## Complete nucleotide sequences of three $V_{\rm H}$ 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<sup>†</sup>

\*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  $V_{\rm H}$  genes (C3, E1, and G4) that hybridize with a murine  $V_{\rm H}$  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  $V_{\rm H}$  genes because (i) the spacer region within the 3' recombination signal sequence is 12 base pairs, typical of  $V_{\kappa}$  genes but not of  $V_{\rm H}$ genes, which possess 22- to 23-base-pair spacers and (ii) a near-perfect  $V_{\rm H}$  recombination signal sequence is present within the intervening sequence that splits the segment encoding the leader. These studies establish  $V_{\rm H}$  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  $V_{\rm H}$  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 immunglobulin  $V_{\rm H}$  (2, 3, 6, 7, 9) and  $V_{\rm L}$  genes (9, 11, 12) and by complementary segments flanking the D,  $J_{\rm H}$ , and  $J_{\rm L}$  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  $V_{\rm H}$  genes largely have been restricted to comparisons between the members of 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 crosshybridization between restriction enzyme-digested caiman genomic DNA and murine  $V_{\rm H}$  probes (23) and demonstrated sequence similarities between caiman and prototypic mammalian  $V_{\rm H}$  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*)  $V_{\rm H}$  genes and identify unique recombination signal sequences associated with one of these genes.

## **MATERIALS AND METHODS**

The construction, amplification, and screening of a caiman- $\lambda$ 47.1 library with S107V, a murine  $V_{\rm H}$  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  $V_{\rm H}^+$  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  $V_{\rm H}$  (and  $V_{\rm L}$ ) 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  $V_{\rm H}$  genes. The G4 donor sequence cannot be accommodated into this particular consensus sequence, but it preserves the general  $A \cdot G/G \cdot T$  structure and is consistent with other functional splice donor sequences (27, 28).

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.

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  $V_{\rm H}^+$  caiman genomic DNA- $\lambda$  clones IVD (a) and VIIIB (b) with BamHI (I), HindIII ( $\phi$ ), Sma I ( $\dot{O}$ ), and Sst I ( $\dot{\Box}$ ) localizing the G4 ( $\Box$ ) and El ( $\boxtimes$ ) (a) and the C3 ( $\Box$ ) (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 ( $5' \rightarrow$ 3') a hyperconserved upstream octamer and the putative promoter for RNA polymerase II transcription (1), interrupted leader ( $\blacksquare$ ), mature coding ( $\Box$ ), and recombination signal sequences ( $\boxtimes$ ) 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 (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 HinfI 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  $V_{\rm H}$ 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



(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,

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  $V_{\rm H}$  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)  $V_{\rm H}$  genes representing different subgroups and families. Overall the three caiman sequences are closest to human  $V_{\rm H}$ 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  $\stackrel{A}{C}$ -T-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  $V_{\rm H}$ 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-

а alignments: ---2:------2:----2:-2:---------6::::::-comm COMMON/consensus: notable: TATaaGG,gggGgacATGCAAATc,c,,,CaG,,t,,gCtgcaTAAAT gCt tg,g g,t,cccaggCagCtgcTG,,c,,aCAgccc ,tGg,gTctG,t cAG Ig C3 alignments: C3 gene mmon bases: commo E1 gene codon phase: Ig E1 common acids: Ig C3 alignments: C3 gene common bases: G4 gene codon phase: Ig G4 -19 Met common acids: Ala Ig E1 alignments: E1 gene common bases: G4 gene codon phase Ig G4 common acids: geg, TGA CCAge , c, Ca, GAGageTeCCteCAGeaCCaG, , aaCATG, , g, ceTGGg, at, attT, CTCtTeCTe, tegeagCe, Tge, AGGTa, GTCatCT, COMMON/consensus: notable: gene segment: <---LR (leader region) segment [start]-----[---IVS-----C3 gene gene segment: 828 840 850 860 870 880 890 900 910 920 930 GGACCTGCCACGTCCACGCCGTGGGAGCCCAGGGAGCACAAAATCCTCACTTGATGTAAAGGCTGATGGAAATCAGCAGCGGATGCCCCTTGGCGC E1 gene inverted repeat notable gene segment: 31 940 950 960 966 CGTGGCTTTGGCAAGCCTGCCAGGACAGAGCCCTGA ---IVS within LR segment------E1 gene gene segment: 453 460 470 480 490 500 510 519 CTCTGGGGCACCAGAGAGACACATCCCCGGTTTGGCTCTGTTTCACAGGCTCCCAGGAGATTTC ---IVS within LR segment-----G4 gene gene segment: Ig C3 alignments common acids: -4 -1 1 || 98 lyAlaTrpSerGln...|...Arg 

Ig G3
-4
-1
1
90
-5
-7

alignments
597
610
620
620
918
930
940
950
961

C3 gene
ATCCTAC C TTCCTTTTCAGGCGCTCTGGTCCCAG...
... ACAGACACAGTGACTCAACACCTATCACGCAATACAAATCC
G4 gene
AC T CTC T CTCTTTGAGGCGTCCTCGGAG...
... ACGACACAG GA AAA C T GC A CAAA C T
GC A CAAA C T
GC A CAAA C T
GC A CAAA C T
SC A CACAG GA AAA C T
GC A CAAA C T
SC A CACAG GA AAA C T
GC A CAAA C T
SC A CACAG GA AAA C T
GC A CAAA C T
SC A CACAG GA AAA C T
GC A CAAA C T
SC A CACAG CAAAACTCCCCTCTGAGGCAAAACT
S4 + 3
S6 800
870
880 884
S60
870
880 884
S60
1
1
98
G4: 23-bp spacer
1
98
G4: 23-bp spacer
1
98
G4: 23-bp spacer
1
1
S6
S6</td C3: 23-bp spacer common acids: codon phase Ig G4 common acids: COMMON/consensus: AtCgT,etC,TttCTtt,CAGGtGeec,gTCecAG || AGg,aCACaGtGag,aAAa,,c,,t,,,,GgC,AgaCAAAAaCc gene segment

FIG. 2. DNA sequence alignments of the caiman C3, E1, and G4 and murine S107 (3)  $V_{\rm H}$  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> c3	caiman Ig C3	1 5 GlnValGlnLeuValGluSe	10 GlyGlyAspValArg	15 LysProGlyAsn: 670	20 SerLeuArgLeuSe 680 6	25 rCysLysAlaSerGlyPh 90 700	30 eThrPheGly 710 716	
Ý	caiman gene C3	CAGGTGCAGCTGGTGGAGTC	CGGAGGAGATGTGAGG	AAACCTGGAAAC	тстттбсосстсто	ĊŤĠĊĂĂĂĠĊĊŤĊĠĠĠĠŦŦ	CACCTTCGGT	
versus E1	common bases: caiman gene El	CAGGTGCAGCTGGTGGAGTC CAGGTGCAGCTGGTGGAGTC 999 1010 10	GGAGG GATGT AGO TGGAGGGGGATGTAAGO 20 1030	AA CCTGGA AC AAGCCTGGAGAC 1040	TCT TGCGCCTCTC TCTCTGCGCCTCTC 1050 1060	CTGCAAAG CTC GG TT CTGCAAAGGCTCCGGCTT 1070 10	CACCTTC G CACCTTCAGC 080 1088	
	caiman Ig E1 common acids:	GlnValGlnLeuValGluSe 1 GlnValGlnLeuValGluSe	rGlyGlyAspValArg 10 rGlyGlyAspValArg	LysProGlyAsp 15 LysProGly	SerLeuArgLeuSe 20 SerLeuArgLeuSe	erCysLysGlySerGlyPt 25 erCysLys SerGlyPt	neThrPheSer 30 neThrPhe	
C3 versus G4	alignments: common bases: caiman gene G4	AG T CAGCTGGTGGAGTC GAGATCCAGCTGGTGGAGTC 550 560 57	CGGAGGAG T AGO CGGAGGAGCCATAAGO 0 580	GAAACCTGGA AC GAAACCTGGAGAC 530	TC TGCGCCTCTC TCCCTGCGCCTCTC 600 610	CCTGCAAAGCCTC GGGT CCTGCAAAGCCTCTGGGT 620 63	TCAC TTC GT TCACTTTCAGT 30 639	
	codon phase: caiman Ig G4 common acids:	<pre><in (+0)="" gluileglnleuvalgluse<="" th=""><th>rGlyGlyAlaIleArg 10 rGlyGly Arg</th><th>gLysProGlyAsp 15 gLysProGly</th><th>SerLeuArgLeuSe 20 SerLeuArgLeuSe</th><th>erCysLysAlaSerGlyPt 25 erCysLysAlaSerGlyPt</th><th>ne Thr Phe Ser 30 ne Thr Phe</th><th></th></in></pre>	rGlyGlyAlaIleArg 10 rGlyGly Arg	gLysProGlyAsp 15 gLysProGly	SerLeuArgLeuSe 20 SerLeuArgLeuSe	erCysLysAlaSerGlyPt 25 erCysLysAlaSerGlyPt	ne Thr Phe Ser 30 ne Thr Phe	
C3 🛔	alignments:			• • • • • • • • • • • • • • • • • • • •			CACCTTC CT	
S107	mouse gene S107	AGGIG AGCIGGIGGA IC GAGGTGAAGCTGGTGGAATC 216 220 230	TGGAGGAGGCTTGGT 240 250	ACAGCCTGGGGGT 260	TCTCTGAGACTCT	280 290	TCACCTTCAGT 300 305	
	common acids:	GluValLysLeuValGluSe 1 5 Val LeuValGluSe	rGlyGlyGlyLeuVa 10 rGlyGly	lGlnProGlyGly 15 ProGly	SerLeuArgLeuS 20 SerLeuArgLeuS	erCysAlaThrSerGlyPi 25 erCys SerGlyPi	neThrPheSer 30 neThrPhe	
common bases	C3,E1,G4 : C3,E1,G4,S107:	AG T CAGCTGGTGGAGTC Ag t Agctggtgga to	GGAGG G T AG GGAGG G T	GAA CCTGGA AC A CCTGG	TC TGCGCCTCT TC TG G CTCT	CCTGCAAAG CTC GG T CCTG A TC GG T	TCAC TTC G TCAC TTC G	
common acids	C3,E1,G4 : C3,E1,G4,S107:	GlnLeuValGluSe LeuValGluSe	rGlyGly Ar. rGlyGly	gLysProGly ProGly	SerLeuArgLeuS SerLeuArgLeuS	erCysLys SerGlyP erCys SerGlyP	heThrPhe heThrPhe	
gene seg	ment (Ig region):	<fr1 (first="" frame<="" th=""><th>work region) se</th><th>gment</th><th></th><th></th><th>&gt;</th><th></th></fr1>	work region) se	gment			>	
C3	31 35 GlyTyrGlyMetPhe 717 310 731 GGCTACGGCATGTTC	36 40 TrpValArgG1nAlaProG 732 740 750 TGGGTCCGCCAGGCTCCTG	45 .yLysGlyLeuAspTr 760 gGAAggggCTggACTg	49 50 pValAla Thr 773 774 GGTGGCT AC	) -IleA snTh rA 780 AATTA ATAC TG	55 sp GlySerSerGlnTr 790 800 AT GGATCCAGCCAGTG	60 pTyrSerProAlaV 810 -GTACTCCCCGGCCG	65 66 (alGinGly 820 824 STTCAGGGG
C3 versus E1	TAC G TG C AATTACTGGCTGGGC 1089 330 1104	TGGGTCCG CAGGCTCC G TGGGTCCGTCAGGCTCCCG 1110 1120	GGAAGG CT GA TG GGAAGGC-CTTGAATG 1130 11	G T CT C GATCTCT GC 40 1145 1	ATT A AC CATTG ACAC CT 150 11	T GG CAGC T CT GGCAGCAGCACCTA 60 1170	TAC CCC G G CTTACATCCCTGGAG	T GGG TGACTGGG 190 1196
	AsnTyrTrpLeuGly 31 35 Tyr	TrpValArgGlnAlaProG 36 40 TrpValArgGlnAlaProG	LyLysAl aLeuAsnG 45 LyLys	(+1) lySerLe uP 48	roLeu ThrP ro 50	Le uAlaAlaAlaProT 55	hrTyrIleProGlyv 60 Tyr Pro V	ValSerGly 65 66 Val Gly
C3 versus G4	G C C G ATG C Gacacctggatggcc 640 650 654	TGGG CCG CAG C CCTG TGGGCCCGGCAGCCCCCTG 655 660 670	GGAAGGGGGCTG A TO GGAAGGGGGCTGCAGTO 680 69	GGT G T GGTTGGT GA 0 696 697	AAT A AT AATCA ATGG GA 70	A A A C C AC TCAGAGACCATCAC 00 710	TA C CC G C ATATGCACCAGAAC 720 73	GT A GG GTGAAAGGT 30 747
	AspThrTrpMetAla 31 35 Met	TrpAlaArgGlnProProG 36 40 Trp ArgGln ProG	lyLysGlyLeuGlnTr 45 lyLysGlyLeu Tr	pValGly G1 49 5 pVal	uIleA snGl y# 0 IleA sn	sn SerGluThrIleAr 55	gTyrAlaProGluv 60 Tyr Pro V	ValLysGly 65 66 Val Gly
C3 versus S107	G T C CATG GATTTCTACATGGAG	TGGGTCCGCCAG CTCC G TGGGTCCGCCAGCCTCCAG	GGAAG G CTGGA TO GGAAGAGACTGGAGTO	GG T GCT C GGATTGCT GC 362 363	15:: AA TA A AC C AAGTAGAAACAAAG 370	G T GAT C A C G T GAT C A C G C T A A T G A T T A T A C A A C A C 390	G GTAC C C C GAGTACAGTGCATCTC 400 410	GT AGGG GTGAAGGGT 419
	AspPheTyrMetGlu 31 35 Met	TrpValArgGlnProProG 36 40 TrpValArgGln ProG	lyLysArgLeuGluTr 45 lyLys Leu Tr	pIleAla Al 49 5 p Ala	aSerArgAsnLys 0	-3> <out (-5)-2<br="">AlaAsnAspTyrThrThr( 55 60</out>	SluTyrSerAlaSer 65 TyrSer	ValLysGly 68 Val Gly
common bases	CGTGC CTG	TGGG CCG CAG C CC G TGGG CCG CAG C CC G	GGAAGG CT A TO GGAAG CT A TO	GGT T GGT T	AT A A A	A C C	TA CCG TA C	GT GG GT GG
common acids	1	TrpValArgGln ProC TrpValArgGln ProC	lyLys lyLys				Tyr Pro Tyr	Val Gly Val Gly
	<-CDR1 segment>	<fr2 (second="" fra<="" th=""><th>mework region)</th><th>segment&gt; <cd< th=""><th>R2 (second com</th><th>nplementarity-dete</th><th>rmining region)</th><th>segment:</th></cd<></th></fr2>	mework region)	segment> <cd< th=""><th>R2 (second com</th><th>nplementarity-dete</th><th>rmining region)</th><th>segment:</th></cd<>	R2 (second com	nplementarity-dete	rmining region)	segment:
C3	↓ ¥	67 70 LysPheThrIleSerArg( 825 830 840 AAATTCACCATCTCCAGA(	75 1 y Asn Ser Gln Asn M 850 8 GCAACTCCCAGAACA	80 etLeuTyrLeuG1 60 870 TGCTGTACCTGCA	85 nMetSerSerLeu 880 IGATGAGCAGCCTC	90 ThrProGluAspThrAla 890 900 ACACCTGAGGACACAGCC	95 ThrTyrTyrCysAla 910 ACGTATTACTGCGCC	98 Arg 920 AGA
C3 versus E1		TTCACCATCTCCAG C CGCTTCACCATCTCCAGGO 1197 1210	CAA CC G C ACAATGCCAGGGCCT 1220 1230	TGCTG ACCTG A TGCTGCACCTGGA 1240	ATGAGC CCT CATGAGCGACCTG 1250	A CCTGAGGACAC G C AGGCCTGAGGACACCGGC 1260 1270	TAT ACTGCG CGATATCACTGCGAG 1280 1	AG AGG 292
		ArgPheThrIleSerArg 67 70 PheThrIleSerArg	spAsnAlaArgAlaL 75 Asn	euLeuHisLeuAs 80 Leu Leu	pMetSerAspLeu 85 MetSer Leu	ArgProGluAspThrGly 90 ProGluAspThr	ArgTyrHisCysGlu 95 Tyr Cys	Arg
C3 versus G4		A A TCACCATCTCCAGA AGACTCACCATCTCCAGA 748 760	CAAC CCCAGAAC ACAACACCCAGAACC 770 780	TGCTGT CCTGC/ TGCTGTTCCTGC/ 790	AGAT AGCAGCCTC Agataagcagcctc 800	A ACC GAGGACACAGCC AAACCCGAGGACACAGCC 810 820	ACGTATTACTG GC ACGTATTACTGTGCA 830	AG AGG 843
		ArgLeuThrIleSerArg, 67 70 ThrIleSerArg	Asn GlnAsn 75 Asn GlnAsn	euLeuPheLeuG 80 Leu LeuG]	InIleSerSerLeu 85 In SerSerLeu	LysProGluAspThrAla 90 ProGluAspThrAla	ThrTyrTyrCysAla 95 ThrTyrTyrCysAla	Arg 98 Arg
C3 versus S107	<b>A</b>	TTCA C TCTCCAGA CGGTTCATCGTCTCCAGA 420 430	G CA TCCCA A CA Gacacttcccaaagca 140 450	T CT TACCT CA TCCTCTACCTTCA 460	AGATGA CCT Agatgaatgccctg 470 48	A A CTGAGGACAC GCC Agagctgaggacactgcc 0 490	A TATTACTG GC ATTTATTACTGTGCA 500 510	AGA AGA 515
		ArgPheIleValSerArg 69 70 Phe SerArg	AspThrSerGlnSerI 75 SerGln	leLeuTyrLeuG 80 LeuTyrLeuG	InMetAsnAlaLeu 85 InMet Leu	ArgAlaGluAspThrAla 90 GluAspThrAla	IleTyrTyrCysAla 95 TyrTyrCysAla	Arg 100 Arg
common bases		TCACCATCTCCAG TCA C TCTCCAG	G CAA CC G C G CA CC C	ТGCTG ССТС Т СТ ССТ	A AT AGC CCT A AT A CCT	A CC GAGGACAC G C A C GAGGACAC G C	TAT ACTG G TAT ACTG G	AG Ag
common acids		ThrIleSerArg SerArg	Asn	Leu Leu Leu Leu	Ser Leu Leu	ProGluAspThr GluAspThr	Tyr Cys Tyr Cys	Arg Arg
		<fr3 (third="" fram<="" td=""><td>nework region) s</td><td>egment</td><td></td><td></td><td></td><td>&gt;</td></fr3>	nework region) s	egment				>

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  $V_{\rm H}$  sequences as pseudogenes (20, 33).

Recombination signal sequences are located 3' to the coding segments of all mammalian  $V_{\rm H}$  and  $V_{\rm L}$  genes. Both C3 and G4 sequences match the consensus recombination 7-mer, C-A-C-A-G- $^{T}_{C}$ -G, possess a 23-bp spacer and have typi-cal 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  $J_{\rm H}$ ,  $J_{\kappa}$ , and  $J_{\lambda}$  recombination elements (9, 35). An identical sequence has been detected 33 bp 3' of the prototypical recombination 7-mer of murine  $V_{\rm H}^{441}$  (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' 9mer (A-C-A-A-A-A-C-C) is identical to the mammalian  $V_{\rm H}$  prototypes; however, the spacer segment is only 12 bp, typical of  $V_{\kappa}$  (and D) but not  $V_{\rm H}$  (or  $V_{\lambda}$ ) recombination elements (9). A spacer segment deletion similar to the El 3' structure has been described for a pseudogene member of the T15 family (37). As noted above (Figs. 1 and 2a), a  $V_{\rm H}$ 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.

		heptamer					21-bp spacer					nonamer			
		<><										><>			
	81	13	8	50					862	2		87	1	ł	879
E 1	IVS	CAC	AGC	ССТ	GI	GG	AG	сс	CAC	GGGA	GCA	AC	ACA.	AAA'	тсс
соп	nmon	CAC	G	С	G	G	AG	cc	CAC	G			ACA.	AAA	CC
E1	3'	CAC	TGT	GCG	G	GA	AG	сс	CAC	G			ACA.	AAA.	ACC
	129	95	13	02				1	31	3	1	31	4	1	322
<><>											>				
	heptamer				12-bp spacer					r	nonamer				
		<-R	<b>s</b> (	rec	on	ıbi	na	ti	on	sig	nal	)	seg	men	t->

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  $\rightarrow D \rightarrow J_{\rm H}$  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, El 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  $V_{\rm H}$  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.).

- 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., Matthyssens, G. & Hamlyn, P. H. (1980) Nature (London) 284, 238-243.
- Kemp, D. J., Tyler, B., Bernard, O., Gough, N., Gerondakis, S., Adams, J. & Cory, S. (1981) J. Mol. Appl. Genet. 1, 245-261.
- Crews, S., Griffin, J., Huang, H., Calame, K. & Hood, L. 6. (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.
- Sakano, H., Kurosawa, Y., Weigert, M. & Tonegawa, S. 10. (1981) Nature (London) 290, 562-565.
- 11. Sakano, H., Hüpi, K., Heinrich, G. & Tonegawa, S. (1979) Nature (London) 280, 288-294.
- Max, E. E., Seidman, J. G. & Leder, P. (1979) Proc. Natl. 12. Acad. Sci. USA 76, 3450-3454.
- Cohen, J. B., Effron, K., Rechavi, G., Ben-Neriah, Y., Zakut, 13. 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.
- Ollo, R., Auffray, C., Sikorav, J.-L. & Rougeon, F. (1981) Nu-17. cleic Acids Res. 9, 4099-4109.
- 18. Ollo, R., Sikorav, J.-L. & Rougeon, F. (1983) Nucleic Acids Res. 11, 7887-7897.
- Matthyssens, G. & Rabbitts, T. H. (1980) Proc. Natl. Acad. 19. 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. Rechavi, G., Ram, D., Glazer, L., Zakut, R. & Givol, D.
- 21. (1983) Proc. Natl. Acad. Sci. USA 80, 855-859.
- Ohno, S., Matsunaga, T. & Wallace, R. B. (1982) Proc. Natl. 22. Acad. Sci. USA 79, 1999–2002.
- Litman, G. W., Berger, L. & Jahn, C. L. (1982) Nucleic Acids 23. Res. 10, 3371-3380.
- Litman, G. W., Berger, L., Murphy, K., Litman, R., Hinds, 24. K., Jahn, C. L. & Erickson, B. W. (1983) Nature (London) 303, 349-352.
- Sellers, P. H. (1980) J. Algorithms 1, 359-373. 25
- 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.
- Mount, S. M. (1982) Nucleic Acids Res. 10, 459-472. 27.
- 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.
- Parslow, T. G., Blair, D. L., Murphy, W. J. & Granner, D. K. 30. (1984) Proc. Natl. Acad. Sci. USA 81, 2650-2654.
- 31. Falkner, F. G. & Zachau, H. G. (1984) Nature (London) 310, 71-74.
- Corden, J., Wasylyk, B., Buchwalder, A., Sassone-Corsi. P.. 32. Kedinger, C. & Chambon, P. (1980) Science 209, 1406-1414.
- 33. Cohen, J. B. & Givol, D. (1983) Eur. Mol. Biol. Organ. J. 2, 1795-1800.
- Siebenlist, U., Ravetch, J. V., Korsmeyer, S., Waldmann, T. 34. & 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.
- Höchtl, J. & Zachau, H. G. (1983) Nature (London) 302, 260-36. 262.
- 37. Huang, H., Crews, S. & Hood, L. (1981) J. Mol. Appl. Genet. 1. 93-101
- Calos, M. P. & Miller, J. H. (1980) Cell 20, 579-595. 38.