One-Megabase Sequence Analysis of the Human Immunoglobulin λ Gene Locus

Kazuhiko Kawasaki, Shinsei Minoshima, Eriko Nakato, Kazunori Shibuya, Ai Shintani, James L. Schmeits, Jun Wang, and Nobuyoshi Shimizu[1]

Department of Molecular Biology, Keio University School of Medicine, Shinjuku, Tokyo 160, Japan

A total of 1,025,415 bases of nucleotide sequence, including the entire human immunoglobulin λ gene locus has been determined. This is the largest contiguous human DNA sequence ever published. The sequence data revealed the organization of 36 potentially active V_λ gene segments, 33 pseudogene segments, and seven J_λ–C_λ gene segments. Among these 69 functional or nonfunctional V_λ gene segments, 32 were newly discovered. These V_λ gene segments are located within five gene-rich clusters and are divided into five clans based on sequence identity. Five potentially active nonimmunoglobulin genes were also detected within the λ gene locus, and two other genes were observed in the upstream region. Sequence organization suggests that large DNA duplications diversified the germ-line repertoire of the V_λ gene segments.

[The sequence information is available through the Advanced Lifescience Information Systems (ALIS) project World Wide Web site (http://www-alis.jst-c.go.jp) of Japan Science and Technology Corporation (JST), as well as DDBJ/GenBank databases (accession nos. D86989–D87024 and D88268–D88271).]

Immunoglobulin molecules are composed of light (L) and heavy (H) chains, each consisting of variable (*V*) and constant (*C*) regions (Tonegawa 1983). There are two types of light chains, κ and λ, each encoded by separate genes on different chromosomes (Lai et al. 1989). During B-cell development, germ-line V_L gene segments are juxtaposed to J_L gene segments and form mature V_L genes. In this *V–J* joining, a large number of distinct *V* gene segments contribute to the diversity of antigen recognition (Tonegawa 1983).

The V gene segment is composed of protein coding regions and *cis*-acting elements. The protein coding regions are divided into a signal peptide coding region and a *V* coding region (Kabat et al. 1991). These regions are encoded by two exons. The first exon encodes a large portion of the signal peptide, and the second exon encodes the carboxyl terminus of the signal peptide and the *V* region (Kabat et al. 1991). The *cis*-acting elements are composed of promoter motifs (Falkner and Zachau 1984), splice sites (Stephens and Schneider 1992), and a recombination signal sequence (RSS; Hesse et al. 1989). Two promoter motifs, an octamer (8-mer) and a TATA box precede the signal peptide coding region (Falkner and Zachau 1984). The *V* coding region is followed by the RSS, which is necessary for *V–J* joining (Hesse et al. 1989). The RSS is composed of three elements, a heptamer (7-mer), a spacer (23 nucleotides long in V_λ genes), and a nonamer (9-mer) (Hesse et al. 1989).

Previously, we isolated 176 cosmid clones and one bacterial artificial chromosome (BAC) clone, which cover the entire λ gene locus located on 22q11 (Kawasaki et al. 1995; Asakawa et al. 1997). Southern hybridization analysis of restriction fragments revealed that 69 V_λ and 7 C_λ DNA segments are located in a 911-kb region (Kawasaki et al. 1995). Using a set of yeast artificial chromosome (YAC) deletion and cosmid contigs, Frippiat et al. (1995) reported that V_λ gene probes hybridize to 55 DNA segments within the V_λ gene locus. Considering that half of these segments contain pseudogene segments as observed in V_H (Matsuda et al. 1993; Cook and Tomlinson 1995) and V_κ (Zachau 1995) loci, the total number of active V_λ gene segments had been estimated to be 30–35 (Kawasaki et al. 1995).

To study the λ gene locus comprehensively, we have completed >1 Mb of nucleotide sequence including the entire λ gene locus using 33 cosmid clones and one BAC clone. This study establishes the genomic organization of the λ gene locus uncovering the entire germ-line repertoire and phylogeny of the V_λ gene segments. The sequence data

[1] **Corresponding author.**
E-MAIL shimizu@dmb.med.keio.ac.jp; FAX 81-3-3351-2370.

7:250–261 ©1997 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/97 $5.00

obtained in this study will also give insight into the evolution of multigene families.

RESULTS

Sequence Data Evaluation and Allelic Variations

The complete 1,025,415-nucleotide sequence (DDBJ/GenBank accession nos. D86989–D87024 and D88268–D88271), including the entire germline λ gene locus, was determined by the shotgun sequencing method (Fig. 1). The sequenced data were 6.5- to 9.1-fold redundant, and ~90% of the sequenced region was determined in both directions (see Methods for details). To evaluate the reliability of the sequence data, we compared sequence differences appearing in a total of 39,236-nucleotide sequence of the two BAC–cosmid overlapping regions (35B9–288A10 and 288A10–50D10 in Fig. 1). These two cosmid clones and the BAC clone are derived from two different DNA sources: 35B9 and 50D10 are from Caucasian and 288A10 is from Japanese origin (Kawasaki et al. 1995). Nucleotide differences were detected at a total of 107 distinct sites. Strict inspections of wave patterns of these 107 sites enabled us to conclude that 100 sites are real differences that are derived from allelic variations reflecting the difference of DNA sources (Table 1) and that the remaining seven sites are apparently originated from sequence editing errors. This implies that the accuracy of our sequence data is >99.98% (7/39,236 nucleotides).

Interestingly, 73 of the 100 differences reside within a 10-kb *Alu*-rich region (3′ half of the 288A10–50D10 overlapping region in Fig. 1; 6,001–16,643 nucleotides in Table 1). These sequence variations are equally detectable within and outside of the *Alu* sequences (data not shown). Among the 73 allelic variations, 42 (58%) are transitions (nucleotide substitutions of a pyrimidine by a pyrimidine, or a purine by a purine), 22 (30%) are transversions (nucleotide substitutions of a pyrimidine by a purine, or a purine by a pyrimidine), and the remaining 9 (12%) are nucleotide insertions or deletions (data not shown). The frequencies of transversions, transitions, and insertions/deletions observed in this study are remarkably similar to those (62:23:14.6) observed in the factor IX gene (Sommer 1992). Thus, the accuracy of our sequence data is high enough to investigate the biological significance further.

Organization of the Immunoglobulin λ Gene Locus

To determine the locations of the V_λ gene segments, genomic sequences were searched for sequence homology against >130 entries of germ-line V_λ gene sequences in a GenBank database and V BASE (I.M. Tomlinson, pers. comm.). As a result, 36 potentially active V_λ gene segments, which maintain the open reading frame and essential amino acids (such as cysteine at amino acid residues 23 and 88) for immunoglobulin protein (Kabat et al. 1991), were identified. Among these 36 V_λ gene segments, 10 were newly discovered in this study (Fig. 2). These newly discovered V_λ gene segments show <97% sequence identity to any known germ-line V_λ gene segments (allelic variations account for 1%–3% sequence difference). A total of 33 pseudogene segments, which contain all essential V_λ gene elements but are disrupted by frameshift and/or stop codon caused by point mutation, small insertion, or deletion, were also detected (Fig. 1). Among these 33 pseudogene segments, 22 were newly identified. In addition, 34 V_λ relics, which have large deletions or insertions, were detectable. Because of the difficulty of aligning these severely disrupted gene sequences to complete V_λ gene segments, all relics were eliminated from this study. All of these V_λ gene segments, including pseudogene segments and relics, are clustered in five regions, I–V (Kawasaki et al. 1995; Fig. 1) and have the same transcriptional polarity as for the J_λ and C_λ gene segments.

Three *Alu*-rich regions were identified (Fig. 1). These *Alu* clusters reside in an intervening region II (itv-region II) between regions II and III, in an upstream half of itv-region III between regions III and IV, and immediately upstream of region V. Minisatellite clusters and α satellite clusters (Vogt 1990) were also identified within itv-region III and itv-region IV, respectively.

Phylogeny of the V_λ Gene Segments

For the 36 potentially active V_λ gene and 33 pseudogene segments, nucleotide sequences were multiple-aligned and a phylogenetic tree was constructed (Fig. 3). As a result, all of these V_λ gene and pseudogene segments are integrated into five distinct groups, which we designate clan 1–clan 5 (see Kirkham et al. 1992 for V_H clans). A clan represents a family or a group of families that were originally defined on the basis of amino acid sequences. For example, clan 1 contains families I, II, VI (Chuchana et al. 1990), and X (Stiernholm and Berinstein 1995). Clan 2 is identical to family III; however, clan 2 is highly diverged and is obviously divided into three different subfamilies, III-1, III-2, and III-3 (Fig. 3). These subfamilies are different

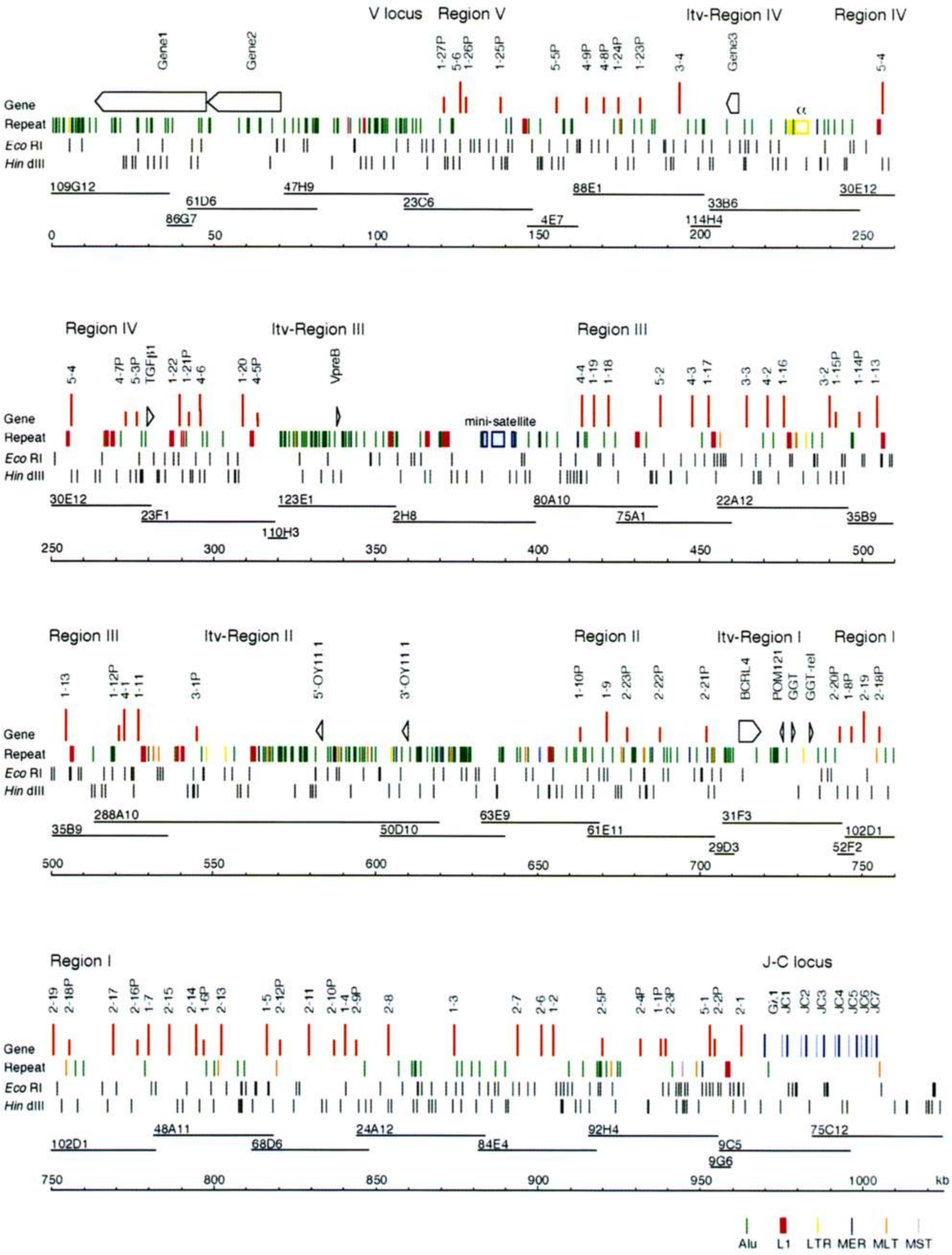


Figure 1 (*See facing page for legend.*)

Table 1. Nucleotide Sequence Variations Detected within BAC–Cosmid Overlapping Regions

35B9/288A10 (no. of nucleotides)	Variations	288A10/50D10 (no. of nucleotides)	Variations
1–2,000	1	1–2,000	7
2,001–4,000	1	2,001–4,000	1
4,001–6,000	1	4,001–6,000	2
6,001–8,000	2	6,001–8,000	15
8,001–10,000	1	8,001–10,000	12
10,001–12,000	5	10,001–12,000	16
12,001–14,000	0	12,001–14,000	12
14,001–16,000	0	14,001–16,000	14
16,001–18,000	3	16,001–16,643	4
18,001–20,000	2		
20,001–22,593	1		
Total	17		83

The 5′ end of each overlapping region (35B9/288A10 and 50D10/288A10) is designated nucleotide 1. The length of the overlapping regions are 22,593 nucleotides for 35B9/288A10 and 16,643 nucleotides for 50D10/288A10, respectively.

from subfamilies IIIa, IIIb, and IIIc, which were determined previously by a serological method (Eulitz et al. 1991). Subfamily III-2 includes both IIIa and IIIc, and subfamily III-1 includes IIIb. The nucleotide sequence of V_λ*4-2* shows 92% identity to a cDNA sequence T1 (Berinstein et al. 1989), which is the sole member of family V (Chuchana et al. 1990). Because three V_λ gene segments, *4-1, 4-2,* and *4-3,* are all closely related, we included these three gene segments in family V. The relationships among clans, families, V_λ gene segments, and locations are summarized (Table 2). All previously identified V_λ families, I–X (Chuchana et al. 1990; Winkler et al. 1992; Deftos et al. 1994; Stiernholm and Berinstein 1995) are identified within clan 1–clan 5.

Expression of the V_λ Gene Segments

Despite an extensive search of GenBank and dbEST (Boguski et al. 1993) databases, 8 of the 36 potentially active V_λ gene segments (*1-9, 1-18, 2-8, 2-15, 4-1, 4-3, 4-6,* and *5-1*) failed to detect cDNA sequences with high sequence identity [≥95%; allelic variations and subsequent hypermutations account for 1%–5% sequence difference (Ch'ang et al. 1994)] (Fig. 4). Considering the large sizes of these databases (>460 entries for rearranged or expressed V_λ genes), lack of cDNA sequences corresponding to these eight V_λ gene segments could be attributable to less active transcription compared to the other V_λ gene segments.

To find a relationship between V_λ gene activity and the *cis*-acting elements that are essential for producing functional immunoglobulin molecules, promoter motifs (Falkner and Zachau 1984), splice sites (Stephens and Schneider 1992), and RSSs (Hesse et al. 1989) are aligned (Fig. 4). The nucleotide sequences of the 8-mer (Falkner and Zachau 1984) and the 7-mer (Hesse et al. 1989) are well conserved. In contrast, the TATA box (Falkner and Zachau 1984) and the 9-mer (Hesse et al. 1989) are rather variable. The consensus sequences of splice donor and acceptor sites (GT and AG) are all perfectly conserved. Three nucleotides 5′ to the splice donors (CAG), as well as two nucleotides adjacent to the acceptors (5′ C and 3′ G), are also highly conserved (Stephens and Schneider 1992). Interestingly, variable sequences within these *cis*-acting elements, as well as the length of the introns, are significantly conserved within families and clans.

Evolution of the λ Gene Locus

The dot matrix analysis (Sonnhammer and Durbin

Figure 1 A comprehensive map of the immunoglobulin λ gene locus. The first row of each group shows names and locations of genes and pseudogenes. The V_λ gene segments are indicated by red lines (half-height lines represent pseudogene segments), and J_λ–C_λ gene segments as well as the Gλ1 gene (Evans and Hollis 1991) are indicated by lavender lines. Nonimmunoglobulin genes and pseudogenes are shown by open triangles or pentagonal boxes, representing the transcriptional polarity. The second row shows the locations of repetitive elements detected by XGRAIL (REPBASE): *Alu* (green), L1 (magenta), LTR (yellow), MER (blue), MLT (orange), and MST (mauve). Minisatellite (Vogt 1990) clusters and α-satellite (Vogt 1990) clusters are indicated by blue and yellow shaded boxes, respectively. The third and fourth rows show the locations of *Eco*RI and *Hin*dIII sites, respectively. The fifth row, above the scale, shows sequenced regions of 33 cosmid clones and a BAC clone (288A10).

1996) of the entire λ gene locus versus itself revealed that regions I and III have large internal duplications (Fig. 5). Broken lines in the dot plot show that the duplicated regions underwent a large number of deletions and/or insertions after duplicating. In region III, four lines, each separated by ~25 kb, are detectable in parallel with the diagonal line, suggesting that five large amplification units are tandemly repeated (Kawasaki et al. 1995). An ~10-kb duplication is also detectable between regions I and II. Smaller duplications (4–6 kb) dispersed throughout the locus represent L1 repeats (Smit et al. 1995). Small duplications corresponding to each J_λ–C_λ unit are detectable in the J_λ–C_λ gene locus (Vasicek and Leder 1990). The dot matrix analysis also shows that there are no obvious short interspersed elements (SINES), such as *Alu* repeats (Batzer et al. 1996) in the vicinity of the J_λ–C_λ gene locus (Vasicek and Leder 1990), whereas a large number of SINES are located in itv-regions II, the upstream half of itv-region III, and upstream of region V.

Other Genes and Pseudogenes within the λ Gene Locus

Two λ-related genes, the VpreB gene (Kawasaki et al. 1995; Frippiat et al. 1995) and the Gλ1 gene (Evans and Hollis 1991), are located within the λ gene locus. The VpreB gene is localized within an *Alu* cluster in itv-region III, and the Gλ1 gene is located upstream of the J_λ–C_λ gene locus (Evans and Hollis 1991; Fig. 1). In addition to these genes, XGRAIL (Uberbacher and Mural 1991) analysis identified five putative genes (labeled Gene1, Gene2, Gene3, 5′-OY11.1, and 3′-OY11.1). A homology search of dbEST of these five sequences showed several different expressed sequence tags (ESTs; Boguski et al. 1993) with high sequence identity (>98%; because of ambiguous sequences in the 3′ end of ESTs, sequence identity is often <100%), except for 3′-OY11.1. Gene1 is identical to a 5.1-kb cDNA sequence (GenBank accession no. D13640), and the deduced amino acid sequence has a protein phosphatase 2C motif (PROSITE accession no. PDOC00792; Wenk et al. 1992). Gene2 has moderate sequence homology to human (46%) and yeast (43%) topoisomerase III (GenBank accession nos. U43431 and M24939; Hanai et al. 1996). Gene3 does not show high sequence homology to any known genes. Two OY11.1 (sheep putative gene, GenBank accession no. U30307)-like sequences (5′-OY11.1 and 3′-OY11.1) were detected in itv-region II. Two ESTs were identified for 5′-OY11.1, but there was no EST corresponding to 3′-OY11.1.

A breakpoint cluster region gene (BCR) (GenBank accession no. Y00661)-like sequence (BCRL4; Frippiat et al. 1995; Kawasaki et al. 1995), a POM121 (GenBank accession no. Z21513; Hallberg et al. 1993)-like sequence, a γ-glutamyl transpeptidase (GGT, GenBank accession no. J04131)-like sequence (Frippiat et al. 1995; Kawasaki et al. 1995), and a GGT-related gene (GenBank accession no. M64099; Heisterkamp et al. 1991)-like sequence were all detected within itv-region I. A 5′-truncated transforming growth factor-β1 (TGF-β1) (GenBank accession no. M60315; Celeste et al. 1990)-like sequence was identified within region IV. These five sequences possess only a part of the original genes (BCR, the GGT-related gene, and TGF-β1), a frameshift (POM121-like sequence), or a defect in one of the splice sites (GGT-like sequence). Accordingly, these sequences would not contribute to the protein-encoding function of this locus.

DISCUSSION

The complete 1,025,415-nucleotide sequence, including the entire human immunoglobulin λ gene locus has been determined with >99.98% sequence accuracy. This sequence is larger than the T-cell receptor β (TCRβ) locus (685 kb), which, until this study, was the largest known human contiguous nucleotide sequence (Rowen et al. 1996). Within the λ gene locus, 36 potentially active V_λ gene segments have been identified. It had been difficult to unambiguously identify a germ-line sequence of a given V_λ sequence in the databases because of the allelic sequence variations or polymorphisms in germ-line immunoglobulin V_λ gene segments and because of the high sequence similarity among individual nonallelic V_λ gene segments. Because this study revealed the entire germ-line repertoire of the V_λ gene segments, these 36 V_λ gene sequences provided us with a standard to assign any unidentified V_λ gene sequences. This standard will be invaluable to examine how each V_λ gene segment participates in immunological responses.

In addition to 36 potentially active V_λ gene segments, 33 pseudogene segments were also identified, and a phylogenetic tree was constructed from the nucleotide sequences of these 69 V_λ gene segments (Fig. 3). This tree is substantiated by a phylogenetic tree generated with amino acid sequences (data not shown). Based on the tree, all 69 V_λ gene and pseudogene segments are divided into five clans, and these five clans are divided further into 10 or more families. We have subdivided (subfamilies III-1, III-2, and III-3) or extended (family V) the

Vλ	Signal peptide	FR1	CDR1	FR2	CDR2	FR3	CDR3	reference
Clan 1								
1-2	MAWALLLLTLLTQGTGSWA	QSALTQPPSASGSPGQSVTISC	TGTSSDVGGYNYVS	WYQQHPGKAPKLMIY	EVSKRPS	GVPDRFSGSKSGNTASLTVSGLQAEDEADYYC	SSYAGSNNF	L27695 (99%)
1-3	MAWALLLLSLLTQGTGSWA	QSALTQPRSVSGSPGQSVTISC	TGTSSDVGGYNYVS	WYQQHPGKAPKLMIY	DVSKRPS	GVPDRFSGSKSGNTASLTISGLQAEDEADYYC	CSYAGSYTF	Z22198 (99%)
1-4	MAWALLLLTLLTQGTGSWA	QSALTQPASVSGSPGQSITISC	TGTSSDVGGYNYVS	WYQQHPGKAPKLMIY	EVSNRPS	GVSNRFSGSKSGNTASLTISGLQAEDEADYYC	SSYTSSSTL	L27693 (100%)
1-5	MAWALLLLTLLTQGTGSWA	QSALTQPPSVSGSPGQSVTISC	TGTSSDVGSYNRVS	WYQQPPGTAPKLMIY	EVSNRPS	GVPDRFSGSKSGNTASLTISGLQAEDEADYYC	SLYTSSSTF	L27689 (100%)
1-7	MAWALLLLTLLTQDTGSWA	QSALTQPASVSGSPGQSITISC	TGTSSDVGSYNLVS	WYQQHPGKAPKLMIY	EGSKRPS	GVSNRFSGSKSGNTASLTISGLQAEDEADYYC	CSYAGSSTF	X14616 (100%)
1-9	MAWALLLLTLLTQGTGSWA	QSALTQPPFVSGAPGQSVTISC	TGTSSDVGDYDHVF	WYQKRLSTTSRLLIY	NVNTRPS	GISDLFSGSKSGNMASLTISGLKSEVEANYHC	SLYSSSYTF	L27687 (100%)
1-11	MAWSPLFLTLITHCAGSWA	QSVLTQPPSVSEAPRQRVTISC	SGSSSNIGNNAVN	WYQQLPGKAPKLLIY	YDDLLPS	GVSDRFSGSKSGTSASLAISGLQSEDEADYYC	AAWDDSLNG	U03900 (99%)
1-13	MAWSPLLLTLLAHCTGSWA	QSVLTQPPSVSGAPGQRVTISC	TGSSSNIGAGYDVH	WYQQLPGTAPKLLIY	GNSNRPS	GVPDRFSGSKSGTSASLAITGLQAEDEADYYC	QSYDSSLSG	M94116 (100%)
1-16	MASFPLLLTLLTHCAGSWA	QSVLTQPPSASGTPGQRVTISC	SGSSSNIGSNTVN	WYQQLPGTAPKLLIY	SNNQRPS	GVPDRFSGSKSGTSASLAISGLQSEDEADYYC	AAWDDSLNG	X59707 (99%)
1-17	MAGFPLLLTLLTHCAGSWA	QSVLTQPPSASGTPGQRVTISC	SGSSSNIGSNYVY	WYQQLPGTAPKLLIY	SNNQRPS	GVPDRFSGSKSGTSASLAISGLRSEDEADYYC	AAWDDSLSG	M94114 (100%)
1-18	MAWSSLLLTLLAHCTGSWA	QSVLTQPPSVSGAPGQRVTISC	TGSSSNIGAGYVVH	WYQQLPGTAPKLLIY	GNSNRPS	GVPDQFSGSKSGTSASLAITGLQSEDEADYYC	KAWDNSLNA	M94112 (100%)
1-19	MTCSPLLLTLLIHCTGSWA	QSVLTQPPSVSAAPGQKVTISC	SGSSSNIGNNYVS	WYQQLPGTAPKLLIY	DNNKRPS	GIPDRFSGSKSGTSATLGITGLQTGDEADYYC	GTWDSSLSA	U03870 (100%)
1-20	MPWALLLLTLLTHSAVSVV	QAGLTQPPSVSKGLRQTATLTC	TGNSNIVGNQGAA	WLQQHQGHPPKLLSY	RNNNRPS	GISERFSASRSGNTASLTITGLQPEDEADYYC	SALDSSLSA	gVLX-4.4 (98%)
1-22	MAWAPLLLTLLAHCTGSWA	NFMLTQPHSVSESPGKTVTISC	TRSSGSIASNYVQ	WYQQRPGSSPTTVIY	EDNQRPS	GVPDRFSGSIDSSSNSASLTISGLKTEDEADYYC	QSYDSSN	M87320 (100%)
Clan 2								
2-1	MAWIPLFLGVLAYCTGSVA	SYELTQPPSVSVSPGQTASITC	SGDKLGDKYAC	WYQQKPGQSPVLVIY	QDSKRPS	GIPERFSGSNSGNTATLTISGTQAMDEADYYC	QAWDSSTA	X57826 (100%)
2-6	MAWTALLLSLLAHFTGSVA	SYELTQPLSVSVALGQTARITC	GGNNIGSKNVH	WYQQKPGQAPVLVIY	RDSNRPS	GIPERFSGSNSGNTATLTISRAQAGDEADYYC	QVWDSSTA	X84947 (100%)
2-7	MAWTPLLLPLLTFCTVSEA	SYELTQPPSVSVSPGQTARITC	SGDALPKKYAY	WYQQKSGQAPVLVIY	EDSKRPS	GIPERFSGSSSGTMATLTISGAQVEDEADYYC	YSTDSSGNH	L26403 (82%)
2-8	MAWTPLLLSLLAHCTGSAT	SYELTQPHSVSVATAQMARITC	GGNNIGSKAVH	WYQQKPGQDPVLVIY	SDSNRPS	GIPERFSGSNPGNTATLTISRIEAGDEADYYC	QVWDSSSDH	X02671 (93%)
2-11	MAWIPLLLPLLTLCTGSEA	SYELTQPPSVSVSLGQMARITC	SGEALPKKYAY	WYQQKPGQFPVLVIY	KDSERPS	GIPERFSGSSSGTIVTLTISGVQAEDEADYYC	LSADSSGTY	S57178 (85%)
2-13	MAWTPLWLTLLTLCIGSVV	SSELTQDPAVSVALGQTVRITC	QGDSLRSYYAS	WYQQKPGQAPVLVIY	GKNNRPS	GIPDRFSGSSSGNTASLTITGAQAEDEADYYC	NSRDSSGNH	X56178 (100%)
2-14	MAWTVLLLGLLSHCTGSVT	SYVLTQPPSVSVAPGQTARITC	GGNNIGSKSVH	WYQQKPGQAPVLVVY	DDSDRPS	GIPERFSGSNSGNTATLTISRVEAGDEADYYC	QVWDSSSDH	M94115 (99%)
2-15	MAWATLLLPLLNLYTGSVA	SYELTQLPSVSVSPGQTARITC	SGDVLGENYAD	WYQQKPGQAPELVIY	EDSERYP	GIPERFSGSTSGNTTTLTISRVLTEDEADYYC	LSGDEDN	X71967 (97%)
2-17	MAWIPLLLPLFTLCTGSEA	SYELTQPPSVSVSPGQTARITC	SGDALPKQYAY	WYQQKPGQAPVLVIY	KDSERPS	GIPERFSGSSSGTTVTLTISGVQAEDEADYYC	QSADSSGTY	S57178 (88%)
2-19	MAWIPLLLPLLILCTVSVA	SYELTQPSSVSVSPGQTARITC	SGDVLAKKYAR	WFQQKPGQAPVLVIY	KDSERPS	GIPERFSGSSSGTTVTLTISGAQVEDEADYYC	YSAADNN	L26403 (85%)
Clan 3								
3-2	MAWTPLFLFLLTCCPGSNS	QTVVTQEPSLTVSPGGTVTLTC	ASSTGAVTSGYYPN	WFQQKPGQAPRALIY	STSNKHS	WTPARFSGSLLGGKAALTLSGVQPEDEAEYYC	LLYYGGAQ	X14614 (100%)
3-3	MAWTPLFLFLLTCCPGSNS	QAVVTQEPSLTVSPGGTVTLTC	GSSTGAVTSGHYPY	WFQQKPGQAPRTLIY	DTSNKHS	WTPARFSGSLLGGKAALTLLGAQPEDEAEYYC	LLSYSGAR	Z22205 (99%)
3-4	MAWMMLLLGLLAYGSGVDS	QTVVTQEPSFSVSPGGTVTLTC	GLSSGSVSTSYYPS	WYQQTPGQAPRTLIY	STNTRSS	GVPDRFSGSILGNKAALTITGAQADDESDYYC	VLYMGSGI	Z22206 (100%)
Clan 4								
4-1	MAWTPLLLLLLSHCTGSLS	QPVLTQPPSSSASPGESARLTC	TLPSDINVGSYNIY	WYQQKPGSPPRYLLY	YYSDSDK	GQGSGVPSRFSGSKDASANTGILLISGLQSEDEADYYC	MIWPSNAS	L27695 (68%)
4-2	MAWTPLLLLFLSHCTGSLS	QAVLTQPSSLSASPGASASLTC	TLRSGINVGTYRIY	WYQQKPGSPPQYLLRY	KSDSDKQQGS	GVPSRFSGSKDASANAGILLISGLQSEDEADYYC	MIWHSSAS	Z22199 (67%)
4-3	MAWNPLLLLFLSHCTGSLS	QPVLTQPTSLSASPGASARLTC	TLRSGINLGSYRIF	WYQQKPESPPRYLLSY	YSDSSKHQGS	GVPSRFSGSKDASSNAGILVISGLQSEDEADYYC	MIWHSSAS	Z22199 (66%)
4-4	MAWTLLLLVLLSHCTGSLS	QPVLTQPSSHSASSGASVRLTC	MLSSGFSVGDFWIR	WYQQKPGNPPRYLLY	YHSDSNKGQGS	GVPSRFSGSNDASANAGILRISGLQPEDEADYYC	GTWHSNSKT	Z22197 (56%)
4-6	MALTPLLLLLLSHCTGSLS	RPVLTQPPSLSASPGATARLPC	TLSSDLSVGGKNMF	WYQQKPGSSPRLFLY	HYSDSDKQL	GPGVPSRVSGSKETSSNTAFLLISGLQPEDEADYYC	QVYESSAN	Z22194 (65%)
Clan 5								
5-1	MAWVSFYLLPFIFSTGLCA	LPVLTQPPSASALLGASIKLTC	TLSSEHSTYTIE	WYQQRPGRSPQYIMKV	KSDGSHSKGD	GIPDRFMGSSSGADRYLTFSNLQSDDEAEYHC	GESHTIDGQVG*	Z22211 (100%)
5-2	MAWAPLLLTLLSLLTGSLS	QPVLTQPPSASASLGASVTLTC	TLSSGYSNYKVD	WYQQRPGKGPRFVMRV	GTGGIVGSKGD	GIPDRFSVLGSGLNRYLTIKNIQEEDESDYHC	GADHGSGNNFV*	X92339 (99%)
5-4	MAWTPLLLLFPLLLHCTGSLS	QPVLTQSSSASASLGSSVKLTC	TLSSGHSSYIIA	WHQQQPGKAPRYLMKLE	GSGSYNKGS	GVPDRFSGSSSGADRYLTISNLQFEDEADYYC	ETWDSNT	U03868 (88%)
5-6	MAWTPLLFLTLLLHCTGSLS	QLVLTQSPSASASLGASVKLTC	TLSSGHSSYAIA	WHQQQPEKGPRYLMKLNSD	GSHSKGD	GIPDRFSGSSSGAERYLTISSLQSEDEADYYC	QTWGTGI	L29806 (100%)

Figure 2 Alignment of deduced amino acid sequences of 36 potentially active V_λ gene segments. Each of these gene segments was named based on sequence identity (clan; designated by the phylogenetic tree in Fig. 3) and the location (numbered from V_λ gene segment proximal to the J_λ–C_λ locus), such as (clan)–(location number). Sequences are aligned for seven domains: a signal peptide, framework regions (FR1–FR3), and complementary determining regions (CDR1–CDR3) according to Kabat et al. (1991). Asterisks at the end of CDR3 for V_λ*5-1* and V_λ*5-2* indicate stop codons. These two genes can be active only if these stop codons are removed by *V–J* joining. (Reference) The GenBank accession number of the germ-line V_λ gene segments with the highest sequence identity. No sequences were detected with >85% sequence identity to V_λ*1-20* in the GenBank; however this gene has 98% sequence identity to gVLX-4.4 (Stiernholm and Berinstein 1995, in V BASE).

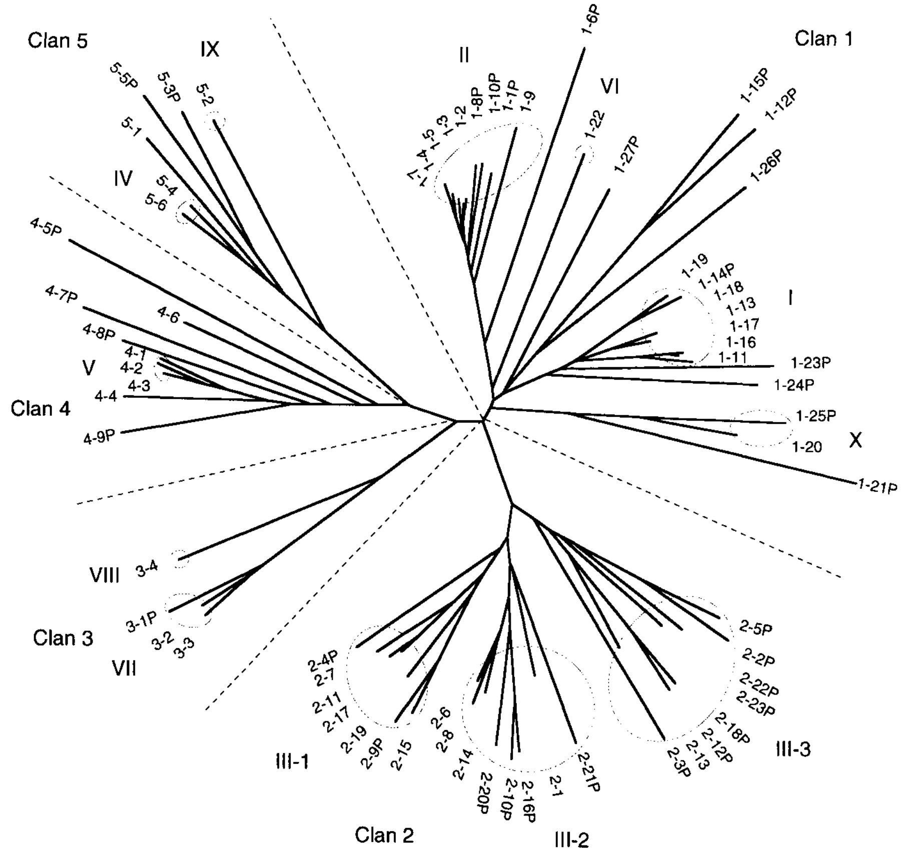


Figure 3 A phylogenetic tree of 69 V_λ gene and pseudogene segments. Nucleotide sequences of the *V* coding region (FR1–CDR3) were aligned and a phylogenetic tree was constructed by the neighbor-joining method (Saitou and Nei 1987). Five distinct groups (separated by broken lines) have been assigned the names clan 1–clan 5. Names of pseudogene segments are indicated with a P at the end. The V_λ families, I, II, and IV–X (Chuchana et al. 1990; Winkler et al. 1992; Deftos et al. 1994; Stiernholm and Berinstein 1995) are circled, and the family names are indicated by Roman numerals. Subfamilies, III-1, III-2, and III-3 are surrounded by broken circles.

previously assigned V_λ families, to illustrate unambiguously the diversity among families.

The three largest V_λ gene families, I–III, are clustered in both regions I and II (family II and III), or region III (family I) (Table 2). The dot matrix analysis shows that regions I and III have large internal duplications (Fig. 5). Large duplications are also detectable between regions I and II. In addition, the V_λ gene segments in regions I and II have small sequence variations and mostly belong to either families II or III (except 1-6P and 5-1 in region I). The high sequence identity between regions I and II suggests that regions I and II have a common ancestor. All of these observations concerning the large duplications and the clustering of the V_λ families suggest that families I, II, and III have been generated by large duplications. Similar, though less complex, duplications are also observed in the TCRβ locus (Rowen et al. 1996), suggesting that large duplications commonly contribute to the generation of diversities among multigene families. Regions IV and V are different in that the V_λ gene segments are highly diverged, and no large duplications other than L1 repeats (Smit et al. 1995) are detected. This suggests that these two regions are devoid of recent dynamic amplifications.

By searching GenBank and dbEST databases, cDNAs for 28 of the 36 potentially active V_λ gene segments were identified; however, the remaining eight V_λ gene segments failed to detect any cDNAs with ⩾95% sequence identity (Fig. 4). This suggests that these eight V_λ gene segments are transcriptionally less active than the other V_λ gene segments. The V_λ*4-6* gene segment has a 4-nucleotide short spacer in the RSS (Fig. 4). This defect presumably renders this gene segment inactive (Hesse et al. 1989). However, the reason for low transcriptional activity of the remaining seven gene segments is unclear. Obviously, V_λ*4-1* has no harmful substitutions in the promoter motifs, splice sites, and the RSS, yet neither rearranged nor expressed V_λ sequence was found among >460 entries in the databases. A large number of active gene segments have sequence substitutions in their 9-mer signals, especially at the third, fourth, and fifth nucleotides (Fig. 4). Apparently, these substitutions do not affect gene activity. V_λ*1-20* has two unique substitutions in the 8-mer promoter; however, this gene segment is shown to be slightly active: ~100 times lower than active gene segments (Stiernholm and Berinstein 1995). Interestingly, substitutions in the *cis*-acting elements (Fig. 4) and introns (data not shown) are well conserved within families or clans as observed in V_H (Matsuda et al. 1993) and V_κ

Table 2. Relationship among Clans, Families, V_λ Gene Segments, and Location

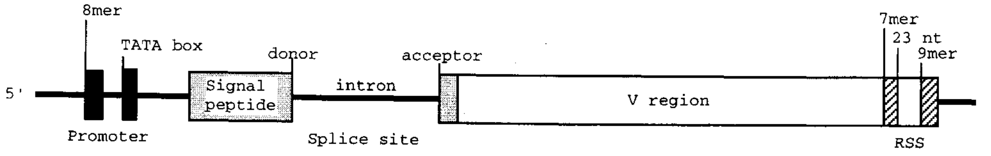
Clan	Family	V_λ gene segments	Region
1	I	*1-11, 1-13, 1-14P, 1-16, 1-17, 1-18, 1-19*	III
1	II	*1-1P, 1-2, 1-3, 1-4, 1-5, 1-7, 1-8P*	I
1	II	*1-9, 1-10P*	II
1	VI	*1-22*	IV
1	X	*1-20*	IV
1	X	*1-25P*	V
1	—	*1-6P*	I
1	—	*1-12P, 1-15P*	III
1	—	*1-21P*	IV
1	—	*1-23P, 1-24P, 1-26P, 1-27P*	V
2	III-1	*2-4P, 2-7, 2-9P, 2-11, 2-15, 2-17, 2-19*	I
2	III-2	*2-1, 2-6, 2-8, 2-10P, 2-14, 2-16P, 2-20P*	I
2	III-3	*2-2P, 2-3P, 2-5P, 2-12P, 2-13, 2-18P*	I
2	III-2	*2-21P*	II
2	III-3	*2-22P, 2-23P*	II
3	VII	*3-1P, 3-2, 3-3*	III
3	VIII	*3-4*	V
4	V	*4-1, 4-2, 4-3*	III
4	—	*4-4*	III
4	—	*4-5P, 4-6, 4-7P*	IV
4	—	*4-8P, 4-9P*	V
5	IV	*5-4*	IV
5	IV	*5-6*	V
5	IX	*5-2*	III
5	—	*5-1*	I
5	—	*5-3P*	IV
5	—	*5-5P*	V

Dashes show unassigned families. The V_λ families were assigned previously (see Chuchana et al. 1990 for families I, II, IV, V, VI, VII; Stiernholm and Berinstein 1995 for family X; Winkler et al. 1992 for family VIII; Deftos et al. 1994 for family IX) or in this study (III-1, III-2, and III-3). Three V_λ gene segments, *4-1, 4-2,* and *4-3*, are assigned to family V in this study. The letter P next to a gene name indicates a pseudogene.

(Zachau 1995) loci. Quantitative analysis of expressed V_λ gene segments, coupled with an analysis of allelic variations of the promoter and RSS sequences, is required further to elucidate the relationship between gene activity and the variations of *cis*-acting elements.

The RSS is also required for J_λ gene segments upon *V–J* joining. The RSS of the J_λ gene segments is composed of a 9-mer signal, a 12-nucleotide spacer, and a 7-mer signal (Hesse et al. 1989). The consensus sequences of the 9-mer (5′-GGTTTTTGT-3′) and 7-mer (5′-CACAGTG-3′) in the J_λ gene segments are complementary to the 9-mer and 7-mer signals in the V_λ gene segments. As shown in Figure 4, the 5′ half of the consensus 23-nucleotide spacer sequence of the V_λ gene segments (5′-acaCAgGCagAT-3′; where lowercase letters indicate less-conserved nucleotides) are also complementary to the consensus sequence of the 12-nucleotide spacer of the J_λ gene segments (5′-ATGAGCCTGTGT-3′), although these are with less fidelity. This complementary nature of the sequence between the 23- and the 12-nucleotide spacer may aid the *V–J* recombination by stabilizing the secondary structure of RSS regions. However, this complementary nature was not detected in the V_κ and J_κ gene segments (GenBank accession no. J00242). Because the available nucleotide sequence of spacer region is limited, studies on immunoglobulin loci of various other organisms will be necessary to clarify the significance of the complementary nature of the sequence.

Of 372 *Alu* repeats (Batzer et al. 1996) detected in this study, 31, 30, and 56 are clustered in the region immediately upstream of region V (42 kb), in an upstream half of itv-region III (37 kb), and in an intermediate region of itv-region II (60 kb), respectively. It is interesting to note that all of these *Alu* clusters occur outside of the V_λ-gene rich, highly duplicated regions. Dense *Alu* clusters have been reported also in the human leukocyte antigen (HLA) class III locus (Iris et al. 1993). Further large-scale sequencing will clarify whether or not this pattern of *Alu* clustering is common throughout the human genome. It is of particular interest that the VpreB gene and two OY11.1-like sequences are embedded in these *Alu* clusters. Considering that the mouse VpreB gene is not included in its own λ gene locus (Carson and Wu 1989) and that both the sheep OY11.1 and the bovine homolog of this sequence are located within Y chromosome repeat regions, these dense *Alu* clusters might have played a role in the translocation of these sequences. Of 20 L1 repeats (Smit et al. 1995), 7 are distributed in region III and 5 are located in region IV. L1 repeats in region III periodically appear every ~25 kb (Fig. 1). This ~25-kb repeat unit is also seen in the dot matrix analysis (Fig. 5). Studies on the distribution of repetitive elements in the λ gene locus will give further insight into the evolution of this locus.



Vλ	Promoter			Splice site			RSS			cDNA entry in databases
	8mer		TATA	donor	intron	acceptor	7mer	23 nt	9mer	
CONSENSUS	ATTTGCAT		gATAAG	CAG GT		TcC AG GgT	CACAGTG	acaCAgGCagATGGGGAAgTGAG	ACAaAAACC	
1-2	--------	29 nt	------	--- --	108 nt	--- -- ---	-------	TTTT-A-TCA---A------A--	-T-------	yes L03633 (99%)
1-3	--------	30 nt	------	--- --	110 nt	--- -- -A-	-------	GTC--A-TTCC------------	--C------	yes L03630 (97%)
1-4	--------	30 nt	------	--- --	110 nt	--- -- ---	-------	GTC--A-TTCC-------C----	--C------	yes L25295 (99%)
1-5	--------	30 nt	------	--- --	109 nt	--- -- -A-	-----A-	GTC--A-TTCC-------C----	--C------	yes Z46852 (98%)
1-7	----T---	30 nt	------	--- --	110 nt	--- -- ---	-------	GTC--A-TTC--------C----	--C------	yes Z37354 (97%)
1-9	--------	29 nt	------	--- --	115 nt	--- -- ---	--T----	GTC--A-TTCC-A-----C----	--C------	no H61392 (83%)
1-11	--------	24 nt	T--G-A	--- --	108 nt	--- -- ---	-------	CTC-----CC-G------CA---	----G----	yes U03898 (95%)
1-13	--------	24 nt	T--G--	--- --	103 nt	--- -- ---	-------	CTC-----CCGG------C----	----G----	yes L26908 (99%)
1-16	--------	24 nt	T--G-A	--- --	108 nt	--- -- ---	-------	CTC-----CA--A-----C----	----G----	yes X57817 (99%)
1-17	--------	24 nt	T--G-A	--- --	108 nt	--- -- ---	-------	CTC-----CC-G------C----	----G----	yes X54446 (99%)
1-18	G---T---	40 nt	T--G-A	--- --	103 nt	--- -- ---	-------	CTC-----CC-G-----G-----	-G--G----	no L26908 (94%)
1-19	--------	24 nt	T----A	--- --	102 nt	--- -- ---	-------	CTC---C-CA--------C----	----G----	yes X14583 (100%)
1-20	----AA--	34 nt	T----A	--- --	105 nt	-A- -- T--	-------	C-T-----CAG------------	-T------T	yes gVLX-4.4 (98%)
1-22	--------	32 nt	------	--- --	118 nt	-G- -- -T-	-------	CTC---A-CC-------------	---G----T	yes Z37375 (99%)
2-1	--------	25 nt	------	--- --	286 nt	-GT -- -A-	-------	-------------C---------	---G-----	yes L25296 (100%)
2-6	--------	25 nt	------	--- --	279 nt	-G- -- -T-	------A	-------------AA--------	---C-----	yes X84941 (99%)
2-7	--------	22 nt	C---G-	--- --	146 nt	-G- -- TC-	-------	----T------------------	---C-----	yes L29164 (100%)
2-8	G-------	25 nt	------	--- --	396 nt	-G- -- -C-	---G---	-----------------------	--------A	no X91131 (91%)
2-11	--------	39 nt	A-----	--- --	147 nt	-G- -- -C-	-------	-----A-G---CA---------A	---T-----	yes S77599 (98%)
2-13	--------	20 nt	------	T-- --	139 nt	-G- -- -T-	----T--	------A----------------	---G-----	yes L35920 (100%)
2-14	--------	25 nt	------	--- --	427 nt	-G- -- -C-	---G---	-------------A---------	--------A	yes X57821 (99%)
2-15	--------	39 nt	A-----	--- --	142 nt	-G- -- -C-	-T-----	-----------------------	---C----T	no L29162 (82%)
2-17	--------	39 nt	A-----	--- --	159 nt	-G- -- -C-	-------	-----A-----CA----------	---T-----	yes L29163 (99%)
2-19	--------	39 nt	A-----	--- --	135 nt	-G- -- TC-	-------	--------------A---A----	---C-----	yes L29162 (100%)
3-2	--------	46 nt	-----A	--- --	80 nt	-T- -- ---	-------	---G-CT--T-A-A----CCA--	---T-----	yes S69332 (95%)
3-3	--------	46 nt	-----A	--- --	82 nt	-T- -- ---	-------	---G-CC--TGA-A----CCA--	---T-----	yes L19893 (99%)
3-4	--------	39 nt	--A---	--- --	92 nt	-TT -- -AG	-------	-TTT-AA-CT---A-------CA	--T------	yes Z18334 (99%)
4-1	--------	30 nt	--A---	--- --	115 nt	-G- -- -T-	-------	-----CA--------------G-	---------	no D01059 (87%)
4-2	--------	29 nt	--A---	--- --	118 nt	--- -- -T-	-------	-----CA--------------G-	---------	yes L28048 (97%)
4-3	--------	29 nt	--A---	--- --	117 nt	--- -- -T-	T------	-----CA--------------G-	--------T	no L28048 (91%)
4-4	--------	33 nt	A-G-G-	--- --	114 nt	-T- -- -T-	-------	CTC---A-CC---A------A--	---------	yes X57820 (99%)
4-6	--------	35 nt	------	--- --	111 nt	-G- -- -T-	-------	-G----AT*--***------CG-	---------	no D01059 (78%)
5-1	--------	34 nt	------	--- --	119 nt	-A- -- -TC	-------	------ATGA-G---A-GTGAG-	C-----C-T	no L39131 (84%)
5-2	--------	36 nt	A-----	--- --	130 nt	--- -- ---	-------	-----------A*----------	---C-----	yes U43928 (99%)
5-4	--------	36 nt	------	--- --	114 nt	-T- -- ---	-------	-T-----------A-------G-	------T--	yes X87950 (98%)
5-6	--------	36 nt	-G----	--- --	113 nt	-T- -- ---	-------	-------------A-------G-	---G-----	yes U03867 (97%)

Figure 4 Alignment of nucleotide sequences of promoters, splice sites, and RSSs. The structure of the V_λ gene segment (Tonegawa 1983) is depicted at the *top*. Dashes indicate identity to the consensus sequence. Lowercase letters appearing in the consensus indicate less-conserved nucleotides. The length of the sequence between the 8-mer and TATA box as well as the intron is indicated. An asterisk in the 23-nucleotide spacer is used as a space to align each sequence to the consensus. By searching GenBank and dbEST databases, expressed V_λ genes were assigned to each germ-line V_λ gene segments (yes, if assigned with ≥95%). The GenBank accession numbers of the expressed V_λ genes with the highest sequence identity are shown. V_λ*1-20* (gVLX-4.4) was shown previously to be active (Stiernholm and Berinstein 1995). V_λ*1-18* shows high sequence identity (94%) to an expressed V_λ gene (GenBank accession no. L26908); however V_λ*1-13* has higher sequence identity (99%) to this sequence, so we concluded that the cDNA of V_λ*1-18* was not identified in the databases.

We have described the entire germ-line repertoire of the V_λ gene segments. To elucidate the expression and the variation of V_λ gene segments, comprehensive analysis of cDNAs and polymorphisms are indispensable. Studies on the λ gene locus of other organisms, especially primates and rodents, will shed light on the evolution of the entire locus as well as on the role of clan 1–clan 5.

METHODS

Cosmid and BAC DNA Sequencing

Cosmid clones and the BAC clone used for nucleotide sequencing have been described previously (Kawasaki et al. 1995). DNA sequences were determined by the shotgun sequencing method, with slight modifications as described (Chissoe et al. 1995). Cosmid and BAC DNAs were extracted by the alkaline–SDS method and purified by equilibrium ultracentrifugation in cesium chloride–ethidium bromide gradients (Sambrook et al. 1989). The purified DNA was sheared using a nebulizer (Okada-izai) at a pressure of 1.0 kg/cm^2 for 1.5 min. DNA termini were then end-repaired using T4 polynucleotide kinase, T4 DNA polymerase, and Klenow fragment, followed by size fractionation using a low-gelling temperature agarose gel. DNA fragments of 2.0–2.5 kb in size were recovered, ligated to a *Sma*I site of the pUC19 plasmid vector (Fermentas MBI), and electroporated to SURE2 (Stratagene) *Escherichia coli* cells. Plasmid DNA was extracted using a plasmid isolation robot PI-100Σ (Kurabo) and used for sequencing

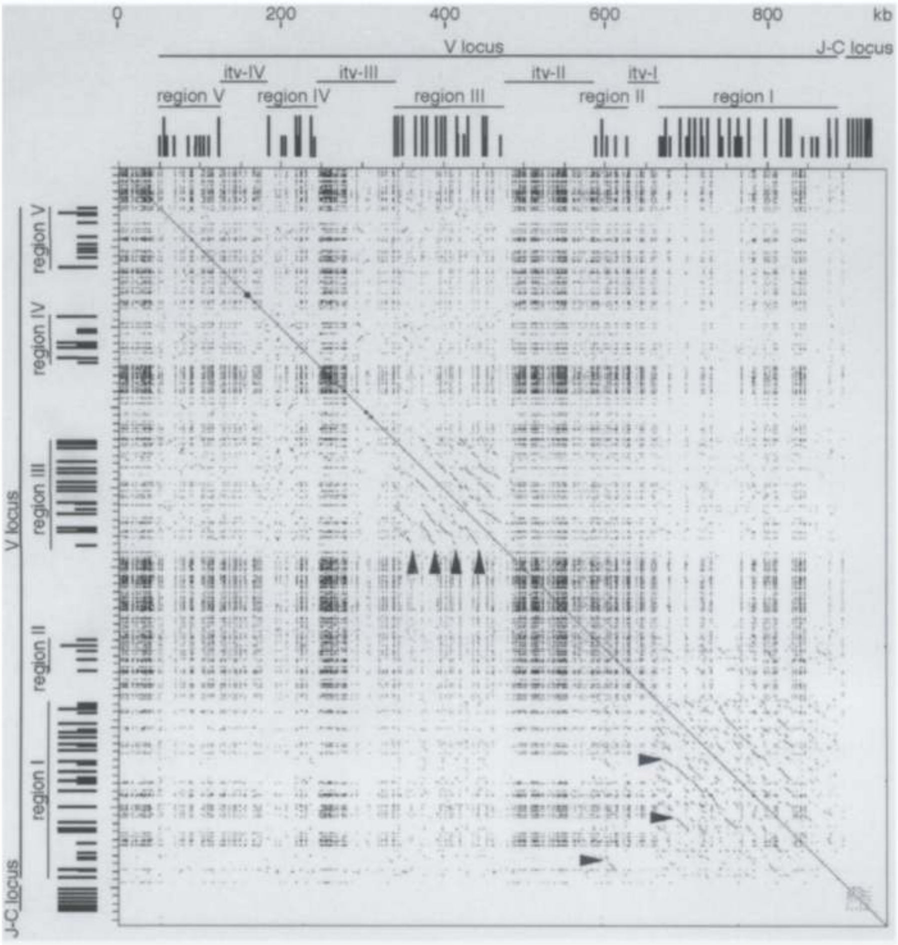


Figure 5 Dot matrix analysis of the entire λ gene locus vs. itself. Locations of the V_λ and C_λ gene segments are aligned at the *top* and the *left* side (see also Fig. 1). Each dot represents a sequence identity between two regions shown *above* and at *left*. The strength of each dot reflects the sequence identity within a 50-nucleotide sliding window (Sonnhammer and Durbin 1996). Lines parallel to the diagonal line represent direct order duplications. The terminus of a large duplication is indicated by an arrowhead.

reactions (ABI Prism dye terminator cycle sequencing ready reaction kit).

Using ABI Prism 377 DNA sequencers, an average of 650-nucleotide sequence from both ends of the inserts was determined for 250–350 shotgun clones (for cosmid clones) or 800 shotgun clones (for the BAC clone) with M13 forward primer and reverse primer. Sequencing data were edited and assembled using the Staden software package (Bonfield et al. 1995). These sequencing conditions provide 6.5–9.1 times redundancy; hence, ~90% of the sequenced regions were determined in both directions. For regions covered by less than three shotgun clones (~10% of the total length), individual nucleotides were confirmed by inspecting wave patterns. When necessary, sequencing primers were designed and used for primer walking to determine ambiguous nucleotides or to fill unsequenced gaps between contigs.

A total of three gaps remained after extensive shotgun sequencing and primer walking (one gap each in 47H9, 33B6, and 123E1 cosmid clones). These three gaps were filled using the nested deletion method. Restriction fragments containing the gaps were subcloned into pBluescript II SK(+) plasmid vector (Stratagene), and nested deletion sets were constructed from both ends of the inserts by digestion with an appropriate restriction enzyme, thioderivative dNTP fill-in, second digestion using another restriction enzyme followed by exonuclease III, and mung bean nuclease digestion as specified by the supplier (Stratagene).

Computer Analyses of DNA and Amino Acid Sequences

Locations of putative exons and repetitive elements were determined using XGRAIL version 1.2 or 1.3c (Uberbacher and Mural 1991). V_λ gene sequences were extracted from the GenBank and V BASE Sequence Directory (I.M. Tomlinson et al., MRC Centre for Protein Engineering, Cambridge, UK) and were searched for sequence similarity using the FASTA program (Pearson and Lipman 1988). Further surveys of sequence homology (BLASTN or BLASTP; Altschul et al. 1990), as well as protein motif analysis (PROSITE) were conducted using GenomeNet WWW server (The University of Tokyo and Kyoto University). The V_λ sequences were multiple-aligned using the CLUSTAL W (v. 1.6) package (Thompson et al. 1994), and a phylogenetic tree was constructed by the neighbor-joining method (Saitou and Nei 1987) using the PHYLIP (v. 3.57c) package (Felsenstein 1989). The dot matrix analysis was accomplished by the DOTTER program (Sonnhammer and Durbin 1996) with dynamic threshold control.

ACKNOWLEDGMENTS

We are grateful to Dr. B.A. Roe at Oklahoma University for help in large-scale sequencing technology, Dr. R. Staden at the Medical Research Council (MRC) for assistance in sequencing data analysis, Dr. I.M. Tomlinson (MRC) for helpful discussion, Dr. P.J. de Jong and the U.S. Department of Energy for providing the cosmid library, and Ms. H. Harigai for preparing the manuscript. This work was supported by the Fund for Human Genome Sequencing Project of Japan Science and Technology (JST) Corporation.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be

hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

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

Asakawa, S., I. Abe, Y. Kudoh, N. Kishi, Y. Wang, R. Kubota, J. Kudoh, K. Kawasaki, S. Minoshima, and N. Shimizu. 1997. Human BAC library: Constraction and rapid screening. *Gene* (in press).

Batzer, M.A., P.L. Deininger, U. Hellmann-Blumberg, J. Jurka, D. Labuda, C.M. Rubin, C.W. Schmid, W. Zietkiewicz, and E. Zuckerkandle. 1996. Standardized nomenclature for Alu repeats. *J. Mol. Evol.* **42:** 3–6.

Berinstein, N., S. Levy, and R. Levy. 1989. Activation of an excluded immunoglobulin allele in a human B lymphoma cell line. *Science* **244:** 337–339.

Boguski, M.S., T.M.J. Lowe, and C.M. Tolstoshev. 1993. dbEST—Database for "expressed sequence tags." *Nature Genet.* **4:** 332–333.

Bonfield, J.K., K.F. Smith, and R. Staden. 1995. A new DNA sequence assembly program. *Nucleic Acids Res.* **23:** 4992–4999.

Carson, S. and G.E. Wu. 1989. A linkage map of the mouse immunoglobulin lambda light chain locus. *Immunogenetics* **29:** 173–179.

Celeste, A.J., J.A. Iannazzi, R.C. Taylor, R.M. Hewick, V. Rosen, E.A. Wang, and J. Wozney. 1990. Identification of transforming growth factor β family members present in bone-inductive protein purified from bovine bone. *Proc. Natl. Acad. Sci.* **87:** 9843–9847.

Ch'ang, L.-Y., C.-P. Yen, L. Besl, M. Schell, and A. Solomon. 1994. Identification and characterization of a functional human Ig $V_{\lambda VI}$ germline gene. *Mol. Immunol.* **31:** 531–536.

Chissoe, S.L., A. Bodenteich, Y.-F. Wang, Y.-P. Wang, D. Burian, S.W. Clifton, J. Crabtree, A. Freeman, K. Iyer, L. Jian, Y. Ma, H.-J. McLaury, H.-Q. Pan, O.H. Sarhan, S. Toth, Z. Wang, G. Zhang, N. Heisterkamp, J. Groffen, and B. Roe. 1995. Sequence and analysis of the human *ABL* gene, the *BCR* gene, and regions involved in the Philadelphia chromosomal translocation. *Genomics* **27:** 67–82.

Chuchana, P., A. Blancher, F. Brockly, D. Alexandre, G. Lefranc, and M.-P. Lefranc. 1990. Definition of the human immunoglobulin variable lambda (IGLV) gene subgroups. *Eur. J. Immunol.* **20:** 1317–1325.

Cook, G.P. and I.M. Tomlinson. 1995. The human immunoglobulin V_H repertoire. *Immunol. Today* **16:** 237–242.

Deftos, M., R. Soto-Gil, M. Quan, T. Olee, and P.P. Chen. 1994. Utilization of a potentially universal downstream primer in the rapid identification and characterization of Vλ genes from two new human Vλ gene families. *Scand. J. Immunol.* **39:** 95–103.

Eulitz, M., C. Murphy, D.T. Weiss, and A. Solomon. 1991. Serologic and chemical differentiation of human λIII light chain variable regions. *J. Immunol.* **146:** 3091–3096.

Evans, R.J. and G.F. Hollis. 1991. Genomic structure of the human Igλ1 gene suggests that it may be expressed as an Igλ14.1-like protein or as a canonical B cell Igλ light chain: Implications for Igλ gene evolution. *J. Exp. Med.* **173:** 305–311.

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

Felsenstein, J. 1989. PHYLIP—Phylogeny Inference Package (version 3.2). *Cladistics* **5:** 164–166.

Frippiat, J.-P., S.C. Williams, I.M. Tomlinson, G.P. Cook, D. Cherif, D. Le Paslier, J.E. Collins, I. Dunham, G. Winter, and M.-P. Lefranc. 1995. Organization of the human immunoglobulin lambda light-chain locus on chromosome 22q11.2. *Hum. Mol. Genet.* **4:** 983–991.

Hallberg, E., R.W. Wozniak, and G. Blobel. 1993. An integral membrane protein of the pore membrane domain of the nuclear envelope contains a nucleoporin-like region. *J. Cell Biol.* **122:** 513–521.

Hanai, R., P.R. Caron, and J.C. Wang. 1996. Human *TOP3*: A single-copy gene encoding topoisomerase III. *Proc. Natl. Acad. Sci.* **93:** 3653–3657.

Heisterkamp, N., E.R.-D. Meyts, L. Uribe, H.J. Forman, and J. Groffen. 1991. Identification of a human γ-glutamyl cleaving enzyme related to, but distinct from, γ-glutamyl transpeptidase. *Proc. Natl. Acad. Sci.* **88:** 6303–6307.

Hesse, J.E., M.R. Lieber, K. Mizuuchi, and M. Gellert. 1989. *V(D)J* recombination: A functional definition of the joining signals. *Genes & Dev.* **3:** 1053–1061.

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

Kabat, E.A., T.T. Wu, H.M. Perry, K.S. Gottesman, and C. Foeller. 1991. *Sequences of proteins of immunological interest.* NIH Publications, Washington, D.C.

Kawasaki, K., S. Minoshima, K. Schooler, J. Kudoh, S. Asakawa, P.J. de Jong, and N. Shimizu. 1995. The organization of the human immunoglobulin λ gene locus. *Genome Res.* **5:** 125–135.

Kirkham, P.M., F. Mortari, J.A. Newton, and H.W. Schroeder, Jr. 1992. Immunoglobulin V_H clan and family identity predicts variable domain structure and may influence antigen binding. *EMBO J.* **11:** 603–609.

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

Matsuda, F., E.-K. Shin, H. Nagaoka, R. Matsumura, M. Haino, Y. Fukita, S. Taka-ishi, T. Imai, J.H. Riley, R. Anand, E. Soeda, and T. Honjo. 1993. Structure and physical map of 64 variable segments in the 3′ 0.8-megabase region of the human immunoglobulin heavy-chain locus. *Nature Genet.* **3:** 88–94.

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

Rowen, L., B.F. Koop, and L. Hood. 1996. The complete 685-kilobase DNA sequence of the human β T cell receptor locus. *Science* **272:** 1755–1762.

Saitou, N. and M. Nei. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406–425.

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

Smit, A.F.A., G. Toth, A.D. Riggs, and J. Jurka. 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J. Mol. Biol.* **246:** 401–417.

Sommer, S.S. 1992. Assessing the underlying pattern of human germline mutations: Lessons from the factor IX gene. *FASEB J.* **6:** 2767–2774.

Sonnhammer, E.L.L. and R. Durbin. 1996. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167:** GC1–10.

Stephens, R.M. and T.D. Schneider. 1992. Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J. Mol. Biol.* **228:** 1124–1136.

Stiernholm, N.B.J. and N.L. Berinstein. 1995. A mutated promoter of a human Ig Vλ gene segment is associated with reduced germ-line transcription and a low frequency of rearrangement. *J. Immunol.* **154:** 1748–1761.

Thompson, J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.

Tonegawa, S. 1983. Somatic generation of antibody diversity. *Nature* **302:** 575–581.

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

Vasicek, T.J. and P. Leder. 1990. Structure and expression of the human immunoglobulin λ genes. *J. Exp. Med.* **172:** 609–620.

Vogt, P. 1990. Potential genetic functions of tandem repeated DNA sequence blocks in the human genome are based on a highly conserved "chromatin folding code." *Hum. Genet.* **84:** 301–336.

Wenk, J., H.-I. Trompeter, K.-G. Pettrich, P.T.W. Cohen, D.G. Campbell, and G. Mieskes. 1992. Molecular cloning and primary structure of a protein phosphatase 2C isoform. *FEBS Lett.* **297:** 135–138.

Winkler, T.H., H. Fehr, and J.R. Kalden. 1992. Analysis of immunoglobulin variable region genes from human IgG anti-DNA hybridomas. *Eur. J. Immunol.* **22:** 1719–1728.

Zachau, H.G. 1995. The human immunoglobulin κ genes. In *Immunoglobulin genes* (2nd ed.) (ed. T. Honjo and F.W. Alt), pp. 173–191. Academic Press Ltd., London, UK.

Received October 21, 1996; accepted in revised form January 27, 1997.