Yoshikazu Ichihara, Hidenori Hayashida<sup>O</sup>, Sanzo Miyazawa<sup>O</sup> and Yoshikazu Kurosawa

Institute for Comprehensive Medical Science, Fujita-Gakuen Health University, Toyoake, Aichi and National Institute of Genetics<sup>O</sup>, Yata, Mishima

# Only $D_{FL16}$ , $D_{SP2}$ , and $D_{Q52}$ gene families exist in mouse immunoglobulin heavy chain diversity gene loci, of which $D_{FL16}$ and $D_{SP2}$ originate from the same primordial $D_H$ gene

In mice, 12 germ-line  $D_H$  genes belonging to three different families ( $D_{Q52}$ ,  $D_{SP2}$  and  $D_{FL16}$ ) have been identified. The  $D_H$  genes other than  $D_{Q52}$  are clustered in the 60 kblong region located between  $V_H$  and  $J_H$  genes. Since there are seven  $D_H$  gene families ( $D_{HQ52}$ ,  $D_{XP}$ ,  $D_A$ ,  $D_K$ ,  $D_N$ ,  $D_M$  and  $D_{LR}$ ) in humans, we tried to identify new  $D_H$  gene families in the 60 kb-long region using human  $D_H$  gene probes. Mouse and human  $D_H$ genes showing the highest similarity were mouse  $D_{FL16}$  genes and human  $D_A$  genes. Southern hybridization of the mouse clones covering the 60-kb region with human  $D_H$ probes did not detect any other  $D_H$  genes. Nucleotide sequence analysis of the 4.0-kb fragment containing the  $D_{FL16.1}$  gene confirmed this conclusion. Comparison of the 12 germ-line  $D_H$  genes and more than 150 somatic  $D_H$  sequences also indicated that there are not more germ-line  $D_H$  genes in the mouse genome. Moreover, comparison of nucleotide sequences of  $D_{FL16.1}$  and  $D_{SP2.2}$  genes and their surrounding regions suggests that both  $D_H$  gene families originate from the same primordial  $D_H$  gene. Using the flanking sequences of both  $D_H$  genes, the divergence date between  $D_{FL16}$  and  $D_{SP2}$ genes was estimated at around 37 million years ago.

## **1** Introduction

The V region of Ig H chain is encoded by three separate genes in the germ-line genome:  $V_H$ ,  $D_H$  and  $J_H$  [1]. Both  $D_H$ - $J_H$  and  $V_{H}$ - $D_{H}$  joinings are necessary to complete an active  $V_{H}$  gene [1]. These DNA rearrangements are mediated by the recombinase which recognizes the heptamers CACTGTG and CACAGTG, and the nonamers GGTTTTTGT and ACAAAAACC [2]. The spacer length separating these oligomers is either 12 or 23 nucleotides [3]. D<sub>H</sub>-coding sequences are bordered by two sets of 12-nucleotide spacer signals. In mouse, 12 germ-line D<sub>H</sub> genes have been identified and they can be classified into three  $D_H$  gene families ( $D_{O52}$ ,  $D_{SP2}$  and  $D_{FL16}$  [4]). The  $D_H$  genes belonging to the  $D_{SP2}$  family are regularly spaced every 5 kb. Although human D<sub>H</sub> genes originally identified by Siebenlist et al. [5] are also regularly spaced every 9 kb, we showed that each 9-kb repeating sequence contains six different  $D_H$  gene families ( $D_{XP}$ ,  $D_A$ ,  $D_K$ ,  $D_N$ ,  $D_M$  and  $D_{LR}$ ; [6]).

In this study we tried to identify new  $D_H$  gene families in the mouse genome using human  $D_H$  gene-containing fragments as probes. Most mouse  $D_H$  genes are clustered in the 60-kb region located between  $V_H$  and  $J_H$  genes. Southern hybridization of the phage DNA covering the 60-kb region indicated that only fragments containing  $D_{FL16}$  weakly cross-hybridized with the human  $D_A$  probe. We determined the nucleotide sequence of the 4-kb DNA fragment containing  $D_{FL16.1}$ . This fragment

[I 7722]

Correspondence: Yoshikazu Kurosawa, Institute for Comprehensive Medical Science, Fujita-Gakuen Health University, Toyoake, Aichi 470-11, Japan does not contain any  $D_H$  gene other than  $D_{FL16.1}$  itself. Comparison of nucleotide sequences of the germ-line  $D_H$  genes and more than 150 somatic  $D_H$  genes indicated that there are not more than 12 germ-line  $D_H$  genes in the mouse genome. We also discuss the evolution of the mouse  $D_H$  gene loci.

### 2 Materials and methods

Six human  $D_H$  probes  $D_{XP}$ ,  $D_A$ ,  $D_K$ ,  $D_N$ ,  $D_M$  and  $D_{LR}$  were described in a previous report [6]. Three mouse  $D_H$  genecontaining clones, RI-2, RI-6, and RP13 were described by Kurosawa and Tonegawa [4]. Southern hybridization was carried out under non-stringent conditions [6, 7]. DNA sequencing was performed by the dideoxynucleotide chain termination method [8].

# **3 Results**

#### 3.1 Identification of putative $D_H$ genes by human $D_H$ probes

Since in the mouse containing clusters of  $D_H$  genes regions consist of highly conserved 5-kb repeats [4], three mouse clones, RI-2, RI-6 and RP13 [4] were used as representatives of mouse  $D_H$  gene-containing clones (Fig. 1). DNA was digested with Eco RI. Six different human  $D_H$  gene-containing fragments, described previously [6], were used as probes for Southern hybridization. Five probes:  $D_{XP}$ ,  $D_N$ ,  $D_M$ ,  $D_K$  and  $D_{LR}$  did not give any distinct signals (data not shown). However, the 4-kb Eco RI fragment in clone RI-2 and the 6.7-kb fragment in clones RI-6 and RP13 gave weak but distinct signals with the  $D_A$  probe as shown in Fig. 2. Southern hybridization of cellular DNA with these six human probes did not give any signal (data not shown). We concluded that if mouse  $D_H$ genes other than  $D_{SP2}$  and  $D_{FL16}$  exist, they should have been on these 4-kb and 6.7-kb fragments.

<sup>\*</sup> Supported by grants from the Ministries of Education, Science and Culture, Health and Welfare, and Agriculture, Forestry and Fisheries in Japan; Fujita-Gakuen Health University.



Figure 1. Organization of mouse D<sub>H</sub> gene loci. Twelve D<sub>H</sub> genes belonging to three families have been identified [4, 9]. Clones RI-2, RI-6, and RP13 were used in this study. Numbers on the second line indicate sizes of Eco RI fragments in kb. The 4-kb fragment containing D<sub>FL16.1</sub> was sequenced.

In the mouse,  $12 D_{\rm H}$  genes have been identified, and they can be classified into  $3 D_{\rm H}$  gene families [4]. In this study, we tried to identify new  $D_H$  gene families in the mouse genome using human  $D_H$  probes, since there are seven human  $D_H$  gene families [6]. However, the only  $D_{H}$  genes detected by human  $D_{\rm H}$  probes were  $D_{\rm FL16}$  genes. Most  $D_{\rm H}$  genes were originally identified by using DNA fragments containing D<sub>H</sub>-J<sub>H</sub> joints [4, 5]. In the case of the mouse system, many  $D_H$ -J<sub>H</sub> fragments have been sequenced, and in all cases published so far, one of the 12  $D_H$  genes identified was involved in such joinings [4, 10, 11]. As shown in this study, the  $D_H$  genes cross-hybridizing with the available  $D_H$  probes belong to the 12 germ-line  $D_H$ genes. Therefore, it is unlikely that new D<sub>H</sub> gene families remain to be found. If so, the 12 germ-line D<sub>H</sub> genes should encode all somatic D<sub>H</sub> sequences known so far.

When Kurosawa and Tonegawa [4] compared germ-line  $D_H$ sequences with somatic  $D_H$  sequences, only 16 somatic sequences were known. Now, more than 200 somatic D<sub>H</sub> sequences are known. It is thus worth comparing once more both germ-line and somatic D<sub>H</sub> sequences. As a source of somatic  $D_H$  sequences, we used the data book (1987) edited by Kabat et al. [12] although more data has since been published. We defined the somatic D<sub>H</sub> segment as the region which is not encoded by either germ-line V<sub>H</sub> or J<sub>H</sub> genes; therefore, N regions are included in somatic D<sub>H</sub> segments [13]. Since all of the germ-line J<sub>H</sub> sequences are known [14], the boundaries between D<sub>H</sub> and J<sub>H</sub> regions can be easily assigned. We tentatively assigned the 94th amino acid residue to the germ-line  $V_{\rm H}$ gene and the region after the 95th residue to the  $D_H$  region (for details see legend of Fig. 5). In the data book [12] 158 somatic  $D_{\rm H}$  sequences are available. As listed in Fig. 5, one fifth of them could not be assigned to any of the three  $D_H$  gene families. Some of them are too short to be assigned. The majority of them are G-rich sequences. Does this mean that there are other germ-line  $D_H$  genes which are rich in G residues? We think that this is not the case because there is no regularity among these sequences. If these G-rich sequences were encoded by germ-line sequences, there should be sequence similarities among them. They are rich in G residues, but seem to be random sequences, and they may be the products of the activity of the terminal transferase as proposed by Alt and Baltimore [13]. The regions encoded by germ-line  $D_H$ genes would have been removed during  $V_H$ -D<sub>H</sub> and D<sub>H</sub>-J<sub>H</sub> joining processes.

Fig. 5 summarizes the assignments of somatic  $D_H$  sequences to germ-line  $D_H$  genes. Classification of somatic  $D_H$  sequences was based on similarities of D<sub>H</sub>-coding regions and coding



[4] with human  $D_A$  probes. (a) Of each phage DNA 0.5 µg was digested with Eco RI, separated by agarose gel electrophoresis and stained with ethidium bromide. (1) RI-2 contains four Eco RI fragments: 5.4, 4.0, 3.8, 1.2 kb. (2) RI-6 contains three Eco RI fragments: 6.7, 5.4, 2.8 kb. (3) RP13 contains three Eco RI fragments: 6.7, 5.2, 5.0 kb. The origin of faint bands is not known. Closed triangles indicate the position of  $\lambda$ -Hind III markers. (b) Southern blots of these separated DNA which were hybridized with the human  $D_A$  probe [6]. The 4.0-kb band in clone RI-2 (1), and the 6.7-kb band in clone RI-6 (2) and clone RP13 (3) gave distinct signals.

## 3.2 Nucleotide sequence of the 4.0-kb fragment containing **D**<sub>FL16.1</sub>

Although the 4-kb and 6.7-kb fragments contained  $D_{FL16}$ genes, we determined the total nucleotide sequence of the 4-kb fragment to find the regions giving positive signals with the  $D_A$  probe. As shown in Fig. 3, there was only one  $D_H$  gene on this fragment. Homology research between nucleotide sequences of the 15-kb human D<sub>H</sub>-containing region [6] and this 4-kb fragment showed that homologous regions are very restricted in  $D_{FL16,1}$  and  $D_{A4}$  genes themselves. Fig. 4 shows the comparison of nucleotide sequences of  $D_{FL16.1}$  and  $D_{A4}$ genes. The signal and coding regions of these two genes showed 85% homology; however, the surrounding regions did not have any distinct homology. Although the 6.7-kb fragment containing D<sub>FL16.2</sub> was not sequenced, it is likely that the region which gave a positive signal with the  $D_A$  probe in the 6.7 kb fragment was the  $D_{FL16.2}$  gene itself.

ECORI <u>GNATTC</u> AGGCGTCAACTGIGACAACCCTCCAGCATTAATTACTATTAATAGATATTTTCTTTACTTAC	120
CATCACCCATCTCCCTGCTCACCAACCCACCCATTCCATAAGAACCAATCAGTACCCCACCCCCCAGAGCTCCCAGGGACTAAACCAGGAACCAGAGAGTACACATGGAAGGACCCATAG	240
CTCCAGCTGCATAGGTAGCACTACTTTGAAAAGGGGGAAGATGATGAAATCATTACTGTGGGGGGAAATGCAAGGAAGATTCCAACACATCTAGGCATCTATGAAGATTTTAAGTCTTCAA	360
ANTCCANANCCACCACANANTTTANANANANANANANANA	480
TTGGAGAAAGTAGAAATAGAGAAAGTGTACAGCCAGTATCCTCTAGCTACTCACATCCAAACAGGGGCCTCCTGACTGCTCTGAGGCCTGTCCTAAGAACAGCAATGATGCCACAGAAATTT	600
TTAGAGTGAGCCCTGAAGGAACTTGAGGCTGGTATGAGCAAGCCAGTCCCAGAGGAAAGGAAAACCCCATAGAGAGAAAACAGGTGAGTTAGTGCATTAAAGGGGCTGAGCAGGGGGGTTCT	720
CATCGCTCCCCAGCACCAGAAATAAGAGCCTCTCCGGAGCTGCTGGGACATGGAATGCAGATGATTCGGACCATCAGCCCCACAGAGACCTTTCCCACTCTGGCTCAGAAAGAGGGCACTG	840
GACCACAGTTGGAGAGGAGAATCGAAAGCTGATATCTCTGTATTCACTTAGCCTGTTACCCACCC	960
CAACTGCCAGTATTAAATACCAGATTTCAGAAGATTGGAAATCACCTCTCTGGTCATTTTTGGGACATGTAAACTGTAACAGGAAACATAGGACCAATTTAAGATGGAGCAGTCCTATAT	1080
CCCTAACCCAGTTGTAAATAAAACATTCAAGAGTGCCATCAGACACCACAGTGGTACAGGAGAGATGAGTGTACCTAGTGCATCAAGAGTTCCCTCACTAGATAAACCAAGATGTAGCCC	1200
CAGGACCACCCAGGCACCTACCAGGACTCCCCTCCAGAGGTCTGAGCCAGTTAGCTCTAGTTCATGTTTCAGACCAAAACATCAGAAACAACAGCATCTCCACTGCAGATGAACC	1320
CCTAAGCCATACAGTGTACCCAAAGGCAGCACCACAGATGGGGGAATGTGGGGGGGG	1440
CACTGATCCCACCAAAAACATAGTAGGACAAGGACCCTTAAATAACATCTGTCAAGGGGAGCTGTCAAATAGCCACTGAGATGGCTCATACGGGGTGATATAGAAAAACAGG	1560
CCAAAGAACCTCCTGTGTTGCAAGCACAAATGGGAAGCTGTGAATCTCCACTACCTAC	1680
GCATTGGACTCTCAGATCATCCAAGGACTAGGGCTAAAGTGGCCATGTGTGGGGAAATACATCCACTTTATAAACCTACCT	1800
ACCTCATGTGGGGTGCTGCGGTGCTGTACATTGAGTCTGGGGGGCATGAGTGTGCCCGGTAAATTCCTTATCACTCAGATGAATTTCCAGTCCACACTCATCACCTTGGAGTAGGAATTTTA	1920
ARAGITAGTGTAGATATAAGTAAAGAGAGGAGGGGGGGGGG	2040
ARACACAGAGCTCAGACAGAACTACCTGGCAATGCGACTGGGCACACTGAAAGCACTGGGCATCAGGACTGAGCCCCCARATATGCACTCAGGATCCTCTGCATAATAATGTGACATAACA	2160
GGAAGGTTAGAACAGGCCAAAAGAGGAAACAGAACAAATGCCCCCAACCAA	2280
TAATGGCCTCAATGCTGAAGCTAGGAAGAACTAGTTAAAAGAAACATGTTCAACGGGATTCCCTGTCACTGGACTTCACAAGCAAAATTCAATCTTTCTGTTAAGGAGATGAGAAGA	2400
GAATATCTGAACCTTGTGTTGACAGTGCCCCACCCCGACTGTCAGGCTGTGGGGAAATGCCAGAGCAATCACTAGGAACACAACGATGAGGGAGACGAGGGTTAGGACACAACCATCAT	2520
GATATCCCACAAGTATGGAAGAAGCAAGAACTTGTAGAGAGCAGAGAATGGCAGACAAAGCAGCATATACATAAGTAGATGGCCAGACTATACAGGAA Bg1II	2617
<u>Agatot</u> acatagoctgtgaggctttctgacagaaaagggcaggcatgtctcaaagcatagcatggcttgggacactgtctgc**********	2728
GAGCTCACAGCAA*AACCACACAGCCTTCCACAAGAGGAGAAGAAAAGGTAGCTTGTCAGTGAGGAAGTCCCCCAGAAACAGACCATTCCAGTAGTTCTTATCATTCCTCCCAAAGCAGC -GAGTC*T-G-A-T-CT-*G-AGGCAGTGCGTTCTCT	2847
CACCATCCAGGCACTGAGAGACCAAAGGCTGTTGGGAGGTCAGCCTAGAGGCAGGC	2965
GGACCTCAACTCAGGATGACAACTGAAACTCAACCGTGCTGCCTGGGCCCCCAATGCTCTCTACACCTGCAAAAACCAGAGACCATACTGGCCAGG TG-ACAG-CCCTG-ACACT-GCCAGG-AA-GTTG-T-TCTTTT-G	3060
DFL16.1 GCTTTTTGT GAAGGGATCTAC TACTGTG TTTATTACTACGGTAGTAGCTAC CACAGTG CTATATCCATCA GCAAAAACC -A C TCTACTATGATTACGAC AG A	3139
DSP2.2 CATTGTGCCCAGCAGACTCTTGAGCTCCAAAAACTGAGTCTAGAAAAGCTGGCATCAC*GGGGTTTATATCCCGAGTCTTGACCACT**GACCCA****TTAATACTATCCAACACAGAG G-ATCCC-GAGGTC-TTAGC*TAGAATTTAGGG	3252
CTCTCCGTCTGCCCACAAAGAAATCCAACCACCCTAAAGTCAGATCCCAA*AGCCTCTCACCCAGAGTCCAGGAAACCCAGCAAAAAAGCACTCAGCAGAAAAACCACATGAACCAATGCC T-AT-AAG-C-GG-GAGAATTTT	3371
BGIII TCCTTTCAATAGGAAGGGCCCCAG <u>AGATCT</u> TTTTCTCAGAAAAACTTGGGAAACACTCCAAGCTGGGCTCAGAGAGACTTCCACAGCCACTTTGAAACCTTTGATTTGC GAAGT	3480
CCCTCTTTACCTGATACCCACAAGACAACACTACCTCAGACACCACAAAAACATGTTCTATTTGTCACTTGTGGACCACAAGGATCCTGTGGTAGGTTCTACAAAGATCAGTGGAAGACT	3600
ATGGGGAGCATGTTGCAGGAACTGAGCCTGACTGCAGCTATCTAAGAACTCCCTGTATCCTGGCAGAGGATCCAAAGAAATGAAGCACCACCTCCCAGGCTACTGCTGACCAGAGAAAGGA	3720
ATTTTGTTTGGCCAGTGTTGGTTTGCATATTTGGGCTCCATTCATCGGACAATGAAGGATCTCTTAAGGGACAGGATGCCAGTGTAACCACACACA	3840
CTTACCATACTCTGGAAGATCAGAGAGAGGCCTAGGCAGAGTCCCCCTCTGACTCCTATTGGGGAAAATTTTGGATATTT <u>GAATTC</u>	3926

Figure 3. Nucleotide sequence of the 4-kb Eco RI fragment containing the  $D_{FL16.1}$  gene. Total nucleotide sequence of the 4-kb Eco RI fragment in RI-2 was determined. For comparison, sequence of the Bgl II fragment containing the  $D_{SP2.2}$  gene [17] is also shown. Bars indicate the same nucleotide as that of  $D_{FL16.1}$ . Stars indicate missing nucleotides.

DFL16.1	GGCCAGG	GCTTTTTGT	GAAGGGATCTAC	TACTGTG	TTTATTACTACGG	TAGTAGCTAC	CACAGTG	CTATATCCATCA	GCAAAAACC	CATTGTG
	• ••	• • • • • • • • • •	•••••	•••••		••••	• • • • • • •	• • • • • •	••••	• •
DA4	CTCAGGG	GCTTTTTGT	GAAGGGTCCTCC	TACTGTG	TGACTAC	AGTAACTAC	CACAGTG	ATGAACCCAGCA	GCAAAAACT	GACCGGA

Figure 4. Comparison of nucleotide sequences between mouse  $D_{FL16.1}$  and human  $D_{A4}$ . Combinations of mouse and human  $D_H$  genes showing the highest similarity were mouse  $D_{FL16}$  genes and human  $D_A$  genes. The sequence of  $D_{A4}$  gene was published in a previous study [6]. Positive signals, with human  $D_A$  probes, from the  $D_H$  gene-containing clones (Fig. 2) should be due to the above homology.

## (A) FL16 family

(B) SP2 family

Frame T	<b>TT TAT</b>	TAC	TAC	GGT	AGT	AGC	TAC							ThC.	~ * *		TAC	GRC			
NL	TT CAT	TAC	TAC	GGC	TAC			NR	JH	re	f.	Frame 1	TC	TAC	TAT	GGT	AAC	TAC			
													сс	TAC	TAT	GGT	TAC	GAC			
TT		c	TAC	GGT	AGT	AGC	т	GG	J4	23	5		00	TAC	TAT	GGT	AAC	TAC			
А		AC	TAC	GGT	AGT			0000001	J2	2	;	Nt.	TC	TAC	TAT	GAT	TAC	GAC	NR	JH	ref.
269		TAC	TAC	GGT	AGT	AG		λ	J2	3(	)										
AGGG		AC	TAC	GIT	AGT	AGG	TAC	GACCC	J3	5:	5,57	GGT			TAT	GGT			GGCCT	J2	25
GICICAA	TT TAT	TAC	TAC	GGT	COL	AGC	GAC	CCT	J3	6	)	GATA	•	TAC	AT TAT	GAT	CAC	т	стст	J2 J3	29
G	AT	TAC	TAC	GGT	AGT				J2	7	5	GGG		IAC	AT	GGT	AAT	TAC	AGGAATTTG	J2	31
c	AT	TAC	TAC	GGT	AGT	AGC	т	cc	J3	8	3	G	;	AC	TAT	GG			G	J3	54
(CCC)	TAT TAT	TAC	TAC	GGT	AGT	AGC		CAT		8	1			TAC	TAT	GGT	AAC	TAC	-	J2 .12	86 108
(000)	161		110	GGT	AGT	AGC	TAC	G	J4	9	5	GAGGAA		TAC	TAT	GAT	TAC	G	CTT	J2	129
TACG	AT	TAC	TAC	GGT	AGT	AGC	TAC		J2	93	2,102	(ATT)			TAT	AGT	AAC	TAC	AAGT	J2	136
	TAT	TAC	TAC	GGT	AGT	AGC			J2	9	5			TAC	TAT	ACG	TAC		CCT	34	150
IACA	TAT	TAC	TAC	GGT	AGT	AGC	TAC		J2	9	, Э			TAC	TAT		TAC		GAGAGGG	J9 J3	151,152,
<i>c</i> c	T TAT	TAC	TAC	GGT	AGT	AGC		CCTTG	J2	1	9	TCGG	;		AT	GG1	TAC	TA		J4	161
TCGGATGG	T	TAC	TAC	G			_	ACTGGTTTG	J3	1	10	GGGGGA	1		TAT	GG			GGGAGTA	J4	166
GGGCAGA	TT TAT	TAC	TAC	GGT	AGT	ACC	T		J2 .T2	1:	12	G	; cc	TAC	TAT	GG		~	AG	J3	171
TCG	IAI	IAC	IAC	GGT	AGT	ABC	TAC	с	J1	12	23	GG	;		TAT		TAC	G	GGG	J3 J3	191
CCCCACCCAT		с	TAC	GGT	AGT	AGC	TAC	-	J4	1	30	60	;		141	GAT	TAC	GAC	GGGG	J3	193
GGG		TAC	TAC	GGT	AGT	AGC			J2	1	32	GATG	; cc	TAC	TAT	AGT	AAC	TAC		J1	194
	TAT	TAC	TAC	GGT	AGT	AGC	TAC		J2	1	33	CCG	;		TAT	GGI	AAC	TAC		J4	200,201
TATG	AI	TAC	TAC	GGT	AGT	AGC	TAC		J2		54	c	$c \alpha$	TAC	TAT	ACC	TAC		CCT	J1	205
AG	T TAT	TAC	TAC	GGT	AGT	AGC	TAC	CGTCCG	J3	: 1. : 1.	41	TGG	3			GGI	AAC	TAC	CC	J4 .74	234
		TAC	TAC	GGT	AGT	AGC	TAC	т	J2	1	45	(AIC)	2	TAC	. 181	GG1	L AAC	TAC	CCTC	J4	253
TACCTC	TAT	TAC	TAC	GGT	AGT	AGC	TAC		J2	1	48	100		TAC	: TA1	AGO	TAC		CCT	J4	254
TACGA	T TAT	TAC	TAC	GG				GT	J3	1	49			TAC	TA1	AG1	г аас	TAC	cc	J4	256
	TAT	TAC	TAC	GGT		100	-	GCTG	J2	: 1	54										
	1A1 TA1	TAC	TAC	GAG	AGT	AGC	1	ĊT	13	1 1	56	···							c		
(CTC)	TAT	TAC	TAC	GTT	AGT	AGC	TAC	G	J3	1	57	Frame II	1	CT A		TG C	STA A	CT A	c		
(CCT)	TAT	TAC	TAC	GG				GGGGG	J1	. 1	58		ć	CT A	ACT 7	TG C	STT A	CGA	c		
(CCT)	TAI	TAC	TAC	GG	_			GG	J1	1	59		c	CT P	CT /	TG C	GTA A	ст а	c		
► GGG			TAC	GGC	T	200	# » C	TAT	J4	1	63 67		0	CT 7	GT J	TG C	GTA A	CT A	.c		
A	A1	TAC	TAC	GGT	AGT	ACC	T		J1	1	72			ICT P	ACT P	ATG 2	ATT A	CG A			
GGA		TAC	TAC	GG				AGGAG	J3	17	3,174	ATGGGGG			CT A	TG	TG G	TT A	C ACCC	J3	59
TC	T	TAC	т <u>с</u> с	GGT	AGT	AGC		с	J3	17	5,176	TCG	3 1	CT A	ic .				CTC	J4	164
GATGCGG		AC	TAC	GGT	AAT	AGC	TAC	TTTG	J1	18	31										
696			<b>T</b> N C	00	AGT	AGC	TAC	GGAG	.73	10	54 55 197 199										
GATGCAGAGGT	T	TAC	TAT	GGT	GGT	AGC	т	СТ	J1	19	9	Frame III	1	CTP	CT	TGO	5 TTA - 788	CGA	C C		
TC	т	TAC	TAC	GGT	AG			cc	J3	20	3		2			TGC	5 TTA	CGA	, c		
A		AC	<u>G</u> AC	GGT	AGT	AGC	TAC	GG	J2	20	94			CT	CTA	TG	TAA	CTA	ċ		
TCCC	AT	TAC	TAT	GGT	<u>G</u> GT	AGC	TAC	G	J2	2:	17		C	CT?	GT?	A TGO	з таа	CTA	C		
TCGGTC		TAC TAC	TAT	GGT	GGT	AGT	TAC	тC	.72	2	18		2	CT7	A CT	TG	A TTA	CGA	c		
GOGAT	T	TAC	TAC	AAT	AGT	AGC	ĨŲĆ	cc	J2	2:	21		~				 		λ	 .73	68.70.77
► GGA	TT CAT	TAC	TAC	GIC	CAC				J2	2:	37,238,239,240	TACAGO	G R			TG	5 TTA		GAGG	J3	135
GAT			TAÇ	GG				G	<b>J</b> 3	2	4,245	G	G		1	A TGO	J TTA	CGA	CGTG	J4	143,144
GAT			TAC	GG				G	J2	2	16					ŤGO	GAA		TC	J4	160
GA	T TAT	TAC	TAC	GGT	AG			GG	<b>J</b> 3	2:	5										
Frame II	TTT AT	T AC	T AC	G GT	A GT.	A GC	т ас					(C) D <sub>Q5:</sub>	2								
	TTC AT	T AC	T AC	G GC	r ac																
											••••••			CAAC	rgggi	AC					
►(CCT)	AI	AC	TAC	GGC				CTTAGAGGGG	J1	1	38	G	AT	C'	rGG					J4	61
GGA		AC	TAC	GGT	GG			GGAGA	J2	1	46	GA	TC	C	rggg	G				J3	63
CCCCCTC	1	T AT	T IC	G IT	A GT	A GC		GG	J4	2	41	T	AT (	CGAC	rggg.	AC G	G			J2	104
	AT	T AC	T AC	G GT	A G				J4	2	51		G	AC	IGGG	T	с			J3	128
												т	CA CC	AAC:	rggg rggg	<b>A</b> C				J2 J1	178.179.
Frame III	т тта	TTA	CTA	CGG (	TAG	TAG	CTA (	2				G	CT	AAC	IGGG.	A T				J3	189
	T TCA	TTA	CTA	CGG (	CTA	с															
(ACC) GA			A	CGG	T	_		TAGGGG	J1	. 1:	22	(D) not	c	las	sif:	ied					
CG		А ттт	CGA	CGG ( CGC	GAG	u.		CGA	J2	14	10	,	9								
TC		×	CTA	CGG				ĞT	J4	1	52	GATAGGGG		J3	32			GA:	TCATGGG	J	3 49
A				GG	CTA			ATGG	J4	10	55	GATGGGGG		J3	33			GA:	TTGG	J	4 50
AAGGGAC		TA	CTA	CGG	т				J4	2	02	GATCGTGGG	_	J3	34			GA:	TCAGGGG	J.	2 51
► <i>C</i> AA			CTA	CGG	CT			с	J3	2	49	GATCGGGGG	6 670	J3 47	35			GA:	CAGGGG	J ,T	4 56
-						_				_		GACAGA	910	. J J . J J	37			GT	AGCTCCGGGG	л С	2 58
Figure 5.	Assign	men	ts of	son	iatic	: D <sub>H</sub>	seq	uences to germ-	-line	eΓ	O <sub>H</sub> genes. The	GATCGGGG		J3	38			GA	TAGG	J	1 65
data book	(page :	508 t	o 51	9) e	dite	d by	Kat	oat et al. [12] wa	s us	sed	as the source	GATGGGTT		J4	39			GGG	3	J	389
of somatic	D <sub>H</sub> see	mer	its. I	Ref.	indi	cate	s the	number used in	1 thi	is ł	ook. Classifi-	GATGGGGA		J4	40			TA	ITG	J	4 96
cation of s	omatic	D	seau	ence	es w	as h	ased	on similarity of	f co	dir	g regions and	GAAGGGG		J4 .73	41 42			GA	- IGGGGCT	.т	3 11/ 3 126
		- n									0.000000000	GUINGCOON	•	0.5							

data book (page 508 to 519) edited by Kabat et al. [12] was used as the source of somatic D<sub>H</sub> segments. Ref. indicates the number used in this book. Classification of somatic D<sub>H</sub> sequences was based on similarity of coding regions and coding frames. Boundaries between  $V_{\rm H}$  and  $J_{\rm H}$  genes were tentatively fixed at the 94th and 95th amino acid residues. N sequences (N<sub>L</sub> at the boundaries between  $V_H$  and  $D_H$ ,  $N_R$  at the boundaries between  $D_H$  and  $J_H$ ) are also written. Since GG, GA, GAT and CC sequences for the 95th residue might

be encoded by germ-line  $V_H$  genes [22], they are shown in italics. Since there  $^{GATGGG}$   $^{J2}$   $^{48}$   $^{TTAGACACCTCCG}$   $^{05}$   $^{250}$  have been no reports showing CGC, CTG, CCT, or ACC at the 94th residue in germ-line  $V_H$  genes [12, 22], they are shown in parentheses. Boundaries between  $D_H$  and  $J_H$  genes were assigned based on germ-line  $J_H$  sequences [14]. When nucleotides at the boundaries can be encoded by germ-line D<sub>H</sub> and J<sub>H</sub> genes, they are indicated in italics. When nucleotides possibly encoded by germ-line D<sub>H</sub> genes are different from the corresponding germ-line  $D_H$  genes, they are underlined. Black triangles in the  $D_{FL16}$  family indicate that  $D_{FL16.2}$  was used; in the other cases,  $D_{FL16.1}$  was used. Bars in (D) indicate that there is no sequence in the somatic  $D_H$  region.

GATCATGGG

GATCGGGG

GATGGGGG

GATGGGG

GATGGG

J2 43

J2 43 J2 44 J3 45 J3 46 J2 47 J2 48

GATCGGG

J3 127 J1 137

J2 206

J2 247 J2 248

J3 250

TATT

AGGGATCTCAGGG

CCGGGGGGTCCC

GACGGGGGA

(CCC) ----TTAGACACCTCCG

Table 1. Nucleotide difference of flanking regions between  $D_{FL16}$  and D<sub>SP2</sub><sup>a)</sup>

	Pos	ition				
	from	to	Ν	М	К	K°
5'-Flanking 3'-Flanking Total	2618 3140	3060 3401	445 266 711	121 65 186	0.2719 0.2443 0.2616	0.3377 0.2956 0.3217

a) 5' and 3'-Flanking sequences of D<sub>H</sub> genes were compared. N is the number of sites compared between D<sub>FL16.1</sub> and D<sub>SP2.2</sub> [17]. Deletion of continuous two to nine nucleotides was assumed to have occurred as a single event. M is the number of sites showing a difference between DFL16 and DSP2. K and K<sup>c</sup> indicate nucleotide difference per site and difference corrected for multiple substitutions  $K^{c} = -\frac{3}{4}\ln(1-\frac{4}{3}K)$  [18, 19], respectively. Using  $K^{c} =$ 

0.3217 for  $D_{FL16}$  and  $D_{SP2}$ , the divergence between rat and mouse would have occurred 17 million years ago [20], and using a K<sup>c</sup> value for rat and mouse of 0.148 [21], the divergence date between  $D_{FL16}$  and  $D_{SP2}$  was estimated to be about 37 million years (17  $\times$  0.322  $\frac{-2}{0.148} = 37$ ).



frames. The following characteristics were observed (a) D<sub>FL16.1</sub> is the most frequently (73/158) used D<sub>H</sub> gene, (b) the codon frame I (TAC TAC GGT and TAC TAT GGT) encoding Tyr-Tyr-Gly is predominantly used in both D<sub>FL16</sub> and D<sub>SP2</sub> genes, 65/77 and 29/38, respectively; (c) in the cases where N sequences were not observed at the boundaries between  $D_H$ and J<sub>H</sub> genes, 1 to 6 nucleotide-long redundancy frequently existed, that is, a few nucleotides such as CTAC can derive either from germ-line D<sub>H</sub> or J<sub>H</sub>. The third point may reflect the repair mechanism taking place after digestion of the ends of  $D_{H}$  and  $J_{H}$  genes with exonuclease. DNA polymerase and ligase might be involved in the joining process of the processed ends. Since DNA polymerase requires a primer for polymerization [15], the ends of the joined fragments should have complementary nucleotides to supply template and primer. These characteristics were already observed in Kurosawa and Tonegawas's study [4], although only 16 somatic sequences were available; now, they can be generalized in mouse somatic  $D_H$  sequences. Since virtually all of the somatic  $D_H$  sequences can be encoded by the 12  $D_H$  genes, we concluded that there are only three D<sub>H</sub> gene families in mouse genome.

D<sub>FL16</sub> family has two members and D<sub>SP2</sub> family has nine members [4]. It is quite obvious that the members belonging to each family were created by a gene duplication mechanism. Moreover, sequences of  $D_{FL16}$  and  $D_{SP2}$  are also homologous to each other, as shown in Fig. 3. The sequence similarity has been found not only in D<sub>H</sub> genes themselves but also in the surrounding regions; therefore, it is likely that both gene families orginate from the same primordial  $D_H$  gene. Using the flanking sequences of both genes, we calculated the divergence date beteen  $D_{FL16}\ \text{and}\ D_{SP2}$  genes as described in Table 1, and concluded that  $D_{FL16}$  and  $D_{SP2}$  genes had diverged around 37 million years ago. Fig. 6 schematically shows the evolutional pathway that created a set of D<sub>H</sub> genes in the mouse genome. A primordial D<sub>H</sub> gene was duplicated around 37 million years ago. Mutations were introduced into both DNA fragments, resulting in  $D_{FL16}$  and  $D_{SP2}$  genes. Both



Figure 6. Phylogenetic relationship between  $D_{FL16}$  and  $D_{SP2}$  gene families.  $D_{FL16}$  and  $D_{SP2}$  genes diverged from a primordial  $D_H$  gene around 37 million years ago. Both genes were duplicated once more. After that, only the D<sub>SP2</sub> gene was multiplied.

genes were duplicated once more. After that, only 5-kb fragments containing the D<sub>SP2</sub> gene were multiplied several times.

The reason why  $D_{FL16.1}$  is the most frequently used  $D_H$  gene is not clear. As long as the usage frequency of  $D_{FL16}$  and  $D_{SP2}$ was observed in  $D_H$ -J<sub>H</sub> joinings,  $D_{SP2}$  and  $D_{FL16}$  genes were equally used [4, 10, 11]. Moreover, judging from the sequence observed in  $D_H$ -J<sub>H</sub> joints [4, 11], not only the codon frame encoding Tyr-Tyr-Gly, but also the other codon frames were used. Selection might have occurred at the cellular level, not at the joining process. The reading frame of D<sub>H</sub> regions has also been discussed by others [16], leading to essentially the same conclusion as ours.

We thank Drs. Y. Takagi, I. Ishiguro and K. Fujita for their encouragement. We are also grateful to M. Yasuda, C. Kato and T. Inoue for their technical assistance, and to Ms A. Nagata for preparing the manuscript.

Received June 11, 1989.

## 5 References

- 1 Tonegawa, S., Nature 1983. 302: 575.
- 2 Sakano, H., Huppi, K., Heinrich, G. and Tonegawa, S., Nature 1979. 280: 288.
- 3 Early, P., Huang, H., Davis, M., Calame, K. and Hood, L., Cell 1980. 19: 981.
- 4 Kurosawa, Y. and Tonegawa, S., J. Exp. Med. 1982. 155: 201.
- 5 Siebenlist, U., Ravetch, J. V., Korsmeyer, S., Waldmann, T. and Leder, P., Nature 1981. 294: 631.
- 6 Ichihara, Y., Matsuoka, H. and Kurosawa, Y., EMBO J. 1988. 7: 4141.
- 7 Southern, E. M., J. Mol. Biol. 1975. 98: 503.
- 8 Sanger, F., Nicklen, S. and Coulson, A. R., Proc. Natl. Acad. Sci. USA 1977. 74: 5463.
- Wood, C. and Tonegawa, S., Proc. Natl. Acad. Sci. USA 1983. 80: 3030.
- 10 Alt, F., Yancopoulos, G. D., Blackwell, T. K., Wood, C., Thomas, E., Boss, M., Coffman, R., Rosenberg, N., Tonegawa, S. and Baltimore, D., EMBO J. 1984. 3: 1209.
- 11 Reth, M. G. and Alt, F. W., Nature 1984. 312: 418.
- 12 Kabat, E. A., Wu, T. T., Reid-Miller, M., Perry, H. M. and Gottesman, K. S., 1987. Sequences of Proteins of Immunological Interest. US Dept Health and Human Services, Washington, DC.
- 13 Alt, F. W. and Baltimore, D., Proc. Natl. Acad. Sci. USA 1982. 79: 418.

- 14 Sakano, H., Maki, R., Kurosawa, Y., Roeder, W. and Tonegawa, S., *Nature* 1980. 286: 676.
- 15 Goulian, M., Proc. Natl. Acad. Sci. USA 1968. 61: 284.
- 16 Kaartinen, M. and Makela, O., Immunol. Today 1985. 6: 324..
- 17 Kurosawa, Y., Von Boehmer, H., Haas, W., Sakano, H., Traunecker, A. and Tonegawa, S., *Nature* 1981. 290: 565.
- 18 Jukes, T. H. and Cantor, C. R. in Munro, H. N. and Allison, J. B., (Eds), *Mammalian Protein Metabolism*, Academic Press, New York 1969, Vol. 2, pp. 21–132.
- 19 Kimura, M. and Ohta, T., J. Mol. Evol. 1972. 2: 87.
- 20 Miyata, T., Hayashida, H., Kikuno, R., Hasegawa, M., Kobayashi, M. and Koike, K., J. Mol. Evol. 1982. 19: 28.
- 21 Hayashida, H. and Miyata, T., Proc. Natl. Acad. Sci. USA 1983. 80: 2671.
- 22 Rechavi, G., Bienz, B., Ram, D., Ben-Neriah, Y., Cohen, J. B., Zakut, R. and Givol, D., *Proc. Natl. Acad. Sci. USA* 1982. 79: 4405.