Immunogenetics (2008) 60:47–55
DOI 10.1007/s00251-007-0270-2

ORIGINAL PAPER

Evolutionary dynamics of the immunoglobulin heavy chain variable region genes in vertebrates

Sabyasachi Das · Masafumi Nozawa · Jan Klein · Masatoshi Nei

Received: 28 August 2007 / Accepted: 10 December 2007 / Published online: 10 January 2008

Abstract Immunoglobulin heavy chains are polypeptides encoded by four genes: variable (*IGHV*), joining (*IGHJ*), diversity (*IGHD*), and constant (*IGHC*) region genes. The number of *IGHV* genes varies from species to species. To understand the evolution of the *IGHV* multigene family, we identified and analyzed the *IGHV* sequences from 16 vertebrate species. The results show that the numbers of functional and nonfunctional *IGHV* genes among different species are positively correlated. The number of *IGHV* genes is relatively stable in teleosts, but the intragenomic sequence variation is generally higher in teleosts than in tetrapods. The *IGHV* genes in tetrapods can be classified into three phylogenetic clans (I, II, and III). The clan III and/or II genes are relatively abundant, whereas clan I genes exist in small numbers or are absent in most species. The genomic organization of clan I, II, and III *IGHV* genes varies considerably among species, but the entire *IGHV* locus seems to be conserved in the subtelomeric or near-centromeric region of chromosome. The presence or absence of specific *IGHV* clan members and the lineage-specific expansion and contraction of *IGHV* genes indicate that the *IGHV* locus continues to evolve in a species-specific manner. Our results suggest that the evolution of *IGHV* multigene family is more complex than previously thought and that several factors may act synergistically for the development of antibody repertoire.

Keywords *IGHV* multigene family · Phylogenetic analysis · Flanking repeat elements analysis · Birth-and-death evolution · Subtelomeric localization

Introduction

The iconic immunoglobulin (IG) molecule is a tetrapartite structure consisting of four polypeptide chains, two identical heavy (H) and two identical light (L) chains (Klein and Hořejší 1997; Lefranc and Lefranc 2001). Both the H and L chains consist of a variable (V) domain and a constant (C) region. The C region is encoded in a *C* gene. The V domain of the H chain is encoded by three kinds of genes, *IGHV*, *IGHJ*, and *IGHD*, each occurring in multiple copies and in different arrangements with the other two kinds of genes, depending on the species. For the formation of a V domain, one copy of each of the three kinds of genes comes together by a special process of genetic recombination. The rearrangement involves recombination signal sequences (RSS) composed of conserved heptamers and less conserved nonamers, separated by 23-bp spacer sequence (Early et al. 1980; Tonegawa 1983). The V domain can further be subdivided into the framework regions (FR) and hypervariable or complementarity-determining regions (CDR) distinguished by the extent of sequence divergence and structural delimitations.

The *IGHV* genes encode the antigen-binding regions of antibodies. Despite a clear sequence homology among *IGHV* sequences from different species, there is a marked plasticity in the organization of the region and in the mechanism for the generation of antibody diversity. In cartilaginous fishes, the *IGHV* genes are organized in cassettes of *IGHV-IGHD-IGHJ-IGHC*, which occur at different chromosomal locations (Litman et al. 1993). This

Electronic supplementary materials The online version of this article (doi:10.1007/s00251-007-0270-2) contains supplementary material, which is available to authorized users.

S. Das (✉) · M. Nozawa · J. Klein · M. Nei
Department of Biology, Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, PA 16802, USA
e-mail: sud13@psu.edu

organization is referred to as the cassette type. In bony fishes and tetrapods, the *IGHV* genes occur in the organization *IGHV*n-*IGHD*n-*IGHJ*n-*IGHC*n (where 'n' stands for multiple copies) and are clustered in a single chromosomal location (Marchalonis et al. 1998). The advantage of this organization is thought to be in the facilitation of combinatorial diversification of antibodies (Litman et al. 1993). The repertoire of *IGHV* genes is produced by the combination of gene duplication and the divergence of duplicate genes (Hughes and Yeager 1997; Ota and Nei 1994). Hence, the evolution of the *IGHV* genes can be explained by two evolutionary processes: the birth-and-death process and diversifying selection (Ota and Nei 1994). In the birth-and-death model, new genes are created by gene duplication. Some of the duplicate genes acquire new functions and remain in the genome, while others become pseudogenes or are eliminated from the genome. The process of diversifying selection serves to increase variation in amino acid sequences of the CDRs by higher rates of non-synonymous compared to synonymous substitutions, without significant changes in the canonical structure of the FR regions (Tanaka and Nei 1989).

On the basis of the degree of sequence identity, mammalian *IGHV* genes have been classified into three major clans (clans I–III) (Kirkham et al. 1992; Kodaira et al. 1986; Kofler et al. 1992; Ota and Nei 1994; Schroeder et al. 1990). The number of *IGHV* genes in these three clans varies among different mammals (Sitnikova and Su 1998). The reason behind the expansion and contraction of the *IGHV* multigene family and the factors affecting the evolution of antibody repertoire in jawed vertebrates are poorly understood. Furthermore, little is known about the evolutionary relationship between mammalian and non-mammalian *IGHV* sequences and the evolutionary dynamics of *IGHV* genes at the chromosomal level, although the structural and functional significance of the genomic location of several genes has been recognized (Linardopoulou et al. 2001). Now that the draft genome sequences of several vertebrate species are available, we have conducted a comparative analysis of *IGHV* genes of 16 vertebrate species. These comparisons are expected to give new insights into the evolution of the *IGHV* multigene family.

Materials and methods

Identification of *IGHV* genes

An exhaustive gene search was conducted to identify all the *IGHV* genes in the draft genome sequences of zebrafish *Danio rerio* (assembly: Zv6, Mar 2006; 6.7× coverage), medaka *Oryzias latipes* (Assembly: HdrR, Oct 2005; 6.7× coverage), stickleback *Gasterosteus aculeatus* (assembly: BROAD S1, Feb 2006; 11× coverage), western clawed frog *Xenopus tropicalis* (assembly: JGI 4.1, Aug 2005; 7.6× coverage), chicken *Gallus gallus* (assembly: WASHUC2, May 2006; 7.1× coverage), platypus *Ornithorhynchus anatinus* (assembly: Ornithorhynchus_anatinus-5.0, Dec 2005; 6× coverage), opossum *Monodelphis domestica* (assembly: MonDom 4.0, Jan 2006; 6.5× coverage), dog *Canis familiaris* (assembly: CanFam 2.0, May 2006; 7.6× coverage), cat *Felis catus* (assembly: Pre Ensembl – release 41, Nov 2006; 2× coverage), mouse *Mus musculus* (assembly: NCBI m36, Dec 2005; 7.7× coverage), rat *Rattus norvegicus* (assembly: RGSC 3.4, Dec 2004; 7.0× coverage), macaque *Macaca mulatta* (assembly: MMUL 1.0, Feb 2006; 5.1× coverage), chimpanzee *Pan troglodytes* (assembly: CHIMP 2.1, Mar 2006; 6× coverage), and human *Homo sapiens* (assembly: NCBI Build 36.2, Sep 2006) from Ensembl Genome Browser. The *IGHV* genes from cow (*Bos taurus*; assembly: Btau 2.0, Oct 2005; 6.2× coverage) were retrieved from NCBI Map Viewer. The sheep (*Ovis aries*) *IGHV* sequences were identified by the sheep-human genome sequence comparison using the Australian Sheep gene mapping web site (http://rubens.its.unimelb.edu.au/%7Ejillm/jill.htm). The human position corresponding to the *IGHV* locus was used to retrieve the sheep *IGHV* genes. For all species except sheep, we performed a two-round TBlastN search (Altschul et al. 1997) with the cutoff *E* value of 10^{-15} against the genome sequences. In the first round, the amino acid sequences of seven functional *IGHV* genes (one from each family previously defined) annotated in the human genome sequence were used as queries. As these seven queries are similar to one another, they hit the same genomic regions. We extracted only non-overlapping sequences given by the best hit (with the lowest *E* value). Taking into account the alignment with the query *IGHV* genes, we manually annotated each retrieved sequence. If the retrieved sequence was aligned with query sequence without any frame shifts or premature stop codons in leader sequence and FR regions (FR1, FR2, and FR3) and has a proper RSS, the sequence was regarded as a potentially functional *IGHV* gene. Other sequences (including truncated sequences) were regarded as *IGHV* pseudogenes. Next, the first round Blast best-hit sequences of a specific organism were used as queries for the second round TBlastN search to find additional *IGHV* sequences and in a similar way non-redundant sequences were retrieved (see Supplementary Table 1 for the list of *IGHV* sequences). The flowchart of the procedure is shown in Supplementary Fig. 1.

For all species, except cat and sheep, the coverage of the genome was >5×. Therefore, the total number of *IGHV* genes identified in the present study appears to be close to the actual numbers. The *IGHV* gene contains one intron between the leader sequence and the *V*-exon, consisting of

Table 1 Number of Immunoglobulin *IGHV* genes in 16 vertebrates

Organism	Functional *IGHV*	Pseudo-*IGHV*	Location	Intraspecies sequence variation[a]
Human	44	60	Chr. 14	0.339 (0.026)
Chimpanzee	35	53	Chr. 14	0.323 (0.026)
Rhesus macaque	55	67	Chr. 7	0.311 (0.024)
Mouse	96	65	Chr. 12	0.370 (0.025)
Rat	115	59	Chr. 6	0.388 (0.026)
Dog	43	37	Chr. 8	0.166 (0.015)
Cat	42	22	–	0.166 (0.014)
Cow	11	6	Chr. 21	0.078 (0.018)
Sheep	7	3	–	0.036 (0.014)
Opossum	26	6	Chr. 1	0.064 (0.011)
Platypus	43	21	–	0.281 (0.023)
Chicken	1	58	–	–
Western clawed frog	39	41	–	0.366 (0.030)
Zebrafish	37	10	Chr. 3	0.457 (0.033)
Medaka	38	15	Chr. 8	0.398 (0.029)
Stickleback	41	11	–	0.395 (0.030)

[a] Intraspecies sequence variation is represented by mean p-distance for *IGHV* sequences and the numbers in parenthesis represent standard error obtained from 1,000 replications

the complementarity-determining regions (CDRs) and framework regions (FRs). The CDRs were excluded from the analysis because they are highly variable and contain many insertions/deletions.

Phylogenetic analysis

The amino acid sequences of FR regions of the functional *IGHV* genes were aligned using CLUSTALW program (Thompson et al. 1994). After elimination of gap sites, p-distances (Nei and Kumar 2000) for amino acid sequences were computed, and phylogenetic trees for functional *IGHV* genes were constructed by the NJ method (Saitou and Nei 1987) using the MEGA4.0 program (Tamura et al. 2007). The p-distance refers to the distance measured by the proportion of amino acid differences between sequences and is known to give phylogenetic trees with higher bootstrap values (Takahashi and Nei 2000). The tree was rooted by using two *IGHV* sequences of elasmobranch species, *Heterodontus francisci* (accession no. S24657 and S24658). The reliability of the tree was assessed by bootstrap resampling with a minimum of 1,000 replications.

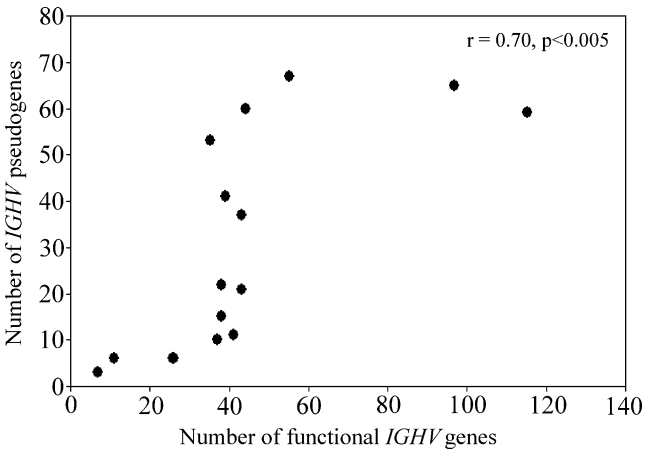

Fig. 1 Relationship between the numbers of functional *IGHV* and *IGHV* pseudogenes

Results

Number of IGHV genes in vertebrates

We determined the number of *IGHV* genes from the draft genome sequences of 16 vertebrate species (Table 1). The total numbers of potentially functional (37–41) and probably nonfunctional (10–13) *IGHV* genes are nearly the same for zebrafish, medaka, and stickleback, although the species belong to different orders. By contrast, the number of *IGHV* genes varies strikingly among the mammalian species. The total number of functional *IGHV* genes in rodents (mouse and rat) is considerably higher than that of the other mammalian species. The numbers of both functional and nonfunctional *IGHV* genes in artiodactyls (cow and sheep) are much smaller than those in other placental mammals. The two non-placental mammals (opossum and platypus) also differ considerably from each other in the number of *IGHV* genes. In chicken, there is a single functional *IGHV* and 58 *IGHV* pseudogenes. As reported previously (Reynaud et al. 1989), most of these *IGHV* pseudogenes had the complete V-exon but lacked the proper leader and/or recombination signal sequence. A few of the *IGHV* pseudogenes were truncated in their 5′ or 3′ ends or contained internal stop codons or frame shift mutations.

There is a significant positive correlation between the number of functional and nonfunctional *IGHV* genes (Fig. 1). Therefore, it seems that the more duplicate genes

occur, the more nonfunctional genes are produced in the *IGHV* multigene family. However, there are some exceptions to this rule. For example, the chicken has a special *IGHV* organization. The *IGHV* pseudogenes in this species are not truly nonfunctional, as they are used to generate immunoglobulin diversity by gene conversion and evolve slowly (Ota and Nei 1995; Reynaud et al. 1994). In earlier studies of *IGHV* evolution from a small number of species, it appeared that the number of *IGHV* genes per genome is roughly the same in different species (Gojobori and Nei 1984; Ota and Nei 1994). The present study, however, shows that the number of *IGHV* genes varies considerably with species. In some species, there are a small number of *IGHV* genes, but in others, the numbers of *IGHV* genes are very large. These differences have apparently arisen independently in different phylogenetic lineages.

IGHV sequence divergence in different species

To determine the pattern of sequence divergence of the *IGHV* genes, we calculated the average p-distance for all pairs of functional *IGHV* sequences in each species (Table 1). The extent of intraspecific *IGHV* sequence variation varied with species, and the average variation was generally higher in bony fishes than in tetrapods. Marked differences in sequence variation exist between mammalian species. Thus, artiodactyls (cow and sheep) show a lower level of variation than primates and rodents (Table 1).

Phylogenetic relationships of IGHV genes in vertebrates

In an earlier study based on a limited number of species, Ota and Nei (1994) classified various *IGHV* genes from vertebrates into five different phylogenetic classes. However, this classification no longer holds when a large number of species are included. On the 50 to 70% condensed phylogenetic trees (Nei and Kumar 2000) based on both fish and tetrapod sequences, no reproducible phylogenetic classification could be obtained when different sets of sequences are used (data not shown). When the trees were made separately for fish or tetrapod sequences, however, reproducible classification was observed in tetrapods (Fig. 2), but not in fishes (data not shown). The absence of clear-cut classification of fish sequences could be due to a high degree of intraspecific sequence divergence (Table 1). In fact, it has been shown that there are at least 11 *IGHV* families in the rainbow trout (Roman et al. 1996) and multiple *IGHV* families in the channel catfish (Ghaffari and Lobb 1999). By contrast, the phylogenetic classification of tetrapod *IGHV* sequences into three clans (I, II, and III) is clearly supported by high (>75%) bootstrap values. These three clans are equivalent

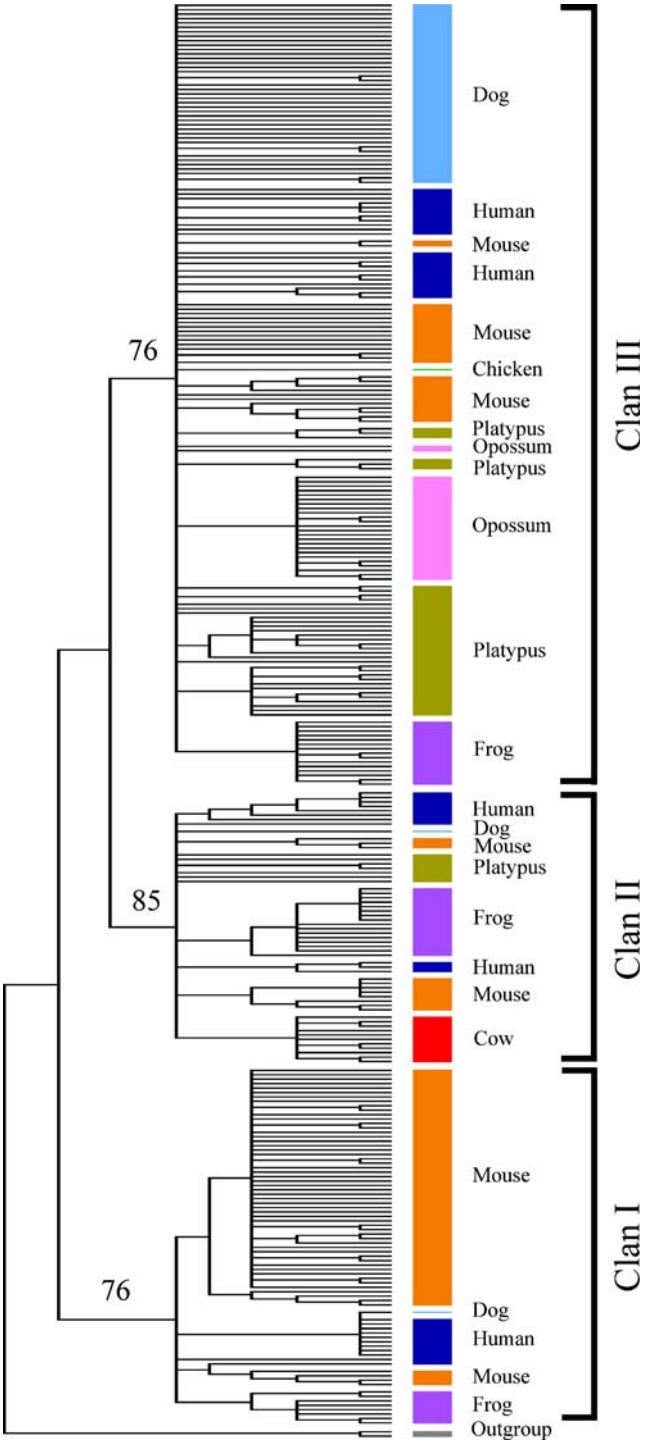

Fig. 2 NJ phylogenetic tree condensed at the 50% bootstrap value level for all functional *IGHV* genes of eight tetrapod species. Two shark *IGHV* sequences were used as the outgroup

to the clans I, II, and III reported previously for the phylogenetic classification of mammalian *IGHV* sequences (Kirkham et al. 1992; Schroeder et al. 1990). The presence of all three clans in the frog *Xenopus* (Fig. 2, Table 2) indicates that their divergence occurred before the radiation

Table 2 Number of functional *IGHV* genes in each tetrapod *IGHV* clan

Organism Name	Clan I	Clan II	Clan III
Human	11	11	22
Chimpanzee	7	9	19
Macaque	7	15	33
Dog	1	1	41
Cat	3	1	38
Mouse	58	11	27
Rat	24	43	48
Cow	0	11	0
Sheep	0	7	0
Opossum	0	0	26
Platypus	0	7	36
Chicken	0	0	1
Frog	8	16	15

of tetrapods. The absence of clans I and II genes in the chicken suggests that they were lost in this species. Whether this situation is representative for all bird species remains to be seen. Similarly, the absence of clan I genes in several additional species (cow, sheep, opossum, and platypus), of clan II genes in opossum, and of clan III in cow and sheep could represent random losses of the *IGHV* genes.

Distribution of the functional *IGHV* sequences at the heavy chain locus of mammals

Some *IGHV* genes of a particular species often cluster together in the phylogenetic trees. For example, the mouse clan I genes form a large cluster in the tree in Fig. 2. To understand how such clustering has evolved, we analyzed the chromosomal distribution of functional *IGHV* genes representing the three phylogenetic clans in the heavy chain locus of human, mouse, rat, and dog whose genome sequences are better assembled than those of others. The chromosomal distribution of the functional *IGHV* genes is different in different species (Fig. 3). In mouse and rat, most of the functional *IGHV* genes proximal to the *IGHD* gene are members of clans III and II. No functional clan I members are found in the first 20 *IGHD*-proximal *IGHV* genes in mice. Similarly, the rat has no clan I gene in the first 68 *IGHD*-proximal functional *IGHV* genes. In the human, however, functional *IGHV* genes of all three clans are intermingled (Lefranc 2001; Lefranc and Lefranc 2001; Matsuda et al. 1998). Most *IGHD*-proximal *IGHV* in the dog belong to the functional clan II genes, and a single functional clan I *IGHV* gene is located in the middle of the clan III genes. In the mouse, 48 functional clan I genes most distant from the *IGHD* gene (Fig. 3), all belong to a specific cluster of the tetrapods *IGHV* tree (see Fig. 2). Therefore, it seems that these genes may have originated by tandem duplication after separation of the mouse lineage. These observations are consistent with the idea that the *IGHV* genes evolve by the birth-and-death process rather than by concerted evolution.

Locations of orthologous sequences in the IGHV locus of humans and chimpanzees

IGHV genes are short and evolve relatively fast so that it is difficult to identify the orthologous genes between mammalian species belonging to different orders. However, this can be done relatively easily between the human and chimpanzee, which diverged about 6 million years ago. We therefore examined the orthologous relationships of human and chimpanzee *IGHV* genes. These relationships were determined primarily by phylogenetic analysis (except truncated *IGHV* pseudogenes). However, this analysis occasionally gave ambiguous results because of the relatively low bootstrap values. We therefore used another method of identification of orthologous and paralogous genes using information about the flanking gene or repeat sequences. In this approach, we first identified the SINE or LINE or other repeat elements flanking the 5′ and 3′ sides of each *IGHV* gene in the human and chimpanzee genome sequences and used this information for identifying the homologous genes between the two species (the names of the repeat elements used for this purpose are given in Supplementary Table 2). We could identify the orthologous and paralogous relationships of about 80% of *IGHV* genes and their chromosomal locations by this method. In the case of truncated *IGHV* pseudogenes, we used the latter method exclusively. One limitation of this analysis was the incompleteness of the chimpanzee genome sequence, and

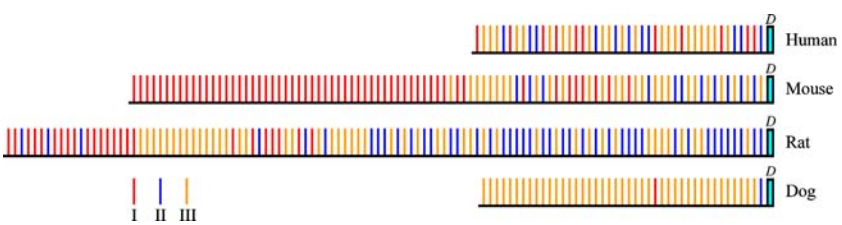

Fig. 3 Distribution of the functional *IGHV* genes in the heavy chain locus of humans, mice, rats, and dogs. The *red*, *blue*, and *yellow* colors represent clan I, clan II, and clan III genes, respectively. The *rectangular box* indicates the *IGHD* gene

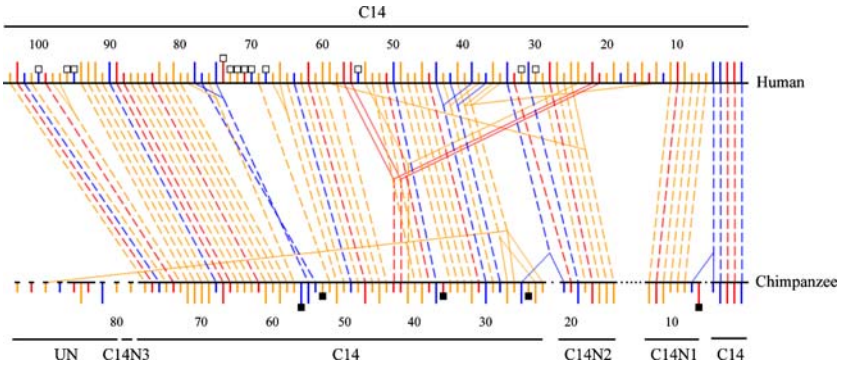

Fig. 4 Location of human and chimpanzee *IGHV* genes and their orthologous and paralogous relationships. Long and short vertical rods represent *IGHV* functional and pseudogenes, respectively. *Broken* and *solid lines* show orthologous and paralogous relationships between *IGHV* sequences, respectively. The *filled* and *open rectangles* indicate the *IGHV* genes whose orthologs were presumably lost in human and chimpanzee lineages, respectively. 'C', 'N', and 'UN' stand for "chromosome", "non-assembled" and "undetermined" regions, respectively. The gene number starts with the first *IGHD*-proximal *IGHV* genes. The *red*, *blue*, and *yellow* colors represent clan I, clan II, and clan III genes, respectively. Due to incompleteness of the chimpanzee genomic sequences (indicated by the *gaps in the lines*), orthologous relationships and exact locations could not be found for some of the *IGHV* genes. By using parsimony principle, 16 *IGHV* sequences have been placed between the 5th and 22nd *IGHV* genes

because of this limitation, certain conclusions remain tentative.

The results of the analysis indicated that the *IGHV* genes and their chromosomal locations are generally conserved in both human and chimpanzee (Fig. 4). There are, however, some scattered events of gene duplication and deletion that have occurred after divergences of the two species. A small scale of sequence inversion and transposition also appears to have occurred. One block duplication involving two functional and three nonfunctional *IGHV* genes that occurred in the human lineage is also identifiable (Fig. 5). These results indicate that the *IGHV* locus has undergone a continuous change in gene copy number.

Chromosomal location of the IGHV multigene family

We determined the chromosomal location of the *IGHV* multigene family in the species in which genome assembly has been completed or nearly completed. In eutherian mammals, the *IGHV* multigene family is always found in the subtelomeric region of a specific chromosome (Fig. 6). By contrast, the gene family in zebrafish (Wallace and Wallace 2003), medaka (Ueda and Naoi 1999), and opossum (Rens et al. 2001) is located near the centromere. To examine whether other genes present in subtelomeric regions of different chromosomes are similarly conserved across eutherian species, we carried out synteny analysis between human, mouse, and dog genomes using the Ensembl genome browser. More than 50% of the subtelomeric regions of one species were found at the interstitial sites of chromosomes in other species (not shown). Earlier studies also indicated that several subtelomeric segments of the human chromosomes have homologous counterparts at interstitial sites of mouse and rat chromosomes (Gibbs et al. 2004). Hence, the subtelomeric conservation of *IGHV* genes does not appear to be a property that applies to other genes located in subtelomeric regions. As the subtelomeric localization of *IGHV* genes is not observed in the opossum, it appears that the subtelomeric conservation of the *IGHV* gene family in eutherian mammals occurred through chromosomal rearrangement after separation of placental and non-placental mammals.

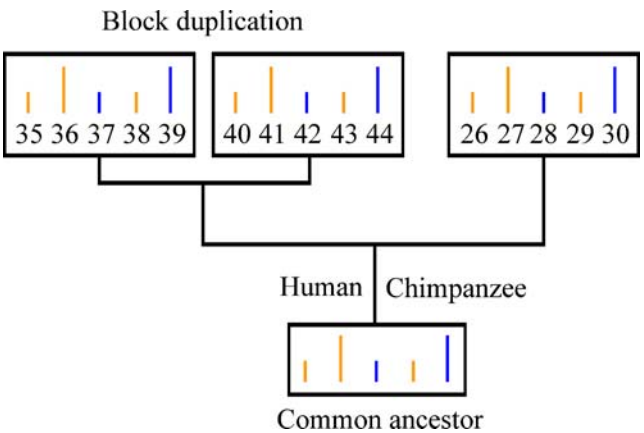

Fig. 5 An example of a block gene duplication event in the human *IGHV* locus (see Fig. 4). The *IGHV* genes are color-coded as in Fig. 4

Discussion

Our analysis of *IGHV* genes from 16 vertebrate genomes presents a more complete picture of the evolutionary dynamics of the multigene family than earlier studies suggested. We found that the number of *IGHV* genes varies considerably among different species of mammals, but the number of *IGHV* genes in teleosts is more or less uniform,

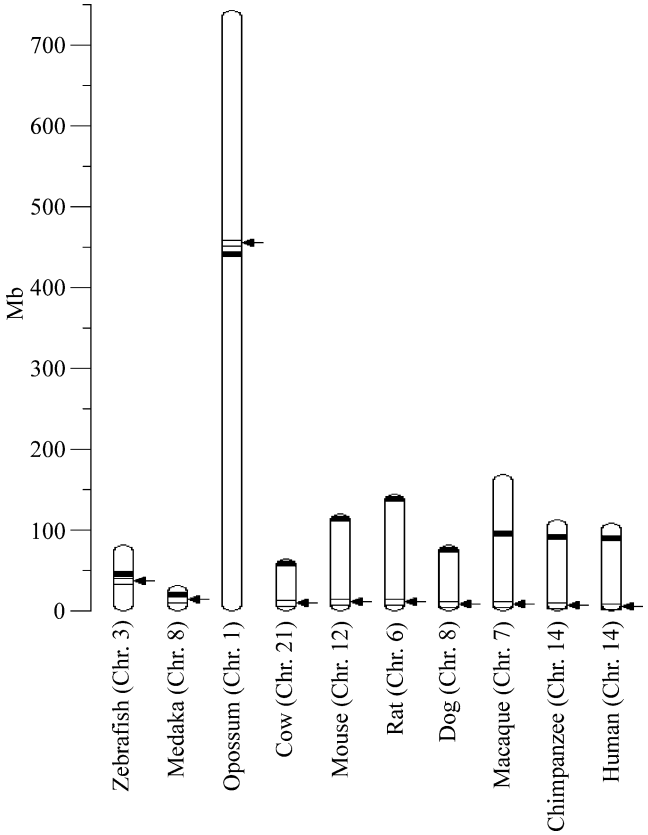

Fig. 6 Chromosomal location of the *IGHV* gene family deduced from the completely annotated or nearly completed genomes. The *IGHV* locus is indicated by an *arrow*. *Black band* indicates the centromeric region. The genomic location of *IGHV* gene family in macaque and cow was determined by synteny analysis, as the sequences are not yet completely assembled

although the three teleosts species studied diverged from their common ancestor about 140 Myr ago, long before the major mammalian radiation (80–100 Myr ago; Hedges and Kumar 2002; Yamanoue et al. 2006). By contrast, the overall intraspecies *IGHV* sequence variation is higher in teleosts than in tetrapods. There are several mechanisms to produce antibody diversity in jawed vertebrates (Klein and Hořejší 1997). After the activation of B cells, somatic hypermutation (SHM) introduces additional diversity and can improve the antigen-binding affinity of the expressed antibodies (Cannon et al. 2004). The enzyme activation-induced cytidine deaminase (AID) is known to be required for inducing SHM (Cannon et al. 2004; Yang et al. 2006). The role of AID in bonyfishes is still unclear. Although a recent study indicates the presence of SHM in fishes, the spectrum of mutational targets is restricted in comparisons to mammals (Yang et al. 2006). Hence, it is possible that in fishes, a high degree of intraspecies variation of *IGHV* sequences may compensate for the apparently poor somatic hypermutation in generation of antibody diversity. In mammals, there is also a marked heterogeneity in the intraspecies sequence divergence. For example, artiodactyls show low levels of sequence variation of *IGHV* genes, whereas primates and rodents exhibit a high degree of variation. The mechanism of antibody diversification in artiodactyls is different from primates and rodents in that somatic gene conversion and hypermutation play a more important role in artiodactyls than in primates and rodents (Reynaud et al. 1991; Sun and Butler 1996). The differences in intraspecies *IGHV* sequence variation between different species might therefore be associated with different mechanisms for the development of antibody repertoires, although several other factors might be involved in a synergistic way.

There is a significant positive correlation between the number of functional and nonfunctional *IGHV* genes. This observation suggests that the more gene duplications occur, the more *IGHV* pseudogenes are generated in the *IGHV* multigene family. Previously, Nei and coworkers (Nei et al. 1997; Nei and Hughes 1992; Nei and Rooney 2005; Nozawa and Nei 2007) showed that in many multigene families, gene duplication often occurs, but because of deleterious mutations, many duplicate genes become nonfunctional and either stay in the genome as pseudogenes or are gradually eliminated from the genome by unequal crossing over. Although the number of deleted *IGHV* pseudogenes cannot be assessed easily, our results suggest that throughout *IGHV* evolution, the numbers of functional and nonfunctional genes are maintained by birth-and-death evolution.

The *IGHV* genes of diverse tetrapod species have been found to fall into three major clans (clan I, II, and III). Hence, these clans must have persisted for about 370 Myr in the tetrapod genomes. Clan III *IGHV* genes have a broader taxonomic distribution than clan I and clan II genes. In cow and sheep (artiodactyls), only clan II sequences are found. By contrast, swine has only clan III *IGHV* genes (Aitken et al. 1997), although it belongs to the same mammalian order, Artiodactyla. Apparently, the ancestors of the extant artiodactyl species had *IGHV* genes belonging to clan II and III and swine lost clan II genes, whereas cow and sheep lost clan III genes after their divergence. In chicken, the single functional *IGHV* gene and all *IGHV* pseudogenes are closely related and belong to clan III (Ota and Nei 1995). A restricted *IGHV* repertoire is also observed in several non-placental and placental mammals. Therefore, it seems that the loss of the entire set or part of specific *IGHV* clan(s) is a relatively common phenomenon during *IGHV* gene evolution. The results of our evolutionary study of *IGHV* genes from different vertebrate species suggest that the great diversity of *IGHV* locus organization have been generated by large-scale birth-and-death evolution or genomic drift (Nei 2007).

Acknowledgments We thank Nikolas Nikolaidis, Dimitra Chalkia, Zhenguo Lin, and Hiroki Goto for their valuable comments and suggestions. This work was supported by NIH grant GM020293-35 to MN.

References

Aitken R, Gilchrist J, Sinclair MC (1997) A single diversified VH gene family dominates the bovine immunoglobulin repertoire. Biochem Soc Trans 25:326S

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Cannon JP, Haire RN, Rast JP, Litman GW (2004) The phylogenetic origins of the antigen-binding receptors and somatic diversification mechanisms. Immunol Rev 200:12–22

Early P, Huang H, Davis M, Calame K, Hood L (1980) An immunoglobulin heavy chain variable region gene is generated from three segments of DNA: VH, D and JH. Cell 19:981–992

Ghaffari SH, Lobb CJ (1999) Structure and genomic organization of a second cluster of immunoglobulin heavy chain gene segments in the channel catfish. J Immunol 162:1519–1529

Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE, Okwuonu G, Hines S, Lewis L, DeRamo C, Delgado O, Dugan-Rocha S, Miner G, Morgan M, Hawes A, Gill R et al. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428:493–521

Gojobori T, Nei M (1984) Concerted evolution of the immunoglobulin VH gene family. Mol Biol Evol 1:195–212

Hedges SB, Kumar S (2002) Vertebrate genomes compared. Science 297:1283–1285

Hughes AL, Yeager M (1997) Molecular evolution of the vertebrate immune system. Bioessays 19:777–786

Kirkham PM, Mortari F, Newton JA, Schroeder HW, Jr. (1992) Immunoglobulin VH clan and family identity predicts variable domain structure and may influence antigen binding. EMBO J 11:603–609

Klein J, Hořejší V (1997) Immunology. Blackwell Science Ltd, Oxford London Malden

Kodaira M, Kinashi T, Umemura I, Matsuda F, Noma T, Ono Y, Honjo T (1986) Organization and evolution of variable region genes of the human immunoglobulin heavy chain. J Mol Biol 190:529–541

Kofler R, Geley S, Kofler H, Helmberg A (1992) Mouse variable-region gene families: complexity, polymorphism and use in non-autoimmune responses. Immunol Rev 128:5–21

Lefranc MP (2001) Nomenclature of the human immunoglobulin heavy (IGH) genes. Exp Clin Immunogenet 18:100–116

Lefranc MP, Lefranc G (2001) The immunoglobulins fact book. Academic, London

Linardopoulou E, Mefford HC, Nguyen O, Friedman C, van den Engh G, Farwell DG, Coltrera M, Trask BJ (2001) Transcriptional activity of multiple copies of a subtelomerically located olfactory receptor gene that is polymorphic in number and location. Hum Mol Genet 10:2373–2383

Litman GW, Rast JP, Shamblott MJ, Haire RN, Hulst M, Roess W, Litman RT, Hinds-Frey KR, Zilch A, Amemiya CT (1993) Phylogenetic diversification of immunoglobulin genes and the antibody repertoire. Mol Biol Evol 10:60–72

Marchalonis JJ, Schluter SF, Bernstein RM, Shen S, Edmundson AB (1998) Phylogenetic emergence and molecular evolution of the immunoglobulin family. Adv Immunol 70:417–506

Matsuda F, Ishii K, Bourvagnet P, Kuma K, Hayashida H, Miyata T, Honjo T (1998) The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. J Exp Med 188:2151–2162

Nei M (2007) The new mutation theory of phenotypic evolution. Proc Natl Acad Sci U S A 104:12235–12242

Nei M, Hughes AL (1992) Balanced polymorphism and evolution by the birth-and-death process in the MHC loci. In: Tsuji K, Aizawa M, Sasazuki T (eds) 11th Histocompatibility Workshop and Conference. Oxford University Press, Oxford

Nei M, Kumar S (2000) Molecular evolution and phylogenetics. Oxford University Press, Oxford

Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 39:121–152

Nei M, Gu X, Sitnikova T (1997) Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci U S A 94:7799–7806

Nozawa M, Nei M (2007) Evolutionary dynamics of olfactory receptor genes in *Drosophila* species. Proc Natl Acad Sci USA 104:7122–7127

Ota T, Nei M (1994) Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. Mol Biol Evol 11:469–482

Ota T, Nei M (1995) Evolution of immunoglobulin VH pseudogenes in chickens. Mol Biol Evol 12:94–102

Rens W, O'Brien PC, Yang F, Solanky N, Perelman P, Graphodatsky AS, Ferguson MW, Svartman M, De Leo AA, Graves JA, Ferguson-Smith MA (2001) Karyotype relationships between distantly related marsupials from South America and Australia. Chromosome Res 9:301–308

Reynaud CA, Dahan A, Anquez V, Weill JC (1989) Somatic hyperconversion diversifies the single Vh gene of the chicken with a high incidence in the D region. Cell 59:171–183

Reynaud CA, Mackay CR, Muller RG, Weill JC (1991) Somatic generation of diversity in a mammalian primary lymphoid organ: the sheep ileal Peyer's patches. Cell 64:995–1005

Reynaud CA, Bertocci B, Dahan A, Weill JC (1994) Formation of the chicken B-cell repertoire: ontogenesis, regulation of Ig gene rearrangement, and diversification by gene conversion. Adv Immunol 57:353–378

Roman T, Andersson E, Bengten E, Hansen J, Kaattari S, Pilstrom L, Charlemagne J, Matsunaga T (1996) Unified nomenclature of Ig VH genes in rainbow trout (*Oncorhynchus mykiss*): definition of eleven VH families. Immunogenetics 43:325–326

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

Schroeder HW, Jr., Hillson JL, Perlmutter RM (1990) Structure and evolution of mammalian VH families. Int Immunol 2:41–50

Sitnikova T, Su C (1998) Coevolution of immunoglobulin heavy- and light-chain variable-region gene families. Mol Biol Evol 15:617–625

Sun J, Butler JE (1996) Molecular characterization of VDJ transcripts from a newborn piglet. Immunology 88:331–339

Takahashi K, Nei M (2000) Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. Mol Biol Evol 17:1251–1258

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. Mol Biol Evol 24:1596–1599

Tanaka T, Nei M (1989) Positive Darwinian selection observed at the variable-region genes of immunoglobulins. Mol Biol Evol 6:447–459

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through

sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Tonegawa S (1983) Somatic generation of antibody diversity. Nature 302:575–581

Ueda T, Naoi H (1999) BrdU-4Na-EDTA-Giemsa band karyotypes of 3 small freshwater fish, *Danio rerio, Oryzias latipes*, and *hodeus ocellatus*. Genome 42:531–535

Wallace BM, Wallace H (2003) Synaptonemal complex karyotype of zebrafish. Heredity 90:136–140

Yamanoue Y, Miya M, Inoue JG, Matsuura K, Nishida M (2006) The mitochondrial genome of spotted green pufferfish *Tetraodon nigroviridis* (Teleostei: Tetraodontiformes) and divergence time estimation among model organisms in fishes. Genes Genet Syst 81:29–39

Yang F, Waldbieser GC, Lobb CJ (2006) The nucleotide targets of somatic mutation and the role of selection in immunoglobulin heavy chains of a teleost fish. J Immunol 176:1655–1667