The EMBO Journal vol.7 no.11 pp.3413-3422, 1988

Diverse organization of immunoglobulin V_H gene loci in a primitive vertebrate

F.Kokubu, R.Litman, M.J.Shamblott, K.Hinds and G.W.Litman

Showa University Research Institute, 10900 Roosevelt Boulevard, St Petersburg,FL 33716, USA

Communicated by R.Perry

The immunoglobulin (Ig) heavy chain variable (V_H) gene family of *Heterodontus francisci* (horned shark), a phylogenetically distant vertebrate, is unique in that V_H, diversity (D_H), joining (J_H) and constant region (C_H) gene segments are linked closely, in multiple individual clusters. The V regions of 12 genomic (liver and gonad) DNA clones have been sequenced completely and three organization patterns are evident: (i) $V_H-D_1-D_2-J_H-C_H$ with unique 12/22 and 12/12 spacers in the respective D recombination signal sequences (RSSs); V_H and J_H segments have 23 nucleotide (nt) spacers, (ii) $V_HD_H-J_H-C_H$, an unusual germline configuration with joined V_H and D_H segments and (iii) $V_HD_HJ_H-C_H$, with all segmental elements being joined. The latter two configurations do not appear to be pseudogenes. Another $V_H-D_1-D_2-J_H-C_H$ gene possesses a D_1 segment that is flanked by RSSs with 12 nt spacers and a D_2 segment with 22/12 spacers. Based on the comparison of spleen, V_H^+ cDNA sequences to a germline consensus, it is evident that both D_H segments as well as junctional and N-type diversity account for Ig variability. In this early vertebrate, the Ig genes share unique properties with higher vertebrate T-cell receptor as well as with Ig and may reflect the structure of a common ancestral antigen binding receptor gene.

Key words: V_H gene organization/VDJ joining/N-diversity/antigen binding receptor gene/evolution of diversity

Introduction

Immune diversity mediated by both immunoglobulins (Igs) and T-cell receptors (TCRs) involves rearrangement of different genetic segments that are separated by varying chromosomal distances (Tonegawa, 1983; Alt *et al.*, 1986). While the variable (V) regions of these genes are encoded by related segmental elements, the structure and function of their constant (C_H) regions differ markedly. The different chromosomal organizations of Ig and TCR gene families may be associated with their unique functions as well as developmental and transcriptional regulation (Yancopoulos *et al.*, 1984; Yancopoulos and Alt, 1986; Chou *et al.*, 1987; Alt *et al.*, 1986). Significant levels of nucleotide (nt) sequence identity (Hedrick *et al.*, 1984; Yanagi *et al.*, 1984), overall organizational homology and similar mechanisms of gene rearrangement (Chien *et al.*, 1984; Siu *et al.*, 1984), however, suggest a common evolutionary origin for the two systems (Patten *et al.*, 1984; Hood *et al.*, 1985).

The mechanisms involved with the evolution of immune receptors have been examined by characterizing these genes in species that represent critical points in the vertebrate radiations. To date, efforts have centered on Ig V_H genes that are closely homologous to their higher vertebrate counterparts (Litman *et al.*, 1985b; Hinds and Litman, 1986). The most phylogenetically distant species that has been studied in terms of Ig gene structure and organization is *Heterodontus francisci*, an elasmobranch, that based on cladistic considerations is representative of the earliest extant lineage in the radiation of the jawed vertebrates. Previous studies showed that the humoral immune response of *Heterodontus* does not undergo affinity maturation and that the hapten-specific antibody response of genetically unrelated animals lacks fine specificity (Mäkelä and Litman, 1980; Litman *et al.*, 1982). While the sequences of Ig V_H genes in *Heterodontus* are related closely to those found in man and mouse, individual V_H, D_H, J_H and C_H segments are linked closely, ~10 kb; (Hinds and Litman, 1986). In order to characterize this system further, different primary and secondary screening strategies have been used to isolate genomic clones that have been subjected to extensive DNA sequence analysis. Based on these data, it is apparent that large numbers of germline genes are organized in an entirely unique manner not seen in any other vertebrate system. Furthermore, the *Heterodontus* Ig loci may reflect the common ancestral organization of both TCRs and Ig V_H genes.

Table I. Organization of V_H^+ *Heterodontus* genomic DNA-λ clones

$V-D_1-D_2-J$	$VD-J$	VDJ
1113[a]	1111	1101
1207[b]	1320	2809
1315	2806	F101
1403	3083	
2807	F301	

Variable (V), diversity (D_1, D_2) and joining (J) segments are indicated. In VD-J and VDJ type genes, D indicates a contribution from D_1 and/or D_2 segments. Numerical designations are individual clones recovered from an adult liver genomic DNA library; designations preceded by F are clones recovered from an adult gonadal DNA genomic library of a different animal. Liver DNA clones were selected with the V_HHXIA probe (Litman *et al.*, 1985b); gonadal DNA clones were selected with a V_H-specific probe derived from λ-2809, using conditions of moderate hybridization-wash stringency as described (Litman *et al.*, 1985b). Southern blot hybridization analyses were used to select clones from the gonadal DNA library that were potentially homologous, i.e. shared restriction sites with liver DNA clones, e.g. F101 and 1101 have identical sequences (see text); F301 and 3083 are closely related sequences (see text).

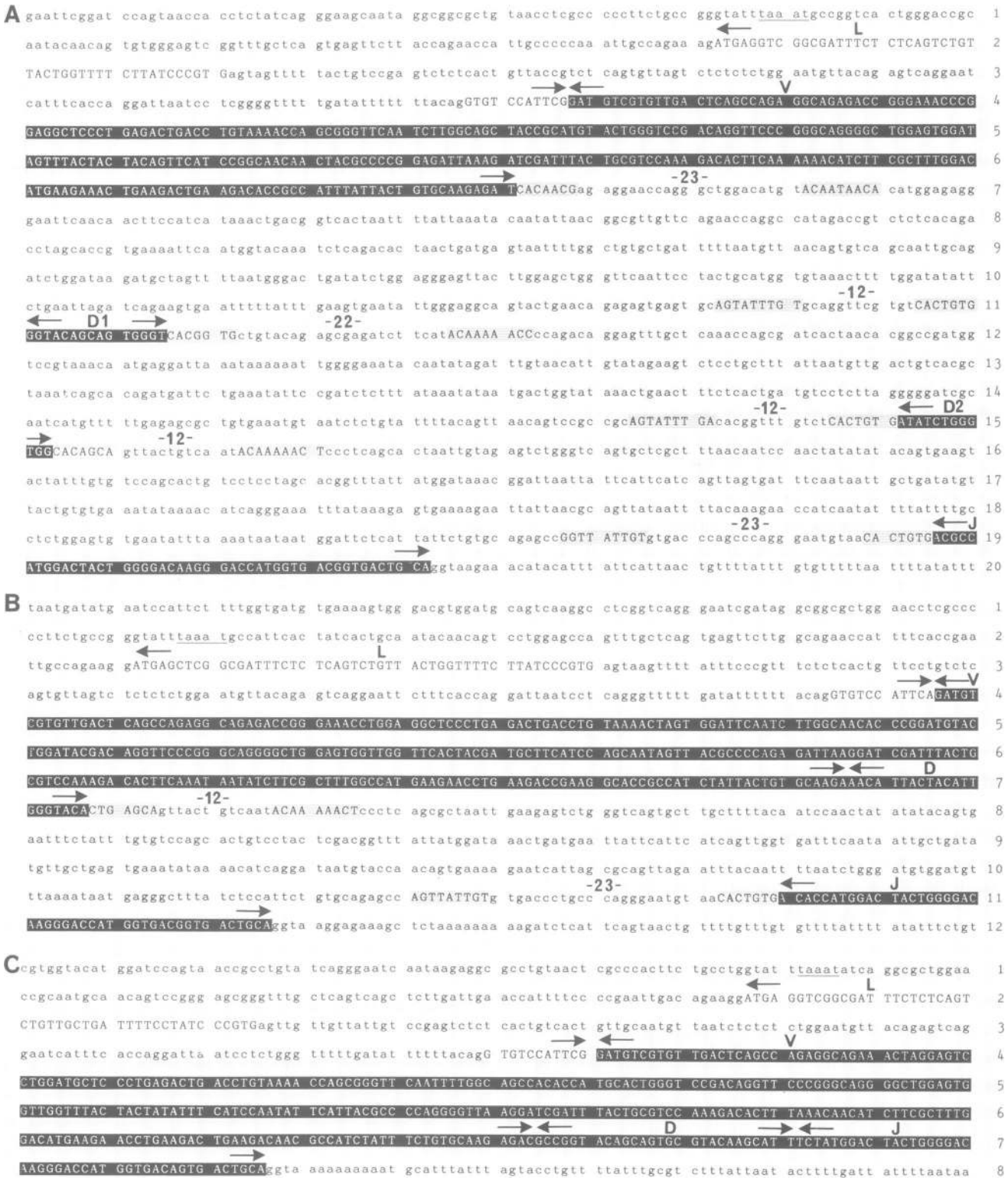
[a]1113 sequence is from amino acid position 19 in FR1 through the 3′ of J_H.

[b]1207 extends only through V_H and 3′ RSS.

Results

Estimation of V_H gene number

A *Heterodontus* V_H probe, V_HHXIA (Litman *et al.*, 1985b), was used to screen a *Heterodontus* liver DNA library ($\sim 6 \times 10^5$ p.f.u.), and 366 clones that hybridized at varying intensities were detected. Restriction mapping of 23 of these isolates showed 19 to be unique. A second probe derived from clone 2809, that hybridizes weakly with V_HHXIA, was used to screen a representative sampling of a second genomic library. Of 45 V_H^{2809+} clones, 33 also hybridized with V_HHXIA; 12 hybridized specifically with the V_H2809 probe. Of these, at least ten represent unique genes based on restriction mapping and Southern blotting. Presumably, additional genes would be detected by extending this type of analysis. These findings, previous genomic Southern blot analyses (Litman *et al.*, 1985b), and estimates from gene titrations (Kokubu *et al.*, 1987) are consistent with a family of V_H genes that most likely contains at least 200 individual members, including the possibility of allelic



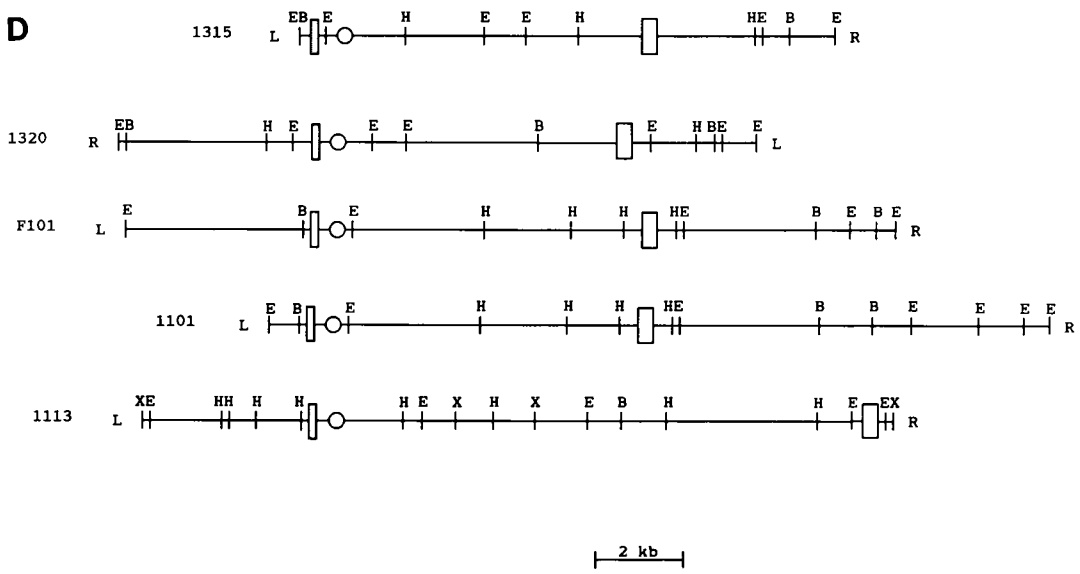


Fig. 1. Nucleotide sequences of representative genes: (**A**) V−D_1−D_2−J, 1315; (**B**) VD−J, 1320; (**C**) VDJ, F101. Predicted coding segments are shown in upper-case letters; the mature V_H coding segment is shown in reversed image lettering to distinguish this sequence from the leader. IVSs are shown in lower-case letters. Assignment of leader splice site is based on analyses of cDNAs and differs from that predicted in Litman *et al.*, 1985b. RSS 7mers and 9mers are indicated in upper-case, shaded lettering; the lengths of spacer segments are shown. 5′ taaat is underlined and major sequence regions are designated, except in (B) and (C) where D_1 and D_2 are joined; junctional boundaries are based on reference to consensus prototypes in Figure 3. Numbers at the right refer to 100 nt strings. (**D**) Restriction maps of genomic DNA-λ clones. The locations of restriction sites are shown; (B) *Bam*H1, (E) *Eco*RI, (H) *Hin*dIII, (X) *Xba*1. The peripheral E and X in 1113 and E in all other maps, are contained in the polylinker site of λ DASH. Coding segments V_H (□), J_H (○), C_H1 (□) are indicated in the center of the restriction fragments where they are localized. Right (R) and left (L) λ vector arms are indicated, scale is shown. F101 and 1101 represent closely related, overlapping clones that possess identical nt sequences in coding and through extensive IVS regions but differ at several peripheral restriction sites.

variants. All data obtained in the course of this and earlier studies (Hinds and Litman, 1986; Kokubu *et al.*, 1987), that now include 114 clones, are consistent with the invariant association of each V_H with J_H, C_H regions and presumably D_H regions, which cannot be probed directly.

Genomic organization of Heterodontus V_H genes

The sequences of all segmental elements and their respective intervening sequences (IVSs) in 12 of the 13 λ clones indicated in Table I have been determined. Clone 1315, Figure 1A, typifies the V−D_1−D_2−J-type segmented gene. The V_H, D_1 and J_H coding segments are homologous to the prototype shark gene HXIA (Litman *et al.*, 1985b; Hinds and Litman, 1986) and to the corresponding structures of mammalian genes. A second, putative D_H element, D_2, flanked by typical RSSs with symmetrical 12 nt spacers (see below), is located between D_1 and J_H. Clones 1403 and 2807 are organized similarly. The V−D_1 IVSs of 1315, 1403 and 2807 are 75% identical, including mismatches that arise from 2−8 nt sequence gaps. The segmental elements of an additional clone, 1113, also are organized as V_H−D_1−D_2−J_H. In this case, RSSs containing symmetrical 12 nt spacers flank the D_1 segment, whereas the D_2 segment is flanked by RSSs with 22 nt (5′) and 12 nt (3′) spacers (see below). With both types of genes, the V−D_1, D_1−D_2 and D_2−J IVS lengths are 306−382 nts. The V_H segment of clone 1207, which has been mapped to a λ arm, is flanked by a RSS with a 7mer−23 bp spacer−9mer RSS, consistent with a V−D_1−D_2−J organization pattern.

Five other clones, represented by gene 1320 in Figure 1B, have unusual extended V regions that lack the V_H 3′ RSS. The V_H segment appears to be joined with a sequence that resembles the D_H coding segments seen in the V−D_1−D_2−J-type genes and possesses a typical 3′ D_2 RSS (see below). Since clones 1111, 1320, 2806 and 3083 were isolated from a liver genomic DNA library, somatic rearrangement, perhaps associated with lymphopoiesis, could account for these structures. A genomic library, constructed from gonadal tissue of another specimen of *Heterodontus* was screened; F301, a V_H^{2809+} clone, recovered from this library was found to differ only at three positions in 1236 nt (extending 167 nt 5′ of the ATG initiation codon to 130 nt 3′ of J_H) from 3083. These differences are: (i) 7 nt 5′ to the conserved promoter region (TATA box) sequence TAAAT, (ii) 16 nt 3′ to the start of the leader region IVS and (iii) two nt 5′ to the J_H 7mer. The VD junctional sequence is identical. The complete sequences of these genes establish a germline VD−J organization and explain the previous failure of a V−D_1 IVS-specific probe to hybridize with a genomic $V_HJ_H^+$ clone (Hinds and Litman, 1986); i.e. this sequence would be deleted in a VD-joined gene.

A second clone, F101 recovered from the gonad DNA library exhibits a third unique form of genomic organization, Figure 1C. In this case, the V_H−D_1, D_1−D_2, D_2−J_H IVSs are deleted; a VDJ joining appears to have occurred. Clone 1101, isolated from the liver library, is identical (from 5′ of the initiation codon to 3′ of J_H) to F101. Clone 2809 which was recovered from the liver library and was used to derive a V_H-specific probe (see above), also is VDJ-joined but differs extensively from F101; see below predicted amino acid sequences. Restriction maps of 1315, 1320, F101, 1101 and 1113 are shown in Figure 1D.

Two types of V−D_1−D_2−J organization

Two different forms of V−D_1−D_2−J genes are distinguished by RSSs associated with D_1 and D_2 segments. Typical RSSs are located 3′ to the V_H segments of four V−D_1−D_2−J genes, Figure 2A. While the sequences of the 7mers differ, the 9mers as well as the 23 bp spacers of


A V
1315 CACAACGagaggaaccagggctggacatgtACAATAACA
1403 CACAGCGagaggaaccagggctggacccgtACAAGAACA
2807 CACAATGagaggaaccagggctggacccgtACAAGAACA
1113 CACTGCGacccaagcaaatcctgggctcgtACAGAAACA
D1
1315 AGTATTTGTgcaggttcgt gt CACTGTGGGTACAGCAGTGGGT CACGGTGctgtacagagcgagatcttcatACAAAAACC
1403 AGTATTTGTacagctttgaagtcCACTGTGGGTACAGCGGTGGGT CACGGTGctgtacagagcgagttcctcatACAAAAACC
2807 AGTATTTGTgctggttcgt gt CACTGTGGGTACAGCAGTGGGT CACGGTGctgtacagaacgagttcctcatACAAAAACC
1113 AGTATTTGTgcgggttcgt gt CACAGTGGGTACTACAGTGGGTATCACAGTGagacacaccgtgTCAAATACTgaggtggactatta
D2
1315 AGTATTTGAcacggtttgtctCACTGTG ATAT CTGGGTGGCACAGCAgttactgtcaatACAAAAACT
1403 AGTATTTGAcacggtttgtctCACTGTG ATAT CTGGATGGCACAGCAgttactgtcagtACAAAAAGT
2807 GGTATTTGAcacggtgtgtctCACTGTG ATATACTGGATGGCACAGCAgttactgtcaatACAAAAAGC
1113 GGTTTTTGTataagaaactcgctctgtacagTACCGTG TTATACTGGGTGGCACAGCAgttactgtcaatACAAAAAGT
J
1315 GGTTATTGTgtgacccagcccagggaatgtaaCACTGTG
1403 GGTTATTGTctgaaccagcccggggaatgtaaCACTGTG
2807 AGTTATTGTctgacccagcccagggaatgtaaCACTGTG
1113 GGTTATTGTgtgacccagcccagggaatgtaaCACTGTG
B ACTTTTT gtattgacagta ACTGCTG TGCCACCCAGTATAA-- CACGGTA ctgtacagagcgagtttcttat ACAAAAACC C 1113IC
AGTATTTGT gcaggttcgtgt CACTGTG GGTACAGCAGTGGGT-- CACGGTG ctgtacagagcgagatcttcat ACAAAAACC C 1315
AGTATTTGT gcgggttcgtgt CACAGTG GGTACTACAGTGGGTAT CACAGTG --agacacaccgtgtcaaatac TGAGGTGGA C 1113
CAGACAGCGGATGTGTTCAAACCTACGATCTCTACCACGGCAGGAACTTCGAAGGAAATGAGTCTTAAATTGAAGTAATTGCTAGAAGTATT 1113IC
CAGACAGGAGTT-TGCTCAAACCAGCGATCACTAACACGGCCGATGGTCCGTAAACAATGAGGATTAAATAAAAA-ATTGGGGAAATACAAT 1315
TATTAACTGTAAAACACAGAGATTAAA--> 1113
AGAGATTTGTAAAAATGTATAAAATTCTTCTGGTTTGTAAGTGTTGACTGACATGCAAAATCTGGACAGATACTTTTAAGAGTATTTCAATC 1113IC
ATAGATTTGTAACATTGTATAGAAGTCTCCTGCTTTATTAATGTTGACTGTCACGCTAAATCAGCACAGATGATTCTGAAA-TATTCCGATC 1315
TCCTTTGCACATATAATGACCAGTGTGAACTGAAATGTCTCACTGATGCACACTTAGGGTAGTCGGAATCAGGTTTCTGAGAACTCTGTGGA 1113IC
TC-TTTATAAATATAATGACTGGTATAAACTGAACTTTCTCACTGATGTCCTCTTAGGGGGATCGCAATCATGTTTTTGAGAGCGCTGTGAA 1315
ATTTAATCTCTGTGTTTTACAGTTAATAGTCCACCTC AGTATTTGA cacggtgtgtct CACTGTG ATAC-CCACTGTA GTACCCA c 1113IC
ATGTAATCTCTGTATTTTACAGTTAACAGTCCGCCGC AGTATTTGA cacggtttgtct CACTGTG ATAT-CTGGGTGG CACAGCA g 1315
<--AACACATCCGCTGTCTGGGGTTTTTGTA TAAGAAACT cgctctgtacag TACCGTG TTATACTGGGTGG CACAGCA g 1113
tgtgacacgaa CCCGCACAA 1113IC
ttactgtcaat ACAAAAACT 1315
ttactgtcaat ACAAAAAGT 1113
C V D1 D2 J
a 7 23 9 9 12 7 7 22 9 9 12 7 7 12 9 9 23 7
b 7 23 9 9 12 7 7 12 9 9 22 7 7 12 9 9 23 7

1403 and 2807 are identical and vary only by a single nt from 1315. The 9mer of gene 1113 differs by two bases and the 23 nt spacer is less related. The 9mer element located 5′ to D_1 is identical in all four genes. The 12 nt spacers of 1315, 2807 and 1113 differ only at a single position and the 5′ D_1 7mer of 1113 differs at a single base from the other 7mers. The 3′ D_1 7mer, which is identical to that of TCR $D_\beta 1.1$ (Siu *et al.*, 1984), and 9mer recombination elements 3′ of D_1 in genes 1315, 1403 and 2807 are identical, and the 22 nt spacers are related closely; however, in 1113 an A-rich, 9mer-like sequence is located 12 nt 3′ to a V_H-type 3′ D_1 7mer. Similarly, although the 5′ D_2 9mers, 12 nt spacers and 7mers in 1315, 1403 and 2807 are related closely, 1113 differs significantly; the 1113 5′ 9mer more closely resembles a J_H 9mer (Tonegawa, 1983) and the 5′ D_2 7mer initiates with a T and lacks the characteristic T at position 4, as does a δ TCR J segment (Chien *et al.*, 1987). The predicted 22 nt, versus 12 nt, spacer is not related to the other D_2 spacers; the 3′ D_2 7mers, spacers and 9mers in all four genes are related closely. The D_1 and D_2 coding, 3′ D_2 RSSs, D_2–J IVSs (not shown) as well as the J RSSs and spacers are related closely in all four genes. The D_1–D_2 IVSs of 1315, 1403 and 2807 are 68–77% related but are not related to 1113. The inverse complement of D_1–D_2 IVS of 1113, including the RSS is, however, ~77% identical to e.g. gene 1315, suggesting that it may have originated through chromosomal inversion, Figure 2B. The different types of V–D_1–D_2–J organization are shown in Figure 2C; both patterns preserve the 12/23 spacer rule (Tonegawa, 1983) that would permit at least three types of joining involving: (i) both D_H elements (all four genes), (ii) D_2 alone, 1315-type, or (iii) D_1 alone, 1113-type.

Additional characteristics distinguish gene 1113 from the other V–D_1–D_2–J as well as VD–J and VDJ genes. Neither the V_H–D_1 IVS of 1113 nor its inverse complement is related to the corresponding IVSs of the other genes. Furthermore, the predicted nucleotide sequence of the V_H coding region that initiates at residue 19 of framework region 1 (FR1) shares only two positions with 16 highly conserved, contiguous residues in FR3 of the other nine sequences in Figure 3A; as of yet, it has not been possible to obtain more 5′ sequence information. Finally, the distance between J_H and C_H is greater in 1113 than in the other clones, Figure 1D. The V_H segment of 1113 possesses, however, the hyperconserved FR3 sequence Tyr-Tyr-Cys-Ala-Arg as well as a typical 3′ RSS. Studies now in progress seek to determine whether additional 1113-type genes are present.

V_H, D_H, J_H diversity

The predicted amino acid sequences of nine different *Heterodontus* V_Hs corresponding to three different types of organization are shown in Figure 3A. *Heterodontus* genes are related closely in FR and, like V_H genes of higher vertebrates, vary extensively in the complementarity determining regions (CDR). At the nucleotide level, the four genes in the V–D_1–D_2–J configuration are related most closely, overall sequence identity = 88–92%, whereas the lowest degree of relatedness, 81%, is observed for the VD–J genes 2806 and 3083. Pairwise comparisons of the nine V_H genes indicate ~86% overall nucleotide sequence identity.

The leader segments of nine V_H genes are closely related (Figure 3B). Only two positions vary by more than two different amino acids at a single position. Evidence exists for strong selective pressure at certain positions, e.g. Leu^{10}, which is encoded by nucleotide triplets that differ at the first and third positions.

D_1 coding segments are related closely to each other as are D_2s; D_1 and D_2 are related partially to each other (Figure 2A). The D_1 segments of 1315 and 2807 have identical 14 nt sequences and vary from gene 1403 by 1 nt; the putative coding segment of gene 1113 is related and is 2 nt longer. The three different D_1 segments could encode eight different sequences and the four D_2 segments could encode eleven different sequences in all three reading frames. The sequences of mammalian Ig D_H segments belonging to the same family are related (Kurosawa and Tonegawa, 1982; Tonegawa, 1983) as are TCR D_β segments (Toyonaga *et al.*, 1985), however, the two TCR D_δ segments differ appreciably (Chien *et al.*, 1987).

The nucleotide and predicted amino acid sequences of the J_H segments of eight different V–D_1–D_2–J genes are shown in Figure 3C. The first two nucleotide positions are shared by all genes and nucleotide sequence variation largely is limited to positions 3–11 that encode the first three amino acids; only 3083 varies outside of this region. The predicted amino acid sequences of the J_H segment of 1111 (VD–J) and 1403 (V–D_1–D_2–J) are identical. Since 5′ J_H nucleotides are deleted frequently during Ig joining (see below), the contributions of germline J_H amino acid sequences to Ig diversity may be relatively low.

Segmental joining

The sequences at the VDJ junctions of six V_HHXIA^+ (Litman *et al.*, 1985b) clones recovered from a *Heterodontus* spleen cDNA library are compared with a consensus sequence derived from four V–D_1–D_2–J genes, Figure 4A. The sequences of cDNAs 12171, 12021 and 12061 exhibit appreciable contiguous homology with a consensus D_H element; less identity is evident with 12022, HC3 and 12423. Varying numbers of nucleotides occur between the assigned D_1 and D_2 segments of certain cDNAs, e.g. 12022 versus HC-3. Like TCRs, some of the genes exhibit short sequence stretches in N regions that are found in coding segments (Huck *et al.*, 1988), e.g. ACTACT in HC3 and GGGG in 12423, are invariant in J_H. While the data suggest contributions from both D_1 and D_2, it must be noted that consensus sequences tend to maximize sequence identity in individual rearrangment products, since alternatives that may not exist in a single parental segment are included. Internal homology between D_1 and D_2 also may complicate

Fig. 2. (A) Comparisons of the nt sequences of V_H, D_1, D_2 and J_H RSSs and D_H coding segments of four V–D_1–D_2–J genes. RSS 7mers and 9mers are shaded; putative coding segments of D_1 and D_2 are bold. Spacer segments are shown in lower-case letters and are extended 4 nt in the case of D_1 1113. The assignments of RSSs in 1113 is based on the relationship of 1113 to other V–D_1–D_2–J genes (B). **(B)**. Comparison of the nucleotide sequence from the 5′ 9mer of D_1 through the 3′ 9mer of D_2 of prototypic gene 1315 to the corresponding segment of gene 1113 or its inverse complement, 1113IC. Recombination elements are enclosed in boxes and the putative coding segments (Figure 2A) are bold (direct sequences only). [] indicates the boundaries of the segment between the D_1 and D_2 coding segments in 1315 that is 77% homologous to 1113IC. The absence of significant sequence identity between 1315 and 1113 extends beyond the region indicated by ---> <---. **(C)**. Schematic representation of two types of V–D_1–D_2–J genes; (a) 1315-type (b) 1113-type. Individual segments are indicated, RSSs are enclosed in boxes and the lengths of the spacers are indicated between 7mers [7] and 9mers [9].

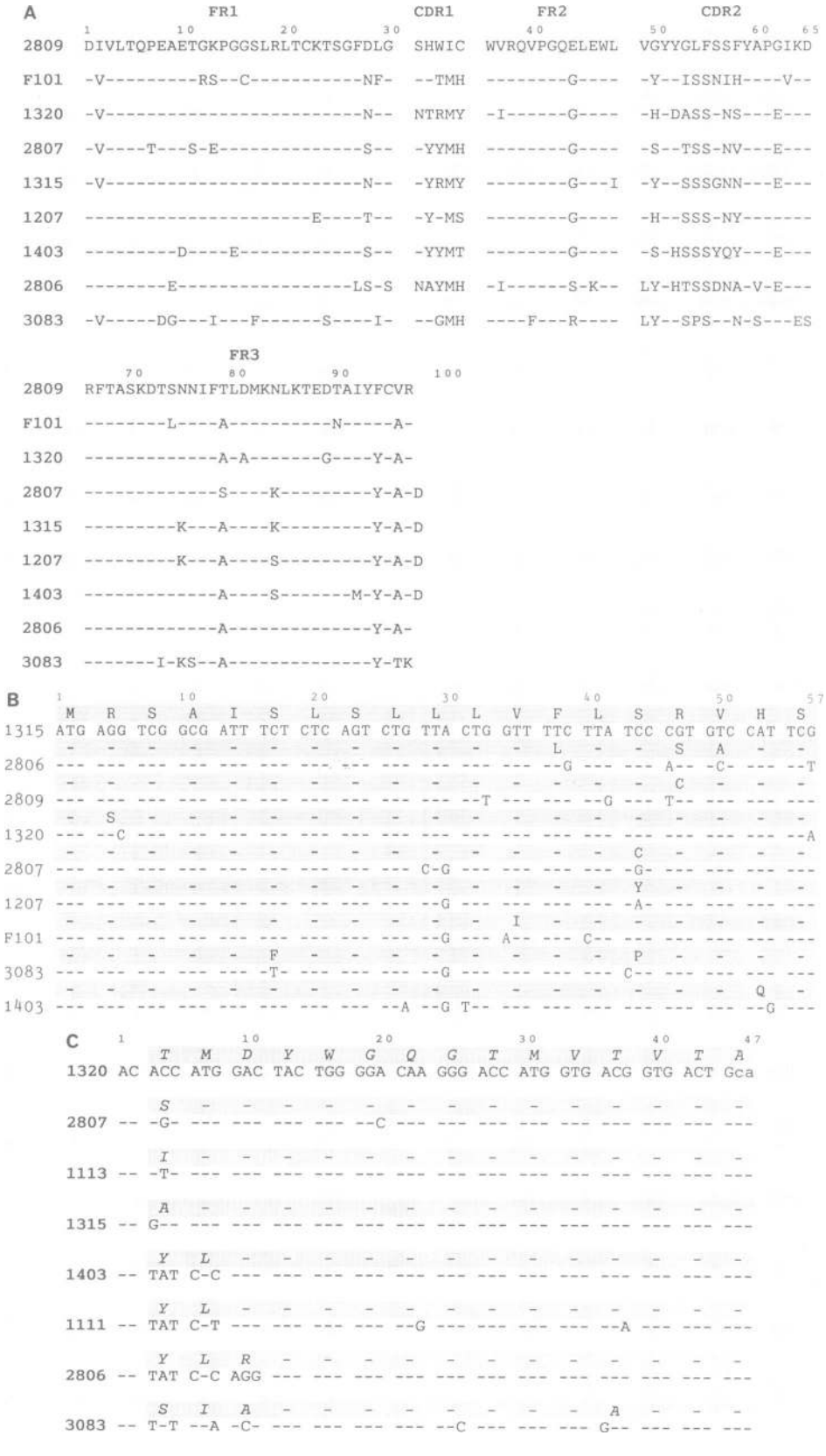


Fig. 3. **(A)** Comparison of the predicted coding regions of nine different *Heterodontus* V_H genes. Sequences are arrayed by the Bionet GENALIGN program that ranks sequences relative to the most related prototype; identity with prototypes in (A), (B), and (C) is indicated by (−). The boundaries of FR1, CDR1, FR2, CDR2 and FR3 are assigned as in Litman *et al.* (1985b), however, the 3′ boundary of FR3 includes sequences up to the 3′ RSS 7mer, where applicable. CDR and FR are not intended to imply a relationship between structural diversity and combining site variation. Complete DNA sequence of these genes will be available in GENBANK. **(B)** Nucleotide and predicted amino acid sequences (shaded) of the leader segments of nine different V_H genes. Sequences are arrayed by the Bionet GENALIGN program; order differs from both V_H and J_H (see below) segments. **(C)** Nucleotide and predicted amino acid sequences (shaded) of the J_H segments of eight different genes arrayed as in (A) and (B). The first 2 nt located immediately 3′ to the RSS 7mer, with some exceptions, are deleted in joining (see text). J_H regions of VDJ-type genes are not shown; the sequence of F101 is identical with the consensus, whereas 2809 varies at the otherwise invariant Met[10].

these assignments. Distinguishing minor variations from the D_1 and D_2 consensus sequences is complicated by the predicted extensive nature of this gene family and the possibility of somatic variation.

Since the eight J_Hs are related closely, a consensus for interpreting potential J_H junctional diversity can be derived. At least five cDNAs possess sequences 3′ to D_2 that do not correspond to J_H or D_H sequences in the present database. Only cDNA 12171, which appears to contain a significant D_2 segment, lacks additional nts at the D_2-J_H junction. Appreciable variation also occurs at the V_H-D_1 junction. cDNAs 12022 and HC-3 exhibit the most significant deletion of germline V_H sequence, and both contain nucleotides not found in the V_H or D_1 consensus sequences. Based on comparisons to these prototypes, four of the six cDNAs exhibit junctional diversity at the V_H-D_1, D_1-D_2 and D_2-J_H boundaries. One cDNA, 12021, exhibits only V_H-D_1 and D_2-J_H junctional diversity and another, 12171, exhibits only V_H-D_1 junctional diversity, with a limited J_H deletion. In all cases, the predicted coding sequences retain a reading frame that is homologous with higher vertebrate Ig prototypes and is consistent with splice donor−acceptor relationships found in genomic sequences (Kokubu *et al.*, 1988).

Joined germline genes

The sequences at the VDJ junction of the germline joined genes F101 and 2809 are shown in Figure 4B. F101 exhibits

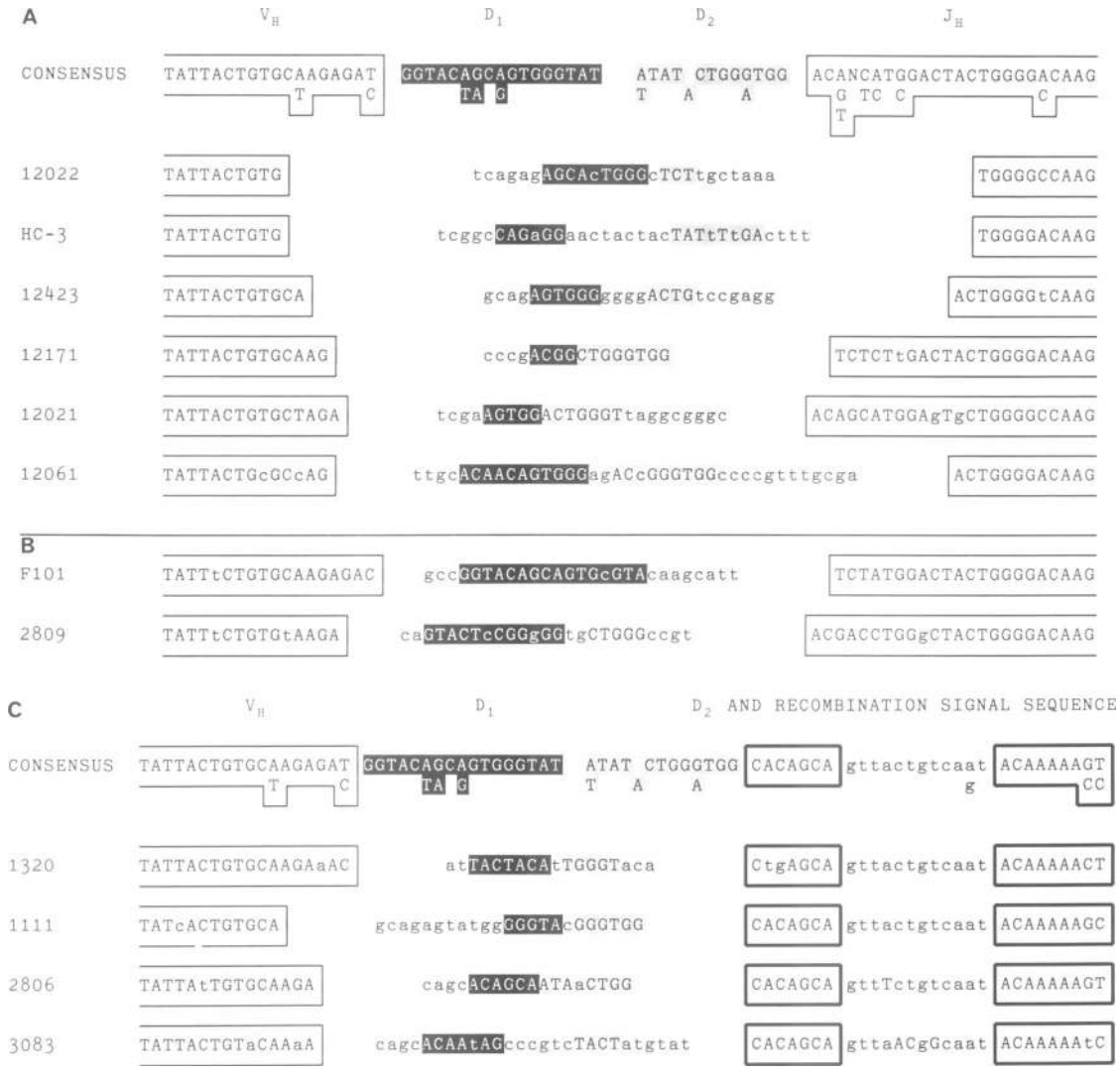


Fig. 4. Analysis of segmental joining in spleen V_HHXIA^+ cDNAs (**A**), germline VDJ (**B**) and germline VD-J (**C**) genes. Except for HC3 (Hinds and Litman, 1986), the cDNAs are derived from the same animal that was used for constructing a genomic DNA library. A consensus was derived from the sequence of four $V-D_1-D_2-J$ genes, the boundaries of D_1 (black) and D_2 (shaded) are defined by the 5′ and 3′ RSS 7mers as are the boundaries of germline V_H and J_H. The order of alternative nts reflects the frequency of their occurrence. The 3′ sequence of consensus D_1 includes the AT that is unique to 1113, Figure 2A. The gap in D_2 indicates the absence of an A at position 5 of some D_2s. Nucleotide identities, of at least three contiguous positions, with the consensus D_1 and D_2 sequences are shown in upper-case letters using appropriate shading; lower-case letters indicate a lack of identity with V_H, D_1, D_2 or J_H segments. A sequence dissimilarity is included in the homology zone (and noted in lower-case letters) if it extends the homology by at least two additional contiguous bases, either 5′ or 3′. In (**C**) the same consensus, including the D_2 3′ RSS is employed. The 7 and 9mer RSS elements are in upper-case letters and are enclosed. Nucleotide sequence differences from the consensus are indicated in lower-case letters. The 12 nt (noncoding) spacers are shown in lower-case letters with dissimilarities shown in upper-case letters. The CDR2 sequences of all of the genes shown in A, B and C differ from each other as well as from the four consensus genes and other VD−J and VDJ-type genes that are not presented.

a correct reading frame with a high degree of sequence identity to the V_H, D_1 and J_H consensus sequences. F101 also contains additional sequences 5′ and 3′ of D_H that are not homologous to V_H, D_2 and J_H consensus sequences. F101 may have arisen without a D_2 contribution, as is possible with a 1113-type gene; the 3′ A in the D_1 homology segment of F101 is consistent with this possibility (Figure 2A). Alternatively, a D_2 segment that differs markedly from the consensus may have been integrated. Gene 2809 also exhibits homology to D_1 and has a 5 nt sequence identity with D_2; additional nucleotides are located at the three predicted junctional boundaries.

The nucleotide sequences of four different VD-J genes are compared to the same consensus sequences as in Figure 4A and B including a consensus D_2 3′ RSS, Figure 4C. The 3′ segments of all four genes correspond to the 3′ D_2 RSSs. Furthermore, these genes possess a high degree of sequence identity with $V-D_1-D_2-J$-type genes in the IVS separating VD− and −J, e.g. the D−J IVS of 1111 (VD−J) is 88% related to that of 1403 ($V-D_1-D_2-J$). All four genes contain sequences between V_H and the RSS that are homologous to the potential coding segments of D_1 and D_2. In gene 1320, which possesses the least characteristic 7mer, the homology may correspond only to D_1 as both identity regions are short and the segment that is homologous to D_2 also is homologous to D_1. The 3′ nt sequence identities in D_2 of 1111 and 2086 are contiguous with the RSS, whereas in 1320 and 3083 additional 3′ nts, not present in the D_2 consensus are present. These sequences may derive from D_2 segments that differ from the consensus or may represent 'somatic' insertions. In 1111 and 2806 the 7mer sequences CACTGTG and CACAGCA are found in V_H and D_1 respectively. Both 2086 and 3083 possess the same, unique CAGC sequence at the V_H-D boundary. The functional relationship of the 2806 D_1 7mer to the 3′ D_2 7mer is not understood. These relationships as well as those between a hyperconserved V_H coding sequence and the RSS 7mer noted previously (Litman *et al.*, 1985b) may be significant. Relative to the consensus, all four VD−J genes exhibit V_H-D junctional diversity.

Absence of hyperconserved V_H regulatory octamer

Heterodontus V_H genes possess *TAAAT*, a conserved TATA-box-like sequence, at −95 (relative to the initiation codon) (Figure 1). Additional highly conserved sequence regions are: (i) $^{-109}$TTCTGCCgGGTAtT*TAAAT*. Nucleotides in upper-case letters occur in 8/8 sequences; nucleotides in lower-case letters occur in 7/8 sequences. (ii) $^{-206}$ATGAgTCcATtCTgTTagTGATGtGgAA and $^{-235}$TTTAaTaATAaC. Nucleotides in upper-case letters occur in 6/6 sequences; lower-case letters indicate nts that occur in 5/6 sequences. The invariant V_H regulatory octamer, ATGCAAAT, located 26−27 nt 5′ to *TAAAT* in higher vertebrate Ig genes (Parslow *et al.*, 1984; Falkner and Zachau, 1984) is not found at this position in the different *Heterodontus* sequences examined thus far. Furthermore, neither this octamer nor its functionally equivalent inverse complement (Falkner and Zachau, 1984; Eaton and Calame, 1987) was detected in additional analyses of five different genes that extend 400−600 nt 5′ of *TAAAT*.

Discussion

Three unique patterns of germline Ig gene organization have been characterized in *Heterodontus*. Complete nucleotide sequences of four different genes from 5′ of V_H through the 3′ of J_H confirm the close linkage of the V_H, D_H and J_H segments described previously (Hinds and Litman, 1986) and define an additional D_H segment. The RSSs associated with D_1 and D_2 differ but maintain the 12/23 spacing rule in both types of $V-D_1-D_2-J$ organization. Variation in the structures of RSSs associated with these genes may influence recombination efficiency (Akira *et al.*, 1987) and in this way could be of regulatory significance. Productive recombination without chromosomal inversion (Malissen *et al.*, 1986) is possible in both types of genes which can recombine one or both but not either D_H element(s). Direct VJ joining that bypasses the D segment(s), is possible with the β TCR (Kavaler *et al.*, 1984) and in at least one reptilian V_H gene (Litman *et al.*, 1985a), but cannot occur in *Heterodontus*, since both V_H and J_H RSSs have 23 nt spacers.

The VD− and VDJ-joined genes are not processed (mRNA) pseudogenes as they possess intact leader and J_H-C_H IVSs. These genes have homologous reading frames, typical splice sites, upstream 'regulatory' and additional 5′ and 3′ sequences that are shared by other genes. The sequences of the coding segments of these genes are related closely to corresponding portions of the $V-D_1-D_2-J$ genes. While lymphopoiesis in a somatic tissue could account for VD− and VDJ-joining, this is unlikely since the same gene has been recovered from both liver and gonad DNA (~90% sperm) of two unrelated animals. Even if such joinings occurred by expression of recombinase activity in the germ cells of the individual *Heterodontus*, absolute joining fidelity would have been required in order to account for the sequence identities through the VD(J) junctional boundaries. In mammals, recombinase activity appears to be expressed in lymphoid cells during early development as well as in the progenitors of myeloid cells but not in several other tissues, including embryonic liver (Lieber *et al.*, 1987). In the relatively few genomes that were screened, detection of large numbers of joined Ig genes and identity of joined genes in different tissues from different animals support the contention that these do not derive from typical somatic rearrangement. It is more reasonable to conclude that a significant portion of *Heterodontus* V_H genes are arranged in the germline in this entirely unique manner. Their functional status, however, remains unclear (see below).

In the absence of lymphoid tumors, cell lines and/or compartmentalized expansion of lymphoid cell progeny (as in the avian bursa, Reynaud *et al.*, 1987), relating a rearrangement product to a specific gene cluster is difficult, particularly given the size and complexity of the *Heterodontus* V_H family. The close relationship (86% nt identity) between all *Heterodontus* V_H mature coding segments further confounds selection of the parental cluster(s). While CDR2-specific oligodeoxynucleotide probes can identify individual genes, their use in direct screenings is complicated by hybridization to interspersed repetitive DNA (unpublished observation), thus necessitating the isolation and characterization (including analyses of allelic polymorphisms, since

Heterodontus may be tetraploid) (Schwartz and Maddock, 1986) of each germline cluster. Although the roles, if any, of somatic mutation and combinatorial diversity cannot be assessed, the cDNA analyses suggest that both junctional and N-type diversity (Tonegawa, 1983) effect somatic change in these genes. The data suggest a role for the D_H element in joining; however, interpreting the nature of these contributions is even more complex than in studies of mammalian Ig joining (Tonegawa, 1983; Kurosawa and Tonegawa, 1982) due to the extent (and close relatedness) of this gene family as well as to the potential for two D_H segments contributing to a single rearrangment event. Limited variation in D_H and J_H segments, evidenced by the detection of identical D_1 segments and near absence of variation in J_H, may be of little significance when weighed against the effects of junctional and N diversity that accompany somatic rearrangement. Comparisons of VD−J and VDJ (rearranged) genes with cDNA (spleen mRNA) sequences suggest that they arose by the same mechanism(s), although there is greater conservation of consensus V_H, J_H and in one case D_H sequences in the germline joined genes than in the cDNAs. At this level of analysis, segmental joining in *Heterodontus* cannot be distinguished from that diversifying mammalian Ig and TCR genes (Alt *et al.*, 1986).

While the organization of *Heterodontus* Ig V_H gene clusters is relatively basic, their large numbers (Hinds and Litman, 1986), alternative germline arrangements, multiplicity of C_H isotypes (Kokubu *et al.*, 1987, 1988) and possibility for utilizing one or both D_H segments qualifies them as the most complexly organized family of antigen binding receptor genes characterized thus far. The gene system may have originated from a 1315-type $V-D_1-D_2-J$ gene. VD−J or VDJ-type genes may have arisen from expression of recombinase activity in germ cells. Duplication and a chromosomal inversion, perhaps during an abortive rearrangement (germline), could account for a 1113-type gene; perhaps the relative close proximity of segmental elements would facilitate such an event. During evolution, it is possible that similar processes may have accounted for changes in the orientation of RSSs associated with other antigen binding receptor genes. Alternatively, the VDJ genes may reflect the evolutionary predecessors of the segmented genes (Sakano *et al.*, 1979; Hood *et al.*, 1985). Transposon-like behavior of a segment containing a RSS(s) and/or inter/intrachromosomal recombination would first give rise to the VD−J and then to the $V-D_1-D_2-J$ genes. Regardless of their relative position in the evolution of this gene system, it is of considerable significance that these genes are potentially functional; i.e. they are not overt pseudogenes. They may represent a means whereby a portion of the multigene family selectively preserves favorable 'joins', thus eliminating a need for chance somatic recombination in certain antibody responses. Alternatively, they may serve highly specialized needs such as developmental stage-specific expression. Even if they are not expressed, joined genes may be substrates for gene conversion, as has been demonstrated at the avian V_λ locus (Reynaud *et al.*, 1987), or may participate in replacement (secondary) recombination (Kleinfield *et al.*, 1986; Reth *et al.*, 1986). The germline-joined genes may lack direct or indirect functional activity but retain coding potential through intense gene correction that is acting on the segmented members of this extensive multigene family. Such a mechanism, however, would have to retain individuality in CDR segments.

This report provides new information concerning the origin and diversification of the antigen-binding receptor genes encoding both B and T cell immunity. The collective sequence data for coding regions presented here and previously (Hinds and Litman, 1986) emphasizes that the V_H and J_H regions, including the CDR segments, are unequivocally of the higher vertebrate Ig-type. Furthermore, the genomic organization and secretory versus transmembrane processing of the *Heterodontus* C_H region is related closely to mammalian μ-type Ig (Kokubu *et al.*, 1988). While the presence or absence of somatic mutation in *Heterodontus* Ig would be important in comparisons to different antigen binding receptor genes, it is not possible as of yet to make such assignments; neither junctional nor N-type diversity distinguish antibody and TCR genes (Hayday *et al.*, 1985; Quertermous *et al.*, 1986; Traunecker *et al.*, 1986; Klein *et al.*, 1987; Huck *et al.*, 1988).

In several regards, *Heterodontus* Ig shares many properties with certain TCR genes including; (i) the D_H, J_H and C_H segments are in close proximity as seen both in the β (Toyonaga *et al.*, 1985; Lai *et al.*, 1987; Lindsten *et al.*, 1987; Chou *et al.*, 1987; Wilson *et al.*, 1988a) and δ (Chien *et al.*, 1987) TCR gene families, (ii) V_H, J_H and C_H are linked closely as with the murine γ TCR genes (Traunecker *et al.*, 1986); and perhaps, like certain TCR γ genes, *Heterodontus* genes may not be isotypically excluded (Heilig and Tonegawa, 1987), (iii) two D segments that are closely linked to J_H also are found with the murine TCR D_δ gene (Chien *et al.*, 1987); nucleotide additions between joined D_H segments occurs in *Heterodontus* Ig as well as in adult but not fetal mammalian T cells (Elliott *et al.*, 1988), and (iv) the hyperconserved regulatory octamer, an invariant feature of all higher vertebrate, teleost (Litman *et al.*, 1985a; Wilson *et al.*, 1988b) and above, Ig V_H (and light chain V region) genes, is absent both in *Heterodontus* V_H and TCR genes (Luria *et al.*, 1987; Lee and Davis, 1988). The apparent absence of affinity maturation, predicted restriction of rearrangement to a single cluster (Traunecker *et al.*, 1986) and limited structural variation deriving from J_Hs (Quertermous *et al.*, 1986; Hayday *et al.*, 1985), represent additional similarities to some TCRs. Thus, the secreted Ig-like molecules found in *Heterodontus* are encoded within a gene complex that more closely resembles the TCRs, both in terms of gene organization and possibly regulation. The *Heterodontus* Ig V_H system may resemble the common ancestral form of both the Ig and TCR genes of higher vertebrates. The presence of large numbers of rearranged V_H gene segments has no counterpart in any other vertebrate system and also may be an essential feature of antigen binding receptor diversity in more primitive species.

Materials and methods

Animals

Adult specimens of *H. francisci* (horned shark) were obtained from Pacific Biomarine Supply Co., Venice, CA. After the animals were killed, all tissues were processed immediately.

DNA libraries

Genomic DNA libraries were constructed in λ DASH (Stratagene) from *Sau*3A-digested high-mol. wt DNA isolated from the liver and gonads (testes) of individual specimens of *H.francisci* essentially as described (Litman *et al.*, 1985b). Both libraries were amplified selectively on P2392 which is a P2 lysogen of LE392. Approximately 6×10^5 recombinant phage, corresponding to ~1.5 haploid genomes and 4.8×10^5recombinant phage corresponding to ~1.2 haploid genomes (assuming proportional representation) were recovered from the liver and gonad libraries respectively. The cDNA library was constructed from spleen mRNA of the same animal that was used for the genomic DNA (liver) library as described (Kokubu *et al.*, 1988).

Probes and library screening

A V_H probe corresponding to the coding region of gene HXIA (V_HHXIA) has been described previously (Litman *et al.*, 1985b). λ genomic clone 2809, that hybridizes weakly with V_HHXIA, was digested with *Eco*RV and *Sca*I and a 293 nt fragment, corresponding exclusively to V_H coding sequence (85% identical to HXIA), was subcloned in pUC12. Probes were labeled by the hexanucleotide random priming method (Feinberg and Vogelstein, 1983) and nitrocellulose lifts were hybridized and washed using the moderate stringency-wash conditions described previously (Litman *et al.*, 1985b).

DNA sequencing and sequence analysis

All subcloning was done in commercially available M13 RFs and the DNA sequences were determined in both directions by the dideoxy method (Sanger *et al.*, 1977) using ^{35}S-label and T7 DNA polymerase, Sequenase (United States Biochemical Corporation). The primary strategy used to extend sequences of specific clones or verify sequences on the opposite strand utilized sequence-specific 18mer extension primers. Routine analyses of DNA sequences primarily used programs available through the Bionet Resource. Alignments were made using IFIND and GENALIGN programs which are copyrighted software products of Intelligenetics, Inc., Palo Alto, CA; GENALIGN was developed by Dr Hugo Martinez of the University of California at San Francisco.

Acknowledgements

The editorial assistance and comments of Mrs Barbara Allen are appreciated. This work was supported by a grant from the National Institues of Health, AI-23338. Computer analyses utilized the Bionet Resource that is supported by grant U41-RR-01685-03.

References

Akira,S., Okazaki,K. and Sakano,H. (1987) *Science*, **238**, 1134–1138.
Alt,F.W., Blackwell,T.K., DePinho,R.A., Reth,M.G. and Yancopoulos, G.D. (1986) *Immunol. Rev.*, **89**, 5–30.
Chien,Y.-h., Gascoigne,N.R.J., Kavaler,J., Lee,N.E. and Davis,M.M. (1984) *Nature*, **309**, 322–326.
Chien,Y.-h., Iwashima,M., Wettstein,D.A., Kaplan,K.B., Elliott,J.F., Born,W. and Davis,M.M. (1987) *Nature*, **330**, 722–727.
Chou,H.S., Nelson,C.A., Godambe,S.A., Chaplin,D.D. and Loh,D.Y. (1987) *Science*, **238**, 545–548.
Eaton,S. and Calame,K. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 7634–7638.
Elliott,J.F.,Rock,E.P., Patten,P.A., Davis,M.M. and Chien,Y.-h. (1988) *Nature*, **331**, 627–631.
Falkner,F.G. and Zachau,H.G. (1984) *Nature*, **310**, 71–74.
Feinberg,A.P. and Vogelstein,B. (1983) *Anal. Biochem.*, **132**, 6–13.
Hayday,A.C., Saito,H. Gillies,S.D., Kranz,D.M., Tanigawa,G., Eisen,H.N. and Tonegawa,S. (1985) *Cell*, **40**, 259–269.
Hedrick,S.M., Nielsen,E.A., Kavaler,J., Cohen,D.I. and Davis,M.M. (1984) *Nature*, **308**, 153–158.
Heilig,J. and Tonegawa,S. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 8070–8074.
Hinds,K.R. and Litman,G.W. (1986) *Nature*, **320**, 546–549.
Hood,L., Kronenberg,M. and Hunkapiller,T. (1985) *Cell*, **40**, 225–229.
Huck,S., Dariavach,P. and LeFranc,M.-P. (1988) *EMBO J.*, **7**, 719–726.
Kavaler,J., Davis,M.M. and Chien,Y.-h. (1984) *Nature*, **310**, 421–423.
Kleinfield,R., Hardy,R.R., Tarlinton,D., Dangl,J., Herzenberg,L.A. and Weigert,M. (1986) *Nature*, **322**, 843–846.
Klein,M.H., Concannon,P., Everett,M., Kim,L.D.H., Hunkapiller,T. and Hood,L. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 6884–6888.
Kokubu,F., Hinds,K., Litman,R., Shamblott,M.J. and Litman,G.W. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 5868–5872.
Kokubu,F., Hinds,K., Litman,R., Shamblott,M.J. and Litman,G.W. (1988) *EMBO J.*, **7**, 1979–1988.
Kurosawa,Y. and Tonegawa,S. (1982) *J. Exp. Med.*, **155**, 201–218.
Lai,E., Barth,R.K. and Hood,L. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 3846–3850.
Lee,N.E. and Davis,M.M. (1988) *J. Immunol.*, **140**, 1665–1695.
Lieber,M.R., Hesse,J.E., Mizuuchi,K. and Gellert,M. (1987) *Genes Dev.*, **1**, 751–761.
Lindsten,T., Lee,N.E. and Davis,M.M. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 7639–7643.
Litman,G.W., Stolen,J., Sarvas,H.O. and Mäkelä,O. (1982) *J. Immunogenet.*, **9**, 465–474.
Litman,G.W., Murphy,K., Berger,L., Litman,R.T., Hinds,K.R. and Erickson,B.W. (1985a) *Proc. Natl. Acad. Sci. USA*, **82**, 844–848.
Litman,G.W., Berger,L., Murphy,K., Litman,R., Hinds,K.R. and Erickson,B.W. (1985b) *Proc. Natl. Acad. Sci. USA.*, **82**, 2082–2086.
Luria,S., Gross,G., Horowitz,M. and Givol,D. (1987) *EMBO J.*, **6**, 3307–3312.
Mäkelä,O. and Litman,G.W. (1980) *Nature*, **287**, 639–640.
Malissen,M., McCoy,C., Blanc,D., Trucy,J., Devaux,C., Schmitt-Verhulst,A.-M., Fitch,F., Hood,L. and Malissen,B. (1986) *Nature*, **319**, 28–33.
Parslow,T.G., Blair,D.L., Murphy,W.J. and Granner,D.K. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 2650–2654.
Patten,P., Yokota,T., Rothbard,J., Chien,Y.-h., Arai,K.-i. and Davis,M.M. (1984) *Nature*, **312**, 40–46.
Quertermous,T., Strauss,W., Murre,C., Dialynas,D.P., Strominger,J.L. and Seidman,J.G. (1986) *Nature*, **322**, 184–187.
Reth,M., Gehrmann,P., Petrac,E. and Wiese,P. (1986) *Nature*, **322**, 840–842.
Reynaud,C-A., Anquez,V., Grimal,H. and Weill,J.-C. (1987) *Cell*, **48**, 379–388.
Sakano,H., Huppi,K., Heinrich,G. and Tonegawa,S. (1979) *Nature*, **280**, 288–294.
Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.
Schwartz,F.J. and Maddock,M.B. (1986) In Uyeno,T., Arai,R., Taniuchi,T. and Matsuura,K. (eds), *Indo-Pacific Fish Biology: Proceedings of the Second International Conference on Indo-Pacific Fishes.* The Ichthyological Society of Japan, Tokyo, pp. 148–157.
Siu,G,., Kronenberg,M., Strauss,E., Haars,R.,.Mak,T.W. and Hood,L. (1984) *Nature*, **311**, 344–350.
Tonegawa,S. (1983) *Nature*, **302**, 575–581.
Toyonaga,B., Yoshikai,Y., Vadasz,V., Chin,B. and Mak,T.W. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 8624–8628.
Traunecker,A., Oliveri,F., Allen,N. and Karjalainen,K. (1986) *EMBO J.*, **5**, 1589–1593.
Wilson,R.K., Lai,E., Concannon,P., Barth,R.K. and Hood,L.E. (1988a) *Immunol. Rev.*, **101**, 149–172.
Wilson,M.R., Middleton,D. and Warr,G.W. (1988b) *Proc. Natl. Acad. Sci. USA*, **85**, 1566–1570.
Yanagi,Y., Yoshikai,Y., Leggett,K., Clark,S.P., Aleksander,I. and Mak,T.W. (1984) *Nature*, **308**, 145–149.
Yancopoulos,G.D. and Alt,F.W. (1986) *Annu. Rev. Immunol.*, **4**, 339–368.
Yancopoulos,G.D., Desiderio,S.V., Paskind,M., Kearney,J.F., Baltimore,D. and Alt,F.W. (1984) *Nature*, **311**, 727–733.

Received on June 23, 1988