Immunogenetics (1998) 47: 355–363

ORIGINAL PAPER

Juan Carlos Almagro · Ismael Hernández
Maria del Carmen Ramírez · Enrique Vargas-Madrazo

Structural differences between the repertoires of mouse and human germline genes and their evolutionary implications

Received: 1 June 1997 / Revised: 6 October 1997

Abstract Although human and mouse antibodies are similar when one considers their diversification strategies, they differ in the extent to which kappa and lambda light chains are present in their respective variable light chain repertoires. While the *Igk-V* germline genes are preponderant in mice (95% or more), they comprise only 60% in humans. This may account for differences in the structural repertoire encoded in the *Igk-V* germline genes of these species. However, this subject has not been properly investigated, partially because a systematic structural characterization of the mouse *Igk-V* germline genes has not been undertaken. In the present study we compiled all available information on mouse *Igk-V* germline genes to characterize their structural repertoire. As expected, comparison with the structural repertoire of human *Igk-V* germline genes indicates differences. The most interesting is that the mouse *Igk-V* germline gene repertoire is more diverse in structural terms than its human counterpart: the mouse encodes seven canonical structure classes (combination of canonical structures in L1 and L3). In contrast, the human encodes only four. Analysis of the evolutionary relationships of human and mouse *Igk-V* germline genes led us to propose that the difference reflects a strategy of mice to compensate for the small lambda chain contribution to the repertoire of their variable light chains.

Key words Mice · Immunoglobulins · Canonical structures *Igk-V* · Evolution

Introduction

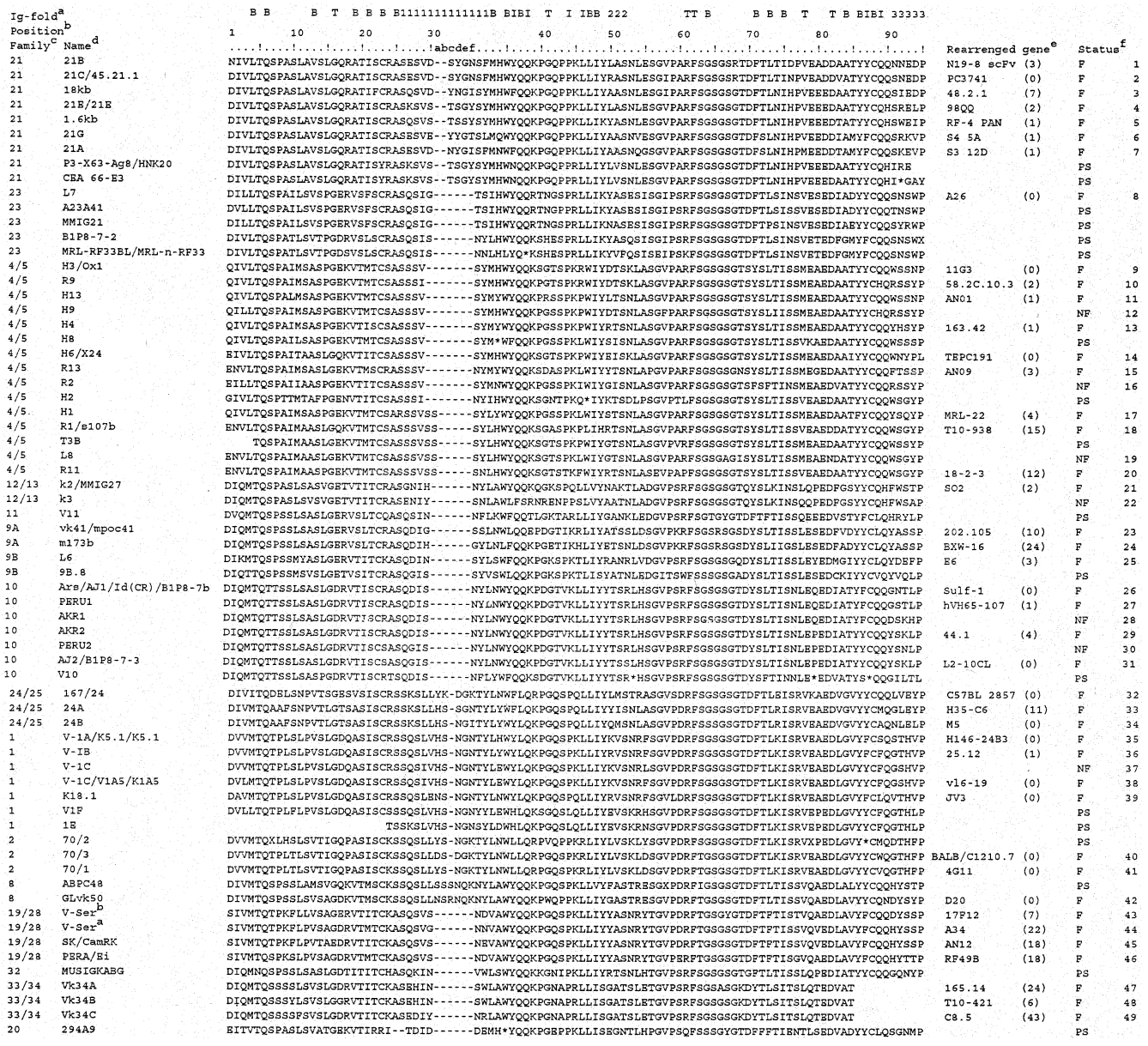
In contrast to other species such as shark, chicken, rabbit, or sheep, the human and mice species generate their antibody diversity in a similar fashion (Weill and Reynaud 1996). Prior to the antigenic challenge, these species produce a primary repertoire through the recombination of multiple germline genes (Berek and Milstein 1988; Neuberger and Milstein 1995; Tonegawa 1983). The variable kappa or lambda light chain is produced by the recombination of the *Igl-V* or *Igk-V* and *Igl-J* or *Igk-J* germline genes, respectively (Tonegawa 1983). The variable heavy chain is caused by a recombination of the *Igh-V* germline genes with two additional germline genes, *Igh-D* and *Igh-J* (Tonegawa 1983). Antibodies thus generated should be capable of interacting with any antigen at least with a low or medium affinity in the primary immune response (Berek and Milstein 1988; Neuberger and Milstein 1995). Upon selection by the antigen, the chosen human or mouse antibodies improve their affinity mainly by somatic hypermutation during a secondary or tertiary immune response (Berek and Milstein 1988; Neuberger and Milstein 1995; Weill and Reynaud 1996).

Despite the similarity that renders humans and mice "equivalent" in their diversification strategies for antibodies, these species possess different proportions of kappa and lambda light chains in their germline genes. In human, roughly 60% of the variable light chain repertoire is kappa [40 functional *Igk-V* germline genes (Klein et al. 1993; Tomlinson et al. 1995) vs 30 functional *Igl-V* germline genes (Williams et al. 1996)]. In mice, kappa preponderates, being as much as 95% or more [fewer than 160 functional *Igk-V* germline genes (Zocher et al. 1995) vs three functional *Igl-V* germline genes (Dildrop et al. 1987; Selsing et al. 1989)]. This difference may account for a divergence in the primary repertoire of human and mouse antibodies. However, a systematic analysis has not been made.

A way to characterize that difference is through a comparison of the repertoire of antigen binding site structures implicit in the variable light chain germline genes of mice and humans. Even though considerable sequence variability exists at the antigen binding site (Kabat and Wu 1971; Wu and Kabat 1970), it has been shown that the hypervariable loops do not adopt a large and unpredictable

J. C. Almagro (✉) · I. Hernández · M. del Carmen Ramírez
Instituto de Biotecnologia, Universidad Nacional Autónoma de México, Cuernavaca, APDO.POSTAL 510–3, Cuernavaca, Morelos 62250, México

E. Vargas-Madrazo
Instituto de Investigaciones Biológicas, Universidad Veracruzana, Xalapa, México



number of conformations: they possess one of a small set of main-chain conformations or canonical structures (Chothia and Lesk 1987; Chothia et al. 1989; Martin and Thornton 1996). On the basis of this, it has recently been reported that of the total number of possible combinations of these canonical structures (denoted canonical structure classes) (Almagro et al. 1996; Chothia et al. 1992; Lara-Ochoa et al. 1996; Tomlinson et al. 1995; Vargas-Madrazo et al. 1995) only a few options effectively exist. The existence of canonical structures and canonical structure classes implies restrictions to a free diversification of hypervariable loops and their combination within the same gene. Therefore, if there are significant differences between the primary repertoires of human and mouse antibodies, they might

Fig. 1 Multiple amino acid sequence alignment of mouse *Igk-V* germline genes. [a]Positions primarily responsible for the variable immunoglobulin fold (*V-Ig*-fold) conserved features (Chothia et al. 1988) and hypervariable loop definition (Chothia and Lesk 1987). *B* Residues buried in the protein; *T* residues in turns; *I* inter-domain residues**.** 1: L1; 2: L2, and 3: L3. [b]Residue numbering as in Chothia and Lesk (1987). [c]*Igk-V* gene family. [d]Name, clone, or sequence access number in GenBank, or name of the sequence in the literature. [e]Name in the Kabat's Database of the closest *Igk-V* rearranged gene and number of amino acid differences between this and the germline gene. [f]Sequence status: *F* sequences with rearranged counterpart (functional); *NF* non-functional sequence because it has no rearranged counterpart or possesses structural defects (SD); *PS* pseudogene. *Numbers* on the *right* stand for the code of each sequence in Fig. 2. The multiple sequences alignment and all the calculations presented therein were done using the VIR package (Almagro et al. 1995)

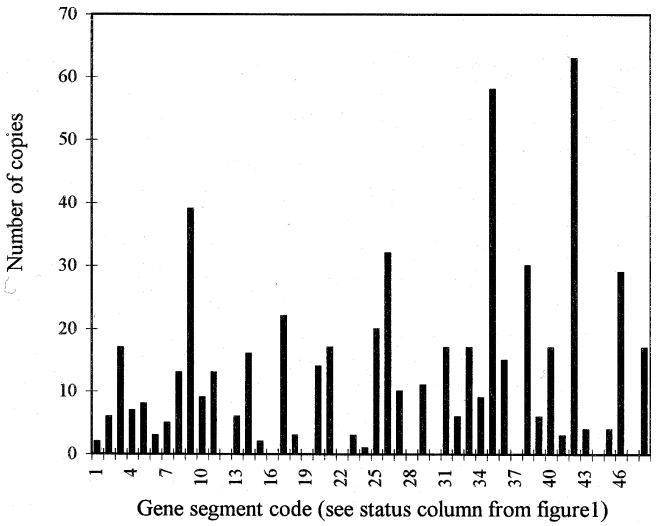


Fig. 2 Usage of *Igk-V* germline genes

become evident when their corresponding structural repertoires are analyzed.

Since the variable light chain repertoire in mice is kappa dominated, our analysis is mainly concerned with mouse *Igk-V* germline genes. The structural repertoire of human *Igk-V* germline genes has been described in detail by Tomlinson and co-workers (1995) but not the mouse. Thus in the first part of this paper we made a systematic characterization of the structural repertoire of mouse *Igk-V* germline genes so that in the second part we could make a comparison with the structural repertoire of human *Igk-V* germline genes. Finally, results are discussed in the light of the evolutionary relationships of human and mouse *Igk-V* genes.

Materials and methods

Mouse Igk-V germline genes

In order to estimate the *Igk-V* repertoire in the germlines of mice, we compiled all of the *Mus musculus Igk-V* genes reported as germline genes or pseudogene sequences in GenBank and LIGM, as well as in the literature published up to December 1996. We found a total of 97 *Igk-V* genes and immediately discarded eight of them because they were fragments of sequences (see web site http://www.ibt.unam.mx/nalmagro for a full description of the sequences).

Having collected 89 useful *Igk-V* genes, those fully identical at the nucleotide level in the coding region were considered to be the same. Similarly, sequences with only one or two nucleotide differences (99.6% and 99.2% identities, respectively) resulting in silent mutations (100% identical at amino acid level) were considered to be the same. This was so decided because they might be alleles in different individuals or in different strains of the mouse.

Sequences in which the nucleotide difference resulted in replacements (different amino acid sequences) were considered different genes. In that way we finally gathered 67 sequences as representative of the mouse *Igk-V* locus.

Table 1 *Igk-V* classification and germline gene repertoire

Igk-V family complexity		
Igk-V gene family[a]	Number of *Igk-V* germline genes (estimated)	Number of *Igk-V* germline genes (found)
21	6–13	9
23	2– 4	5
4/5	25–50	15
12–13	2– 8	2
11	4– 6	1
9A	4– 9	2
9B	2	2
10	2– 3	7
24–25	6	3
1	4– 6	7
2	1– 6	3
8	5–16	2
19–28	4– 6	4
38C	–	–
RF	0– 1	–
22	1– 2	–
20[b]	5– 7	1
32[c]	4– 8	1
33/34[d]	1– 3	3
Total	79–156	67

[a] *Igk-V* gene family nomenclature according to Strohal and co-workers (1989)
[b] Gene family described by Shefner and co-workers (1990)
[c] Gene family described by D'Hoostelaere and Klinman (1990)
[d] Two groups (D'Hoostelaere and Klinman 1990; Valiante and Caton 1990) have independently described this family and termed it *Igk-V33* and *Igk-V34*, respectively. To avoid confusion, it has been renamed *Igk-V33/34* (Kofler and Helmberg 1991)

Classification of the known mouse Igk-V germline genes in Igk-V families

On the basis of nucleotide comparisons, mouse *Igk-V* genes have been classified into 19 *Igk-V* families: 16 defined by Strohal and co-workers (1989) and three more defined by Valiante and Caton (1990), Shefner and co-workers (1990), and D'Hoostelaere and Klinman (1990). Therefore, in order to cluster the 67 *Igk-V* genes we had found into the 19 established *Igk-V* families, we followed the criteria established by the aforementioned authors. The resulting alignment of sequences, organized by families, is given in Fig. 1 and can be retrieved in a computer-ready format from web site: http:\\www.ibt.unam.mx/nalmagro.

The functional Igk-V germline genes of mice and their structural repertoire

Of the 67 *Igk-V* genes depicted in Fig. 1, 18 have been reported as pseudogenes in databases or in the literature (see status column of Fig. 1). This led us to assume they had serious genetic defects and were not taken into account to determine the functional repertoire of mouse *Igk-V* germline genes.

The remaining 49 *Igk-V* genes reported as germline and potentially functional were examined to observe their expression *in vivo.* This was done by assigning the amino acid sequence of each of them to their closest rearranged functional *Igk-V* sequence in a database of 574 Igk-V amino acid sequences compiled from the Kabat's Database on-line service (Kabat et al. 1991; web site: http://immuno.bme.nwu.edu). We chose those *Igk-V* rearranged sequences having a reported specificity to avoid non-productive rearrangements, while guaranteeing the assignment of functional *Igk-V* genes only. Since those segments not expressed *in vivo* might have minor genetic or structural defects hindering the formation of a stable three-dimensional *V* domain, they were considered to be non-functional (Klein et al. 1993; Tomlinson et

Fig. 3 Structural repertoire of the functional *Igk-V* germline gene segments of mice. V_κCSC Canonical structures classes of light chain *Igk-V* genes. *?* A hypervarible loop which does not fit the canonical structure pattern. Residues responsible for mismatch are *underlined*. *U* Unknown

Position[b] Family[c]	Name[d]	L1							L3							V_kCSC
		B	1	1	1	1	1	1	1	1	1	1	B	3	3	3	3	3	3	
2	25	30						33	71				90					95		
.	!	.	.	.	.	!	a	b	c	d	e	f	.	!	.	.	.	!		
21	21B	I	A	S	E	S	V	D	-	-	S	Y	G	N	S	F	M	F	Q	N	N	E	D	P	5-1
21	21C/45.21.1	I	A	S	E	S	V	D	-	-	S	Y	G	N	S	F	M	F	Q	S	N	E	D	P	5-1
21	18kb	I	A	S	Q	S	V	D	-	-	Y	N	G	I	S	Y	M	F	Q	S	I	E	D	P	5-1
21	21E/21E5	I	A	S	K	S	V	S	-	-	T	S	G	Y	S	Y	M	F	H	S	R	E	L	P	5-1
21	1.6kb	I	A	S	Q	S	V	S	-	-	T	S	S	Y	S	Y	M	F	H	S	W	E	I	P	5-1
21	21G	I	A	S	E	S	V	E	-	-	Y	Y	G	T	S	L	M	F	Q	S	R	K	V	P	5-1
21	21A	I	A	S	E	S	V	D	-	-	N	Y	G	I	S	F	M	F	Q	S	K	E	V	P	5-1
23	L7	I	A	S	Q	S	I	G	-	-	-	-	-	T	S	I		F	Q	S	N	S	W	P	2-1
4/5	H3(Ox1)	I	A	S	S	S	V	-	-	-	-	-	S	Y	M			Y	Q	W	S	S	N	P	1-1
4/5	R9	I	A	S	S	S	I	-	-	-	-	-	S	Y	M			Y	Q	R	S	S	Y	P	1-1
4/5	H13	I	A	S	S	S	V	-	-	-	-	-	S	Y	M			Y	Q	W	S	S	N	P	1-1
4/5	H4	I	A	S	S	S	V	-	-	-	-	-	S	Y	M			Y	Q	Y	H	S	Y	P	1-1
4/5	X24/H6	I	A	S	S	S	V	-	-	-	-	-	S	Y	M			Y	Q	W	N	Y	P	L	1-2
4/5	R13	N	A	S	S	S	V	-	-	-	-	-	N	Y	M			Y	Q	F	T	S	S	P	?-1
4/5	H1	I	A	R	S	S	V	S	S	-	-	-	-	S	Y	L		Y	Q	Y	S	Q	Y	P	?-1
4/5	R1/s107b	N	A	S	S	S	V	S	S	-	-	-	-	S	Y	L		Y	Q	W	S	G	Y	P	6-1
4/5	R11	N	A	S	S	S	V	S	S	-	-	-	-	S	N	L		Y	Q	W	S	G	Y	P	6-1
12/13	k2/MMIG27	I	A	S	G	N	I	H	-	-	-	-	-	N	Y	L		Y	H	F	W	S	T	P	2-1
9A	vk41/mopc41	I	A	S	Q	D	I	G	-	-	-	-	-	S	S	L		Y	Q	Y	A	S	S	P	2-1
9B	L6	I	A	S	Q	D	I	N	-	-	-	-	-	S	Y	L		Y	Q	Y	D	E	F	P	2-1
10	Ars/AJ1/Id(CR)/B1P8-7b	I	A	S	Q	D	I	S	-	-	-	-	-	N	Y	L		Y	Q	G	N	T	L	P	2-1
10	PERU1	I	A	S	Q	D	I	S	-	-	-	-	-	N	Y	L		Y	Q	G	S	T	L	P	2-1
10	AKR2	I	A	S	Q	D	I	S	-	-	-	-	-	N	Y	L		Y	Q	Y	S	K	L	P	2-1
10	AJ2/B1P8-7-3	I	A	S	Q	G	I	S	-	-	-	-	-	N	Y	L		Y	Q	Y	S	K	L	P	2-1
24/25	167/24	I	S	S	K	S	L	L	Y	K	-	D	G	K	T	Y	L	F	Q	L	V	E	Y	P	4-1
24/25	Vk24	I	S	S	K	S	L	L	H	S	-	N	G	I	T	Y	L	F	Q	M	L	E	R	P	4-1
24/25	24B	I	S	S	K	S	L	L	H	S	-	N	G	I	T	Y	L	F	Q	N	L	E	L	P	4-1
1	V-1A/K5.1/K5.1	V	S	S	Q	S	L	V	H	S	-	N	G	N	T	Y	L	F	Q	S	T	H	V	P	4-1
1	V-IB	V	S	S	Q	S	L	V	H	S	-	N	G	N	T	Y	L	F	Q	G	T	H	V	P	4-1
1	V-1C/V1A5/K1A5	V	S	S	Q	S	I	V	H	S	-	N	G	N	T	Y	L	F	Q	G	S	H	V	P	4-1
1	18.1/K18,1	A	S	S	Q	S	L	E	N	S	-	N	G	N	T	Y	L	F	Q	V	T	H	V	P	?-1
2	70/3	V	S	S	Q	S	L	L	D	S	-	D	G	K	T	Y	L	F	Q	G	T	H	F	P	4-1
2	70/1	V	S	S	Q	S	L	L	Y	S	-	N	G	K	T	Y	L	F	Q	G	T	H	F	P	4-1
8	GLvk50	I	S	S	Q	S	L	L	N	S	R	N	Q	K	N	Y	L	F	N	D	Y	S	Y	P	3-1
19/28	V-Ser[c]	I	A	S	Q	S	V	S	-	-	-	-	-	N	D	V		F	Q	D	Y	S	S	P	2-1
19/28	SK/CamRK	I	A	S	Q	S	V	S	-	-	-	-	-	N	E	V		F	Q	H	Y	S	S	P	2-1
19/28	PERA/Ei	I	A	S	Q	S	V	S	-	-	-	-	-	N	D	V		F	Q	H	Y	T	T	P	2-1
33/34	Vk34B	I	A	S	E	H	I	N	-	-	-	-	-	S	W	L		Y							2-U

al. 1995). The database with the 574 Igk-V amino acid sequences is available from the authors on request.

Characterization of the canonical structures in L1 and L3

The patterns of residues determining the different canonical structures for L1 and L3 have been described in detail (Chothia and Lesk 1987; Chothia et al. 1989; Tomlinson et al. 1995), and have recently been reviewed by Martin and Thornton (1996). This information is summarized at web site: Antibody Structure-function (http: //www.biochem.-ucl.ac.uk/~martin/antibodies.html:Chothia.dat.auto). Using these patterns, we analyzed the functional *Igk-V* germline genes of mice. It must be noted that L2 was not considered in the analysis because it has only one canonical conformation and does not influence the structural variability in antibodies.

Results

Known mouse Igk-V germline genes

Although the exact number of *Igk-V* germline genes in the mouse genome is currently unknown, several estimations exist (Cory et al. 1981; Kofler et al. 1992; Zeelon et al. 1981; Zocher et al. 1995). Early proposals ranged from 90–320 (Cory et al. 1981) up to 2000 (Zeelon et al. 1981). However, on the basis of restriction fragment length polymorphism (RFLP) criteria, as well as on current established knowledge about *Igk-V* germline gene sequences and expressed *Igk-V* sequences, it has been estimated that the entire *Igk-V* germline repertoire may not much exceed 160 (Kofler et al. 1992). Recently, cloning experiments have facilitated the proposal that the final number of genes in the mouse *Igk-V* locus is notably smaller than 160 (Zocher et al. 1995). In agreement with this, we found 89 *Igk-V* germline genes, of which 67 turned out to be unique (Fig. 1).

In Table 1 a comparison between the established number of *Igk-V* genes within the individual *Igk-V* families and those we found is shown. Such a comparison indicates that nine *Igk-V* gene families (*21*, 23, *12–13*, *9B*, *10*, *1*, *2*, *19–28*, and *33–34*) are well represented in our compilation. However, in seven *Igk-V* families (*4/5*, *11*, *9A*, *24–25*, *8*, *32*, *20*) we found fewer *Igk-V* genes than estimated (see Table 1). For the *38C*, *RF*, and *22* gene families, no *Igk-V*

Fig. 4 Structural repertoire of the functional human *Igk-V* germline gene segments. V_KCSC, Canonical structures classes of light chain *Igk-V* genes. *?*, A hypervarible loop which does not fit the canonical structure pattern. Residues responsible for mismatch are u*nderlined*

		L1			L3	
		B1111111111111B	B		33333	V_kCSC
Position[b]		2 25 30 33	71	90	95	
Family[c]	Name[d]	. !....!abcdef...	.	!	....!	
I	012/DPK9	I ASQSISS------YL	F	Q	SYSTP	2-1
I	02/DPK9	I ASQSISS------YL	F	Q	SYSTP	2-1
I	018/DPK1	I ASQDISN------YL	F	Q	YDNLP	2-1
I	08/DPK1	I ASQDISN------YL	F	Q	YDNLP	2-1
I	A20/DPK4	I ASQGISN------YL	F	K	YNSAP	2-?
I	A30	I ASQGIRN------DL	F	Q	HNSYP	2-1
I	L14/DKP2	I ARQGISN------YL	F	Q	HNSYP	2-1
I	L1	I ASQGISN------YL	F	Q	YNSYP	2-1
I	L15/DPK7	I ARQGISS------WL	F	Q	YNSYP	2-1
I	L4	I ASQGISS------AL	F	Q	FNSYP	2-1
I	L18	I ASQGISS------AL	F	Q	FNSYP	2-1
I	L5/DPK5	I ASQGISS------WL	F	Q	ANSFP	2-1
I	L19/DPK6	I ASQGISS------WL	F	Q	ANSFP	2-1
I	L8/DPK8	I ASQGISS------YL	F	Q	LNSYP	2-1
I	L23	I ASQGISS------YL	Y	Q	YYSTP	2-1
I	L9	I ASQGISS------YL	F	Q	YYSYP	2-1
I	L24/DPK10	I MSQGISS------YL	F	Q	YYSFP	?-1
I	L11/DPK3	I ASQGIRN------DL	F	Q	DYNYP	2-1
I	L12	I ASQSISS------WL	F	Q	YNSYS	2-?
II	011/DPK13	I SSQSLLDSDDGNTYL	F	Q	RIEFP	3-1
II	01/DPK13	I SSQSLLDSDDGNTYL	F	Q	RIEFP	3-1
II	A17/DPK18	V SSQSLVYS-DGNTYL	F	Q	GTHWP	4-1
II	A1/DPK19	V SSQSLVYS-DGNTYL	F	Q	GTHWP	4-1
II	A18/DPK28	I SSQSLLHS-DGVTYL	F	Q	GTHLP	4-1
II	A2/DPK12	I SSQSLLHS-DGKTYL	F	Q	SIQLP	4-1
II	A19/DPK15	I SSQSLLHS-NGYNYL	F	Q	ALQTP	4-1
II	A3/DPK15	I SSQSLLHS-NGYNYL	F	Q	ALQTP	4-1
II	A23/DPK16	I SSQSLVHS-DGNTYL	F	Q	ATQFP	4-1
III	A27/DPK22	I ASQSVSSS-----YL	F	Q	YGSSP	6-1
III	A11/DPK20	I ASQSVSSS-----YL	F	Q	YGSSP	6-1
III	L2/DPK21	I ASQSVSS------NL	F	Q	YNNWP	2-1
III	L16	I ASQSVSS------NL	F	Q	YNNWP	2-1
III	L6	I ASQSVSS------YL	F	Q	RSNWP	2-1
III	L20	I ASQGVSS------YL	F	Q	RSNWH	2-?
III	L25/DPK23	I ASQSVSSS-----YL	F	Q	DYNLP	6-1
IV	B3/DPK24	I SSQSVLYSSNNKNYL	F	Q	YYSTP	3-1
V	B2	T ASQDIDD------DM	F	Q	HDNFP	2-1
VI	A26/DPK26	I ASQSIGS------SL	F	Q	SSSLP	2-1
VI	A10/DPK26	I ASQSIGS------SL	F	Q	SSSLP	2-1
VI	A14/DPK25	V ASEGIGN------YL	F	Q	GNKHP	2-1

genes were detected and consequently these families remained empty. But they contain very few members and in some cases, as in the *38C* gene family, the sequences belonging to these *Igk-V* families might represent highly mutated genes from other families (Strohal et al. 1989). Therefore, we considered that they made only a marginal contribution to the diversity of the entire *Igk-V* repertoire.

The functional mouse Igk-V germline genes and their structural repertoire

Analysis of the expression in vivo of the 49 *Igk-V* genes reported as *Igk-V* germline (see status column of the Fig. 1) suggests that 42 of them are functional (Fig. 2). Within the seven *Igk-V* genes not expressed in vivo, and therefore defined as non-functional, four are seen to posses structural defects when compared with antibodies of known three-dimensional structure (data not shown). The remaining three *Igk-V* genes were considered to have minor genetic defects and were not analyzed further.

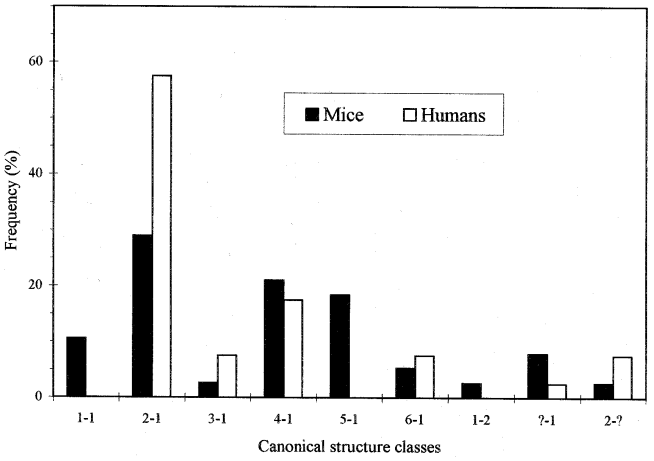


Fig. 5 Comparison of the use frequency of the canonical structural repertoire of mouse and human *Igk-V* germline genes

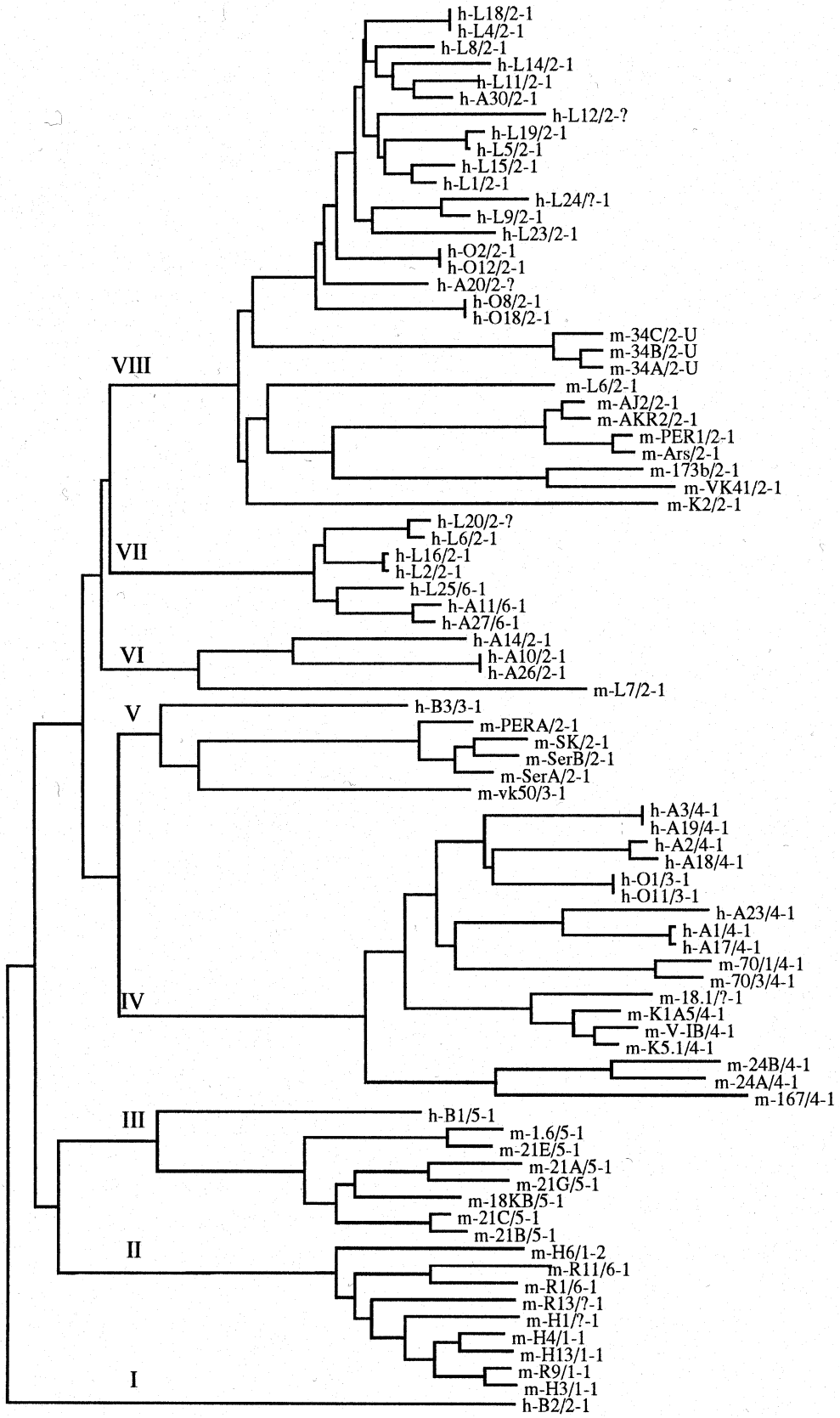


Fig. 6 Evolutionary relationship between human and mouse *Igk-V* germline genes. The evolutionary tree is based on the nucleotide similarities of mouse and human *Igk-V* germline genes shown in Figs. 3 and 4, respectively. Mouse and human genes are distinguished by an "m" or "h" before the name of each gene. After the name, separated by a *slash*, is the canonical structure class encoded in each *Igk-V* germline gene. Being a human pseudogene, *B1* is not included in Fig. 4 and was added to the analysis for completeness. The *B1* sequence was obtained from V-base (web site: http://www.mrc-cpe.cam.ac.uk). The tree was obtained with the CLUSTALW program (Thompson et al. 1994)

Of the 42 functional mice *Igk-V* germline genes, 39 have patterns compatible with some canonical structure in L1 (Fig. 3). In L3, all sequences present canonical structures, with the exception of three *Igk-V* genes (*34A*, *34B*, and *34C*) for which this loop has not been sequenced (Valiante and Caton 1990) and could not be assigned any structure at all.

Inspection of the structural repertoire of the mouse *Igk-V* germline genes (Fig. 3) indicates that the mouse encodes seven canonical structure classes. Classes 2–1 and 4–1 are the most frequent (31% and 19%, respectively) followed by classes 5–1 and 1–1 (17% and 10%, respectively). Classes 6–1, 3–1, and 1–2 are poorly represented in the mouse *Igk-V* germline genes (5%, 2%, and 2%, respectively).

Comparison between the structural repertoires implicit in mouse and human Igk-V germline genes

To compare the mouse and human structural repertoires, the canonical structures classes implicit in the 40 functional human *Igk-V* germline genes are depicted in Fig. 4. In contrast to mice, the entire repertoire of human *Igk-V* germline genes encodes only four canonical structure classes. Canonical structure classes 5–1, 1–1, and 1–2 implicit in the functional *Igk-V* germline genes of mice were found not to have a human counterpart.

In addition, differences exist in the degree to which these species encode class 2–1 (Fig. 5). This class is encoded in mice by 13 of 42 genes (~30%), while in humans it is encoded by 23 of 40 genes (~60%). As can be seen, in humans just one class contributes ~60% of the structural repertoire, whereas in mice this class contributes ~30%. Classes 4–1 (19%), 5–1 (17%,), and 1–1 (10%) play an important role in the mouse, but the last two classes are not present in humans. Thus, the structural repertoire in mice is not only more diverse than in humans, it is also more heterogeneous in the sense of being less skewed or biased towards a particular class.

Discussion

Because the mice variable light chain repertoire is essentially, composed of kappa, while in humans the kappa chain is only ~60%, in the preceding section we compared the structural repertoires of mouse and human *Igk-V* germline genes. Results indicated that in structural terms mouse is more diverse than human. That suggests an evolutionary strategy of mice to compensate for the small lambda chain contribution in its repertoire of variable light chains. To expedite further examination of this suggestion, in Fig. 6 a phylogenetic tree relating the *Igk-V* germline genes of humans and mice is shown.

Tree topology is similar to that proposed by Kroemer and co-workers (1991), who classified human and mouse *Igk-V* genes into groups of *Igk-V* families or clans (Roman numerals in the figure). This classification represents the common ancestral human and mouse *Igk-V* germline genes or *Igk-V* gene families (Kroemer et al. 1991). As is evident from the tree, the structural repertoire implicit in functional *Igk-V* germline genes is far from being randomly distributed within *Igk-V* gene families or even within clans: namely, canonical structures classes are family- and clan-specific. This suggests that the structural repertoire was established 1) prior to the divergence of humans and mice, and 2) has been preserved despite the diversification of human and mouse *Igk-V* germline genes.

On the basis of these suggestions, we followed the evolutionary pathway of the structural repertoire and noticed that clans II and III encode almost exclusively the mouse-specific canonical structure classes 5–1, 1–1, and 1–2. Interestingly, clans II and III have no human counterpart except for sequence B1 (Lorenz et al. 1988) which possesses canonical structure 5–1. However, B1 is a pseudogene due to its having a modified start codon (Klein et al. 1993) and, being non-functional, does not contribute to actual human repertoire variability (we added this sequence to the tree for completeness). Therefore it seems reasonable to propose that *Igk-V* germline genes belonging to these clans were deleted or never developed in humans.

In contrast to the absence of the human functional *Igk-V* germline genes in clans II and III, we found that the mouse *Igk-V* gene families belonging to these clans are two of the three more complex families in that species. Clan II contains the *Igk-V* 4/5 family, the most complex mouse *Igk-V* gene family (see Table 1). Likewise, clan III consists of the mouse *Igk-V* 21 family, which is its third largest (see Table 1). The large number of *Igk-V* germline genes in these families suggested that there were influences to expand them, such as a demand to complement the "poor" diversity encoded by the remaining *Igk-V* gene families. Since expansion of the *Igk-V* 4/5 and *Igk-V* 21 families implied development of the canonical structure classes 1–1, 1–2, and 5–1, which are not present in the functional genes of humans, it can be assumed that these classes developed in mice to supplement the poor structural diversity inherited from the human and mouse ancestors: namely, the remaining four classes, 2–1, 6–1, 4–1, and 3–1.

In humans, the lambda chain might have furnished the additional structural diversity to set aside canonical structure classes 1–1, 1–2, and 5–1. In accordance with this assumption, in structural terms the repertoire of human *Igl-V* germline genes is more diverse than that of mice: human encodes nine canonical structure classes, while mouse only encodes two (Williams et al. 1996). This accounts for the converse relationship we found in the *Igk-V* germline genes and supports the suggestion that human *Igl-V* germline genes supply structural diversity to compensate for the lack of canonical structure classes 1–1, 1–2, and 5–1. This, together with the analysis of the *Igk-V* germline genes, indicates that the ultimate reason for the major diversification of the mouse *Igk-V* structural repertoire is related to the poor lambda chain contribution. In other words, the possible deletion of an important part of the *Igl-V* locus in mice might have forced its *Igk-V* locus to be structurally more diverse. The hypothesis that mice lost an

important part of the *Igl-V* locus is supported by phylogenetic analysis showing that the human *Igl-V* genes were originated at early stages of vertebrate evolution, prior to the divergence of humans and mice (Haire et al. 1996).

These results have interesting implications for the evolution of the *Ig-V* genes. To explain their evolution it has been suggested that the *Igh-V, Igk-V,* and *Igl-V* loci evolved by stochastic processes, where no positive darwinian selection operated to retain the complexity of the *Ig-V* loci and the coherence within *Ig-V* gene families (Tutter and Riblet 1989). Alternatively, it has been proposed that some environmental selection pressures retain the complexity of the *Ig-V* loci (Kirkham et al. 1992; Schroeder et al. 1990). In the case of the *Igh-V* locus, this latter suggestion is supported by analysis of the mice and human sequences (Kirkham et al. 1992; Schroeder et al. 1990). It shows that although the number of *Igh-V* genes might vary between species, the genes can be clustered into *Igh-V* families and *Igh-V* clans, which share structural features like the framework regions 1 and 3 (Kirkham et al. 1992; Schroeder et al. 1990). This has been extended to species that diverged from human 200 million years ago (Anderson and Matsunaga 1995) or more (Ota and Nei 1994). The structural features preserved throughout evolution within the *Igh-V* gene families and *Igh-V* clans suggest a reflection of those environmental selection pressures (Kirkham et al. 1992; Schroeder et al. 1990).

Human and mouse *Igk-V* genes can also be clustered into families and clans. However, the nature of the environmental selection pressures operating to retain the complexity and coherence within *Igk-V* gene families remains speculative (Kroemer et al. 1991). Here we found that the evolutionary diversification of the *Igk-V* germline genes preserves the structural repertoire. Moreover, the divergence of the mouse *Igk-V* gene families from those of human could be explained in terms of the structural repertoire diversification strategies. Therefore, we suggest that the conservation of a "basic" structural repertoire, on the one hand, and its diversification to furnish a minimum of structural diversity, on the other, operate as opposite selective pressures to retain the complexity of the *Igk-V* locus and coherence within *Igk-V* gene families.

Similar observations have been made regarding the *Igl-V* locus. It is interesting to note that, for example, horses, in which the lambda chain predominates, have developed more canonical structure classes than have humans (Williams et al. 1996). Thus, further analysis of other species whose kappa and lambda light chain contributions differ would help test the consistency and generalization of an evolutionary model for the *Igl-V* and *Igk-V* loci based on the diversification and conservation of structural repertoires.

Acknowledgments We thank I. A. Tomlinson for kindly providing the sequences of the 40 functional human *Igk-V* gene segments, and H. Ceceña and B. Levin for revision of the submitted manuscript. E.V. was supported by CONACyT grant 1833 PN. This work was partially supported by grant DGAPA-UNAM IN213796.

References

Almagro, J.C., Vargas-Madrazo, E., Zenteno-Cuevas, R., Hernandez-Mendiola, V., and Lara-Ochoa, F. VIR: a computational tool for analysis of immunoglobulin sequences. *BioSystems 35:* 25–32, 1995

Almagro, J. C., Domínguez-Martinez, V., Lara-Ochoa, F., and Vargas-Madrazo, E. Structural repertoire in human *VL* pseudogenes of immunoglobulins: comparison with functional germline genes and amino acid sequences. *Immunogenetics 43:* 92–96, 1996

Anderson, A. and Matsunaga, T. Evolution of immunoglobulin heavy chain variable region genes: a *VH* family can last for 150–200 million years or longer. *Immunogenetics 41:* 18–28, 1995

Berek, C. and Milstein, C. The dynamic nature of the antibody repertoire. *Immunol Rev 105:* 5–26, 1988

Chothia, C. and Lesk, A. M. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol 196:* 901–917, 1987

Chothia, C., Boswell, D. R., and Lesk, A. The outline structure of the T-cell αβ receptor. *EMBO J 7:* 3745–3755, 1988

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D., Tulip, W. R., Colman, P. M., Spinelli, S., Alzari, P. M., and Poljak, R. J. Conformations of immunoglobulin hypervariable regions. *Nature 342:* 877–883, 1989

Chothia, C., Lesk, A. M., Gherardi, E., Tomlinson, I. M., Walter, G., Marks, J. D., Llewelyn, M. B., and Winter, G. Structural repertoire of the human V_H segments. *J Mol Biol 227:* 799–817, 1992

Cory, S., Tyler, B. M., and Adams, J. M. Sets of immunoglobulin V kappa genes homologous to ten cloned V kappa sequences: implications for the number of germline V kappa genes. *J Mol Appl Genet 1:* 103–116, 1981

D'Hoostelaere, L. A. and Klinman, D. Characterization of new mouse V kappa groups. *J Immunol 145:* 2706–2712, 1990

Dildrop, R., Gause, A., Muller, W., and Rajewsky, K. A new V gene expressed in lambda-2 light chains of the mouse. *Eur J Immunol 17:* 731–734, 1987

Haire, R. N., Ota, T., Rast, J. P., Litman, R. T., Chan, F. Y., Zon, L. I., and Litman, G. W. A third Ig light chain gene type in *Xenopus laevis* consists of six distinct VL families and is related to mammalian lambda genes. *J Immunol 157:* 1544–1550, 1996

Kabat, E. A. and Wu, T. T. Attempts to locate complementarity determining residues in the variable positions of light and heavy chains. *Ann N Y Acad Sci 190:* 382–383, 1971

Kabat, E. A., Wu, T. T., Perry, H. M., Gottesman, K. S., and Foeller, C. *Sequences of Proteins of Immunological Interest (5th edn)*, Public Health Service. N.I.H. Washington. D.C.,1991

Kirkham, P. M., Mortari, F., Newton, J. A., and Schroeder, H. W. Jr. Immunoglobulin VH clan and family identity predicts variable domain structure and may influence antigen binding. *EMBO J 11:* 603–609, 1992

Klein, R., Jaenichen, R., and Zachau, H. G. Expressed human immunoglobulin *k* genes and their hypermutation. *Eur J Immunol 23:* 3248–3271, 1993

Kofler, R. and Helmberg, A. Comment to the article "A new Igk-V gene family in the mouse". *Immunogenetics 34:* 139–140, 1991

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

Kroemer, G., Helmberg, A., Bernot, A., Auffray, C., and Kofler, R. Evolutionary relationship between human and mouse immunoglobulin kappa light chain variable region genes. *Immunogenetics 33:* 42–49, 1991

Lara-Ochoa, F., Almagro, J. C., Vargas-Madrazo, E., and Conrad, M. Antibody-antigen recognition: a canonical structure paradigm. *J Mol Evol 43:* 678–684, 1996

Lorenz, W., Schable, K. F., Thiebe, R., Stavnezer, J., and Zachau, H. G. The J kappa proximal region of the human K locus contains three uncommon V kappa genes which are arranged in opposite transcriptional polarities. *Mol Immunol 25:* 479–484, 1988

Martin, A. C. and Thornton, J. M. Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. *J Mol Biol 263:* 800–815, 1996

Neuberger, M. S. and Milstein, C. Somatic hypermutation. *Curr Opin Immunol 7:* 248–254, 1995

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

Schroeder, H. W. Jr., Hillson, J. L., and Perlmutter, R. M. Structure and evolution of mammalian VH families. *Int Immunol 20:* 41–50, 1990

Selsing, E., Durdik, J., Moore, M. W., and Persiani, D. ML. *In* T. Honjo, F.W. Alt, and T.H. Rabbits (eds.): *Immunoglobulin Genes (2nd edn),* p. 111, Academic Press, New York, 1989

Strohal, R., Helmberg, A., Kroemer, G., and Kofler, R. Mouse *Vk* gene classification by nucleic acid sequence similarity. *Immunogenetics 30:* 475–493, 1989

Thompson, J. D., Higgins, D. G., and Gibson, T. J. 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, 1994

Tomlinson, I. A., Cox, J. P., Gherardi, E., Lesk, A. M., and Chothia, C. The structural repertoire of the human V kappa domain. *EMBO J 14:* 4628–4638, 1995

Tonegawa, S. Somatic generation of antibody diversity. *Nature 302:* 575–581, 1983

Tutter, A. and Riblet, R. Conservation of an immunoglobulin variable-region gene family indicates a specific noncoding function. *Proc Natl Acad Sci USA 86:* 7460–7464, 1989

Valiante, N. M. and Caton, A. J. A new *Igk-V* gene family in the mouse. *Immunogenetics 32:* 345–350, 1990

Vargas-Madrazo, E, Lara-Ochoa, F., and Almagro, J. C. Canonical structure repertoire of the antigen-binding site of immunoglobulins suggests strong geometrical restrictions associated to the mechanism of immune recognition. *J Mol Biol 254:* 497–504, 1995

Weill, J.-C. and Reynaud, C.-A. Rearrangement/hypermutation/gene conversion: when, where and why? *Immunol Today 17:* 92–97, 1996

Williams, S. C., Frippiat, J.-P., Tomlinson, I. A., Ignatovich, O., Lefranc, M.-P., and Winter, G. Sequence and evolution of the human germline Vλ repertoire. *J Mol Biol 264:* 220–232, 1996

Wu, T. T. and Kabat, E. A. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J Exp Med 132:* 211–250, 1970

Zeelon, E. P., Bothwell, A. L. M., Kantor, F., and Schechte, I. An experimental approach to enumerate the genes coding for immunoglobulin variable-regions. *Nucleic Acids Res 9:* 3809–3820, 1981

Zocher, I., Roschenthaler, F., Kirschbaum, T., Schable, K. F., Horlein, R., Fleischmann B., Kofler, R., Geley, S., Hameister, H., and Zachau, H. G. Clustered and interspersed gene families in the mouse immunoglobulin kappa locus. *Eur J Immunol 25:* 3326–3331, 1995