Identification of the 48-base-long primordial building block sequence of mouse immunoglobulin variable region genes

(framework and hypervariable region/ultimate ancestor/repetitious origin)

Susumu Ohno, Takeshi Matsunaga, and R. Bruce Wallace

City of Hope Research Institute, Duarte, California 91010

Contributed by Susumu Ohno, December 7, 1981

ABSTRACT Mouse immunoglobulin heavy-chain variable region (Ig VH) genes apparently arose from the ∼600-base-pair-long (=12 tandem repeats of the 48-base-long primordial building block sequence TTC-AGC-ACC-CTG-ACT-GGA-TAT-GAC-CTG-GAG-TGG-ACT-TAC-TGC-GCA-AGA) that in the original reading frame specified the amino acid sequence Phe-Ser-Ser-Leu-Thr-Gly-Tyr-Asp-Leu-Glu-Trp-Thr-Tyr-Cys-Ala-Arg. The previously identified, shorter prototype building blocks merely represented particular portions of the above primordial sequence. Even today, the direct descendant in toto of this primordial sequence specifies the last one-sixth of each VH coding sequence: the 83rd to 98th amino acid residues. Furthermore, its four truncated derivatives specify the 4th to 14th, 17th to 23rd, 29th to 37th, and 38th to 48th amino acid residues. Accordingly, all three relatively invariant—therefore, conserved—framework regions (FW-1, FW-2, and FW-3) of VHs are specified by recognizable—therefore, conserved—descendants of the primordial sequence.

Evidence presented in our last two papers (1,2) indicated that both the genes for the serum albumin family of proteins (1) and those for immunoglobulin heavy chain variable regions (Ig VH) evolved from tandem repeats of two or more different kinds of 14- to 21-base-long prototype building block sequences. Nevertheless, we were fully aware that these readily identifiable prototype building blocks were likely to represent segments of the much longer primordial building block sequence.

Even if the entire gene (internal and adjacent noncoding sequences included) started eons ago as repeats of the one relatively long primordial sequence, subsequent sustained random base substitutions, deletions, and insertions by now should have diversified the base sequence of individual copies to such an extent that recognizable homology would remain with regard only to various segments of the original primordial building block sequence.

In view of the above, we have reexamined all the published (including in press) DNA base sequences of mouse Ig VH at our disposal in the hope of identifying the primordial building block sequence of considerable length that was ancestral to all Ig VHs.

Germ-line DNA sequence of anti-NP\(^{186-2}\) IgVH

As with Ig VH pCH 108A (3) and MOPC 141 (4) that were examined in detail in our previous paper (2), the Ig VH anti-NP\(^{186-2}\) germ-line DNA sequence (5) (Fig. 1) contained recognizable copies of the two previously identified prototype building blocks. Following the rule, the relatively invariant first and second framework (FW-1 and FW-2) regions—the 7th to 13th and 41st to 47th amino acid residues—and the first hypervariable (HV-1) region—the 30th to 36th amino acid residues—were specified by three copies of the 21-base-long prototypetype sequence ACT-GGA-TAT-GAC-CTG-GAG-TGG. The 14-base-long prototype building block idiosyncratic to anti-NP\(^{186-2}\) VH was identified as AC-AGT-TAC-TGC-GCA and its nine recognizable copies, demonstrating >60% base sequence homology to their prototype, were found in the coding and adjacent noncoding sequences alike. Nevertheless, the previously established rule was again confirmed because one copy of this free-wheeling prototype specified the most invariant Tyr-Tyr-Cys portion of FW-3 (Fig. 1). As previously noted (2), anti-NP\(^{186-2}\) VH (5) was closely related to pCH 108A VH (3) as evidenced by their 79.6% amino acid sequence homology. Fig. 1 shows three recognizable copies as well of a 15-base-long prototype idiosyncratic to pCH 108A, AAG-TCC-ATG-CTG-AGT.

In addition, the presence of the previously overlooked 14-base-long prototype building block sequence TTC-AGC-ACC-CTG-GG came to light. Because its recognizable copies were found in abundance in every mouse Ig VH examined, this prototype was conservative, as opposed to being idiosyncratic. A copy of this 14-base-long conserved prototype tended to occur in tandem with the 21-base-long, equally conserved, prototype, partially overlapping the 5′ side of the latter (Fig. 1). Another important fact that had previously escaped our attention was that, when the minimal allowable homology was decreased from 57.1% (12 of 21) to 52.4% (11 of 21), two additional copies of the 21-base-long conserved prototype building block invariably occupied the fixed positions within every Ig VH specifying the 17th to 24th and 87th to 93rd amino acid residues.

Identification of the 48-base-long ultimate primordial sequence

As a result of these newly acquired insights, we came to realize that the last one-sixth of each VH encoding sequence was invariably represented by one copy each of the three different kinds of prototype building blocks arranged in the following order: the newly identified 14-base-long conserved prototype, the 21-base-long equally conserved prototype, and the 14- to 15-base-long idiosyncratic prototype (Fig. 1). Viewed in another light, the above indicated that the region of VH DNA that specified the 83rd to 98th amino acid residues was represented by a sole surviving copy in toto of the ultimate 48-base-long, primordial sequence.

The ultimate ancestor, eons ago, of Ig VH genes might have been the ∼600-base-pair-long stretch of DNA composed of approximately 12 tandem repeats of such a 48-base-long primordial building block sequence, subsequent mutational diversification of individual copies eventually yielding not only the Ig VH encoding sequence per se but also its attendant hydrophobic leader coding sequence, an intervening sequence between the two, and the 5′ and 3′ terminal regions of intergenic spacers.

On the one hand, the hypothetical 48-base-long DNA se-

**Abbreviation:** Ig VH, immunoglobulin heavy chain variable region.
Fig. 1. The ~620-base-long germ-line DNA segment of the mouse that contained the coding sequence for anti-NP\(^{b}\) (186-2) Ig V\(_{H}\) and its attendant hydrophobic leader coding sequences (5). Only the DNA strand that corresponds to the transcript is shown. The canonical A-G-C-T-G and cocannonal G-G-G-T-G pentamers are enclosed in boxes. Amino acid residues of V\(_{H}\) sequences not homologous with those specified by pCH 108A (3) analyzed in our previous paper (2) are shown in italics. The nine recognizable copies demonstrating >64.3% base sequence homology to the idiosyncratic 14-base-long prototype building block ACAA-GTT-ACT-GAG-CA are marked by solid bars, and bases homologous with the prototype are identified by asterisks. Only a single base insertion or deletion is allowed for maximization of the homology. The six copies demonstrating >52.4% base sequence homology to the longer 21-base-long prototype building block are marked by hatched bars. The three recognizable copies of the 15-base-long prototype idiosyncratic to the closely related pCH 108A (2, 3) are so labeled. The seven recognizable copies of the new 14-base-long prototype building block are marked by open bars labeled \(\sim\). The beginning and the end of the two hypervariable regions are indicated by HV; those of each of the three framework regions are designated by FW.

Sequence identified in Fig. 2 as the primordial building block of Ig V\(_{H}\) readily yields, by fragmentation, the three different types of shorter prototype building blocks identified in the previous (2) as well as present paper. This point is shown at the top of Fig. 2. On the other hand, the consensus sequence of the region specifying the 53rd to 98th amino acid residues still demonstrates 70.8% base sequence homology with this primordial V\(_{H}\) building block. The consensus sequence identified as the modern intact building block in Fig. 2, demonstrates >70% base sequence homology with all the published sequences of this region.

In Fig. 2, the sequence of the mouse Ig V\(_{H}\) anti-phosphorylcholine (PCh) 5-15 (V1) germ-line gene (6) is used as a representative of the published sequences. Corresponding regions of PCH 108A (3), MOPC 141 (4) shown in the previous paper (2) and anti-NP\(^{b}\) (186-2) V\(_{H}\) (5) shown in Fig. 1 demonstrate
Fig. 2. Base sequence of the deduced 48-base-long ultimate primordial V<sub>H</sub> building block and the amino acid sequence specified by it are presented in the box so marked. The four shorter prototype building blocks, from Fig. 1, are arranged above the box in a manner to indicate their derivations from different segments of the primordial V<sub>H</sub> building block. Below the box, the consensus sequences of the five coding regions for mouse Ig V<sub>H</sub> are shown as the direct derivatives of the ultimate primordial V<sub>H</sub> building block in the box. Degrees of base sequence homologies to the corresponding regions in the box are indicated. These consensus sequences are marked "modern intact building block," and "1st, 2nd, 3rd, and 4th modern truncated building blocks." The positions of the first and last amino acid residues in the V<sub>H</sub> polypeptide specified by each consensus sequence are shown by large numbers. Below each consensus sequence is the corresponding region of the anti-phosphorylcholine (PCh) T-15 (VI) gene (6). Nonhomologous amino acid residues are shown in italics. Bases homologous with those in the consensus sequence were marked by asterisks; the degree of homology to the consensus sequence is shown as a percentage in parentheses.

90.9%, 72.7%, and 84.8% base sequence homology with this 48-base-long consensus sequence specifying the 83rd to 98th amino acid residues.

The consensus sequences of the four other coding regions specifying the 4th to 14th, 17th to 23rd, 29th to 37th, and 38th to 48th amino acid residues were identified as the first, fourth, third, and second modern truncated building blocks (Fig. 2). This is because each is so clearly derived from the primordial V<sub>H</sub> building block by an internal deletion and the 3'-terminal deletion. The first and second truncated building blocks were derived from the primordial building block by losing the 9th to 12th bases as well as 12 bases beyond the 36th. Deletions are more extensive in the case of the third and fourth modern truncated building blocks, the internal deletion involving the 4th to 12th bases. Yet these consensus sequences still demonstrated >63.0% base sequence homology with the corresponding portions of the primordial V<sub>H</sub> building block. Of special interest is the fourth modern truncated building block that specifies the 17th to 23rd amino acid residues. Translation of the conserved base sequence of this region is done one base out of phase with the original reading frame.

At any rate, it is now clear that all the relatively invariant—therefore, conserved—descendants of the 48-base-long ultimate primordial building block sequence.

Specific properties embodied in the primordial building block sequence

V<sub>H</sub> belong to the family of polypeptides roughly 100 amino acid residues long characterized by the possession of one intrachain
disulfide bridge. It is not surprising then to find the cysteine codon incorporated in the primordial building block sequence. Indeed, the downstream cysteine (96th residue) for an intrachain disulfide bridge of each $V_H$ is specified by a direct descendant of this codon (Fig. 2). The codon for the upstream cysteine (22nd residue), on the other hand, is supplied by the fourth truncated prototype derived from the primordial sequence which is translated one base out of phase with the original reading frame. The 48-base-long primordial sequence also contains the codons for tyrosine and tryptophan. These two amino acids are often thought of as the residues directly involved in the antigen binding.

Intergenic spacers of Ig genes are rich in the canonical A-G-C-T-G and cocanonical G-G-G-T-G pentamers (7, 8). It has recently been pointed out that the decameric sequence (A-G-C-T-G)-(G-G-G-T-G) can readily yield the Chi recombination signal sequence (-G-C-T-G)-(G-G-T-G-G) (9). Fig. 2 shows that the 48-base-long primordial sequence of mouse Ig $V_{H}\beta$ is abounds with single and two-base variants of the canonical and cocanonical pentamers. In fact, the 24th to 33rd bases of the primordial $V_H$ building block C-CTG-GAG-TGG becomes the Chi recombination signal by a replacement of the first base with G and deletion of the sixth base. This fact might have caused frequent internal deletions and insertions, thus contributing to the transformation of the simple repeats to $V_H$ genes.

As already noted, the domain consisting of roughly 100 amino acid residues that are held together in the form of a loop by one intragenic disulfide bridge characterizes not only variable ($V_H$, $V_L$, and $V_J$) and constant (C_H1, C_H2, C_J1, C_J2, and C_J3) regions of Ig heavy and light chains but also the two regions of each major histocompatibility antigen—H-2D, H-2K, and H-2L of the mouse and HLA-A, HLA-B, and HLA-C of man (10). The common ancestry of this family of domains is generally sought in $\beta_2$-microglobulin (10). This ancestor/descendant relationship, however, is not clearly evident in their amino acid sequences (10) or in their coding base sequences (11, 12). It may be that their kinship is more indirect, being due to their original choices of similar 48-base-long sequences as their primordial building block. The evolutionary relationship between the ancestral tandem repeats of the 48-base-long primordial building block and the present Ig $V_H$ gene is illustrated in Fig. 3.

This work was supported in part by National Institutes of Health Grant RO1 AT5620 and a research grant from the Wakunaga Pharmaceutical Company.