The Complete Nucleotide Sequence of the Human Immunoglobulin Heavy Chain Variable Region Locus

By Fumihiko Matsuda,* Kazuo Ishii,* Patrice Bourvagnet,‡ Kei-ichi Kuma,§ Hidenori Hayashida,† Takashi Miyata,§ and Tasuku Honjo*

From the *Department of Medical Chemistry, Kyoto University Graduate School of Medicine, Kyoto 60601, Japan; ‡Centre National de la Recherche Scientifique Institute National de la Santé et de la Recherche Médicale, Montpellier 34033, France; §Department of Biophysics, Kyoto University Graduate School of Science, Kyoto 60601, Japan; and †Statistics and Analysis of Information, Nara Medical University, Nara 634, Japan

Summary

The complete nucleotide sequence of the 957-kb DNA of the human immunoglobulin heavy chain variable (V_{H}) region locus was determined and 43 novel V_{H} segments were identified. The region contains 123 V_{H} segments classifiable into seven different families, of which 79 are pseudogenes. Of the 44 V_{H} segments with an open reading frame, 39 are expressed as heavy chain proteins and 1 as mRNA, while the remaining 4 are not found in immunoglobulin cDNAs. Combinatorial diversity of V_{H} region was calculated to be \( \sim 6,000 \). Conservation of the promoter and recombination signal sequences was observed to be higher in functional V_{H} segments than in pseudogenes. Phylogenetic analysis of 114 V_{H} segments clearly showed clustering of the V_{H} segments of each family. However, an independent branch in the tree contained a single V_{H}, V4-44.1P, sharing similar levels of homology to human V_{H} families and to those of other vertebrates. Comparison between different copies of homologous units that appear repeatedly across the locus clearly demonstrates that dynamic DNA reorganization of the locus took place at least eight times between 133 and 10 million years ago. One nonimmunoglobulin gene of unknown function was identified in the intergenic region.

Key words: immunoglobulin gene • physical map • antibody repertoire • phylogenetic tree • DNA duplication

During the vertebrate immune response, Ig and TCR play central roles in antigen recognition. The NH2-terminal portion of their subunits is called the V region because of its diverse amino acid sequence required for interaction with a diverse spectrum of antigens. Generation of the primary V-region repertoire depends on the common genetic basis and molecular mechanisms characteristic of these antigen-receptor molecules (1–3). First, the V regions are encoded by two or three genetic segments, namely V, D, and J segments, each of which comprises multiple copies and provides the repertoire before somatic mutation. Second, during the ontogeny of lymphocytes, each one of these segments is chosen to undergo a somatic recombination event called V-(D)-J recombination, giving rise to the combinatorial and junctional amino acid diversity. Upon encounter with antigens, further diversification and refinement of the Ig repertoire is accomplished by a process known as affinity maturation, which includes somatic hypermutation, receptor editing, somatic gene conversion, and clonal selection (1, 2). In contrast, the V-region diversity of TCR is fixed through the selection process in the thymus and maintained without any modification (3). Although these molecules are likely to be derived from a common ancestral receptor molecule, much more complex molecular mechanisms are used for the refinement of the V-region repertoire of Ig than TCR after maturation of lymphocytes.

The Ig molecule is encoded by three independent gene loci, namely Ig\( _{k} \) and Ig\( _{\lambda} \) genes for the L chain and IgH genes for the H chain, which are located on chromosome 2 (4, 5), chromosome 22 (5, 6), and chromosome 14 (7), respectively. Each of these loci spans a large DNA region of from one to a few megabases (Mb)\(^1\) (8–12). Although antibody function is determined by the complementation of L and H chains, accumulating evidence suggests that the major contribution to the generation of the diversity and specificity of Ig is from the H-chain molecule. Existence of an

\(^1\)Abbreviations used in this paper: Mb, megabase; ORF, open reading frame; RSS, recombination signal sequences; YAC, yeast artificial chromosome.
additional set of gene segments, namely D segments, and their involvement in V-D-J recombination increases enormously the sequence variability of the V\(H\) region. Receptor editing by rearrangement of the silent alelle has not been reported for the H-chain locus (13, 14), possibly indicating a critical role of the H chain for the antigen specificity of the Ig molecule.

It is, therefore, important to have the complete structure of the human V\(H\) locus in order to understand the origin and behavior of the human immune repertoire. In addition, such studies will be useful in designing humanized antibodies. One of the best examples is the establishment and analysis of the xenomouse, which has deletions of the endogenous \(J_H\) and \(J_K\) loci but carries human V\(H\) and V\(K\) segments as transgenes to produce known human antibodies (15). Knowledge of the number and organization of germ-line V\(H\) and V\(K\) segments is essential to test the correlation between the germline repertoire and B lymphocyte repertoire formation in vivo.

Comparison of nucleotide sequences of the 5' regulatory region of V\(H\) segments may tell us how human V\(H\) segments are transcriptionally regulated. Because the recombination signal sequences (R SS) flanking the germline gene segments play a key role in V-D-J rearrangement, it is interesting to test the correlation between the usage of individual V\(H\) segments and the sequence variation within the R SS. Existence of a novel V\(H\) family may provide additional V-region diversity. Isolation of polymorphic markers along the locus will greatly facilitate IgH haplotyping and subsequent systematic genetic analyses to examine the possible association between polymorphisms of the V\(H\) locus and susceptibility to immune disorders. It is also feasible to search for somatic gene conversions that have not yet been demonstrated in humans, the most critical test of which would be the extensive sequence comparison between germ-line and rearranged V\(H\) segments.

From an evolutionary point of view, nucleotide sequence comparison between different parts of the locus will enable us to trace evolution of this multigenic family by DNA reorganization. It would also be very interesting to clarify the origin and nature of the translocated V\(H\) loci on chromosomes 15 and 16 (16, 17). The existence of many V\(H\) pseudogenes suggests frequent gene conversions during the evolution of the V\(H\) locus (18). Moreover, comparative analysis of the structure and organization of the human V\(H\) locus with those of other species or with other multigene loci (V\(K\), V\(\lambda\), and TCR) will provide clues for further understanding the molecular mechanisms that govern the evolution of multigene families. Finally, the V\(H\) locus that lies adjacent to the 14q telomere may provide a suitable candidate for the study of the structure of a human telomere.

To address the above questions, earlier studies on the organization of the human V\(H\) locus have resulted in the completion of the physical map of the entire locus by isolation of yeast artificial chromosome (YAC) clones (8, 9). Here, we report the determination of the complete nucleotide sequence of the 957-kb DNA encompassing the human V\(H\) locus consisting of 123 V\(H\) segments. This permitted the classification of the V\(H\) segments according to their structure and utilization into 39 functional, 1 transcribed, 4 open reading frame (ORF), and 79 pseudogenes. Both frequent DNA reorganization after mammalian divergence and high levels of repetitive elements were revealed. We also identified a putative ancestral V\(H\) segment that is distantly related to V\(H\) segments of other vertebrates as well as to those of humans. A single exon-encoded nonimmunoglobulin gene of unknown function was mapped in the J\(H\) proximal part.

### Materials and Methods

**Isolation of the Distal Part of the V\(H\) Locus.** The J\(H\)-distal region contains one member each of the V\(H\)2 and V\(H\)5 families, which can serve as markers of missing DNA because they contain relatively few numbers (8). Probes and primers specific to these two families were used for the initial screening of cosmid (19) and IC RF YAC libraries (20) by the method as described (21). A contig of 125-kb DNA consisting of four cosmids (M146, U22-1, U22, and M83) and one YAC clone (13.3) which does not overlap with the J\(H\)-proximal 0.8-Mb region was obtained. The remaining gaps between YAC 24/6Y and YAC 13.3 was filled with the P1 clones A1 and H10, obtained by screening a human bacteriophage P1 library (22) with the primers corresponding to the 5' terminus of Y24 and the coding sequence of the V2-70 segment. The probe representative for the human telomere repeat was synthesized as described previously (23) and used for hybridization.

**Nucleotide Sequencing of the V\(H\) Locus.** Two different methods were used to determine the nucleotide sequence. 637-kb regions whose plasmid subclones were available were sequenced by a primer walking method. The remaining DNAs were sequenced by a combination of shotgun and primer walking methods as follows: (a) insert DNA (average size ~3-4 kb) for the shotgun library was obtained from cosmid and P1 clones either by partial digestion with Sau3AI or by mechanical shirring and subsequent fractionation by agarose gel electrophoresis; (b) plasmid DNA of 96 shotgun clones from each cosmid or P1 clone was used for the first round sequencing analysis by using vector primers of both ends. The 192 sequences obtained were assembled to generate contigs by Sequencher software (Gene Codes Corporation, Inc., Ann Arbor, MI). (c) The remaining gaps between contigs were then filled by primer walking using the plasmid DNA of the shotgun clones that bridge different contigs as a template. Accuracy of the nucleotide sequence was estimated to be 99.98% by comparison of the two sequences of 23-kb DNA between the V6-1 segment and D gene cluster from two independent cosmid clones.

**Identification of Nonimmunoglobulin and Repetitive Sequences.** Eight nonimmunoglobulin genes were identified by BLASTN and analyzed in detail using GENE TRY X-MAC version 9.0 (Software Development Co. Ltd., Tokyo, Japan). Content and distribution of genome-wide repetitive sequences were extensively searched by CENSOR (24) at the Genetic Information Research Institute (Palo Alto, CA) as well as by dot matrix analysis.

**Molecular Evolutionary Analysis.** Optimal alignment of nucleotide sequences was obtained by visual inspections maximizing the sequence homology between any pair of the V\(H\) segments. Introns sequences were not included for the analysis. The evolutionary distance \(k_s\) was calculated by the simple Poisson model correction as \(k_s = -\ln[1 - (4/3)K_s]\) (25), where \(K_s\) represents the nonsynonymous substitution per site between sequences compared. The usage of \(K_s\) (synonymous substitution per site) for the alignment is not appropriate because \(K_s\) is saturated in many
pairs of the V_{H} segments compared. The evolutionary tree was inferred by the neighbor-joining method (26).

To estimate the divergence time between V_{H}3/V_{H}4 units, optimal alignment of spacer sequences was obtained by the methods as described (27, 28) together with visual inspections, and the tree was constructed by the neighbor-joining method (26). The divergence time of duplicated copies $T$ was estimated by the equation $T = k/2v$, where $k = -ln[1 - (4/3)K ]$. $K$ represents the simple nucleotide difference between sequences compared. The evolutionary rate $v$ was calculated as $1.4 \times 10^{-9}$ per site per year by comparison of spacer DNA sequences among primate β-globin gene clusters (human, orangutan, Old World monkey, and New World monkey) (data not shown).

**Results and Discussion**

Complete Mapping of the Human V_{H} Locus. Previously, we isolated and analyzed the J_{H}-distal 0.8-Mb region of the human V_{H} locus. 64 V_{H} segments (V6-1 to V3-64) have been completely sequenced and were categorized into 33 structurally functional and 31 pseudo V_{H} segments (8). In this study, we further extended the region by screening and characterization of human YAC, P1, and cosmid clones (Fig. 1). The newly isolated region encompasses the 170-kb DNA upstream of the V3-64 segment and its J_{H}-distal end hybridized with a human telomere repeat probe. Physical mapping and Southern blot analysis identified 16 V_{H} segments and additional 9 DNA fragments that weakly hybridize with human V_{H} probes within the 170-kb region (Fig. 1). Subsequent comparison of the physical map with that of ylgh6 (9) revealed that the content and organization of V_{H} segments are almost identical except that ylgh6 carries additional 7-kb DNA at the telomeric end. Since the ylgh6 clone was isolated by the activity of the human telomere in yeast (9), it may extend to the 14q terminus.

Figure 1. Organization of the human immunoglobulin V_{H} locus. The 957-kb DNA is represented by the four collections of thick horizontal lines with the 3' end at the bottom right corner. V_{H} segments belonging to different V_{H} families are indicated by vertical lines of different colors with their names on the upper row. Pseudogenes and newly identified V_{H} segments are indicated with a P and an asterisk at the end of the name, respectively. Full height vertical lines represent V_{H} segments without truncation while those containing truncation at the 5'-, 3'-, and both 5'- and 3'-portions are indicated by half-height upper, lower, and middle lines, respectively. An enlarged physical map of the 39-kb DNA of the human D gene cluster is also shown. Locations of D segments of six families are shown by diamonds of different colors with their names.

Eight nonimmunoglobulin genes are shown with their names by short arrows of different colors indicating the transcriptional orientation. 13 locus-specific homology units are indicated by boxes of different colors in the middle row. Different classes of sequences are shown in the lower row: (a) Alu (red), MIR (magenta); (b) LINE1 (green), LINE2 (dark green); (c) retrotransposons (yellow), retoviral and other LTRs (blue); (d) DNA transposons (black); (e) medium reiteration frequency repetitive sequences (purple); and (f) simple repeats (cyan). DNA clones covering the locus are shown at the bottom. The YAC clone Y13.3, cosmid clones M146, U22-1, U22, and M83, as well as P1 clones H10 and A1 were newly isolated in this study whereas the others have been described previously (8). The nucleotide sequence was deposited in DDBJ/GenBank/EMBL database under the accession number AB019437-AB019441.
Complete Nucleotide Sequence of the Human V_H Locus.

The complete sequence of 957,090 bp between the J_H segment and the telomeric part of chromosome 14q was determined. The region contains a total of 123 V_H segments of 7 different families (Fig. 1 and Table 1). The V_H segments are localized in the 883-kb DNA between 73 and 956 kb upstream of the J_H cluster. The 5' most V_H segment, V3-82P, is located only 1,480 bp downstream of the 5' terminus of YAC 13.3. Highly interspersed organization of the V_H segments belonging to seven different families was confirmed (19). The lengths of intergenic regions are quite variable; the average distance between neighboring V_H segments is ~6.8 kb, with the longest being 41.4 kb (V1-2/V4-1.1P) and the shortest 418 bp (V3-67.2P/V4-67.1P). However, a clustered distribution of V_H segments is not evident, unlike the human Igk locus where five V_k clusters are physically separated by long spacer DNA (11, 12). The transcriptional polarities of the 123 V_H segments are the same as that of the J_H segments, unlike the human Igk locus in which distal 36 V_k segments are in an inverted transcriptional orientation because of the gross inversion of 440-kb DNA (10).

In the J_H-distal 170-kb region, the existence of 17 V_H segments (V3-65P to V7-81) was suggested by earlier physical mapping studies (9). We identified 16 of them at the positions proposed (9) and classified them into 7 structurally functional V_H segments and 9 pseudogenes. However, we failed to identify the V_H sequence corresponding to the V7-77 segment even though the physical maps of the corresponding portions are exactly identical.

Many DNA fragments of YACs and cosmid clones in the V_H locus weakly hybridized with V_H probes (although such hybridization was not detectable by Southern blotting of human genomic DNA), suggesting the presence of additional V_H-related sequences including possible novel human V_H families. Indeed, the mouse Q52 family does not have human counterparts that show >66% nucleotide sequence homology (18). We identified 43 such V_H-related sequences in the total locus and classified them into 19 V_H3, 22 V_H4, and 1 V_H7 segments according to the homology to known 7 V_H families, making the total number of the human V_H segments to be 123 (Tables 1 and 2). Unfortunately, all of these 43 newly identified V_H segments have defects (Table 2) and thus are categorized as pseudogenes, excluding the possibility of novel V_H families in humans. Of note, only three V_H segments (V3-30.2P, V3-33.2P, and V7-34.1P) have the basic V_H structure while the other 40 V_H segments contain the truncation.

The D region gene cluster consists of 26 D segments within the 39-kb DNA between 53 and 14 kb upstream of the J_H segments (Fig. 1). Because all of the 26 D segments belong to some of the known six families, we named each D segment based on the family and localized duplication unit, thus following the nomenclature of earlier studies (29, 30). The organization and numbering of D segments are in accordance with that proposed in earlier studies (29, 30) and confirmed by nucleotide sequencing analysis (31) in that the D cluster consists of four copies of a 9-kb element containing a set of six D family segments in the order 5'-D_L15-D_L10-D_L5-D_L0-D_L5-D_L10-D_L15-3'. An extensive analysis of D segment usage in V_H cDNAs successfully classified the 27 D segments, including the unique D_{052} segment that is located in the J_H gene cluster (32), into 25 functional and 2 pseudogenes (31).

The Total Number of the Functional Human V_H Segments. The definition of a functional V_H segment is important and essential for determining their number in the human V_H locus because there is some discrepancy regarding the classification of V_H segments into functional or pseudogene segments, a discrepancy which is in part attributable to the incomplete nucleotide sequences of some V_H segments (8, 9, 33). However, given the complete sequences of all the V_H segments, we can propose the following criteria for the functional V_H segment. The functional V_H segment should have an intact exon-intron structure, a complete ORF, and no fatal defects in RSS. In addition, expression of the V_H segment should be confirmed by identification of the given V_H sequence in data bases of full-length V_H cDNA. Identification of a partial cDNA sequence is not sufficient because a part of the V3-47P sequence is found in the V_H cDNA database even though V3-47P must be a pseudogene because of a point mutation at the initiation codon (ATG to AGG) (18). Transcription of a rearranged TCR V_beta segment which carries a defect in splicing signal sequence has also been demonstrated (34). Needless to say, the best proof for the functional V_H segment is to identify its sequence in the IGH protein database, although some V_H segments might be difficult to identify because of hypermutation.

Therefore, we looked for the full-length V_H cDNAs and proteins that correspond to the 40 structurally functional V_H segments mentioned above. 37 of them fulfill the requirement for the functional V_H segment since they are utilized for H-chain polypeptides (Tables 1 and 2). The V4-28 segment shares >97% homology with partial V_H cDNA sequences in the database. Although it appears to be transcribed, its translation product remains to be identified. Hence, it is classified as the second group, transcribed.

Table 1. Summary of the Human V_H Segments

| Classification  | VH family | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Total |
|-----------------|-----------|---|---|---|---|---|---|---|------|
| Functional      |           | 9 | 3 | 19| 6 | 1 | 1 | 0 | 39   |
| Transcribed     |           | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1    |
| ORF             |           | 0 | 0 | 3 | 0 | 0 | 0 | 1 | 4    |
| Pseudogene      | Point mutation | 3 | 1 | 21| 2 | 0 | 0 | 2 | 29   |
|                 | Truncation | 2 | 0 | 22| 23| 1 | 0 | 2 | 50   |
| Total           |           | 14| 4 | 65| 32| 2 | 1 | 5 | 123  |
Table 2. Summary of the Human V<sub>\alpha</sub> Segments

| N. Name | bp from J<sub>\alpha</sub> | 5' regulatory region | RS | 7mer | Defects in the pseudogenes |
|---------|-----------------|-------------------|-----|-----|-----------------|
| H-1     |                 |                   |     |     |                 |
| 1-2     | 121362          | CTCATGA           | 2   | 19  | TAAATAC 82 + +  CACAGTG 23 TCAAGAAACC |
| 1-3     | 139937          | CTCATGA           | 2   | 8   | TGAATAT 77 + +  CACAGTG 23 TCAAGAAACC |
| 1-8     | 207771          | CTCATGA           | 2   | 19  | TAAATAT 81 + +  CACAGTG 23 TCAAGAAACC |
| 1-18    | 310253          | TTCATGA           | 2   | 12  | TATAGAT 76 + +  CACAGTG 23 TCAAGAAACC |
| 1-24    | 401835          | CTCATGA           | 2   | 19  | TAAATAC 80 + gc/ag CACAGTG 23 TCAAGAAACC |
| 1-45    | 631622          | CTCATCA           | 2   | 19  | TAAATAT 81 + +  CACAGTG 23 TCAAGAAACC |
| 1-46    | 623540          | CTCATGA           | 2   | 19  | TAAATAT 81 + +  CACAGTG 23 TCAAGAAACC |
| 1-58    | 747064          | CTCATGA           | 2   | 19  | TAAATAT 81 + +  CACAGTG 23 TCAAGAAACC |
| 1-69    | 838623          | TTCATGC           | 2   | 19  | TAAATAT 81 + +  CACAGTG 23 TCAAGAAACC |
| 2-5     | 162833          |                   | --  | 26  | TGGAAAA 41 + +  CACAAAG 23 ACAAAACC |
| 2-26    | 426348          |                   | --  | 26  | TGGAAAA 41 + +  CACAAAG 23 ACAAAACC |
| 2-70    | 847518          |                   | --  | 26  | TGGAAAA 41 + +  CACAAAG 23 ACAAAACC |
| 3-7     | 187109          |                   | --  | 18  | ATGAAA 100 + +  CACAGTG 23 ACAAAACC |
| 3-9     | 220985          |                   | --  | 18  | ATGAAA 100 + +  CACAGTG 23 ACAAAACC |
| 3-11    | 241936          |                   | --  | 18  | ATAAA 101 + +  CACAGTG 23 ACAAAACC |
| 3-13    | 258443          |                   | --  | 18  | ATGAAA 101 + +  CACAGTG 23 ACAAAACC |
| 3-15    | 279028          |                   | --  | 18  | ATGAAA 101 + +  CACAGTG 23 ACAAAACC |
| 3-20    | 336289          |                   | --  | 17  | ATGAAA 100 + +  CACAGTG 23 ACAAAACC |
| 3-21    | 360380          |                   | --  | 18  | ATGAAA 100 + +  CACAGTG 23 ACAAAACC |
| 3-23    | 393910          |                   | --  | 18  | ATGAAA 100 + +  CACAGTG 23 ACAAAACC |
| 3-30    | 459712          |                   | --  | 18  | ATGAAA 100 + +  CACAGTG 23 ACAAAACC |
| 3-33    | 484429          |                   | --  | 18  | ATGAAA 100 + +  CACAGTG 23 ACAAAACC |
| 3-43    | 594900          |                   | --  | 18  | ATGAAA 101 + +  CACAGTG 23 ACAAAACC |
| 3-48    | 662523          |                   | --  | 18  | ATGAAA 100 + +  CACAGTG 23 ACAAAACC |
| 3-49    | 681653          |                   | --  | 18  | ATGAAA 101 + +  CACAGTG 23 ACAAAACC |
| 3-53    | 727376          |                   | --  | 18  | ATGAAA 98 + +  CACAGTG 23 ACAAAACC |
| 3-64    | 782450          |                   | --  | 18  | ATGAAA 101 + +  CACAGTG 23 ACAAAACC |
| 3-66    | 799737          |                   | --  | 18  | ATGAAA 100 + +  CACAGTG 23 ACAAAACC |
| 3-72    | 867647          |                   | --  | 18  | ATGAAA 101 + +  CACAGCG 23 ACAAAACC |
| 3-73    | 879647          |                   | --  | 18  | ATGAAA 101 + +  CACAGTG 23 ACAAAACC |
| 3-74    | 887385          |                   | --  | 18  | ATGAAA 90 + +  CACAGTG 23 ACAAAACC |
| 4-4     | 146795          |                   | --  | 39  | TTTAATT 59 + +  CACAGTG 23 ACAAAACC |
| 4-31    | 473900          |                   | --  | 38  | TTTAATT 59 + +  CACAGTG 23 ACAAAACC |
| 4-34    | 498280          |                   | --  | 39  | TTTAATT 59 + +  CACAGTG 23 ACAAAACC |
| 4-39    | 545311          |                   | --  | 39  | TTTAATT 58 + +  CACAGTG 23 ACAAAACC |
| 4-59    | 759841          |                   | --  | 39  | TTTAATT 59 + +  CACAGTG 23 ACAAAACC |
| 4-61    | 763817          |                   | --  | 39  | TTTAATT 59 + +  CACAGTG 23 ACAAAACC |
| 5-51    | 703418          |                   | --  | 18  | ATCTTA 79 + +  CACAGTG 23 CTAAACC |
| 6-1     | 74312           |                   | --  | 19  | TTAAAT 78 + +  CACAGTG 23 ACAAAACC |
| Transcribed ORF | | | | | |
| 4-28    | 449201          |                   | --  | 38  | TTAAAT 59 + +  CACAGTG 23 ACAAAACC |
| 3-16    | 290601          |                   | --  | 18  | ATGAAA 94 + +  TCTGTG 23 ACAAAACC |
| 3-35    | 534033          |                   | --  | 18  | ATAAA 95 + +  CACTGAG 23 ACAAAACC |
| 3-38    | 535112          |                   | --  | 18  | ATCAGA 95 + +  CACAGTG 23 ACAAAACC |
| 7-81    | 951482          | TTTCATGA          | 2   | 18  | TTTAATT 79 + +  CACAGTG 23 TCAAGAAACC |
| Pseudogene with point mutation(s) | | | | | |
| 1-17P   | 299723          | CTCATGA           | 2   | 19  | TAAATTT 79 + +  CACAGTG 23 TCAAGAAACC |
| 1-67P   | 805315          | CTCATGA           | 2   | 16  | TACAGAT 77 + +  CACAGTG 23 TCAAGAAACC |
| 1-68P   | 828559          | CTCATGA           | 2   | 18  | TAAATAT 76 + gl/gg CACAGTG 23 TCAAGAAACC |
| 2-10P   | 228236          |                   | --  | 26  | TGGAAAA 44 + +  CACAGTG 23 ACAAAACC |
| 3-6P    | 180487          |                   | --  | 18  | ATGAAA 98 + +  TGCTGA 23 ACAAAACC |
| 3-22P   | 383077          |                   | --  | 18  | ATGAAA 101 + +  CACAGTG 23 ACAAAACC |
| 3-25P   | 434329          |                   | --  | 18  | ATAAA 100 + +  CACAGTG 23 ACAAAACC |
| 3-29P   | 456071          |                   | --  | 14  | TGTAAA 101 + +  CCAAGT 23 ACAAAATT |
| 3-30.2P | 469013          |                   | --  | 18  | ATGAAA 101 + +  CCAAGT 23 ACAAAATT |
| 3-32P   | 480784          |                   | --  | 18  | ATGAAA 101 + +  CCAAGT 23 ACAAAATT |
| 3-33.2P | 493733          |                   | --  | 18  | ATGAAA 101 + +  CCAAGT 23 ACAAAATT |

(continued)
| Name     | bp from JH1 | 5' regulatory region | R SS | Defects in the pseudogenes |
|----------|-------------|----------------------|------|---------------------------|
| 3-37P    | 521282      | --                   | --   | CATGGT -- 23 CACAGACC    | 522(2), 1 bp-D(16,21), 2 bp-D(163), 10 bp-D(90-93) |
| 3-41P    | 567750      | --                   | --   | CATGGT -- 23 CACAGACC    | 54(7), 1 bp-D(71) |
| 3-47P    | 643217      | --                   | --   | AGG -- 23 ATACAAATT      | ATG to AGG (--19), 5'T-T (108 bp upstream of --19) |
| 3-50P    | 690799      | --                   | --   | CCAATG -- 23 CACACAAATT  | 5'--2,33,36,47,58,92, 1 bp-D(16), ATG to ATA (--19) |
| 3-52P    | 711066      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5(9) |
| 3-54P    | 726051      | --                   | --   | CATGGA -- 23 CACACAAATT  | 53(37,52,52A,94) |
| 3-60P    | 755914      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5(66), 1 bp-D(16) |
| 3-62P    | 767844      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5(46) |
| 3-64P    | 776951      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5(33) |
| 3-65P    | 790304      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5(132) |
| 3-71P    | 851113      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5(59) |
| 3-75P    | 900633      | --                   | --   | CATGGA -- 23 CACACAAATT  | 522(66), 1 bp-D(80), 1 bp-D(22,23) |
| 3-76P    | 904794      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5(53,38,51), 2 bp-D(52A) |
| 3-79P    | 945560      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5(34) |
| 4-55P    | 750317      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5(34) |
| 7-34.1P  | 505202      | TTCTAGA 2            | --   | CATGGA -- 23 CACACAAATT  | 5(34) |
| 7-36P    | 754662      | TTCTAGA 2            | --   | CATGGA -- 23 CACACAAATT  | 5(34) |
| 4-14P    | 271076      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5(34) |
| 1-12P    | 247559      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5(34) |
| 3-2.1P   | 136244      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(50) |
| 3-5.1P   | 164252      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(85) |
| 3-5.2P   | 170347      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(12) |
| 3-13.1P  | 267558      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(6) |
| 3-16.1P  | 296096      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(51) |
| 3-19P    | 321917      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(11) |
| 3-26.1P  | 434239      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(11) |
| 3-35.1P  | 524512      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(11) |
| 3-38.1P  | 587869      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(11) |
| 3-47.1P  | 655704      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(11) |
| 3-51.1P  | 708556      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(11) |
| 3-57P    | 743949      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(11) |
| 3-67.2P  | 812748      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(11) |
| 3-67.3P  | 812761      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(11) |
| 3-67.4P  | 819680      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(11) |
| 3-82P    | 956317      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(11) |
| 3-22.2P  | 389722      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(11) |
| 3-36P    | 537398      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(11) |
| 3-44P    | 602655      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(11) |
| 3-51.1P  | 708556      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(11) |
| 3-67.2P  | 812748      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(11) |
| 3-67.4P  | 819680      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(11) |
| 3-76.1P  | 905628      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(11) |
| 4-11P    | 79503       | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(13) |
| 4-15.1P  | 288747      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(10) |
| 4-20.1P  | 338221      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(10) |
| 4-22.1P  | 388477      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(10) |
| 4-26.2P  | 439482      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(4) |
| 4-28.1P  | 454194      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(4) |
| 4-30.1P  | 687131      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(4) |
| 4-31.1P  | 478899      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(4) |
| 4-33.1P  | 491850      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(4) |
| 4-44.1P  | 610336      | --                   | --   | CATGGA -- 23 CACACAAATT  | 5'T(4) |
V3-35 and V7-81 segments did not correspond to any V<sub>H</sub> cDNAs. To be conservative, however, we allowed their possible usage in the V-D-J rearrangement and classified them as the third ORF group (Tables 1 and 2) (discussed below).

Previous classifications of V1-24, V1-58, V3-16, V3-38, and V5-78P were corrected in this study. The V1-24 and V1-58 segments containing a complete ORF except for the abnormal splicing signal GC/AG (18) were found in a recently published IgH protein and therefore categorized as pseudogene. We include the V3-16 and V3-38 segments in the ORF group even though they contain highly diverged RSS heptamers (discussed below). The V5-78P segment, which completely loses RSS by replacement of an unknown sequence, was not identified in cDNAs and therefore is classified as pseudogene. Conversely, none of the 79 pseudogenes that have defects in the V<sub>H</sub> gene structure itself were found in full-length V<sub>H</sub> cDNAs. Taken together, the corrected numbers of the functional, transcribed, ORF, and pseudogenes are 39, 1, 4, and 79, respectively (Table 1).

This raises the issue of how many functional V<sub>H</sub> segments are required for the full antibody repertoire in humans. The immune response of transgenic mice carrying human Ig YAC clone (xenomice) gives some hints on this (15). The xenomouse II strains with 35 functional V<sub>H</sub> and 18 functional V<sub>K</sub> segments develop human adultlike antibody repertoires with high levels of mature B lymphocytes and high-affinity human antibodies against diverse antigens while the xenomouse I strains bearing 5 functional V<sub>H</sub> and 3 functional V<sub>K</sub> segments are capable of only modest immune responses. This strongly suggests the importance of the complete germline V gene repertoire in the highly diverse human antibody response.

A analysis of the 5'-regulatory Sequences. The 5' flanking region of V<sub>H</sub> segments plays an important role in the regulation of H-chain gene expression. Already, we have shown the striking conservation of the upstream sequences of V<sub>H</sub> segments in a family-specific manner (18). This region contains two 5's acting elements, namely the octamer motif, which is essential for correct transcription of Ig genes, and the TATA box required by the general transcription machinery (35). In this study, 500 bp of 5'-flanking sequences from the 79 V<sub>H</sub> segments without 5' truncation were aligned to identify and compare these two motifs, as well as unknown conserved sequences across V<sub>H</sub> families that might also act as 5's acting elements. We found that 40 out of the 44 functional, transcribed, or ORF V<sub>H</sub> segments contain an octamer sequence 100% identical to the consensus (ATGCAAAT) (Table 2). Slightly less conserved are ATCCAAAT in V3-53, AGGCAAAT in V6-1, and ATGCAGGT in V3-20 although these three V<sub>H</sub> segments have been shown to be transcribed. The V3-38 segment, an ORF group member, completely loses the octamer and TATA motifs due to 5'-truncation. Because this V<sub>H</sub> segment also contains a point mutation in a critical site of the RSS (discussed below), it might not be capable of rearrangement or transcription. Interestingly, the octamer sequence of pseudogenes appears much less conserved. In the 33 V<sub>H</sub> pseudogenes with the octamer motif in their 5'-flanking regions, as many as 15 have diverged octamer motifs, as well as unknown conserved sequences across VH families that might also act as 5's acting elements.

In contrast, the distance between the octamer sequence and the TATA consensus, as well as the sequence of the TATA box itself, are well conserved within the same V<sub>H</sub> family but vary between different V<sub>H</sub> families. Another motif, the heptamer, is reported to be essential for
full $V_H$ promoter activity in mouse lymphoid cells (36), is found 2 bp upstream of the octamer in the $V_H1$ and $V_H7$ family members only, confirming our previous observation (18) that the heptamer is not essential for the expression of $H$-chain genes in humans. We could not find any other conserved nucleotide motifs across the seven $V_H$ families by nucleotide sequence alignment. However, such novel acting elements may be identified by investigation of the correlation between promoter activity and nucleotide sequence variation of the 5′-flanking region.

RSS and V-D-J rearrangement. The RSS of $V_H$ segments is located immediately downstream of the coding region sequence and is composed of conserved heptamer (CACAGTG) and nonamer (ACAATAATC) sequences, which are separated by 23-bp spacer nucleotides. Recent in vitro analysis of RSS (37, 38) clearly demonstrated that the first three positions of the heptamer and the fifth and sixth positions of the nonamer are critical for efficient V-D-J recombination. Among the 123 $V_H$ segments identified in this study, 108 have RSS heptamer and nonamer signals (Table 2). All 40 of the functional or transcribed $V_H$ segments maintain the first three nucleotides of the heptamer signal (CAC) intact, although five of them are slightly different from the consensus in the four 3′ nucleotides (AGTG).

Slightly more variation can be seen in the RSS nonamers of these 40 $V_H$ segments as follows: (a) the fifth and sixth positions are highly conserved except for two $V_H2$ segments; (b) C is more frequently used than A at the fourth position; (c) the $V_H1$ segments have a family-specific nonamer, TCAGAAAGC, capable of V-D-J rearrangement. The G nucleotide at the fifth position of V2-26 and V2-70 appears to maintain recombination efficiency as shown in the human $V_L$ genes (12). In the ORF group, V3-35 and V7-81 contain mutated RSS heptamers (CAGTGG and CAC -CATG, respectively) although their third heptamer nucleotides retain the CAC consensus. It is not clear whether this might affect the efficiency of the V-D-J recombination. In contrast, the first three positions of the heptamer signal of V3-16 (TCC) and V3-38 (TAC) have diverged even though their nonamer signals remain well conserved. This might be the reason why these two $V_H$ segments cannot be found in functional V-D-J rearrangements. In all 44 $V_H$ segments, the spacer length is strictly maintained at 23 nucleotides (Table 2). Given the number of the human $V_H$, $D$, and $J_H$ segments now, the combinatorial diversity of the human $V_H$ genes can be calculated as $40\text{(functional/transcribed } V_H \text{ segments)} \times 25\text{(functional } D \text{ segments)} \times 6\text{( } J_H \text{ segments)} = 600$. Of course, this value is only approximate, as the number of $V_H$ and $D$ segments shows allelic variation.

The RSS of the 64 pseudogenes appear much more diverged (Table 2). First, 26 $V_H$ pseudogenes carry mutations at one or more of the five critical positions (A to G mutation at the fifth nucleotide of the nonamer is excluded). In addition, V3-36P and V3-76.1P have lost the nonamer signal due to the truncation in the 23-bp spacer. Second, the spacer length is not well conserved; the 7, 6, 2, and 1 $V_H$ segments have spacer lengths of 662.0x792.0

2158 Complete Sequence of Human $V_H$ Locus
cluence based on amino acid comparisons of only a limited number of \( V_{H} \) segments from other species.

Dot matrix analysis of the 957-kb sequence against itself failed to find the large scale genome duplication. However, we found 13 DNA sequences of variable length (4–24 kb) that appear repeatedly across the \( V_{H} \) locus (Fig. 1). These homologous units constitute 67% of the entire locus and contain the DNA fragments previously shown to cross-hybridize with 14 intergenic probes by Southern blot analysis (42). Of note is the DNA sequence that appears 11 times in the region between 380 and 955 kb upstream of the \( J_{H} \) segments (indicated by red boxes in Fig. 1) and contains a \( V_{H} \), segment with the 5’ truncation flanked by a \( V_{H} \), segment at its upstream end. Among them, the nucleotide sequence of the spacer DNA between the two \( V_{H} \), segments is highly conserved in the 10 different \( V_{H} \),/\( V_{H} \) units. The spacer sequences were aligned to estimate the divergence time of these \( V_{H} \),/\( V_{H} \) units.

As shown in Fig. 2B, nine DNA duplication events took place between 132 and 10 million years ago. Of note, seven events occurred after the mammalian divergence 75 million years ago (43), demonstrating the recent high frequency reorganization of the human \( V_{H} \) locus. The 48-kb DNA ranging between the \( V_{H}-33.2P \) and \( V_{H}-28 \) segments consists of four copies of the \( V_{H} \),/\( V_{H} \) units, each of which contains 2 \( V_{H} \),/\( V_{H} \) units, the 48-kb region was considered to be generated by tandem duplication of the 24-kb DNA (8). A similar score in the divergence time was obtained between the corresponding pairs: 13 million years ago between \( V_{H}-33.2P/V_{H}-31.1P \) and \( V_{H}-29P/V_{H}-28.1P \), and 10 million years ago between \( V_{H}-33.2P/V_{H}-31.1P \) and \( V_{H}-30.2P/V_{H}-30.1P \) (Fig. 2B). This strongly suggests the initial internal DNA duplication within the copy (73 million years ago) and subsequent recent gross DNA duplication (∼10–13 million years ago). Clustering of two \( V_{H} \),/\( V_{H} \) units is also seen in the 19-kb DNA between \( V_{H}-3-54P \) and \( V_{H}-51.2P \). However, in this cluster the upstream \( V_{H}-35P/V_{H}-53.1P \) pair is nearest to the \( V_{H}-33.2P/V_{H}-33.1P \) and the \( V_{H}-29P/V_{H}-28.1P \) pairs (39 million years ago) while the \( V_{H}-35P/V_{H}-51.2P \) pair is most distantly related to the other nine copies (132 million years ago). This excludes the possibility of another gross duplication. The most recent duplication, which took place 10 million years ago between \( V_{H}-33.2P/V_{H}-33.1P \) and \( V_{H}-30.2P/V_{H}-30.1P \), suggests the existence of the both pairs in gorilla and chimpanzee but not in gibbon (44, 45). A similar calculation was performed between DNA regions containing the truncated \( V_{H} \) segments, \( V_{H}-67.3P/V_{H}-67.2P \) and \( V_{H}-5.2P/V_{H}-5.1P \) and the divergence time was found to be 61 million years ago, again after the divergence between mouse and human (data not shown).

Identification and Characterization of Nonimmunoglobulin Genes. We identified eight DNA sequences in the 957 kb that are highly homologous to known DNA sequences in the databases. Three of them were mapped within the \( V_{H} \), rare downstream part (Fig. 1). The 7,883-bp cDNA of KIAA0125 (46) displayed 99.8% identity to the DNA sequence between the \( V_{H}-6 \) segment and the \( D \) gene cluster. KIAA0125 is encoded by a single exon and its transcriptional orientation is in the opposite direction to that of the \( V_{H} \) segments. This cDNA has several interesting features, including an extremely short putative protein coding region (77 amino acid residues) and, in contrast, very long 5′- and 3′-untranslated regions (1,289 and 6,087 nucleotides, respectively) (46). In addition, its 3′-untranslated region contains two tandem repeats of 68- and 48-bp units. Moreover, the expression of KIAA0125 is limited to lymphoid organs (46). It is interesting to investigate its physiological roles because these characteristics are often found in imprinted genes including the H19 gene whose transcripts work as an RNA component of ribonucleoprotein particle (47–49).

We also found two processed pseudogenes within the largest spacer DNA between the \( V_{1} \)-2 and \( V_{4} \)-1P segments (Fig. 1). The 681-bp DNA segment located ∼105 kb upstream of the \( J_{H} \) segments is 94.9% homologous to the human ribosomal protein S8 cDNA (50). Another DNA segment of 2,348 bp located ∼133 kb upstream of the \( J_{H} \) cluster shows 89.9–91.4% homology to a series of cDNAs of the metalloprotease-like, disintegrin-like, cysteine-rich protein family of Macaca (51) (Fig. 1). Two copies of the 1.7-kb sequence showing the 77% homology to the 3′-half of human leukemia virus receptor 1 cDNA were identified in the spacer DNA between \( V_{H}-18/V_{H}-17P \) and \( V_{H}-47.1P/V_{H}-67 \) (52). These two distant DNA segments are 90.4% homologous and contain the common 5′-truncation, suggesting the integration of reverse transcribed human leukemia virus receptor 1 mRNA followed by the truncation and DNA duplication. Similarly, three copies of the DNA segment that show >86% nucleotide sequence homology to the 3′-most 500 bp of the human golgin-245 cDNA are also scattered within the locus (53) (Fig. 1).
Structure of the Human 14q Subtelomeric Region. Physical mapping studies (9) suggest that the 14q terminus is located several kilobases upstream of the 5′ end of the YAC clone 13.3. Indeed, we could not find the complete repeat of human telomere-specific hexanucleotides CCCTAA at the distal end. However, the 5′-most 873 nucleotides contained a divergent telomeric repeat array of 181 bp which shares 69.1% homology to poly-(CCCTAA) sequence. Interestingly, this 873-bp DNA, which is unique within the VH locus, showed striking similarity to the telomeric regions of human chromosome 4p (92.4%), 4q (93.0%), and 22q (94.3%) that have been deposited in GenBank/EMBL/DDJB databases. Since the telomeric region is hyper-recombinogenic and the telomeric region of one chromosome in an individual often corresponds to that of another chromosome in others (54), these might represent alleles. It is reported that the divergent telomeric repeat array appears several kilobases downstream of the complete telomere repeat (54). In chromosome 4p and 4q, the homologous DNAs to the above 873 bp are located 5 and 13 kb, respectively, downstream of the authentic telomere repeat. Taken together, the distance between chromosome 14q terminus and the 5′ end of our contig would be ~10 kb.

GC Content and Genome-wide Repetitive Elements. Cytogenetically, band 14q32.33 is an early replicating and GC-rich R band. Other characteristics of R band are being rich in housekeeping genes, having G + C-rich third coding bases, and SINE-rich/LINE-poor genome composition. However, studies on the DNA replication classified the human VH locus as a G-bandlike gene locus that replicates at the late stage of S phase, whereas the CH locus was classified as R-bandlike (55). The third position in codons of VH segments is A + T rich, again inconsistent with the cytogenetical observation (Hayasida, H., unpublished observations). We found that in this locus 893 kb upstream is A + T predominant (average 58.4%) while the 65 kb downstream is rich in G + C nucleotides (average 58.6%). The high G + C percentage at the JH-proximal part appears to continue toward the CH gene region. Existence of polypurine/polypyrimidine tracts has been reported at the boundary of A + T-rich class II and G + C-rich class III gene clusters of human HLA locus (56). In the VH locus, however, such tracts are not evident at the boundary. Nonetheless, the boundary may contain a switch point for DNA replication timing and scaffold-associated regions.

We looked for the content and distribution of various kinds of genome-wide repeats and identified 722 genome-wide repetitive elements, which correspond to as much as 41.8% of the entire locus. They are categorized into 135 SINE elements (133 Alu and 3 MIR), 340 LINE elements (338 LINE1 and 2 LINE2), 213 LTR elements (82 LTR-retrotransposon or MAlR, 69 retroviral LTR, and 62 retrovirus-like other LTR), 5 DNA transposons, and 25 medium reiteration frequency repetitive sequences (Fig. 1). LINE1 element is the largest contributor, constituting 23.2% of the locus while the number of Alu elements is much less than that expected by random distribution (239 copies) and constitutes a relatively small fraction (3.4%). Of note, only 2 copies out of 338 LINE1 elements contain the complete LINE1 structure of ~6 kb whereas 278 copies are <1 kb in size.

Identification of a much larger number of LINE1 element in this study than in previous analysis by Southern blot (42) (44 Alu and 11 LINE1 hybridizing DNA fragments in the JH-proximal 730-kb DNA) is due to the usage of the probe in the previous study, which corresponds to the conserved portion between LINE1 subfamilies, resulting in the failure in detection of smaller copies lacking the conserved portion. In the case of Alu elements, the difference in the number mainly reflects the multiple Alu elements in a single restricted DNA fragment. In the human TCR Vβ locus, LINE1-rich/Alu-poor structure is consistent with its chromosomal location at G band (34). Discordance of the results between nucleotide and cytogenetical analyses in the human VH locus may be attributed to the extraordinary chromosome structure of subtelomeric region. This locus is also rich in LTR elements (13.3% in total). Possible involvement of retroelement in gross changes of genome structure has been suggested recently (57). Abundant LTR elements may explain in part the dramatic difference in the organization of the VH loci between humans and mice.

We thank Dr. Chris T. Amemiya for screening P1 library; Dr. Ted Choi for kind donation of the YAC 13.3 clone; Mr. Hiroshi Suga, Drs. Akira Shimizu, N obuo Nomura, Yoshimichi Ikemura and Fuyuki Ishikawa for valuable comments; Dr. Melvin Cohn for critical reading of the manuscript; Dr. Jean Thierry-Mieg for extensive BLAST and repetitive analysis; Dr. Masazumi Takahashi for computer scripts; and Ms. Hiroe Ohori-Kurooka for technical assistance.

This work was supported in part by grants from the Ministry of Education, Science, Sports, and Culture of Japan and from the Science and Technology Agency of Japan.

Address correspondence to Taku H Ono, Department of Medical Chemistry, Kyoto University Graduate School of Medicine, Yoshida, Sakyo-ku, Kyoto 60601, Japan. Phone: 81-75-753-4371; Fax: 81-75-753-4388; E-mail: honjo@mfour.med.kyoto-u.ac.jp

F. Matsuda's current address is Centre National de Genotypage, BP191-2, rue Geston Cremieux, 91000.
References

1. Tonegawa, S. 1983. Somatic generation of antibody diversity. Nature. 302:505–581.

2. Honjo, T., and S. H. abu. 1985. Origin of immune diversity: genetic variation and selection. Annu. Rev. Biochem. 54:803–830.

3. Davis, M. M. 1990. T cell receptor gene diversity and selection. Annu. Rev. Biochem. 59:475–496.

4. Malcolm, S., P. Barton, C. Murphy, M. A. Ferguson-Smith, D. L. Bentley, and T. H. Rabbits. 1982. Localization of human immunoglobulin kappa light chain variable region genes to the short arm of chromosome 2 by in situ hybridization. Proc. Natl. Acad. Sci. USA. 79:4957–4961.

5. McBride, O. W., P. A. I. ter, G. F. Hollis, D. Swan, M. C. O'ney, and P. Led. 1982. Chromosomal location of human immunoglobulin lambda light chain constant region genes. J. Exp. Med. 155:1480–1490.

6. Erikson, J., J. M. Martinis, and C. M. Croce. 1981. Assignment of the genes for human kappa immunoglobulin chains to chromosome 22. Nature. 294:173–175.

7. Croce, C. M., M. Shander, J. Martinis, L. Cirecure, G. G. D'Ancora, T. W. Dolby, and H. K. Koprowski. 1979. Chromosomal location of the genes for human immunoglobulin heavy chains. Proc. Natl. Acad. Sci. USA. 76:3416–3419.

8. Matsuda, F., E. K. Shin, H. Nagaoka, R. Matsumura, M. Haino, Y. Fukita, T. H. Ichish, T. Imai, J. H. R. Iley, R. Anand, et al. 1993. Structure and physical map of 64 variable segments in the 3' 0.8-megabase region of the human immunoglobulin heavy-chain locus. Nat. Genet. 3:88–94.

9. Cook, G. P., I. M. Tomlinson, G. W. Carter, N. P. Carter, L. Buluwela, G. Winter, and T. H. Rabbits. 1994. A map of the human immunoglobulin V\(\kappa\) locus completed by analysis of the telomeric region of chromosome 14q. Nat. Genet. 7:162–168.

10. Zachau, H. 1995. The human immunoglobulin \(\kappa\) genes. In Immunoglobulin Genes. 2nd ed. T. Honjo and F. W. Alt, editors. Academic Press, London. 173–192.

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

12. Kawaski, K., S. Nishimura, K. Shibuya, A. Shintani, J. L. Schmelts, J. Wang, and N. Shimizu. 1997. O ne-megabase sequence analysis of the human immunoglobulin \(\lambda\) gene locus. Genome Res. 7:250–261.

13. Radi, M. Z., and M. Zouali. 1996. Receptor editing, immune diversification, and self-tolerance. Immunity. 5:505–511.

14. Han, S., S. R. Dillon, B. Z. Zheng, M. Shimoda, M. S. Schissel, and G. Kelsoe. 1997. V(D)J recombination activity in a subset of germinal center B lymphocytes. Science. 278:301–306.

15. Menezes, M. J., L. L. Green, J. R. Corvalan, X. C. Jia, C. E. M. aynard-Currie, X. D. Yang, M. L. Gallo, D. M. Louise, D. V. Lee, K. L. Erickson, et al. 1997. Functional transplants of megabase human immunoglobulin loci recapitulates human antibody response in mice. Nat. Genet. 15:146–156.

16. Nagaoka, H., K. Oizawa, F. Matsuda, H. Hayashida, R. Matsumura, M. Haino, E. K. Shin, Y. Fukita, T. Imai, R. Anand, et al. 1994. Recent translocation of variable and diversity (D) segment of the human immunoglobulin heavy chain from chromosome 14 to chromosomes 15 and 16. Genomics. 22:189–197.

17. Tomlinson, I. M., G. P. Cook, N. P. Carter, R. Elaswarapu, S. Smith, G. Walter, L. Buluwela, T. H. Rabbits, and G. Winter. 1994. Human immunoglobulin V\(\kappa\) and D segments on chromosomes 15q11.2 and 16p11.2. Hum. Mol. Genet. 3:853–860.

18. Haino, M., H. Hayashida, T. Miyata, E. K. Shin, F. Matsuda, H. Nagaoka, R. Matsumura, S. Takaishi, Y. Fukita, J. Fujikura, and T. Honjo. 1994. Comparison and evolution of human immunoglobulin V\(\kappa\) segments located in the 3' 0.8-megabase region: evidence for unidirectional transfer of segment gene sequences. J. Biol. Chem. 269:2619–2626.

19. Kodaira, M., T. Kinashi, I. Umemura, F. Matsuda, T. Ono, Y. O. no, and T. Honjo. 1986. Organization and evolution of variable region genes of the human immunoglobulin heavy chain. J. Mol. Biol. 190:529–541.

20. Larin, Z., A. P. Monaco, and H. Lehrbach. 1991. Yeast artificial chromosome libraries containing large inserts from mouse and human DNA. Proc. Natl. Acad. Sci. USA. 88:4123–4127.

21. Matsuda, F., and T. Honjo. 1993. Physical mapping of variable region gene locus of human immunoglobulin heavy chain. In Methods in Molecular Genetics. Vol. 2. K. Adolph, editor. Academic Press, Orlando, FL. 226–245.

22. Ioannou, P. A., C. T. Amemiya, J. Darn, P. K. Kroisel, H. Shizuya, C. Chen, M. A. Batzer, and P. J. de Jong. 1994. A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. Nat. Genet. 6:68–91.

23. Moyzis, R. K., J. M. Buckingham, L. S. Cram, M. Dani, L. L. Deaven, M. D. Jones, J. M. Sheth, L. R. Raliff, and J. R. Wu. 1988. A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. Proc. Natl. Acad. Sci. USA. 85:6622–6626.

24. Jurka, J., P. Klontowski, V. Dagman, and P. Pelton. 1996. CEN SO R — a program for identification and elimination of repetitive elements from DNA sequences. Comput. Chem. 20:119–121.

25. Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. In Mammalian Protein Metabolism III. H. N. Munro, editor. Academic Press, New York. 21–132.

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

27. Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48:443–453.

28. Berger, M. P., and P. J. Munson. 1991. A novel randomized iterative strategy for aligning multiple protein sequences. Comput. Appl. Biosci. 7:479–484.

29. Siebenlist, U., J. Y. Ravetch, S. Korsmeyer, T. Waldmann, and G. Kelsoe. 1997. V(D)J recombination activity in a subset of germinal center B lymphocytes. Science. 278:301–306.

30. Kodaira, M., T. Kinashi, I. Umemura, F. Matsuda, T. Ono, Y. O. no, and T. Honjo. 1986. Organization and evolution of variable region genes of the human immunoglobulin heavy chain. J. Mol. Biol. 190:529–541.

31. Larin, Z., A. P. Monaco, and H. Lehrbach. 1991. Yeast artificial chromosome libraries containing large inserts from mouse and human DNA. Proc. Natl. Acad. Sci. USA. 88:4123–4127.

32. Jurka, J., P. Klontowski, V. Dagman, and P. Pelton. 1996. CEN SO R — a program for identification and elimination of repetitive elements from DNA sequences. Comput. Chem. 20:119–121.

33. Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. In Mammalian Protein Metabolism III. H. N. Munro, editor. Academic Press, New York. 21–132.

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

35. Needleman, S. B., and C. D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48:443–453.

36. Berger, M. P., and P. J. Munson. 1991. A novel randomized iterative strategy for aligning multiple protein sequences. Comput. Appl. Biosci. 7:479–484.
30. Ichihara, Y., H. Matsuoka, and Y. Kurosawa. 1988. Organization of human immunoglobulin heavy chain diversity gene loci. EMBO J. (Eur. Mol. Biol. Org.) 7:4141–4150.

31. Corbett, S.J., I.M. Tomlinson, E.L.L. Sonnhammer, D. Buck, and G. Winter. 1997. Sequence of the human immunoglobulin diversity (D) segment locus: a systematic analysis provides no evidence for the use of DIR segments, inverted D segments, “minor” D segments or D-D recombination. J. Mol. Biol. 270:587–597.

32. Ravetch, J.V., U. Siebenlist, S. Korsmeyer, T. Waldmann, and P. Leder. 1981. Structure of the immunoglobulin m locus: characterization of embryonic and rearranged J and D genes. C. Cell. 27:583–591.

33. Matsuda, F., and T. Honjo. 1996. Organization of the human immunoglobulin heavy-chain locus. Adv. Immunol. 62:1–29.

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

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

36. Eaton, S., and K. Calame. 1987. Multiple DNA sequence elements are necessary for the function of an immunoglobulin heavy chain promoter. Proc. Natl. Acad. Sci. USA. 84:7634–7638.

37. Ramsden, D.A., J.F. McCleane, D.C. van Gent, and M. Gelbert. 1996. Distinct DNA sequence and structure requirements for the two steps of V(D)J recombination signal cleavage. EMBO J. (Eur. Mol. Biol. Org.) 15:3197–3206.

38. Coumo, C.A., C.L. Mundy, and M.A. Oettinger. 1996. DNA sequence and structure requirements for cleavage of V(D)J recombination signal sequences. Mol. Cell. Biol. 16: 5683–5690.

39. Kabat, E.A., T.T. Wu, H.M. Perry, K.S. Gottesman, and C. Foeller. 1991. Sequences of Proteins of Immunological Interest. 5th ed. N. Inst. Health, Washington D.C. 1137 pp.

40. Reichavi, G., D. Ram, L. Glazer, R. Zakut, and D. Givol. 1983. Evolutionary aspects of immunoglobulin heavy chain variable region (VH) gene subgroups. Proc. Natl. Acad. Sci. USA. 80:855–859.

41. Lee, K.H., F. Matsuoka, T. Kinashi, M. Kodaira, and T. Honjo. 1987. A novel family of variable region genes of the human immunoglobulin heavy chain. J. Mol. Biol. 195:761–768.

42. Matsuzuka, R., F. Matsuoka, H. Nagaoa, J. Fujiwara, E.K. Shin, Y. Fukita, M. Hano, and T. Honjo. 1994. Structural analysis of the human VH locus using nonrepetitive intergenic probes and repetitive sequence probes. Evidence for recent reshuffling. J. Immunol. 152:660–666.

43. Dayhoff, M.O. Atlas of Protein Sequence and Structure. Vol. 5 (Suppl. 2). 1976. National Biomedical Research Foundation, Georgetown University Medical Center, Washington D.C. 1–8.

44. Sibley, C.G., and J.E. Ahlquist. 1984. The phylogeny of the hominoid primates, as indicated by DNA–DNA hybridization. J. Mol. Evol. 20:2–15.

45. Hori, S., Y. Satta, K. Hayakawa, R. Kondo, T. Inoue, T. Ishida, S. Hayashi, and N. Takahata. 1992. Man’s place in hominoida revealed by mitochondrial DNA genealogy. J. Mol. Evol. 35:32–43.

46. Nagase, T., N. Seki, A. Tanaka, K. Ishikawa, and N. Nomura. 1995. Prediction of the coding sequences of unidentified human genes. IV. The coding sequences of 40 new genes (KIAA0121-KIAA0160) deduced by analysis of cDNA clones from human cell line KG-1. DNA Res. 2:167–174.

47. Efstratiadis, A. 1994. Parental imprinting of autosomal mammalian genes. Curr. Op. Genet. Dev. 4:265–280.

48. Neumann, B., P. Kubicka, and D.P. Barlow. 1995. Characteristics of imprinted genes. Nat. Genet. 9:12–13.

49. Hurst, D., G. McVean, and T. Moore. 1996. Imprinted genes have few and small introns. Nat. Genet. 12:234–237.

50. Davies, B., and M. Fried. 1993. The structure of the human intron-containing S8 ribosomal protein gene and determination of its chromosomal location at 1p32-p34.1. Genomics. 15:68–75.

51. Perry, A.C., R. Jones, and L. Hall. 1995. Analysis of transcripts encoding novel members of the mammalian metalloprotease-like, disintegrin-like, cysteine-rich (MDC) protein family and their expression in reproductive and non-reproductive monkey tissues. Biochem. J. 312:239–244.

52. O’Hara, B., S.V. Johann, H.P. Klinger, D.G. Blair, H. Rubin, K.J. Dunn, P. Saa, S.M. Vitek, and T. Roberts. 1990. Characterization of a human gene conferring sensitivity to infection by gibbon ape leukemia virus. Cell Growth Differ. 1:119–127.

53. Erlich, R.P., P.A. Gleeson, P. Campbell, E. Dietzsch, and B. Toh. 1996. Molecular characterization of trans-Golgi p230. A human peripheral membrane protein encoded by a gene on chromosome 6p12–22 contains extensive coiled-coil alpha-helical domains and a granin motif. J. Biol. Chem. 271:8328–8337.

54. Brown, W.R.A., P.J. MacKinnon, A. Vellasante, N. Spurr, V.J. Buckle, and M.J. Dobson. 1990. Structure and polymorphism of human telomere-associated DNA. C. Cell. 63:119–132.

55. Holmquist, G.P. 1989. Evolution of chromosome bands: molecular ecology of noncoding DNA. J. Mol. Evol. 28:469–486.

56. Tenzen, T., T. Yamagata, T. Fukagawa, K. Sugaya, A. Ando, H. Inoko, T. Gojobori, A. Fujiyama, K. O’kumura, and T. Ikemura. 1997. Precise switching of DNA replication timing in the GC content transition area in the human major histocompatibility complex. Mol. Cell. Biol. 17:4043–4050.

57. O’Neill, R.J., M.J. O’Neill, and J.A. Graves. 1998. Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid. Nature. 393:68–72.