

# Complete nucleotide sequences of three $V_H$ genes in Caiman, a phylogenetically ancient reptile: Evolutionary diversification in coding segments and variation in the structure and organization of recombination elements

(heterologous cross-hybridization/reptilian immunoglobulin genes/metric analysis of DNA sequences)

G. W. LITMAN\*, K. MURPHY\*, L. BERGER\*, R. LITMAN\*, K. HINDS\*, AND B. W. ERICKSON†

\*Memorial Sloan-Kettering Cancer Center, Walker Laboratory, Rye, NY 10580; and †The Rockefeller University, New York, NY 10021

Communicated by Edward A. Boyse, September 26, 1984

**ABSTRACT** Complete nucleotide sequences are described for three caiman (*Caiman crocodylus crocodylus*) immunoglobulin  $V_H$  genes ( $C3$ ,  $E1$ , and  $G4$ ) that hybridize with a murine  $V_H$  probe. The  $E1$  and  $G4$  genes are physically linked (intergenic distance,  $\approx 6.5$  kilobases) in the same transcriptional orientation but are not directly contiguous with the  $C3$  gene. When the coding segments, including both framework and complementarity-determining regions, of these genes and the murine probe sequences are compared by metric analysis, it is apparent that the caiman genes are only slightly more related to each other than to the mammalian sequence, consistent with significant preservation of nucleotide sequence over an extended period of phylogenetic time. Based on the presence of transcriptionally critical 5' sequences and the absence of terminator codons, frameshift mutations, or other recognizable alterations, the genes do not appear to be pseudogenes. The  $E1$  gene, however, is distinguished from other  $V_H$  genes because (i) the spacer region within the 3' recombination signal sequence is 12 base pairs, typical of  $V_{\kappa}$  genes but not of  $V_H$  genes, which possess 22- to 23-base-pair spacers and (ii) a near-perfect  $V_H$  recombination signal sequence is present within the intervening sequence that splits the segment encoding the leader. These studies establish  $V_H$  gene multiplicity in a species that arose prior to mammalian radiation and provide a description of differences in the configuration and location of recombination elements associated with an otherwise potentially functional gene.

Based on studies in mammals, it is apparent that multiple germ line  $V_H$  genes contribute to the overall diversity of the humoral immune response that is amplified further by somatic events such as mutation and segmental joining (1–10). This latter process is mediated by relatively short DNA segments located 3' to the coding segments of immunoglobulin  $V_H$  (2, 3, 6, 7, 9) and  $V_L$  genes (9, 11, 12) and by complementary segments flanking the  $D$ ,  $J_H$ , and  $J_L$  segments (2, 3, 9–12) located 5' of the constant region genes. Understanding the evolution of the V-region gene families and their associated recombination mechanisms is essential for understanding the developmental control of antibody expression and other genetic processes involving somatic changes in DNA. Since it is likely that significant numbers of V-region genes are not subject to direct selection during the lifetime of an individual, the processes that govern the phylogenetic development, diversification, and stabilization of this multigenic family are important and may be unique.

To date, evolutionary studies of  $V_H$  genes largely have been restricted to comparisons between the members of

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

structurally related families, identified in inbred mouse strains (6, 7, 13–18), as well as between murine and human sequences (19–22). In earlier reports, we described cross-hybridization between restriction enzyme-digested caiman genomic DNA and murine  $V_H$  probes (23) and demonstrated sequence similarities between caiman and prototypic mammalian  $V_H$  genes (24). In this report, we compare the coding as well as the 5' and 3' flanking segments of three different caiman (*Caiman crocodylus crocodylus*)  $V_H$  genes and identify unique recombination signal sequences associated with one of these genes.

## MATERIALS AND METHODS

The construction, amplification, and screening of a caiman-λ47.1 library with S107V, a murine  $V_H$  probe (3), have been described (24). Standard phage purification and subcloning approaches were used. Sequences were determined using the dideoxynucleotide termination method and compared by the mathematical methods of metric analysis (25, 26).

## RESULTS AND DISCUSSION

Fig. 1 *a* and *b* illustrates partial restriction maps of two  $V_H$  recombinant phages, IVD and VIIIB. Comparison of the maps and additional restriction mapping data (not shown) indicates that the linked  $E1$  and  $G4$  (intergenic distance is  $\approx 6.5$  kb) are neither directly contiguous with nor most likely allelic to  $C3$ . Fig. 1 *c–e* identifies the functional segments of the three caiman genes and illustrates the primary strategies used in determining their nucleotide structure.

The complete nucleotide sequences of the three caiman genes and their respective noncoding 5' and 3' segments are presented in Fig. 2 *a* and *b*. The sequences of the gene segments adjacent to these (Fig. 1 *c–e*) will be described at a later date. All three genes encode homologous leader regions interrupted by an intervening sequence (IVS), characteristic of mammalian  $V_H$  (and  $V_L$ ) genes (2, 3, 19). The putative splice donor sequence of  $C3$  and  $E1$  and acceptor sequences of  $C3$ ,  $E1$ , and  $G4$  conform to the consensus sequence inferred from nuclear and viral genes (27) and are typical of immunoglobulin  $V_H$  genes. The  $G4$  donor sequence cannot be accommodated into this particular consensus sequence, but it preserves the general A·G/G·T structure and is consistent with other functional splice donor sequences (27, 28).

**Abbreviations:**  $V_H$ , variable region of immunoglobulin heavy chain;  $V_L$ , light chain variable region;  $V_{\kappa}$ ,  $\kappa$  (light) chain variable region;  $V_{\lambda}$ ,  $\lambda$  (light) chain variable region;  $V$ , variable;  $D$ , diversity segment;  $J_H$ , heavy chain joining segment;  $J_L$ , light chain joining segment;  $J_{\kappa}$ ,  $\kappa$  light chain joining segment; CDR, complementarity-determining region; FR, framework region; IVS, intervening sequence (segment); kb, kilobase(s); bp, base pair(s).

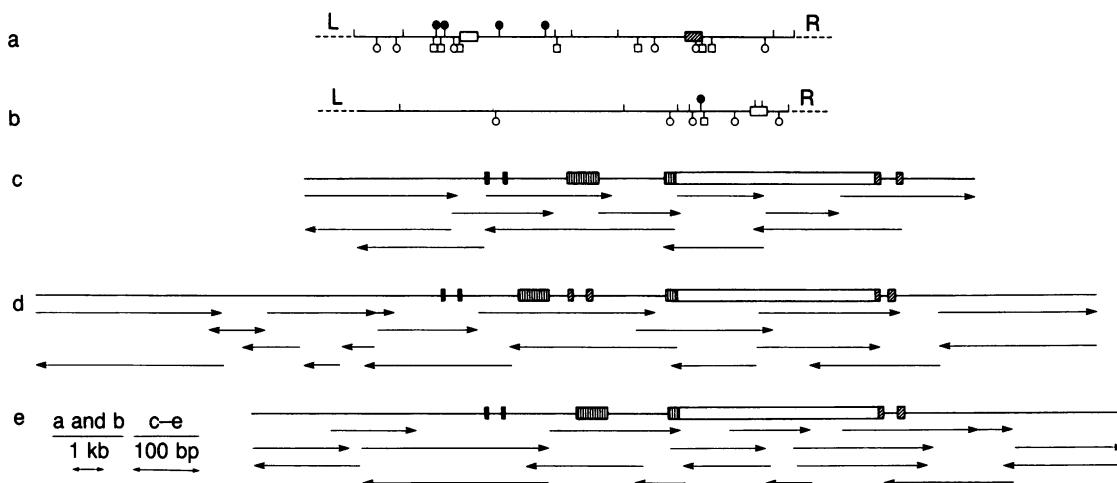


FIG. 1. Partial restriction endonuclease mapping of  $V_H^+$  caiman genomic DNA- $\lambda$  clones IVD (a) and VIIIB (b) with *Bam*HI (I), *Hind*III (●), *Sma* I (○), and *Sst* I (□) localizing the *G4* (□) and *E1* (■) (a) and the *C3* (□) (b) genes. The enclosed distances correspond to the initiation codon through the recombination signal sequence of the respective genes, L and R are the left and right areas of the recombinant phage (----), and the scales for a-e are indicated directly as kilobases (kb) or 100 base pairs (bp). The essential organizational features of the genes, including (5' → 3') a hyperconserved upstream octamer and the putative promoter for RNA polymerase II transcription (I), interrupted leader (III), mature coding (□), and recombination signal sequences (II) of *G4*, *E1*, and *C3* are illustrated in c-e along with the sequencing strategies for each of these genes. The first two rows of arrows indicate sequence determination on the (+) strand, the last two rows correspond to the (-) strand. Double-headed arrows correspond to (same direction) extension of a given sequence (typically achieved by alternative cloning strategy) overlapping the primary sequence by at least 50 bp or (opposite direction) a single clone sequenced in both directions. Dideoxynucleotide sequencing confirmed the restriction map as well as the sequence reported earlier for *C3* (24) and identified several additional *Hinf*I and *Dde* sites located within 25 bp (below the detection limits) of the map positions previously assigned for these enzymes.

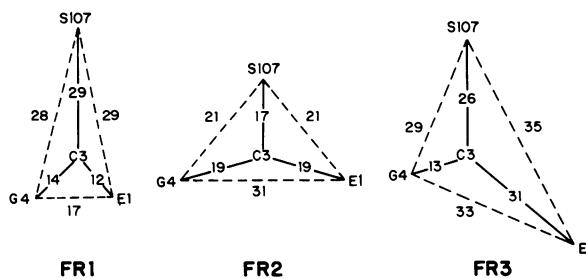
The intron length is 98 bp for both *C3* and *G4* but is 171 bp for *E1*, which contains a sequence resembling an immunoglobulin recombination element (see below). The coding region of each gene is followed by a two-nucleotide spacer and a recombination signal segment. Nucleotide identity between the three genes extends 5' and 3' of the coding segments.

Fig. 2b compares the mature coding segments of the three caiman genes and *S107*, a murine  $V_H$  III prototype (3). These can be related most efficiently by reference to the lines labeled "common acids" and "common bases" that compare the caiman gene *C3* to *E1*, *G4*, or the murine sequence. The nucleotide identities are concentrated within the framework regions FR1, FR2, and FR3 and are significantly lower in the complementarity-determining regions CDR1 and CDR2. At the level of the inferred amino acids, no residues are common to all three caiman sequences in CDR1 (residues 31–35) and the first 12 positions of CDR2 (residues 50–61). The three FR regions of all four protein sequences share 37 amino acid residues. Thus, 49% of the 76 FR residues are conserved identically in the caiman and murine prototype genes. The distances between pairs of FR gene segments are illustrated in the following spanning trees, where each distance

*C3* is closest to *S107*. Finally for FR3, *C3* is closer to murine *S107* than to caiman *E1* but closest to caiman *G4*. Only comparison of the FR1 segments distinguishes among these particular caiman and murine  $V_H$  genes.

The three caiman sequences have also been aligned metrically with 17 additional murine (2, 6, 7, 14, 15, 17, 18) and human (19–21)  $V_H$  genes representing different subgroups and families. Overall the three caiman sequences are closest to human  $V_H$  III prototypes—e.g., FR1, FR2, and FR3 of *G4* are 20, 12, and 20% *bc/b*, respectively, from the human *H11* gene (20) vs. 28, 21, and 29% *bc/b* for the equivalent comparisons of *G4* to the murine probe sequence illustrated above. Although it is tempting to speculate as to the various selective influences and correction processes that may have given rise to these patterns, it is important to emphasize that relaxed criteria were used in selecting the caiman genes and that members of a specific subset were not isolated. Given the apparent high degree of intraspecies variation within these families, individual gene segments from two phylogenetically divergent species may be more related to each other than are equivalent gene segments identified within the same species.

Metric analyses of sequences 5' to the initiation codons indicate several regions of extended nucleotide identity. A potential promoter region for RNA polymerase II, A-T-A-A-A-T (15, 29), is located 5' of ATG in *C3*, *E1*, and *G4*, respectively (Fig. 2a). All three caiman genes also possess the sequence A-T-G-C-A-A-T located 26–27 bp further 5'. This conserved octamer presumably is involved in transcriptional regulation (30, 31). Assuming that the start of transcription is 26–30 bp downstream of A-T-A-A-A-T, each of the three caiman genes possesses a C-A sequence flanked by deoxynucleotides with a pyrimidine/purine content (32) equivalent to functionally mapped transcriptional start sites of murine  $V_H$  genes (15, 29). By these criteria, it appears that all three caiman genes would support transcriptional activity. As indicated above, the presence of leader regions, typical splice sites, and uninterrupted reading frames also are consistent with functional status for these three genes. Similar consid-



(edge length) is the number of base changes per 100 bases (% *bc/b*) between two segments (vertices). Since *C3* is used as the reference sequence in Fig. 2b, it has been used for the primary comparisons (solid line) in the spanning trees. For FR1, caiman *C3* is farthest from murine *S107*, but for FR2,

alignments: | 329 340 350 360 370 380 390 395 400 405 412 413 430 434  
C3 gene: TATAAGGCGGGGACATGCAAACTCACAGCACGGCTGCCGTATAATCTCTGCGGCTGCCAGGACCTGCACACACAAACACAGCCCTGACGTC  
common bases: TAT AGG GGGGA ATGCAAACT C G C ATAAT CT GGCA C CTG C ACA G T G T -AC  
E1 gene: TATAGGGGGGGGGAGATGCAAACTTAGTGTGCTGGAGGCTCTAAAT-GCT GGCACACTGCGCTCACAGGGG-TGGGGTGTG-T-AC  
; 644 650 660 670 680 690 694 695 700 705 712 720 730

alignments: | 329 340 350 360 370 380 390 395 400 405 412 420 430 434  
C3 gene: TATAAGGCGGGGACATGCAAACTCACAGCACGGCTGCCGTATAATCTCTGCGGCTGCCAGGACCTGCACACACAAACACAGCCCTGACGTC  
common bases: TAT AGG G G CATGCAAACT C AG T C G T AAAT C TG G C T C C C G C T G C C A C C G C T G C A C G  
G4 gene: TATATGGAGAAGATCATGCAAAATAGCCCAGTCTGCG-AAAAT-GCG-TGAG-GCTTCACC-CCTGCTGTCAGCCC-TGGCTGCTG  
; 259 270 280 290 300 308 321 330 331 340 351 351

alignments: | 644 650 660 670 680 690 694 695 700 705 712 720 730  
E1 gene: TATAGGGGGGGGGAGATGCAAACTTAGTGTGCTGGAGGCTCTAAAT-GCT GGCACACTGCGCTCACAGGGG-TGGGGTGTG-T-AC  
common bases: TAT GG G G ATGCAAACT C G GCT C TAAAT GC C CTG TG CAG TGG TT G AC  
G4 gene: TATATGGAGAAGATCATGCAAAATAGCCCAGTCTGCG-AAAAT-GCG-TGAG-GCTTCACC-CCTGCTGTCAGCCC-TGGCTGCTG  
; 259 270 280 290 300 308 321 330 331 340 351 351

COMMON/consensus: TATAggG,ggGgacATGCAAAc,c,,CaG,,t,,gCtgcaTAAAT gct tg,g,t,ccaggCagCtgcTG,,c,,aCAgccc,tGg,gTctG,t cAG  
notable: ##### ##### #####

---

Ig C3: | -19 -15 -10 -5  
alignments: | -3:--- -6:--- -2:--- -3:--- -2:--- -4:--- -4:--- -5:  
C3 gene: 435 440 450 460 470 480 490 500 510 520 530 537  
GCCCTGA CCACCAACACAGAGCTTCCCCCAGAACGACAGGACATG-GGACTCTGGCTCACCTGCTGCGGCCAGCTCGCAAGGTATGCACTTC  
common bases: G TGA CCA C GAG T CC CAGC CCAG A CATG G C TG G A T TC TCC T C C T G C AGGT GTC CT  
E1 gene: G-GGTG 731 740 750 760 770 780 790 800 810 820 827  
codon phase: Ig E1: <-> <-out (-1)> <-in (+0)>  
common acids: MetLeuSerTrpGly t yrPheLeuPheLeuProThrLeuProG  
-19 -15 -10 -5  
Met Leu Leu Leu  
Ig C3: | -19 -15 -10 -5  
alignments: | -3:--- -3:--- -3:--- -3:--- -3:--- -3:--- -3:--- -5:  
C3 gene: 435 440 450 460 470 480 490 500 510 520 530 537  
GCCCTGA CCACCAACACAGAGCTTCCCCCAGAACGACAGGACATG-GGACTCTGGCTCACCTGCTGCGGCCAGCTCGCAAGGTATGCACTTC  
common bases: G TGA CCA C C CA GAGAGCT CC CCAG ACC G ACATG TGG A T TC TCT C G AGC T AGGT GTC CT  
G4 gene: -CGTTGACTCCAGC-CCCCAGGAGAGCTCCCTCCAGC-ACAG-GAAGACATGACCCACTGGCTATGTTACTCTTCTCT-CTAGCGCTCAAGGGTGGTCATCTC  
352 360 370 380 390 400 410 420 430 440 450 452  
codon phase: Ig G4: <-> <-out (-1)> <-in (+0)>  
common acids: MetThrHisTrpLeuPheThrLeuAlaLeuValAlaValArg  
-19 -15 -10 -5  
Met Trp Ala G  
Ig E1: | -19 -15 -10 -5  
alignments: | -3:--- -3:--- -3:--- -3:--- -3:--- -3:--- -3:--- -5:  
E1 gene: 731 740 750 760 770 780 790 800 810 820 827  
G CGTCA-CGAGG TGTCT-GAGCCATCCTCGAGC CCAG CATTGATCTGCTGGGGAT-ATTGCTCTTCTCCCTCCACCTGGCAAGGTAGTC-CITG  
common bases: G TGA CAG C GAG T CCT C G G C T G G A T T T C T C T C G T C C T G C T AGGT GTC CT  
G4 gene: -CGTTGACTCCAGC-CCCCAGGAGAGCTCCCTCCAGC-ACAG-GAAGACATGACCCACTGGCTATGTTACTCTTCTCT-CTAGCGCTCAAGGGTGGTCATCTC  
352 360 370 380 390 400 410 420 430 440 450 452  
codon phase: Ig G4: <-> <-in (+0)> <-out (-1)> <-in (+0)>  
common acids: MetThrHisTrpLeuPheThrLeuAlaLeuValAlaValArg  
-19 -15 -10 -5  
Met Trp Ala G  
COMMON/consensus: gcg,TGA CCAgC ,c,Ca,GAGagCcCtcCAGcaCcAG,,aaCATG,,g,ccTGGat,attT,CTCtCtc,tcgcagCc,Tgc,AGGTa,GCatCT,  
notable: #####  
gene segment: <-LR (leader region) segment [start]->-[IVS->

---

C3 gene: | 538 550 560 570 580 590 596  
gene segment: CCCACAGGTGGGGAGGAATCTTCCACTGATGCTGCCATAATACTACATCTCTTG  
IVS (intervening segment) within LR (leader region) segment

E1 gene: | 828 840 850 860 870 880 890 900 910 920 930  
inverted repeat: GGACCTGCCACGTTCCACAGGCTGTGGAGGCCAGGGAGCAACACAAAACTTCACTGTGATGTAAGGCTGATGAAATCAGCAGGAATGCCCTTGGCGC  
notable: <-><-><-><-><->  
heptamer 21-spacer nonamer  
<-RS (recombination signal) segment->

gene segment: ---IVS within LR segment-----

E1 gene: | 931 940 950 960 966  
gene segment: CGTGGCTTGGCAAGGCTCAGGAGACGCCCTGA  
---IVS within LR segment-----

G4 gene: | 453 460 470 480 490 500 510 519  
gene segment: CTCTGGGGACCAGAGAGACATCCCCGGTTGGCTCTGTTCAACAGGCTCAGGGAGGATTC

---

Ig C3: | -4 -1 1 98  
alignments: | -4:--- -4:--- -4:--- -4:--- -4:--- -4:--- -4:--- -5:  
C3 gene: 597 610 620 629 918 930 940 950 960 961  
ATCCTAC-C-TTGCTTTCAGGTCTGGTCCCGAGAGACAGCTGACTCAACCCATTACCGGCAATACAAATTC  
common bases: ATC T C C T T C T GAGTCG GTCCAG AG C G T G A A C C A C C A C C  
E1 gene: ATCGTGTCTTTCCCGGGCTCCCTGGAGAGCTCCAGCAGAGCTACTGCTGCGGGAAAGCC  
967 980 990 1001 1290 1300 1311 1312 1322  
codon phase: Ig E1: <-in (+0)> -->  
Ig E1: 1yAlaInSerGln.. Arg  
-4 -1 1 98  
common acids: 1yAla SerGln .. Arg

Ig C3: | -4 -1 1 98 C3: 23-bp spacer  
alignments: | -4:--- -4:--- -4:--- -4:--- -4:--- -4:--- -4:--- -5:  
C3 gene: 597 610 620 629 918 930 940 950 960 961  
ATCCTAC C TTGCTTTCAGGTCTGGTCCCGAGAGACAGCTGACTCAACCCATTACCGGCAATACAAATTC  
common bases: A C T C T C T T C G A G G T C A G  
G4 gene: ACCGTT TC TCTTTGGAGGGCTCCCTGGAGAGCTCCAGCAGAGCTACTGCTGCGGGAAAGCC  
520 530 540 552 841 850 860 870 880 884  
codon phase: Ig G4: <-in (+0)> -->  
Ig G4: 1yValueLeuSerGln.. Arg  
-4 -1 1 98 G4: 23-bp spacer  
common acids: 1y Ser .. Arg

Ig E1: | -4 -1 1 98 E1: 12-bp spacer  
alignments: | -2:--- -2:--- -2:--- -2:--- -2:--- -2:--- -2:--- -5:  
E1 gene: 967 980 990 1001 1290 1300 1311 1312 1322  
ATCGTGTCTTTCCCGGGCTCCCTGGAGAGCTCCAGCAGAGCTACTGCTGCGGGAAAGCC  
common bases: A CGT T C T C G G C C T C A G  
G4 gene: ACCGTT -TC TCTTTGGAGGGCTCCCTGGAGAGCTCCAGCAGAGCTACTGCTGCGGGAAAGCC  
520 530 540 552 841 850 860 870 880 884  
codon phase: Ig G4: <-in (+0)> -->  
Ig G4: 1yValueLeuSerGln.. Arg  
-4 -1 1 98  
common acids: 1y Ser .. Arg

COMMON/consensus: AtGdT,ctC,TttCTt,CAGGtGcccAG .. AgG,acAcaGtGag,aAaa,,c,,t,,,Ggc,Aga,aaaaaCc  
notable: #####  
gene segment: ---IVS----->-[SEE FIG. 3]><-><-><-><->  
heptamer spacer nonamer  
S (spacer) segment <-RS (recombination signal) segment->

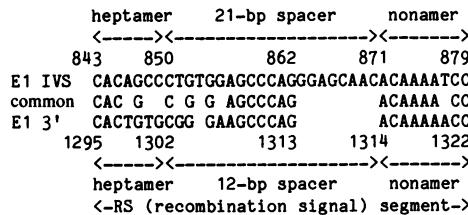
FIG. 2. DNA sequence alignments of the caiman *C3*, *E1*, and *G4* and murine *S107* (3)  $V_H$  genes. The data in *a* illustrate the sequences 5' and 3' of the predicted coding regions of the caiman genes and include the LR, S, and RS IVS segments. The mature coding segments are shown in *b*. Position 480 of *C3* corresponds to position 1 of the published *C3* sequence (24). Metric analysis (25, 26) was used to obtain the best pairwise

<b>b</b>	<b>C3</b>	caiman Ig C3;	1    5    10    15    20    25    30	GlnValGlnLeuValGluSerGlyGlyAspValArgLysProGlyAsnSerLeuArgLeuSerCysLysAlaSerGlyPheThrPheGly
		▼ caiman gene C3;	627    640    650    660    670    680    690	700    710    716
		alignments:	CAGGTGCAGCTGGTGGAGTCGCCGGAGAGATGAGGAAACCTGAAACTCTTGCGCTCTCCGCAAAGCCTCGGGCTCACCTTCGGT	
		common bases:	CAGGTGCAGCTGGTGGAGTCGCCGGAGAGATGAGGAAACCTGAAACTCTTGCGCTCTCCGCAAAGCCTCGGGCTCACCTTCGGT	
		caiman gene E1;	CAGGTGCAGCTGGTGGAGTCGCCGGAGAGATGAGGAAAGCCTGGAAGACTCTTGCGCTCTCCGCAAAGCCTCGGGCTCACCTTCGGT	
		codon phase:	<-in (-0)-> 999    1010    1020    1030    1040    1050    1060    1070    1080    1088	
		caiman Ig E1;	GlnValGlnLeuValGluSerGlyGlyAspValArgLysProGlyAsnSerLeuArgLeuSerCysLysAlaSerGlyPheThrPheSer	
		common acids:	1    5    10    15    20    25    30	
		GlnValGlnLeuValGluSerGlyGlyAspValArgLysProGly    SerLeuArgLeuSerCysLys    SerGlyPheThrPhe		
	<b>C3</b>	versus	alignments:	----- AG T CAGCTGGTGGAGTCGCCGGAGAGATGAGGAAACCTGAAACTCTTGCGCTCTCCGCAAAGCCTCGGGCTCACCTTCGGT
	<b>G4</b>		common bases:	AG GAGATCAGCTGGTGGAGTCGCCGGAGAGATGAGGAAAGCCTGGAAGACTCTTGCGCTCTCCGCAAAGCCTCGGGCTCACCTTCGGT
		caiman gene G4;	550    560    570    580    590    600    610    620    630    639	
		codon phase:	<-in (-0)-> GluLeuGlnLeuValGluSerGlyGlyAspValArgLysProGlyAsnSerLeuArgLeuSerCysLysAlaSerGlyPheThrPheSer	
		caiman Ig G4;	1    5    10    15    20    25    30	
		common acids:	GlnLeuValGluSerGlyGly    ArgLysProGly    SerLeuArgLeuSerCysLysAlaSerGlyPheThrPhe	
	<b>C3</b>	versus	alignments:	----- AGCTG ACCTGGTGGAGTCGCCGGAGAGATGAGGAAACCTGAAACTCTTGCGCTCTCCGCAAAGCCTCGGGCTCACCTTCGGT
	<b>S107</b>		common bases:	AGCTG AAAGCTGGTGGAGACCTGGCTGGGAGGCTTCGAGCTGGGGCTCTGAGACTCTCTCTGGCAAACCTCTGGGTACCTTCAGT
		mouse gene S107;	216    220    230    240    250    260    270    280    290    300    305	
		codon phase:	<-in (-0)-> GluValGlnLeuValGluSerGlyGlyLeuValGlnProGlyGlySerLeuArgLeuSerCysAlaThrSerGlyPheThrPheSer	
		mouse Ig S107;	1    5    10    15    20    25    30	
		common acids:	Val    LeuValGlnSerGlyGly    ProGly    SerLeuArgLeuSerCys    SerGlyPheThrPhe	
common	C3,E1,G4;	:	AG T CAGCTGGTGGAGTCGCCGGAGAGATGAGGAAACCTGAAACTCTTGCGCTCTCCGCAAAGCCTCGGGCTCACCTTCGGT	
bases	C3,E1,G4,S107:	:	AG T CAGCTGGTGGAGTCGCCGGAGATGAGGAAAGCCTGGAAGACTCTTGCGCTCTCCGCAAAGCCTCGGGCTCACCTTCGGT	
common	C3,E1,G4;	:	GlnLeuValGluSerGlyGly	
acids	C3,E1,G4,S107:	:	ArgLysProGly    SerLeuArgLeuSerCys    SerGlyPheThrPhe	
gene segment (Ig region):			<-FR1 (first framework region) segment>	
	<b>C3</b>	31    35	36    40    45    49    50	55    60    65    66
	GlyTyrGlyMetPhe	717    730    731    732	740    750    760	773    780    790    800
	▼ GGCATACGGCATGTT	TGGGTCGCCAGCGCTCTGGAGGGCTGGACTGGGTGGCT	ACATTA ATAC TGAT	790    810    820    824
	<b>C3</b>	versus	----- TAC G TG C	----- C ATT A AC
	<b>E1</b>		TGGGTCGCCAGCGCTCTGGAGGGCTGGACTGGGTGGCT	----- T GG CAGC T
	1089    330    1104	1110    1120    1130	1140    1145	1150    1160    1170
				1180    1190    1196
			<-in (-0)-> AsnTyrTrpLeuGly	<-out (+1)-> uProLeu ThrP
			31    35	role uAlaAlaAlaProThrTyrIleProGlyValSerGly
			Tyr	Tyr    Pro    Val    Gly
			TrpValArgGlnAlaProGlyLys	
	<b>C3</b>	versus	----- G C C G A T G C	----- AAT A AT
	<b>G4</b>		TGGG CCG CAG CCTGGAGGGCTG A TGCT G T	----- A A C G TA C CC G GT A GG
	640    650    654	655    660	670	680
			TGGGCCGGAGCCCCCTGGAGGGCTGGACTGGGTGGCT	TCAGAGACCATAG ATATGCCAACGAGCTGAAAGGT
			640    650    654	655    660
			AspThrTrpMetAla	AspTyrAlaProGluValLysGly
			31    35	36    40    45    48    50
			TrpAlaArgGlnProGlyLysGlyLeuGlnTrpValGly	GluIleA snGly Asn
			Met	55    56    60    65    66
			Trp ArgGln ProGlyLysGlyLeu TrpVal	IleA sn
				Tyr    Pro    Val    Gly
	<b>C3</b>	versus	----- G T C CATG	----- 15:::-----
	<b>S107</b>		TGGGTCGCCAG CTCCC GGGAG C CTGGG TG G T GCT	CAA TA A AC G T GAT C A G GTAC C C GT AGG
	306    310    320	321    330    340	350	GCAAGTAGAACAAAGCTTAATACACAGAGTACAGTGATCTGAAAGGT
			306    310    320	321    330    340
			AspPheTyrMetGlu	TrpValArgGlnProProGlyLysArgLeuGlnTrpIleAla
			31    35	36    40    45    49    50
			Met	AlaSerArgAsnAlaAsnAspTyrThrGluTyrSerAlaSerValLysGly
				60    65    68
				TyrSer    Val    Gly
common	C G TG C	TGGG CCG CAG C CC GGGAGG CT A TGG T T	AT A	A C    TA CC G GT GG
bases	C   TG	TGGG CCG CAG C CC GGGAGG CT A TGG T T	A A	C   TA C GT GG
common				Tyr    Pro    Val    Gly
acids		TrpValArgGln ProGlyLys		Tyr    Pro    Val    Gly
		TrpValArgGln ProGlyLys		
<-CDR1 segment>			<-FR2 (second framework region) segment>	<-CDR2 (second complementarity-determining region) segment>
	<b>C3</b>	67    70	75    80    85    90	95    98
	LysPheThrIleSerArgGlyAsnSerGlnAsnMetLeuTyrLeuGlnMetSerLeuThrProGluAspThrAlaThrTyrTyrCysAlaArg	825    830	840    850    860	870    880    890    900
	AAATTACCATCTCCAGAGGGCAACTCCAGAGCATGCTGACAGTGGAGGACACGGGGATATACTGGGAGG			910    920
	<b>C3</b>	versus	----- TTACCACTCTCA G CAA CC G C T GCTG ACCTG A ATGAGC C CT A CCTGAGGACAC G C	----- TAT ACTCGG AG
	<b>E1</b>		CGCTTACCATCTCCAGAGGCACTCCAGAGGAGCTCTGCTGACAGTGGAGGACACGGGGATATACTGGGAGG	CGCTTACCATCTCCAGAGGCACTCCAGAGGAGCTCTGCTGACAGTGGAGGACACGGGGATATACTGGGAGG
	1197	1210    1220    1230	1240	1250    1260    1270    1280
				1292
			<-1> <-3> <-out (-5)-> <-in (-6)->	
			ArgPheThrIleSerArgAspAsnAlaArgAlaLeuLeuHisIleAspMetSerAspLeuArgProGluAspThrGlyArgTyrHisCysGluArg	67    70    75    80    85    90    95    98
			PheThrIleSerArg Asn Leu Leu MetSer Leu ProGluAspThr	Tyr    Cys    Arg
	<b>C3</b>	versus	----- A A T CACCATCTCCAGAG CAAC CCCAGAAC TCTGCTG ACCTG A ATGAGC C CT A CCTGAGGACAC G C	----- >
	<b>G4</b>		AGACTCACCATCTCCAGAGCAACCCAGAGCTCTGCTGAGCATGAGGCTGGAGGACACGGGGATATACTGGGAGG	AGACTCACCATCTCCAGAGCAACCCAGAGCTCTGCTGAGCATGAGGCTGGAGGACACGGGGATATACTGGGAGG
	748	760    770    780	790	800    810    820    830
				843
			<-ArgLeuThrIleSerArgAspAsnThrGlnAsnLeuLeuPheLeuGlnIleSerLeuLysProGluAspThrAlaThrTyrTyrCysAlaArg>	67    70    75    80    85    90    95    98
			ThrIleSerArg Asn GlnAsn Leu LeuGln SerSerLeu ProGluAspThrAlaThrTyrTyrCysAlaArg	
	<b>C3</b>	versus	----- TTCA C TCTCCAGAG CA TCCC A CAT CT TACCT CAGATGA CCT A A CTGAGGACAC G CCA	----- TATTACTG GC AGA
	<b>S107</b>		CGGTTTACATGCTCTCACAGAGCACCTCCAAAGCATCTTACCTICAGATGAATGCGCTGAGGCTGGAGGACACTGCGATTTACTGCGGAGA	CGGTTTACATGCTCTCACAGAGCACCTCCAAAGCATCTTACCTICAGATGAATGCGCTGAGGCTGGAGGACACTGCGATTTACTGCGGAGA
	420	430    440    450	460	470    480    490    500
				510    515
			<-ArgPheValSerArgThrSerGlnAsnLeuLeuPheLeuGlnMetAsnAlaLeuArgAlaGluAspThrAlaIleTyrTyrCysAlaArg>	
			69    70    75    80    85    90    95    100	
			Phe SerArg SerGln Leu TyrLeuGlnMet Leu GluAspThrAla TyrTyrCysAlaArg	
common			TCACCATCTCCAG G CAA CC G C T GCTG CCTG A AT AGC CCT A CC GAGGACAC G C	----- TAT ACTCG G AG
bases			TCA C TCTCCAG G CAA CC C T C G C T A A T A CCT A C G GAGGACAC G C	----- TAT ACTCG G AG
common			ThrIleSerArg Asn Leu Leu Ser Leu ProGluAspThr Tyr Cys Arg	
acids			SerArg	
			<-FR3 (third framework region) segment>	

alignments. The symbol for each identical nucleotide pair is repeated between the two sequences. Each nonidentical nucleotide pair costs 1 base change and each nucleotide aligned with a null ("—"), corresponding to an insertion-deletion event, costs 2 base changes (26). The *alignments* line shows a dash ("—") for each position present in all metric alignments; positions having a blank space are not metrically aligned. Aligned stretches having a number of equally best alignments are indicated by that number followed by a string of colons (e.g., 4::: denotes four alternative alignments in a stretch of six alignment positions). The *codon phase* line indicates when 2 nucleotide sequences are in or out of phase in the single metric alignment shown. In *a*, the *Common/consensus* line indicates positions having all three (*N*), two (*n*), or no (.) shared nucleotides. In *b*, *common bases* and *common acids* refer to nucleotides and amino acids shared by the caiman sequences (top row) and all four sequences (bottom row). In *a*, functionally (transcription or recombination) significant (*notable*) segments are noted by asterisks. In the IVS of *E1*, an inverted repeat and a RS segment (heptamer/spacer/nonamer) are noted. Extended deletions are present in the 5' segments of *E1* (after 694 and 1311) and *G4* (after 330).

erations have led to classification of  $\approx 40\%$  of mammalian  $V_H$  sequences as pseudogenes (20, 33).

Recombination signal sequences are located 3' to the coding segments of all mammalian  $V_H$  and  $V_L$  genes. Both  $C3$  and  $G4$  sequences match the consensus recombination 7-mer, C-A-C-A-G-C-G, possess a 23-bp spacer and have typical dA>dC nonamers. The  $E1$  3' 7-mer, C-A-C-T-G-T-G, is an inverse complement of the mammalian  $C3$ ,  $G4$  prototype. In addition, the sequence is identical to 7-mers located at the 5' side of  $D$  elements (9, 10, 34, 35) and matches the consensus for human  $J_H$ ,  $J_K$ , and  $J_\lambda$  recombination elements (9, 35). An identical sequence has been detected 33 bp 3' of the prototypical recombination 7-mer of murine  $V_H^{441}$  (17) and is present at the site of an aberrant joining of a murine  $J_{\kappa}1$  to the nonimmunoglobulin gene segment  $L10$  (36). The  $E1$  3' 9-mer (A-C-A-A-A-A-C-C) is identical to the mammalian  $V_H$  prototypes; however, the spacer segment is only 12 bp, typical of  $V_K$  (and  $D$ ) but not  $V_H$  (or  $V_\lambda$ ) recombination elements (9). A spacer segment deletion similar to the  $E1$  3' structure has been described for a pseudogene member of the  $T15$  family (37). As noted above (Figs. 1 and 2a), a  $V_H$ -like recombination element has also been detected within the IVS of  $E1$ . Southern blotting of restriction endonuclease-digested parent phage and plasmid subclones relative to genomic DNA isolated from several caiman genes (data not illustrated) indicates that neither 5' nor 3' structures arose as cloning artifacts. The close relationship of these sequences is illustrated.



The two recombination elements could result in either atypical recombination (5') or inability to undergo typical somatic reorganization (3'). The 5' element could facilitate 5' leader  $\rightarrow D \rightarrow J_H$  joining (assuming that these structures exist in caiman), although the abbreviated, 21 bp, spacer of the 5' element may not be functionally active. According to the 12/23 recombination rule, the 3' spacer segment deletion would preclude typical  $V-D-J$  joining; however,  $E1$  could participate in a direct  $V-J$  (light chain-like) interaction bypassing  $D$  joining and preserving the spacing rule.

Near-perfect direct repeats (831–839:946–954; 843–852:956–965) flanking the presumably nonfunctional 5' recombination signal segment suggest that a recent transposition-like event (38) may have occurred within the IVS. In addition, the 7-mer is part of the 14-bp inverted repeat and contributes to one of the direct repeats (843–852) (Fig. 2a). The 3' segment may have originated through homologous recombination involving caiman equivalents of  $V$ ,  $D$ , or  $J$  segments.

Taken together these data emphasize the extended evolutionary history of this gene family and its associated reorganization mechanism. Evolution appears to preserve a core sequence and favors considerable diversification within the coding segments. The presence of atypical recombination sequences in one of the genes suggests that different alternatives to  $V-D-J$  joining may operate within this gene family and that these elements may be capable of recombining in previously unanticipated fashions.

We thank J. Sekulski, R. Ho, and P. H. Sellers for assistance in devising and programming the metric algorithms and L. Hood for providing the murine  $V_H$  probe. This work was supported by National Institutes of Health Grants CA 08748 and GM 32106 (to G.W.L.) and GM 32622 (to B.W.E.).

1. Davis, M. M., Calame, K., Early, P. W., Livant, D. L., Joho, R., Weissman, I. L. & Hood, L. (1980) *Nature (London)* **283**, 733–739.
2. Sakano, H., Maki, R., Kurosawa, Y., Roeder, W. & Tonegawa, S. (1980) *Nature (London)* **286**, 676–683.
3. Early, P., Huang, H., Davis, M., Calame, K. & Hood, L. (1980) *Cell* **19**, 981–992.
4. Rabbitts, T. H., Matthysse, G. & Hamlyn, P. H. (1980) *Nature (London)* **284**, 238–243.
5. Kemp, D. J., Tyler, B., Bernard, O., Gough, N., Gerondakis, S., Adams, J. & Cory, S. (1981) *J. Mol. Appl. Genet.* **1**, 245–261.
6. Crews, S., Griffin, J., Huang, H., Calame, K. & Hood, L. (1981) *Cell* **25**, 59–66.
7. Bothwell, A. L. M., Paskind, M., Reth, M., Imanishi-Kari, T., Rajewsky, K. & Baltimore, D. (1981) *Cell* **24**, 625–637.
8. Kim, S., Davis, M., Sinn, E., Patten, P. & Hood, L. (1981) *Cell* **27**, 573–581.
9. Tonegawa, S. (1983) *Nature (London)* **302**, 575–581.
10. Sakano, H., Kurosawa, Y., Weigert, M. & Tonegawa, S. (1981) *Nature (London)* **290**, 562–565.
11. Sakano, H., Hüpi, K., Heinrich, G. & Tonegawa, S. (1979) *Nature (London)* **280**, 288–294.
12. Max, E. E., Seidman, J. G. & Leder, P. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 3450–3454.
13. Cohen, J. B., Effron, K., Rechavi, G., Ben-Neriah, Y., Zakut, R. & Givol, D. (1982) *Nucleic Acids Res.* **10**, 3353–3370.
14. Givol, D., Zakut, R., Effron, K., Rechavi, G., Ram, D. & Cohen, J. B. (1981) *Nature (London)* **292**, 426–430.
15. Kataoka, T., Nikaido, T., Miyata, T., Moriwaki, K. & Honjo, T. (1982) *J. Biol. Chem.* **257**, 277–285.
16. Loh, D. Y., Bothwell, A. L. M., White-Scharf, M. G., Imanishi-kari, T. & Baltimore, D. (1983) *Cell* **33**, 85–93.
17. Ollo, R., Auffray, C., Sikorav, J.-L. & Rougeon, F. (1981) *Nucleic Acids Res.* **9**, 4099–4109.
18. Ollo, R., Sikorav, J.-L. & Rougeon, F. (1983) *Nucleic Acids Res.* **11**, 7887–7897.
19. Matthysse, G. & Rabbitts, T. H. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 6561–6565.
20. Ram, D., Benneriah, Y., Cohen, J. B., Zakut, R. & Givol, D. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 4405–4409.
21. Rechavi, G., Ram, D., Glazer, L., Zakut, R. & Givol, D. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 855–859.
22. Ohno, S., Matsunaga, T. & Wallace, R. B. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 1999–2002.
23. Litman, G. W., Berger, L. & Jahn, C. L. (1982) *Nucleic Acids Res.* **10**, 3371–3380.
24. Litman, G. W., Berger, L., Murphy, K., Litman, R., Hinds, K., Jahn, C. L. & Erickson, B. W. (1983) *Nature (London)* **303**, 349–352.
25. Sellers, P. H. (1980) *J. Algorithms* **1**, 359–373.
26. Erickson, B. W. & Sellers, P. H. (1983) in *Time Warps, String Edits, and Macromolecules: Theory and Practice of Sequence Comparison*, eds. Sankoff, D. & Kruskal, J. B. (Addison-Wesley, Reading, MA), pp. 55–91.
27. Mount, S. M. (1982) *Nucleic Acids Res.* **10**, 459–472.
28. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. & Chambon, P. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 4853–4857.
29. Clarke, C., Berenson, J., Goverman, J., Boyer, P. D., Crews, S., Siu, G. & Calame, K. (1982) *Nucleic Acids Res.* **10**, 7731–7749.
30. Parslow, T. G., Blair, D. L., Murphy, W. J. & Granner, D. K. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 2650–2654.
31. Falkner, F. G. & Zachau, H. G. (1984) *Nature (London)* **310**, 71–74.
32. Corden, J., Wasylky, B., Buchwalder, A., Sassone-Corsi, P., Kedinger, C. & Chambon, P. (1980) *Science* **209**, 1406–1414.
33. Cohen, J. B. & Givol, D. (1983) *Eur. Mol. Biol. Organ. J.* **2**, 1795–1800.
34. Siebenlist, U., Ravetch, J. V., Korsmeyer, S., Waldmann, T. & Leder, P. (1981) *Nature (London)* **294**, 631–635.
35. Ravetch, J. V., Siebenlist, U., Korsmeyer, S., Waldmann, T. & Leder, P. (1981) *Cell* **27**, 583–591.
36. Höchtl, J. & Zachau, H. G. (1983) *Nature (London)* **302**, 260–262.
37. Huang, H., Crews, S. & Hood, L. (1981) *J. Mol. Appl. Genet.* **1**, 93–101.
38. Calos, M. P. & Miller, J. H. (1980) *Cell* **20**, 579–595.