Evolution of Immunoglobulin Kappa Chain Variable Region Genes in Vertebrates

Tatyana Sitnikova and Masatoshi Nei
Institute of Molecular Evolutionary Genetics, and Department of Biology, The Pennsylvania State University

The major source of immunoglobulin diversity is variation in DNA sequence among multiple copies of variable region (V) genes of the heavy- and light-chain multigene families. In order to clarify the evolutionary pattern of the multigene family of immunoglobulin light κ chain V region (V\(_k\)) genes, phylogenetic analyses of V\(_k\) genes from humans and other vertebrate species were conducted. The results obtained indicate that the V\(_k\) genes so far sequenced can be grouped into three major monophyletic clusters, the cartilaginous fish, bony fish and amphibian, and mammalian clusters, and that the cartilaginous fish cluster first separated from the rest of the V\(_k\) genes and then the remaining two clusters diverged. The mammalian V\(_k\) genes can further be divided into 10 V\(_k\) groups, 7 of which are present in the human genome. Human and mouse V\(_k\) genes from different V\(_k\) groups are intermingled rather than clustered on the chromosome, and there are a large number of pseudogenes scattered on the chromosome. This indicates that the chromosomal locations of V\(_k\) genes have been shuffled many times by gene duplication, deletion, and transposition in the evolutionary process and that many genes have become nonfunctional during this process. This mode of evolution is consistent with the model of birth-and-death evolution rather than with the model of concerted evolution. An analysis of duplicate V\(_k\) functional genes and pseudogenes in the human genome has indicated that pseudogenes evolve faster than functional genes but that the rate of nonsynonymous nucleotide substitution in the complementarity-determining regions of V\(_k\) genes has been enhanced by positive Darwinian selection.

Introduction

Immunoglobulin molecules are composed of two identical heavy chains and two identical light chains. There are two different types of light chains: kappa (κ) and lambda (λ) chains. Both heavy and light chains consist of two domains: variable (V) and constant (C) domains. The V domains, which interact with foreign antigens, can further be divided into the complementarity-determining regions (CDRs) and the framework regions (FRs) (Kabat et al. 1991). CDRs are antigen-binding sites and are known to be highly variable. The V domains of heavy chains are encoded by variable (V\(_H\)) diversity (D\(_H\)), and joining (J\(_H\)) segment genes, whereas the light-chain V domains are encoded by variable (V\(_L\)) and joining (J\(_L\)) segment genes (Tonegawa 1983).

Immunoglobulin diversity is generated primarily by sequence variation among many copies of V\(_H\) and V\(_L\) genes in the genome, although in some organisms somatic mutation and somatic gene conversion play an important role (Knight and Tunyaaplin 1995; Weill and Reynaud 1995). For this reason, a number of authors have studied the diversity and evolution of V\(_H\) and V\(_L\) genes (e.g., Hood, Campbell, and Elgin 1975; Ohta 1980; Gojobori and Nei 1984; Kurth, Mountain, and Cavalli-Sforza 1993; Ota and Nei 1994; Rast et al. 1994). There are two main hypotheses for explaining the generation and maintenance of immunoglobulin V gene diversity in vertebrates. One is concerted evolution, in which the genetic diversity at a locus or a set of loci is assumed to be generated by introduction of new variants from different loci in the same V\(_H\) or V\(_L\) multigene family (Gally and Edelman 1972; Hood, Campbell, and Elgin 1975; Ohta 1980, 1983). The other is birth-and-death evolution, in which new genes are generated by repeated gene duplication, and some duplicate genes are maintained in the genome for a long time but others are deleted or become nonfunctional by deleterious mutations (Ota and Nei 1994). Taking advantage of the large amount of DNA sequences generated by recent genome projects, Nei, Gu, and Sitnikova (1997) conducted an extensive phylogenetic analysis of the V\(_H\) multigene family and reached the conclusion that this gene family is primarily subject to birth-and-death evolution rather than to concerted evolution.

The purpose of this paper is to extend this analysis to the kappa variable region (V\(_k\)) gene family and examine the pattern of evolution of V\(_k\) genes. We will first conduct a phylogenetic analysis of human V\(_k\) genes, for which the complete sequence is available for all 76 genes of the family. We will then examine the evolutionary relationships of V\(_k\) genes from various organisms to study the general pattern of evolution.

Materials and Methods

Background Information

The entire coding region (2 Mb) of the human immunoglobulin κ chain gene family has been sequenced (Schäble and Zachau 1993; Zachau 1995). It contains 76 V\(_k\) genes, 5 J\(_k\) genes, and 1 C\(_k\) gene. Of the V\(_k\) genes, 40 are functional, 9 do not have obvious defects but are not expressed, and 27 are pseudogenes (Tomlinson et al. 1996). A remarkable feature of the human κ gene family is that a large portion of the V\(_k\) region is duplicated, so that 33 of 76 V\(_k\) genes are present in two copies, and 10 are present in single copies (Schäble and Zachau 1993). This duplication seems to have occurred quite recently because the homologous pairs of V\(_k\) genes have
Fig. 1.—Genomic structure of the human \( \kappa \) chain gene family (adapted from Schäble and Zachau 1993). Open or filled rectangles represent \( \kappa \) genes, solid lines represent J\( \kappa \) genes, and a hatched rectangle represents the C\( \kappa \) gene. \( \kappa \) genes from the O, A, and L regions, which are shown by rectangles with different shadings, are present in both the p and d contigs. The \( \kappa \) genes from the B region, which are shown by filled rectangles, exist only in the p contig. A gene homologous to L8 includes only a leader exon, the rest of the gene being deleted from the d contig, and is designated by L22*. Gene L22, which does not have a homolog in the p contig, includes only exon II (see Huber et al. 1993a).

The roman numeral below each gene indicates the human \( \kappa \) family. The transcriptional direction is shown by arrows. Open triangles indicate deletions that occurred after the duplication in the p or d contig.

A sequence similarity of 95%–100%, and this duplication has not been found in the chimpanzee, gorilla, or orangutan (Ermert et al. 1995). From information on the overall sequence divergence between the two sets of duplicate genes (~1%), Schäble and Zachau (1993) suggested that the duplication occurred about 1 MYA. The copy of the \( \kappa \) region adjacent to J\( \kappa \) genes is called the p (proximal) contig, whereas the other copy of this region is called the d (distal) contig (fig. 1). A population genetic study of the \( \kappa \) region has shown that about 5% of human haplotypes do not have the d contig, and that the absence of the d contig does not seem to affect the fitness of its carrier even when the individual is homozygous for the absence of the d contig (Pargent, Schäble, and Zachau 1991; Schäible et al. 1993).

The \( \kappa \) region is often subdivided into four smaller regions, called the O, A, L, and B regions, which correspond to the DNA blocks used for nucleotide sequencing (Zachau 1995). These four regions are present in the p contig, whereas the d contig contains regions O and A and a large part of region L (fig. 1). \( \kappa \) genes are named according to their region of location on the chromosome (with some exceptions) and are numbered consecutively in each region. Thus, gene O1 is the first gene in region O. Some genes on the boundaries between two regions are named according to the contig where they were first found (e.g., A30 from region L; see Huber et al. 1993d).

The human \( \kappa \) genes are subdivided into seven families according to sequence similarity (Schäble and Zachau 1993). (To be consistent with the classification for \( \lambda \)\( \kappa \) genes, Schäble and Zachau’s “subgroups” are called “families” in this paper.) The \( \kappa \) genes that are similar to one another and belong to the same family are not necessarily located contiguously on the chromosome. Rather, they are intermingled with \( \kappa \) genes from different families. Therefore, evolution by gene duplication followed by transposition has been suggested to explain the arrangement of \( \kappa \) genes (Pech and Zachau 1984).

**Sequences Used for Analysis**

**Human \( \kappa \) Genes**

To examine the evolutionary relationships of human \( \kappa \) genes, we used germline sequences of all \( \kappa \) genes located on the p contig except for genes A22, A24, A28, O13, and O15, which contain large deletions. Duplicate genes from the d contig were not used in the analysis, because the homologous genes in the p and d contigs are closely related to each other. However, we included gene A14 from the d contig in the analysis, because its homolog in the p contig has been deleted (see fig. 1). The GenBank accession numbers for all human \( \kappa \) sequences used are given in Schäble and Zachau (1993).

**Vertebrate \( \kappa \) Genes**

In this analysis, we used several representative \( \kappa \) genes from each human gene family. We also used 22 mouse \( \kappa \) sequences, of which 19 represent 19 mouse \( \kappa \) families according to Strohal et al.’s (1989) and Koeller and Helmberg’s (1991) classification. Two additional sequences from family 4/5 and one sequence from family 23 were included in the analysis, because these mouse gene families are not shared by other species. Our preliminary phylogenetic analysis of all available mouse \( \kappa \) sequences (kindly provided by Reinhard Koeller) showed that the \( \kappa \) genes belonging to different families produce separate clusters. The only exception was a sequence from family 22, which clustered with the family 8 genes.

We also analyzed all complete nucleotide sequences of functional \( \kappa \) genes from other mammals that were available as of January 1997. (However, we did not use rat sequences YTH345HL and IR2, because they were closely related to rat sequence NCS. The horse cDNA
sequence k42 was also excluded, because it was very close to the horse germline sequence k1.) Since light L1 (p) chain V genes from amphibians, V₁,1 genes from bony fishes, and type III Vᵢ genes from sharks have been shown to be more closely related to Vᵣ than to Vᵢ genes (Zezza, Stewart, and Steiner 1992; Daggfeldt, Bengten, and Pilström 1993; Rast et al. 1994; Lundqvist et al. 1996; Partula et al. 1996), we included representative sequences of these Vᵢ-like genes in our analysis.

We used the germline sequences whenever they were available, but otherwise we used cDNAs. (cDNAs are marked by asterisks.) However, this should not affect our results very much, because somatic mutations are generally concentrated on CDRs rather than on FRs (see Gojobori and Nei 1986; Reynaud et al. 1995; Klein and Zachau 1995 for studies on mouse, sheep, and human variable-region genes, respectively), and we used FRs in the present study.

The Vᵣ and Vᵢ-like sequences used in this study (except for human sequences) are as follows. The sequence names together with their GenBank accession numbers are given after species names and are separated by semicolons.

Mouse (Mus musculus): 12/13, MUSIGKVA; 11*, MUSIGKBE; RF*, MUSIGRF2L; 38c*, J04577; 9A, MUSIGKVC; 9B, MUSIG69; 10, MUSIGKID; 32*, MUSIGKCE; 33/34*, MUSIGKABH; 1, MUSIGKVA; 2*, MUSIGKCLM; 24, MUSIGKMH; 8*, MUSIGKCSF; 22*, Kwan et al. 1981, no GenBank entry found; 28, MUSIGKAM2; 20*, MUSIGK85; 21, MUSIGKVR1; 4/5a, 4/5b, and 4/5c (Vᵣ 17, 18, and 28 from R. Kofler's database, respectively); 23a, MUSIG71; 23b* (Vᵢ 14 from R. Kofler's database).

Rat (Rattus norvegicus): 56R-3*, RATIGKVL; 56R-7*, RATIGKVDL; 57R-1*, RATIGKVEL; IR162*, RATIGKAC; NCS*, RATIGKNCSL; NGF*, RATIGNGLV; Y3-Ad1.2.3*, RNIGKY3; 53R-1*, RATIGKVCL; 53R-2*, RATIGKVDL; 57R-1*, RATIGKVEL; IR162*, RATIGKAC; NCS*, RATIGKNCSL; NGF*, RATIGNGLV; Y3-Ad1.2.3*, RNIGKY3.

Rat rat (Rattus rattus): CD25*, RATIGCD25L.

Rabbit (Oryctolagus cuniculus): 18a, OCI06; 18b, OCI07; 19a, OCI08; 19b, OCI09; b95r*, OCVKVR; b95g. OCVKCDR3; b4*, RABIGKAA; b5*, OCI03; k20, RABIGKVA; b9*, RABIGKAC; bas*, RABIGKAD.

Chimpanzee (Pan troglodytes): B3, PTIGK1VR.

Gorilla (Gorilla gorilla): B3, GIGIGKVR.

Orangutan (Pongo pygmaeus): B3, PPICIVR.

Macaque (Macaca mulatta): B3, MMIGKREG.

Sheep (Ovis aries): SK*, OAIKGCL.

Horse (Equus caballus): k1, ECVK1.

Frog (Xenopus laevis) p chain: p-V1, XELIGLVA; p-V2, XELIGLVB; p-V3, XELIGLVC.

Trout (Oncorhynchus mykiss): c10*, OMIGLAG; c3*, OMIGLAI.

Channel catfish (Ictalurus punctatus): G*, ICTIGLSTD; F*, IPU25705.


Shark (Heterodontus francisci): IIIa*, HEFIGCVF; IIb*, HEFIGCVF.

In trees for both human and vertebrate sequences, the human germline sequences for the λ chain variable region genes 3a.119B (3a), 4a,366F5 (4a), and 8a.88E1 (8a) (Tomlinson et al. 1996) were used as outgroups. In this study, we used only the FRs of Vᵣ sequences (Kabat et al. 1991), because the CDRs are highly variable and difficult to align. All three nucleotide positions of codons were used in the study of evolutionary relationships of human Vᵣ sequences and the relationships of rabbit sequences with other mammalian sequences. In the analysis of vertebrate Vᵣ sequences, however, only first and second codon positions were used, since nucleotide changes at third codon positions seem to have saturated when distantly related sequences were compared. All sequences were aligned by using the SEQED computer program (Zharkikh et al. 1991).

Phylogenetic Study

Phylogenetic trees were constructed by the neighbor-joining method (NJ; Saitou and Nei 1987). Evolutionary distances between sequences were computed by using Kimura’s (1980) two-parameter model. The reliability of tree topologies was evaluated by two different tests: the bootstrap test (Felsenstein 1985) and the interior-branch test (Rzhetsky and Nei 1992; Sitnikova, Rzhetsky, and Nei 1995; Sitnikova 1996). The bootstrap probability (BP) and the confidence probability (CP) values were computed for each interior branch of the tree, and the interior branches with BP ≥ 95% and CP ≥ 95% were considered statistically significant.

The computer programs MEGA (Kumar, Tamura, and Nei 1993), METREE (Rzhetsky and Nei 1994), TreePack (I. Belyi, personal communication), and TREEVIEW (K. Tamura, personal communication) were used in this study.

Results

Phylogenetic Analysis of Human Vᵣ Genes

The phylogenetic tree for human Vᵣ genes is presented in figure 2. There are four major clusters (I, II, III, and VI) of Vᵣ’s, which are statistically supported, and three singleton sequences (IV, V, and VII), which do not form any significant cluster. These seven groups of sequences exactly correspond to the human Vᵣ families defined by Schäble and Zachau (1993) according to sequence similarity. The evolutionary relationships of the seven Vᵣ families are not well resolved, since none of the branches connecting the families is statistically significant. This observation is different from that for VH genes, for which seven human VH families can be clearly divided into three groups (Schroeder, Hillson, and Perlmutter 1990; Kabat et al. 1991; Honjo and Matsuura 1995; Nei, Gu, and Sitnikova 1997).

Within the families I and II, the human Vᵣ genes located in region L on the chromosome (see fig. 1) tend to form their own clusters in the tree (fig. 2). This is in agreement with the results of sequence similarity analysis reported by Schäble and Zachau (1993). However,
FIG. 2.—Phylogenetic tree of 36 human Vκ sequences. The sequences used are named according to Schäble and Zachau (1993). The symbols Ψ and n in brackets each indicate a pseudogene and a gene which has an open reading frame but is not expressed (Tomlinson et al. 1996). Seven human families are indicated by roman numerals. Three human Vκ sequences were used as outgroups. Although the standard length of the FRs is 210 nucleotides, the number of nucleotides used was 201, because we ignored all sites containing deletions in some Vκ pseudogenes. Only bootstrap probability (BP) > 50% and confidence probability (CP) > 50% are shown above and below each interior branch, respectively. CP is usually higher than BP, but it is not shown for interior branches which are unimportant in our discussion. The BP and CP values that show significant statistical support for the clusters of human Vκ sequences from the same human family are shown in bold letters.

unlike their results, no other closely located genes form a cluster.

The human Vκ genes include 27 pseudogenes, i.e., genes that contain stop codons, frameshift mutations, or incomplete exons. These pseudogenes generally show a longer branch compared with the branch leading to a functional gene. The higher rate of nucleotide substitution in pseudogenes is apparently due to relaxation of purifying selection (see also Ota and Nei 1994). Interestingly, the branches leading to the family III pseudogenes are somewhat longer than those leading to the pseudogenes of other families, possibly due to their relatively ancient loss of functionality.

Synonymous and Nonsynonymous Sequence Divergence for Duplicate Genes

A unique feature of the human κ gene family is the recent block gene duplication that encompasses most of the original Vκ gene region (Schäble and Zachau 1993). We took advantage of this recent simultaneous duplication of many functional genes and pseudogenes to study the mechanisms of evolution of FRs and CDRs and the relative substitution rates of functional genes and pseudogenes. The number of functional gene pairs used was 12, because we used only duplicates that have been shown to be expressed (Tomlinson et al. 1996). However, in this study, we did not use the gene pairs A30/L14, L1/L15, and L9/L24, because they did not form a cluster in phylogenetic analysis when the whole sequence or a portion of the Vκ gene was used for the analysis, and thus they might have been affected by gene conversion (see also Huber et al. 1993b). Eight pseudogene pairs were used in which both members did not contain large insertions or deletions.
We concatenated the nucleotide sequences of the functional genes for the p and d contigs separately and computed the number of synonymous differences per synonymous site \( (p_S) \) and the number of nonsynonymous differences per nonsynonymous site \( (p_N) \) by using Nei and Gojobori’s (1986) method. This method does not take into account different rates of transitional and transversional substitutions but seemed to be appropriate in this case, because the transition/transversion ratio was low. Similarly, all pairs of the pseudogenes from the p and d contigs were concatenated separately, and the overall nucleotide difference \( (p) \) for duplicate pseudogenes was computed. To test the null hypothesis that synonymous and nonsynonymous substitutions in functional genes occur with the same rate, we used Fisher’s exact test, because we could count all synonymous and nonsynonymous substitutions unambiguously. Furthermore, the total number of substitutions was so small that we could not use large-sample tests such as the likelihood ratio or chi-square test (see Sokal and Rohlf 1995). Thus, the estimate of the time of duplication in the human \( V_k \) region. We estimated the rate of synonymous substitutions using the \( V_k \) human gene B3 and its homolog from chimpanzee assuming that humans and chimpanzees diverged about 5 MYA. (These genes are likely to be orthologous, because they are very closely related, and B3 is a single gene belonging to the human family IV [Ermert et al. 1995].) Thus, the estimate of the time of duplication in the human \( V_k \) region becomes 1–2 MYA, depending on the region used for estimation, i.e., FRs only or both FRs and CDRs. This estimate is still consistent with Schäble and Zachau’s (1993) estimate (1 MYA).

### Table 1
Numbers of Nucleotide Differences per Site \( \times 10^3 \) for Duplicate Human \( V_k \) Functional Genes and Pseudogenes, with the Actual Numbers of Substitutions for the Entire Concatenated Sequence Given in Parentheses

<table>
<thead>
<tr>
<th>Number of Gene Pairs</th>
<th>FRs</th>
<th>CDRs</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional genes,  ( p_N ) ......</td>
<td>12</td>
<td>2.1 ± 1.1 (4)( ^b )</td>
<td>10.6 ± 3.8 (8)( ^b )</td>
</tr>
<tr>
<td>Functional genes,  ( p_s ) ......</td>
<td>12</td>
<td>7.8 ± 3.5 (5)</td>
<td>0.0 (0)</td>
</tr>
<tr>
<td>Pseudogenes,  ( p ) ..........</td>
<td>8</td>
<td>8.9 ± 2.3 (15)</td>
<td>4.6 ± 2.7 (3)</td>
</tr>
</tbody>
</table>

\( ^a \) The framework regions (FRs) of a \( V_k \) gene included 70 codons, whereas the length of complementarity-determining regions (CDRs) varied with gene from 25 to 31 codons because of the length variation in CDR1. The average number of nucleotides used for a pseudogene was 289.

\( ^b \) \( p_s \) in the FRs is significantly lower than \( p \) for pseudogenes at the 1% level according to Fisher’s exact test.

\( ^c \) \( p_N \) in CDRs is significantly higher than \( p_N \) in FRs at the 1% level.

In order to examine the relationships of human \( V_k \) genes with those from other species, we conducted a phylogenetic analysis of \( V_k \) and \( V_{\text{amph}} \)-like genes from various vertebrate species. The results obtained (fig. 3) show that the mammalian \( V_k \) sequences form a cluster separate from the \( V_k \) genes from sharks, bony fishes, and amphibians (fig. 3). The first split of \( V_k \) sequences obviously occurred between shark sequences and others. The cluster of shark sequences is statistically supported, and their divergence from other sequences is likely to have occurred at the time or before cartilaginous fishes and other vertebrates diverged. The second split within the rest of the sequences separates the mammalian \( V_k \) sequences from the sequences of bony fishes and amphibians. The cluster of the amphibian and bony fish sequences is not statistically supported, suggesting that these sequences may not represent a monophyletic cluster.
Fig. 3.—Phylogenetic tree of 64 V_k sequences from vertebrate species (see Materials and Methods). Ten mammalian V_k groups are indicated by letters A–I. First and second codon positions were used, because divergent sequences are included. The number of codons used was 69 because of the lack of one codon in the V_l genes compared to the V_k genes. The bootstrap probability and confidence probability values showing significant statistical support for the clusters of mammalian sequences from the same V_k group are shown in bold letters.

The cluster of mammalian V_k sequences is also not very reliable, because the bootstrap value was low (<50%). However, the mammalian sequences were monophyletic in the phylogenetic analysis of amino acid sequences for the same genes (data not shown) and in the two previous phylogenetic studies (with the exception of the rabbit sequences discussed below) (Lundqvist et al. 1996; Partula et al. 1996). In contrast, in the phylogenetic studies conducted by Daggefeldt, Bengten, and Pilstrom (1993), Rast et al. (1994), and Haire et al. (1996), V_k sequences from mammals were intermingled with V_k-like genes from other vertebrate species, possibly due to the error in the tree reconstruction caused by a poor representation of V_k genes.

We call a cluster of V_k sequences a V_k group if the cluster is statistically supported or nearly supported by at least one of the two statistical tests and remains stable when all available mouse sequences are included in the analysis as well as when amino acid rather than nucleotide sequences are used in the analysis (data not shown). The group Db is not statistically supported but remains stable when other sequences are included. Groups Da and Db were classified as separate groups, because they may represent different clusters when more sequences are added. The mammalian V_k groups defined in our analysis are generally in agreement with those obtained by Kroemer et al. (1991) from the phylogenetic analysis of human and mouse V_k sequences. However, they classified human V_k III (group C in our tree), human V_k A10/14 (group F), and mouse V_k 23 (group H).
into one family and human V_κ IV (group Da) and mouse V_κ 8, V_κ 19/28, and V_κ 22 (group Db) into another (see fig. 3 in Krömer et al. 1991).

Most mammalian V_κ groups include sequences from humans as well as from other mammalian species. Interestingly, all three single-gene human families (IV, V, and VII) have their counterparts in other mammals (groups Da, E, and G). However, two V_κ groups (C and F) consist of only human sequences, and three groups (Db, H, and I) are not shared by humans and contain exclusively mouse sequences. As in the case of human V_κ families, the evolutionary relationships of mammalian V_κ groups are not statistically resolved.

**Representation of Mammalian V_κ Groups in Different Species**

The tree for the V_κ sequences from vertebrate species (fig. 3) shows variation in the representation of V_κ groups in various species. The human, mouse, and rat genomes contain representatives of many V_κ groups, whereas the sheep and horse sequences are restricted to one group, although it is not clear how many V_κ groups actually exist in the latter organisms, because no extensive survey of V_κ genes has been made (see Ford, Home, and Gibson 1994).

Another species which is interesting in this aspect is the rabbit. In phylogenetic trees constructed from amino acid V_κ sequences of various mammalian species, the rabbit sequences often cluster separately from other mammalian sequences (e.g., Hohman, Schluter, and Marchalonis 1992; Lundqvist et al. 1996; Partula et al. 1996). Because it seemed unlikely that the rabbit retained genes from a V_κ group that was lost from the genomes of all other mammals, we examined the phylogenetic relationships of all available rabbit V_κ sequences together with some representatives of mammalian V_κ groups (fig. 4).

![Phylogenetic tree for 48 mammalian V_κ sequences](image)

**Fig. 4.**—Phylogenetic tree for 48 mammalian V_κ sequences. Thirty sequences, including rabbit sequences, belong to V_κ group A, whereas nine pairs of other sequences represent nine V_κ groups (B–I).
(Schäble and Zachau 1993) and about 46 of the 140 \( V_k \) genes estimated in mice (Kirschbaum, Jaenichen, and Zachau 1996) belong to group A. In this respect, the group A \( V_k \) genes are equivalent to the group C \( V_H \) genes, which exist in the genomes of many mammals, reptiles, birds, amphibians, and bony fishes (Tutter and Riblet 1989; Ota and Nei 1994; Nei, Gu, and Sitnikova 1997).

**Discussion**

**Long-Term Evolution of \( V_k \) Genes**

As mentioned earlier, two models have been invoked to describe the evolution of immunoglobulin multigene families: birth-and-death evolution and concerted evolution. In the first model, the birth of genes occurs by gene duplication with subsequent divergent evolution and the death of genes is caused by either gene deletion or loss of functionality (Nei, Gu, and Sitnikova 1997). This model predicts the generation of a diverse repertoire of genes in the gene family, including nonfunctional genes. The second model assumes frequent occurrence of gene conversion and recombination between alleles of the same locus as well as between those of different loci (Ohta 1980, 1983). This mode of evolution therefore has a tendency to homogenize the entire multigene family.

Our phylogenetic analysis of immunoglobulin light-chain \( V_k \) genes from the human and other vertebrate species indicates that the long-term evolution of \( V_k \) genes is consistent with the model of birth-and-death evolution rather than with the model of concerted evolution. First, there is a high level of \( V_k \) gene diversity within the mammalian genome. We showed that \( V_k \) genes from mammalian species form 10 different groups that apparently emerged before the mammalian radiation. Second, a substantial number of genes are nonfunctional. For example, 27 (about 36%) of the 76 \( V_k \) genes in the human are pseudogenes (Tomlinson et al. 1996). Although conversion-like events may occur occasionally among \( V_k \) genes, they do not seem to affect the long-term evolution of the genes very much. Third, a block gene duplication involving a large number of genes may occur, and some of the duplicate genes may then be deleted, as in the case of human \( V_k \) genes. Probably, the DNA region involved in a duplication event is generally smaller than that observed in the human \( V_k \) region, but gene duplication and deletion seem to have occurred repeatedly, because different \( V_k \) gene families are intermingled with respect to chromosomal location. This strongly supports the idea of birth-and-death evolution.

The birth-and-death process plays an important role in the evolution of other immunoglobulin gene families as well (Ota and Nei 1994; Nei, Gu, and Sitnikova 1997). The immunoglobulin heavy-chain \( V_H \) genes from species of tetrapods form three \( V_H \) gene groups that have persisted in the genome for more than 400 Myr (Tutter and Riblet 1989; Schroeder, Hillson, and Perlmutter 1990; Ota and Nei 1994). The immunoglobulin \( \lambda \) chain \( V \) (\( V_{\lambda} \)) gene family in vertebrates comprises three major groups, one group being more closely related to \( V_{\lambda} \) genes than to the other two groups, suggesting that \( V_{\lambda} \) genes are paraphyletic in relation to \( V_k \) genes (Hayzer 1990; Zezza, Stewart, and Steiner 1992; Rast et al. 1994; Haire et al. 1996). The divergence of light-chain \( V \) genes apparently started before the emergence of cartilaginous fishes (i.e., \( >470 \) MYA), since the shark genomes contain \( V_L \) genes closely related to mammalian \( V_k \) genes (type III \( V_L \) genes) and two other \( V_L \) groups that cannot be classified as either \( V_k \)'s or \( V_{\lambda} \)'s (Rast et al. 1994; Haire et al. 1996).

The presence of divergent gene groups in the mammalian genome for such a long period of time cannot be explained by a random birth-and-death process. It seems that some sort of selection is operating to maintain the diverse gene groups in the genome. In recent time, it has been shown that certain antigens called superantigens are preferentially bound by the antibodies encoded by \( V_H \) genes belonging to a specific human \( V_H \) family (Silverman 1995; Zouali 1995). The interaction of antibodies with superantigens involves the third framework region (FR3) as well as CDRs. FR3 has been shown to be conserved among the genes belonging to the same \( V_H \) family (Kirkham et al. 1992). No such antigen specificity has been reported for light chains, but different groups of \( V_L \) genes are likely to be adapted to cope with different groups of antigens.

**Chromosomal Location of Human \( V_k \) Genes**

The human \( V_k \) genes belonging to the same family are not necessarily located on the chromosome contiguously. A similar observation was made for the human \( V_H \) and \( V_k \) genes (Honjo and Matsuda 1995; Williams et al. 1996). This suggests that both \( V_H \) and \( V_L \) genes undergo frequent gene reshuffling. By contrast, the mouse \( V_H \) and \( V_k \) families seem to be located in a more clustered manner, although details of the gene map in mice remain unclear (Honjo and Matsuda 1995; Kirschbaum, Jaenichen, and Zachau 1996).

Human \( V_k \) genes B3, B2, and B1 are unique in several aspects. They represent single-gene families IV, \( V \), and VII, respectively, and are located next to the \( J_k \) gene region. Genes B2 and B3 have a transcriptional direction opposite that of all other \( V_k \) genes from the \( p \) contig and \( J_k \) and \( C_k \) genes. Genes homologous to these genes are found in the great apes, where families \( V \) and VII are represented by a single gene (Ermert et al. 1995). The counterparts of these families have also been found in the mouse (see fig. 3). Interestingly, the members of mouse family 21, which is homologous to human family VII, are all clustered on the chromosome and are located proximal to the \( J_k \) region (Kirschbaum, Jaenichen, and Zachau 1996). Similarly, the member genes of the mouse families 8, 22, 19/28, which are closely related to human family IV, are clustered on the chromosome and are located second next (after the family 21) to the \( J_k \) region. Therefore, the genes located near the \( J_k \) region seem to have been conserved for a long time, suggesting that they are vital for immune response. Similarly, human \( V_{\lambda} \) gene 6–l, the only representative of \( V_H \) family 6, is the most proximal to the \( D_H \) region.
This gene is quite conserved and is the only member of \( V_H \) family 6 in great apes and New World monkeys (Meek, Eversole, and Capra 1991).

Diversity of \( V_H \) and \( V_L \) Genes in Different Species

The level of diversity among \( V \) genes varies with species, and it is usually related to the mechanism of generating antibody diversity. In the human and mouse, all \( V_H, V_V, \) and \( V_L \) genes can be divided into a number of families that emerged long before the divergence of mammals, and the involvement of various \( V \) genes in somatic DNA rearrangement is essential for the production of diverse antibodies. By contrast, \( V \) gene families in some species are quite homogeneous, and somatic gene conversion and somatic hypermutation are employed to generate antibody diversity.

In chickens, both \( V_H \) and \( V_L \) (\( V_L \)) gene families include one functional gene and a large number of pseudogenes, and all of these genes belong to a single family of heavy- and a single family of light-chain \( V \) genes (Reynaud et al. 1987, 1989; Ota and Nei 1995). (There are no \( V_L \) genes in chickens.) Thus, the diversity of expressed antibodies is generated by somatic gene conversion of the functional gene by pseudogenes. A similar situation exists for rabbit \( V_H \) genes, for which one \( V_H \) gene is preferentially expressed and undergoes sequence diversification by somatic gene conversion using many other donor genes, all of which belong to a single family (Knight and Tunyaplin 1995). The rabbit \( \kappa \) chain genes have not been studied as extensively as the heavy-chain genes, but our phylogenetic analysis shows that all published rabbit \( V \kappa \) sequences are closely related to one another and form one cluster (see also Mage 1987). This suggests that, as in the case of rabbit heavy-chain genes, there is a mechanism for somatic diversification of rabbit \( V \kappa \) genes. Similarly, cattle, sheep, and pigs are known to have a restricted amount of \( V_H \) diversity (Sun et al. 1994; Sinclair and Aitken 1995; Dufour, Malinge, and Nau 1996). Although \( V_L \) sequences for these organisms are not well studied, it is possible that they also have a restricted amount of \( V \) gene diversity. Further studies on this point will help us to understand the mechanism of evolution of immunoglobulin genes.

Acknowledgments

We thank Reinhard Koffer for sending us a database of mouse \( V \kappa \) sequences and Igor Belyi and Koichiro Tamura for making their computer programs available for phylogenetic analysis. We would like to thank Andrey Rzhetsky, Yasuo Ima, George Zhang, and Chen Su for their comments. This study was supported by grants from NIH and NSF to M.N.

LITERATURE CITED

DAGGELEDT, A., E. BENGTEN, and L. PILSTROM. 1993. A cluster type organization of the loci of the immunoglobulin light chain in Atlantic cod (Gadus morhua) and rainbow trout (Onocorhynchus mykiss Walbaum) indicated by nucleotide sequences of cDNAs and hybridization analysis. Immunogenetics 38:199–209.


Sitnikova and Nei


NARUYA SAITOU, reviewing editor

Accepted September 24, 1997