the rates of variation or from comparing haplotypes (chromosomes) with markedly different ages of descent. Two large insertion-deletion polymorphisms that affect three V_β elements [V_β13S2 (6-2), V_β7S2 (4-3), and V_β9S2 (3-2)] and two trypsinogen genes (T6 and T7), respectively, exhibit allele frequencies of 0.37 (insertion) and 0.61 (deletion) for the V_β polymorphism and 0.54 (insertion) and 0.46 (deletion) for the trypsinogen polymorphisms, respectively (*36*). These high frequencies of deleted alleles might arise from founder effects or from some type of unknown selection (for example, loss of autoimmune tendencies with the V_β deletion). It is not clear why the loss of the functional trypsinogen T6 gene would confer a selective advantage.

Genome-Wide Repeats

Most interspersed repeats in the human genome are derived from four classes of transposable elements: (i) short interspersed nucleotide elements (SINEs) [Alu and mammalian-wide interspersed repeat (MIR) sequences] (*37*), (ii) long interspersed nucleotide elements (LINEs) (*38*), (iii) long terminal repeat (LTR) elements, including mammalian apparent LTR-retrotransposons (MaLRs) (*39*) and retroviral sequences (*40*), and (iv) DNA transposons (*41*). These sequences constitute 30% of the β TCR locus, with LINE1 (13.0%) and Alu sequences (4.9%) representing the largest contributors (Table 2). The high concentration of LINE1 and relatively low concentration of Alu is consistent with the location of the β locus at chromosome 7q35, a late-replicating "G band" that is theoretically LINE1-rich and Alu-poor (*42*). In addition to a large number of LINE1 elements, the β TCR locus contains a relatively high number of LTR elements. Among the 43 putative DNA transposon fossils is one full-length mariner element (*41*). Included in the unclassified repeats is a 3.2-kb region of mitochondrial DNA located between V_β15S1 (24-1) and V_β11S1 (25-1) (*43*). The four classes of genome-wide repeats are widely distributed across the entire locus (Fig. 1E), with the exception of the constant region (location, 640 to 660 kb), which is low in repeats.

Genome-wide repeats may serve important genomic functions. They may catalyze evolutionary change (expansion or contraction of gene families) through homologous but unequal crossing-over, or they may catalyze translocations or retroposon-like rearrangements through viral-like behavior. (In the β locus, they do not appear to catalyze unequal crossing-over.) Moreover, some may play a role in shifting the regulatory properties of single genes or batteries of genes. Several new unclassified repeats were discovered from the analysis of this sequence (Table 2) (*43*). As longer contiguous stretches of human DNA sequence (contigs) are produced, and as comparisons with other species (for example, mouse) become more extensive, the percentage of genomic DNA explainable by interspersed repeats is likely to increase, raising anew the functional DNA versus junk DNA debate.

Systems Analysis of the β TCR Family

The complete sequence of the β locus provides powerful information for studying the response of the entire β family of genes to developmental and antigenic signals. For example, PCR primers (from unique sequences in the V_β element and C_β gene) can be designed to investigate the changing concentrations of each V_β transcript during, for example, development, immunization, or tolerization. A systematic analysis of the β family response to these signals should provide valuable insights into how the β gene family, as a whole, functions.

Studies have suggested that polymorphisms in the human β TCR locus may correlate with autoimmune diseases (*44*).

Table 2. Distribution of genome-wide interspersed repeats. Genome-wide repeats were identified by comparing the βTCR sequence to a database of interspersed repeat consensus sequences (an expanded version of the publicly available REPBASE at ncbi.nlm.nih.gov) using cross_match, a modified Smith-Waterman comparison algorithm (*4*). Eighteen unusual interspersed repeats were discovered by comparing the screened locus with GenBank by means of the program BLASTN (*46*).

Type	Copies	Fraction of locus (%)
SINEs		
Alu	119	4.9
MIR	66	1.4
LINEs		
LINE1	126	13.0
LTR elements		
MaLRs	39	3.0
Retroviral	11	1.8
Other LTRs	8	1.3
DNA transposons	43	1.2
Unclassified	49	3.2
Total	461	29.8

Data obtained from different laboratories, however, are inconsistent, possibly because this locus behaves as if it has multiple hot spots of recombination, each isolating an island of a few V genes from the others. A genetic marker in each of the islands of V genes would be required to analyze comprehensively whether any V region polymorphism correlates with autoimmune disease. To this end, we have identified 29 simple sequence repeats (microsatellites) spread across the β locus, each of which has more than eight consecutive repeat units. Of these repeats, 19 are polymorphic in the Centre d'Etude du Polymorphisme Humaine (CEPH) families (*45*), covering all but ~15 V_β elements in the locus. Single-base polymorphisms will provide the remaining markers needed to investigate all of the islands of V_β elements. With a more complete set of markers, systematic genetic studies can be undertaken to correlate V_β polymorphisms with disease associations. Alternatively, unique PCR primers can be selected in the 5′ and 3′ flanking sequences for each functional V_β gene segment to identify gene polymorphisms and assess the role that they may play in human disease.

Summary

Our study has resulted in seven important observations: (i) The organization and structures of all of the human β TCR elements are characterized, and the 19 pseudogenes and 22 relics can be readily differentiated from the 46 functional genes by computational approaches. (ii) Comparative analyses of cDNA and genomic sequences have provided new insights into the identification of pseudogenes and the expression and diversification of this gene family. (iii) Eight trypsinogen genes are embedded in the locus, raising a question as to whether their association arises from functional or regulatory constraints, or is inadvertent. (iv) A portion of the β TCR locus has been duplicated and translocated from chromosome 7 to chromosome 9, which suggests a possible mechanism for the creation of new multigene families. (v) Eight locus-specific repeats (homology units) have duplicated in a tandem or dispersed manner, providing insights into the molecular archaeology of this locus. (vi) The extent of sequence variation (polymorphism) within the locus is high. (vii) At least 30% of this locus is composed of genome-wide interspersed repeats, but these do not appear to facilitate the duplication of locus-specific repeats.

The β TCR locus spans almost 685 kb, the longest contiguous stretch of DNA analyzed to date in humans. Sequence analysis of this locus has readily demonstrated the

power of large-scale DNA sequencing in delineating the varieties of information present in a complex multigene family. The ability to analyze the entire V_β response to various signals (tolerance, immunization) and the complex process of T cell development provides new opportunities to explore the biological complexities of immunity.

REFERENCES AND NOTES

1. L. Hood, J. Campbell, S. Elgin, *Annu. Rev. Genet.* **9**, 305 (1975).
2. For example, see R. Wilson *et al.*, *Nature* **368**, 32 (1994); R. D. Fleischmann *et al.*, *Science* **269**, 496 (1995).
3. M. D. Adams, C. Fields, J. Venter, Eds., *Automated DNA Sequencing and Analysis* (Academic Press, London, 1994).
4. P. Green, personal communication.
5. M. M. Davis and P. J. Bjorkman, *Nature* **334**, 395 (1988); J. L. Jorgensen, P. A. Reay, E. W. Ehrich, M. M. Davis, *Annu. Rev. Immunol.* **10**, 835 (1992).
6. M. R. Lieber, *FASEB J.* **5**, 2934 (1991).
7. For example, see P. Concannon, A. A. Pickering, P. Kung, L. Hood, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 6598 (1986); N. Kimura, B. Toyonaga, Y. Yoshikai, R.-P. Du, T. W. Mak, *Eur. J. Immunol.* **17**, 375 (1987); J. P. Tillinghast, M. A. Behlke, D. Y. Loh, *Science* **233**, 879 (1986).
8. J. L. Slightom, D. R. Siemieniak, L. C. Sieu, B. F. Koop, L. Hood, *Genomics* **20**, 149 (1994).
9. G. Siu, E. C. Strauss, E. Lai, L. E. Hood, *J. Exp. Med.* **164**, 1600 (1986); B. Toyonaga, Y. Yoshikai, V. Vadasz, B. Chin, T. W. Mak, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 8624 (1985); R. K. Wilson *et al.*, *Immunol. Rev.* **101**, 149 (1988); S. Wei, P. Charmley, M. A. Robinson, P. Concannon, *Immunogenetics* **40**, 27 (1994); T. M. Zhao, S. E. Whitaker, M. A. Robinson, *J. Exp. Med.* **180**, 1405 (1994).
10. Y. Li, P. Szabo, D. N. Posnett, *J. Exp. Med.* **174**, 1537 (1991).
11. Candidate cosmids encompassing the β TCR locus were identified by screening libraries constructed from yeast artificial chromosome (YAC) DNA or total genomic DNA with probes prepared from V gene segments, β locus–specific sequence-tagged sites, or human repetitive DNA (for cosmids derived from YACs). These cosmids were derived from the DNA of six different individuals (libraries), representing 12 possible distinct haplotypes. A physical map of 139 cosmid clones was constructed by a combination of Southern (DNA) blot hybridization, restriction digest analysis, and alignment of DNA sequence from the ends of cosmid inserts with completed sequence contigs (*33*). This cosmid array represents a sevenfold average coverage of the locus. The complete DNA sequence of the locus was obtained by applying a high-redundancy (eightfold coverage) shotgun method to 23 cosmids (*47*) and a primer-walking directed method to three cosmids (*8*). The sequence and its annotation are deposited in the Genome Sequence Data Base (accession numbers L36092, L36190, and U03115). On the basis of resequencing >60 kb of DNA, we estimate that the sequence contains, on average, one error per 5 kb. For specific details of the sequencing strategy (including chemistry, assembly of repeats, and precision estimates), see (*47*).
12. Non-TCR genes were identified by Xgrail (version 1.2) and BlastX. Similarity analyses were performed with the Smith-Waterman comparison algorithm on a Masspar Supercomputer (MpSRCH; Intelligenetics), with the Inherit analysis program (Applied Biosystems), or with cross_match (*4*).
13. B. Arden, S. P. Clark, D. Kabelitz, T. W. Mak, *Immunogenetics* **42**, 455 (1995).
14. Assuming that all V gene segment names begin with a TCRBV prefix, the correspondence between the proposed new nomenclature and that of Arden *et al.* (*13*) is: 1 (new gene: 27S1P); 2 (22S1A2N1T); 3-1 (9S1A1T); 3-2 (9S2A2PT); 4-1 (7S1A1N2T); 4-2 (7S3A2T); 4-3 (7S2A1N4T); 5-1 (5S1A1T); 5-2 (new gene: 31S1P); 5-3 (5S5P); 5-4 (5S6A3N2T); 5-5 (5S3A2T); 5-6 (5S2); 5-7 (5S7P); 5-8 (5S4A2T); 6-1 (13S3); 6-2 (13S2A1T; 13S2); 6-3 (13S2A1T; 13S9); 6-4 (13S5); 6-5 (13S1); 6-6 (13S6A2T); 6-7 (13S8P); 6-8 (13S7); 6-9 (13S4); 7-1 (6S7P); 7-2 (6S5A1N1); 7-3 (6S1A1N1); 7-4 (6S8A2T); 7-5 (6S9P); 7-6 (6S3A1N1T); 7-7 (6S6A2T); 7-8 (6S2A1N1T); 7-9 (6S4A1); 8-1 (new gene: 30S1P); 8-2 (new gene: 32S1P); 9 (1S1A1N1); 10-1 (12S2A1T); 10-2 (12S3); 10-3 (12S1A1N2); 11-1 (21S1); 11-2 (21S3A2N2T); 11-3 (21S2A2); 12-1 (8S4P); 12-2 (8S5P); 12-3 (8S1); 12-4 (8S2A1T); 12-5 (8S3); 13 (23S1A2T); 14 (16S1A1N1); 15 (24S1A3T); 16 (25S1A2PT); 17 (26S1P); 18 (18S1); 19 (17S1A1T); 20-1 (2S1A1); 21-1 (10S1P); 22-1 (new gene: 29S1P); 23-1 (19S1P); 24-1 (15S1); 25-1 (11S1A1T); 26 (new gene: 28S1P); 27 (14S1); 28 (3S1); 29-1 (4S1A1T); 30 (20S1A1N2). Shorthand versions of the Arden nomenclature are used throughout the text. For example, TCRBV4S1A1T (29-1) is referred to as V_β4S1 (29-1).
15. T. W. Mak and Y. Yanagi, *Immunol. Rev.* **81**, 221 (1984).
16. H. D. Royer and E. L. Reinherz, *Proc. Natl. Acad. Sci. U.S.A.* **84**, 232 (1987); S. J. Anderson, S. Miyake, D. Y. Loh, *Mol. Cell. Biol.* **9**, 4835 (1989).
17. B. F. Koop *et al.*, *Genomics* **19**, 478 (1994); J. E. Hesse, M. R. Lieber, K. Mizuuchi, M. Gellert, *Genes Dev.* **3**, 1053 (1989).
18. D. N. Posnett *et al.*, *J. Exp. Med.* **179**, 1707 (1994).
19. The data in GenBank should be given a qualitative rather than a quantitative interpretation, as much of it has been obtained by assaying specific antigenic responses. However, the generalized conclusion of differential expression of V_β genes holds true for the quantitative or statistical representation studies that have been performed. For additional expression data, see P. J. Doherty *et al.*, *Mol. Immunol.* **28**, 607 (1991); W. M. C. Rosenberg, P. A. H. Moss, J. I. Bell, *Eur. J. Immunol.* **22**, 541 (1992); M. A. Hall and J. S. Lanchbury, *Hum. Immunol.* **43**, 207 (1995).
20. U. Malhotra, R. Spielman, P. Concannon, *J. Immunol.* **149**, 1802 (1992).
21. R. Jores and T. Meo, *ibid.* **151**, 6110 (1993).
22. F. J. Bollum and L. M. S. Chang, *Adv. Cancer Res.* **47**, 37 (1986).
23. M. Levitt, personal communication.
24. M. D. Adams *et al.*, *Nature* **377** (suppl.), 3 (1995).
25. M. Emi *et al.*, *Gene* **41**, 305 (1986).
26. T. Tani, I. Kawashima, K. Mita, Y. Takiguchi, *Nucleic Acids Res.* **18**, 1631 (1990).
27. U. Wiegand, S. Corbach, A. Minn, J. Kang, B. Muller-Hill, *Gene* **136**, 167 (1993).
28. C. L. M. J. Verlinde and W. G. J. Hol, personal communication.
29. J. Roach, personal communication.
30. L. Rowen, unpublished data.
31. K. Wang, personal communication.
32. T. Hunkapiller and L. Hood, *Adv. Immunol.* **44**, 1 (1989).
33. For details of the construction of the map for the β TCR locus (for which fluorescence in situ hybridization was performed by H. Massa and B. Trask), see L. Rowen *et al.*, in preparation.
34. Six of the translocated V genes have been identified by M. A. Robinson *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 2433 (1993).
35. For review of coding region polymorphisms, see (*13*). For examples of restriction fragment length polymorphisms, see M. A. Robinson and T. J. Kindt, *Hum. Immunol.* **14**, 195 (1985); P. Concannon *et al.*, *Cold Spring Harbor Symp. Quant. Biol.* **51**, 785 (1986); P. Charmley, P. Concannon, R. A. Gatti, *Tissue Antigens* **35**, 157 (1990); P. Charmley, K. Wang, L. Hood, D. Nickerson, *J. Exp. Med.* **177**, 135 (1993); S. W. Funkhouser, P. Concannon, P. Charmley, D. L. Vredevoe, L. Hood, *Arthritis Rheum.* **35**, 465 (1992). The number of sequence variations obtained from our analysis of overlapping cosmids is too small to draw generalizations about the frequency of coding versus noncoding polymorphisms in the germline sequence.
36. E. Seboun, M. A. Robinson, T. J. Kindt, S. I. Hauser, *J. Exp. Med.* **170**, 1263 (1989).
37. P. L. Deininger and M. A. Batzer, *Evol. Biol.* **27**, 157 (1993); A. F. A. Smit and A. D. Riggs, *Nucleic Acids Res.* **23**, 98 (1995).
38. A. F. A. Smit, G. Toth, A. D. Riggs, J. Jurka, *J. Mol. Biol.* **246**, 401 (1995).
39. A. F. A. Smit, *Nucleic Acids Res.* **21**, 1863 (1993).
40. D. A. Wilkinson, D. L. Mager, J. A. C. Leong, in *The Retroviridae*, J. A. Levy, Ed. (Plenum, New York, 1994) vol. 3, pp. 465–535.
41. A. F. A. Smit and A. D. Riggs, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 1443 (1996).
42. G. P. Holmquist, *Am. J. Hum. Genet.* **51**, 17 (1992).
43. A. Smit, personal communication.
44. P. Concannon, *Manual of Clinical Immunology* (American Society for Microbiology, Washington, DC, 1992), pp. 885–889.
45. P. Charmley, P. Concannon, L. Hood, L. Rowen, *Genomics* **29**, 760 (1995).
46. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
47. L. Rowen *et al.*, in preparation.

48. We thank A. Smit for extensive analysis of genome-wide interspersed repeats in this sequence; P. Green for cross_match, a sequence comparison program; X. Huang for his multiple sequence alignment program (MAP); D. Gordon and T. Smith for computer scripts; K. Cattell for the MpSRCH analysis; Applied Biosystems and Masspar for allowing us to use early test software; technicians at the California Institute of Technology and the University of Washington for performing the DNA sequencing; P. Charmley for helpful discussion; and M. E. Ahearn for technical assistance. B.F.K. thanks Canadian Genome Analysis and Technology and the Natural Sciences and Engineering Research Council of Canada for their support. Supported by grants from the Department of Energy and NIH.